Technical report:
Debugging tasks supported either by manual or automatic test cases
Analysis of two replications: Trento and Milan

Mariano Ceccato[1], Cu Nguyen Duy[1], Alessandro Marchetto[1], Leonardo Mariani[2], Paolo Tonella[1]
[1]Fondazione Bruno Kessler–IRST, Trento, Italy
[2]University of Milan Bicocca, Milano, Italy

1 Analysis of accuracy

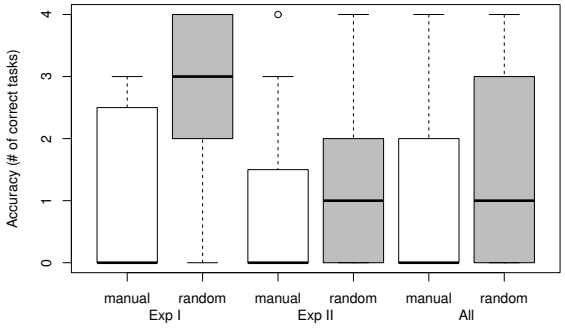


Figure 1. Boxplots of accuracy.

	exp	random.mean	random.median	random.sd	manual.mean	manual.median	manual.sd	p.value
1	I	2.60	3.00	1.67	1.14	0.00	1.46	0.15
2	II	1.39	1.00	1.38	0.80	0.00	1.28	0.11
3	All	1.65	1.00	1.50	0.89	0.00	1.31	0.04

Table 1. Unpaired analysis of accuracy (Mann-Whitney's test).

	exp	N	diff.mean	diff.median	diff.sd	p.value
1	I	4	1.25	1.00	2.22	0.42
2	II	16	0.38	0.00	0.62	0.04
3	All	20	0.55	0.00	1.10	0.03

Table 2. Paired (random-manual) analysis of correctness (Wilcoxon's test).

2 Analysis of efficiency

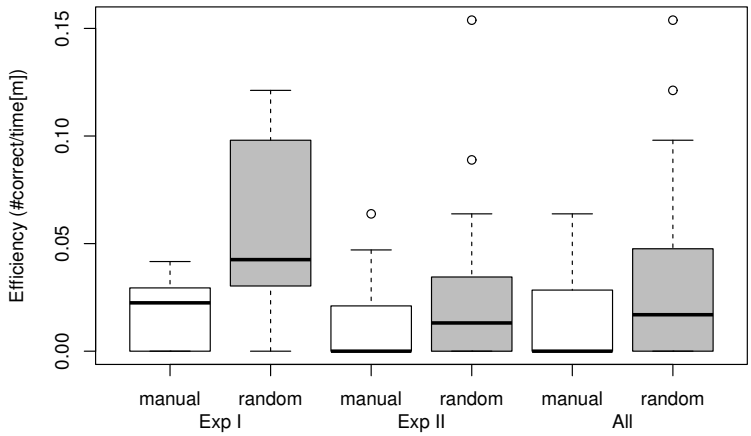


Figure 2. Boxplots of fixing efficiency.

	exp	random.mean	random.median	random.sd	manual.mean	manual.median	manual.sd	p.value
1	I	0.06	0.04	0.05	0.02	0.02	0.02	0.08
2	II	0.03	0.01	0.04	0.01	0.00	0.02	0.09
3	All	0.03	0.02	0.04	0.01	0.00	0.02	0.04

Table 3. Unpaired analysis of efficiency (Mann-Whitney's test).

	exp	N	diff.mean	diff.median	diff.sd	p.value
1	I	4	0.05	0.05	0.06	0.12
2	II	16	0.01	0.01	0.03	0.03
3	All	20	0.02	0.01	0.04	0.01

Table 4. Paired (random-manual) analysis of efficiency (Wilcoxon's test).

3 Analysis test case complexity

	Autogen tests		Manual tests	
	Autogen IDs	User-defined IDs	Autogen IDs	User-defined IDs
		JTopas		
F1	20	4	0	36
F2	18	9	0	59
F3	19	8	0	26
F4	61	22	0	16
		XML-security		
F1	7	3	0	9
F2	63	27	0	18
F3	13	5	0	20
F4	23	7	0	21

Table 5. Occurrences of auto/user generated identifiers in the test cases.

	id.type	N	diff.mean	diff.median	diff.sd	p.value
1	autogen	8	28.00	19.50	21.55	0.01
2	userdef	8	-15.00	-14.50	19.35	0.07

Table 6. Paired (random-manual) analysis of Autogen/User defined indentifiers (Wilcoxon's test).

Fault	MeLOC		McCabe		Exec. methods		Exec. LOCs	
	Ran	Man	Ran	Man	Ran	Man	Ran	Man
	Jtopas							
F1	21	34	3	4	47	80	224	436
F2	14	51	2	7	40	83	170	488
F3	14	17	2	1	12	10	47	28
F4	41	8	2	1	30	22	94	60
	XML-security							
F1	5	5	1	1	17	104	78	593
F2	48	26	1	1	50	142	177	836
F3	13	12	2	1	15	132	98	688
F4	16	14	1	3	16	385	48	2282

Table 7. Metrics on test cases.

	Metric	N	diff.mean	diff.median	diff.sd	p.value
1	MeLoc	8	0.62	0.50	21.12	1.00
2	McCabe	8	-0.62	0.00	2.07	0.59
3	Methods	8	-91.38	-65.00	120.77	0.04
4	LOCs	8	-559.38	-416.50	725.33	0.04

Table 8. Paired analysis on test case size and complexity (Wilcox's test).

	metric	Autogen.TP	Autogen.FN	Manual.FP	Manual.TN
1	Exec. Methods	6	2	2	6
2	Exec. LOCs	8	0	2	6

Table 9. Nearest Neighbor classifier predictor of the test case type (autogen vs. manual) based on executed methods or LOCs.

	metric	Precision	Recall	Accuracy	F.measure
1	Exec. Methods	0.75	0.75	0.75	0.75
2	Exec. LOCs	0.80	1.00	0.88	0.89

Table 10. Prediction performance metrics for the test case based Nearest Neighbor classifier on executed methods or LOCs.

4 Analysis co-factors — Accuracy

	cof	I	II	All
1	Treatment	0.04	0.06	0.03
2	Ability	0.01	0.00	0.00
3	Treatment:Ability	0.61	0.55	0.75
4	Treatment		0.13	0.04
5	Experience		0.00	0.00
6	Treatment:Experience		0.36	0.19
7	Treatment	0.18	0.19	0.06
8	System	0.50	0.90	0.76
9	Treatment:System	0.82	0.28	0.54
10	Treatment	0.14	0.16	0.06
11	Lab	0.19	0.24	0.11
12	Treatment:Lab	0.53	0.07	0.23

Table 11. Summary of ANOVA of Accuracy by Treatment & co-factor C_i

	cof	I	II	All
1	Treatment	0.87	0.77	0.34
2	Fault(JT)	0.07	0.45	0.00
3	Treatment:Fault(JT)	0.95	0.70	0.15
4	Treatment	0.00	0.52	0.02
5	Fault(XS)	1.00	0.90	0.71
6	Treatment:Fault(XS)	1.00	0.07	0.64

Table 12. Summary of ANOVA of Correctness by Treatment & Fault

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	6.19	6.19	5.72	0.0438
Ability	1	15.08	15.08	13.92	0.0058
Treatment:Ability	1	0.31	0.31	0.29	0.6051
Residuals	8	8.67	1.08		

Table 13. ANOVA of Accuracy by Treatment & Ability (Experiment I).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	3.29	3.29	3.66	0.0648
Ability	2	33.66	16.83	18.73	0.0000
Treatment:Ability	2	1.08	0.54	0.60	0.5546
Residuals	32	28.74	0.90		

Table 14. ANOVA of Accuracy by Treatment & Ability (Experiment II).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	7.24	7.24	5.04	0.0298
Ability	2	29.89	14.95	10.41	0.0002
Treatment:Ability	2	0.82	0.41	0.29	0.7529
Residuals	44	63.17	1.44		

Table 15. ANOVA of Accuracy by Treatment & Ability (Experiment All).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	3.29	3.29	2.38	0.1318
Experience	1	15.45	15.45	11.21	0.0020
Treatment:Experience	1	1.19	1.19	0.86	0.3595
Residuals	34	46.84	1.38		

Table 16. ANOVA of Accuracy by Treatment & Experience (Experiment II).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	7.24	7.24	4.67	0.0359
Experience	1	19.83	19.83	12.80	0.0008
Treatment:Experience	1	2.79	2.79	1.80	0.1863
Residuals	46	71.27	1.55		

Table 17. ANOVA of Accuracy by Treatment & Experience (Experiment All).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	6.19	6.19	2.20	0.1761
System	1	1.39	1.39	0.50	0.5014
Treatment:System	1	0.16	0.16	0.06	0.8158
Residuals	8	22.50	2.81		

Table 18. ANOVA of Accuracy by Treatment & System (Experiment I).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	3.29	3.29	1.82	0.1859
System	1	0.03	0.03	0.01	0.9045
Treatment:System	1	2.16	2.16	1.20	0.2811
Residuals	34	61.29	1.80		

Table 19. ANOVA of Accuracy by Treatment & System (Experiment II).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	7.24	7.24	3.58	0.0647
System	1	0.18	0.18	0.09	0.7647
Treatment:System	1	0.75	0.75	0.37	0.5447
Residuals	46	92.95	2.02		

Table 20. ANOVA of Accuracy by Treatment & System (Experiment AII).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	6.19	6.19	2.70	0.1391
Lab	1	4.70	4.70	2.05	0.1904
Treatment:Lab	1	0.99	0.99	0.43	0.5294
Residuals	8	18.37	2.30		

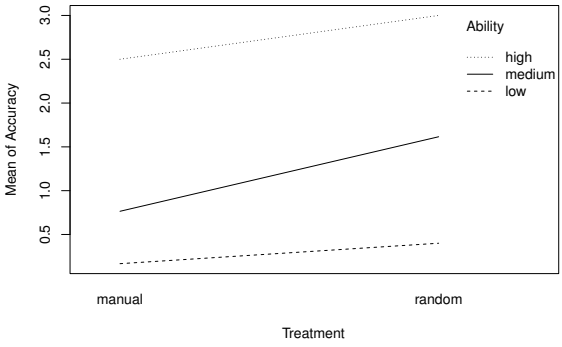
Table 21. ANOVA of Accuracy by Treatment & Lab (Experiment I).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	3.29	3.29	2.02	0.1647
Lab	1	2.34	2.34	1.44	0.2390
Treatment:Lab	1	5.75	5.75	3.53	0.0688
Residuals	34	55.38	1.63		

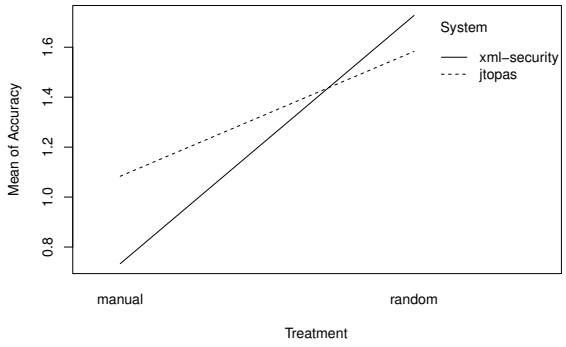
Table 22. ANOVA of Accuracy by Treatment & Lab (Experiment II).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	7.24	7.24	3.86	0.0555
Lab	1	4.87	4.87	2.60	0.1138
Treatment:Lab	1	2.78	2.78	1.48	0.2293
Residuals	46	86.23	1.87		

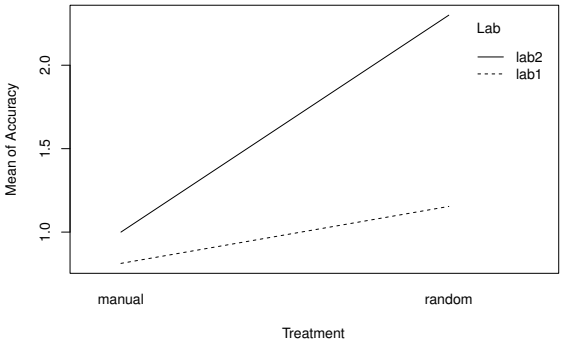
Table 23. ANOVA of Accuracy by Treatment & Lab (Experiment AII).



(a) Influence of Ability

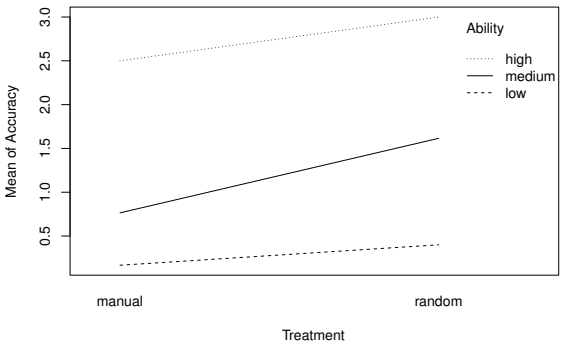


(b) Influence of System

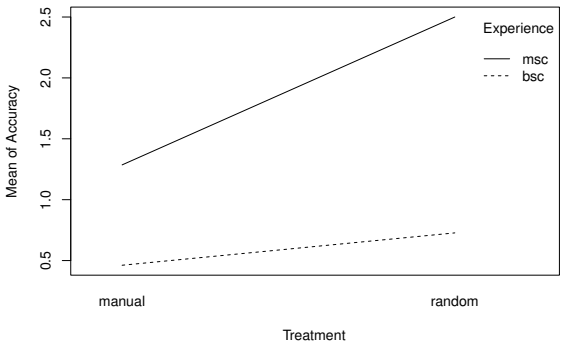


(c) Influence of Lab

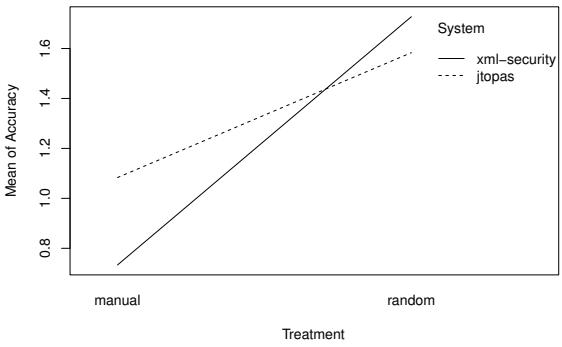
Figure 3. Interatcion plot for fixing accuracy (Experiment I).



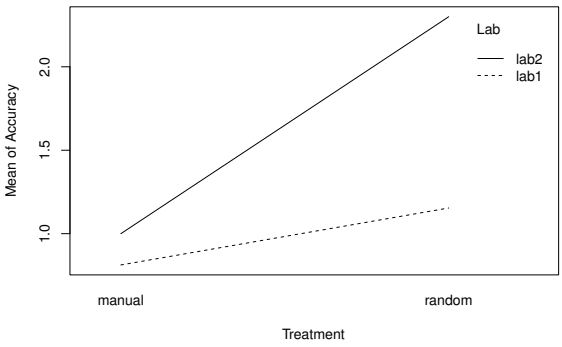
(a) Influence of Ability



(b) Influence of Experience

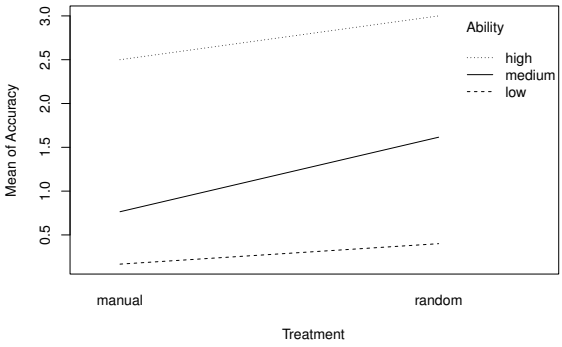


(c) Influence of System

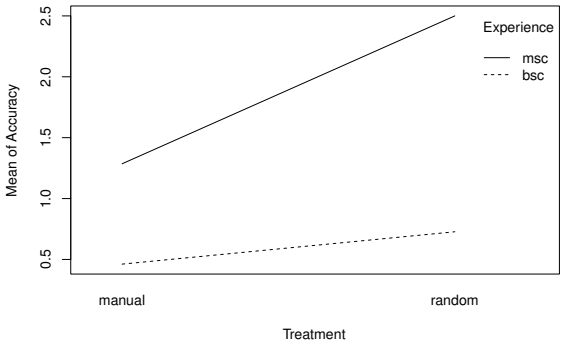


(d) Influence of Lab

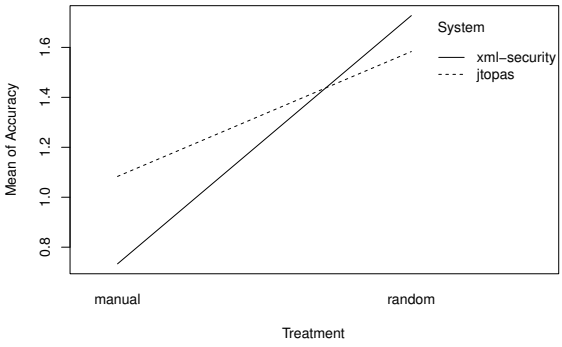
Figure 4. Interatcion plot for fixing accuracy (Experiment II).



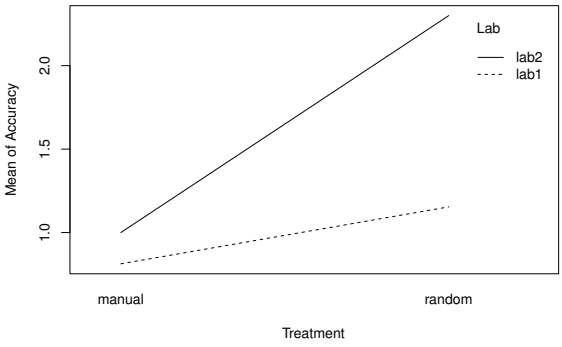
(a) Influence of Ability



(b) Influence of Experience

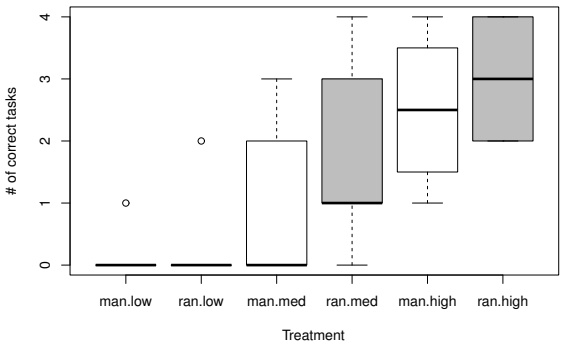


(c) Influence of System

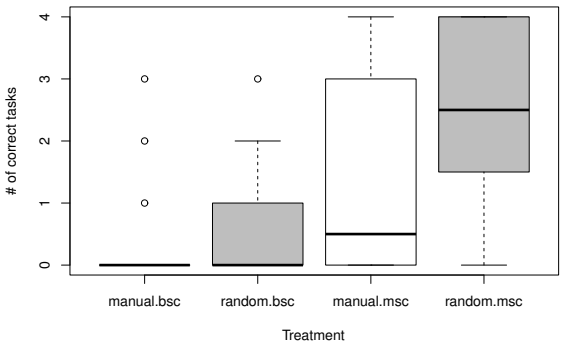


(d) Influence of Lab

Figure 5. Interatcion plot for fixing accuracy (Experiment I and II).



(a) Influence of Ability



(b) Influence of Experience

Figure 6. Boxplot of the effect of Ability and Experience on fixing accuracy.

5 Analysis co-factors — Efficiency

	cof	I	II	All
1	Treatment	0.01	0.05	0.01
2	Ability	0.01	0.00	0.00
3	Treatment:Ability	0.15	0.36	0.55
4	Treatment		0.08	0.01
5	Experience		0.01	0.00
6	Treatment:Experience		0.17	0.07
7	Treatment	0.09	0.11	0.02
8	System	0.61	0.30	0.59
9	Treatment:System	0.55	0.17	0.19
10	Treatment	0.09	0.09	0.02
11	Lab	0.41	0.13	0.10
12	Treatment:Lab	0.71	0.02	0.04

Table 24. Summary of ANOVA of Efficiency by Treatment & co-factor C_i

	cof	I	II	All
1	Treatment	0.26	0.04	0.90
2	Fault(JT)	0.00	0.00	0.08
3	Treatment:Fault(JT)	0.38	0.11	0.12
4	Treatment	0.03	0.92	0.01
5	Fault(XS)	0.33	0.00	0.01
6	Treatment:Fault(XS)	0.43	0.60	0.59

Table 25. Summary of ANOVA of Time by Treatment & Fault

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.00	0.00	9.94	0.0135
Ability	1	0.01	0.01	13.58	0.0062
Treatment:Ability	1	0.00	0.00	2.47	0.1549
Residuals	8	0.00	0.00		

Table 26. ANOVA of Efficiency by Treatment & Ability (Experiment I).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.00	0.00	4.25	0.0474
Ability	2	0.01	0.01	12.44	0.0001
Treatment:Ability	2	0.00	0.00	1.07	0.3563
Residuals	32	0.02	0.00		

Table 27. ANOVA of Efficiency by Treatment & Ability (Experiment II).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.01	0.01	6.79	0.0125
Ability	2	0.01	0.01	7.48	0.0016
Treatment:Ability	2	0.00	0.00	0.60	0.5516
Residuals	44	0.04	0.00		

Table 28. ANOVA of Efficiency by Treatment & Ability (Experiment All).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.00	0.00	3.23	0.0813
Experience	1	0.01	0.01	8.85	0.0054
Treatment:Experience	1	0.00	0.00	1.92	0.1744
Residuals	34	0.03	0.00		

Table 29. ANOVA of Efficiency by Treatment & Experience (Experiment II).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.01	0.01	6.81	0.0122
Experience	1	0.01	0.01	10.96	0.0018
Treatment:Experience	1	0.00	0.00	3.41	0.0713
Residuals	46	0.04	0.00		

Table 30. ANOVA of Efficiency by Treatment & Experience (Experiment All).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.00	0.00	3.59	0.0949
System	1	0.00	0.00	0.28	0.6114
Treatment:System	1	0.00	0.00	0.40	0.5454
Residuals	8	0.01	0.00		

Table 31. ANOVA of Efficiency by Treatment & System (Experiment I).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.00	0.00	2.67	0.1116
System	1	0.00	0.00	1.09	0.3033
Treatment:System	1	0.00	0.00	1.92	0.1749
Residuals	34	0.03	0.00		

Table 32. ANOVA of Efficiency by Treatment & System (Experiment II).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.01	0.01	5.43	0.0243
System	1	0.00	0.00	0.29	0.5916
Treatment:System	1	0.00	0.00	1.80	0.1865
Residuals	46	0.05	0.00		

Table 33. ANOVA of Efficiency by Treatment & System (Experiment All).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.00	0.00	3.68	0.0913
Lab	1	0.00	0.00	0.76	0.4085
Treatment:Lab	1	0.00	0.00	0.15	0.7130
Residuals	8	0.01	0.00		

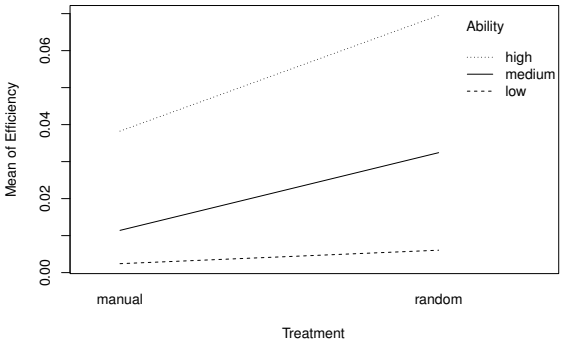
Table 34. ANOVA of Efficiency by Treatment & Lab (Experiment I).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.00	0.00	3.03	0.0908
Lab	1	0.00	0.00	2.45	0.1269
Treatment:Lab	1	0.00	0.00	5.57	0.0241
Residuals	34	0.03	0.00		

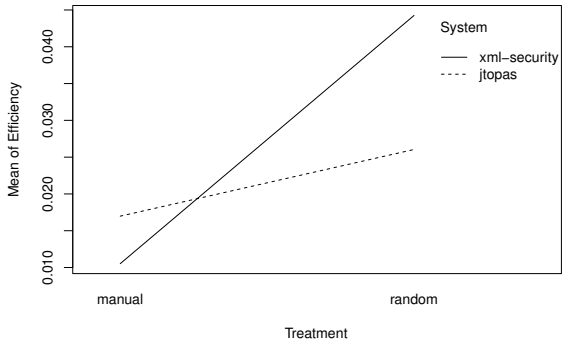
Table 35. ANOVA of Efficiency by Treatment & Lab (Experiment II).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	0.01	0.01	6.03	0.0179
Lab	1	0.00	0.00	2.81	0.1005
Treatment:Lab	1	0.00	0.00	4.67	0.0360
Residuals	46	0.04	0.00		

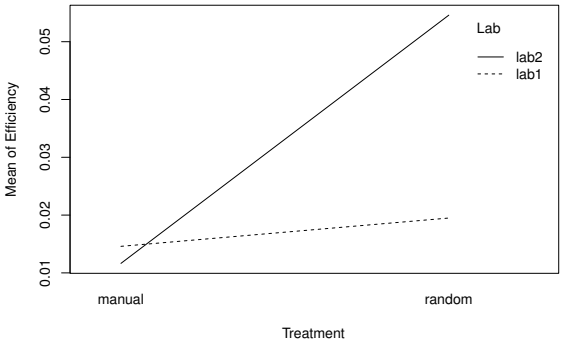
Table 36. ANOVA of Efficiency by Treatment & Lab (Experiment All).



(a) Influence of Ability

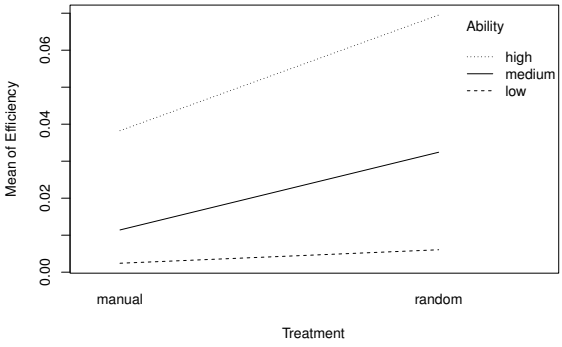


(b) Influence of System

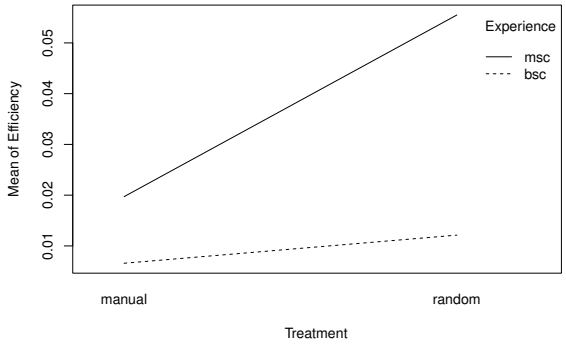


(c) Influence of Lab

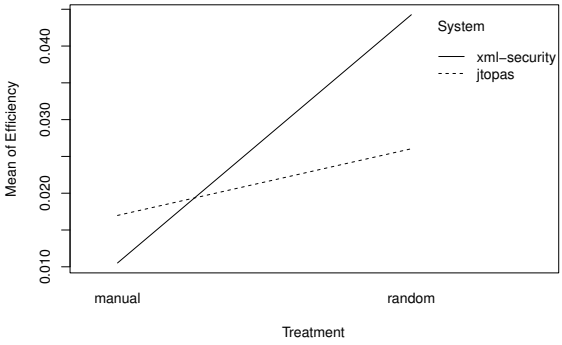
Figure 7. Interatcion plot for fixing efficiency (Experiment I).



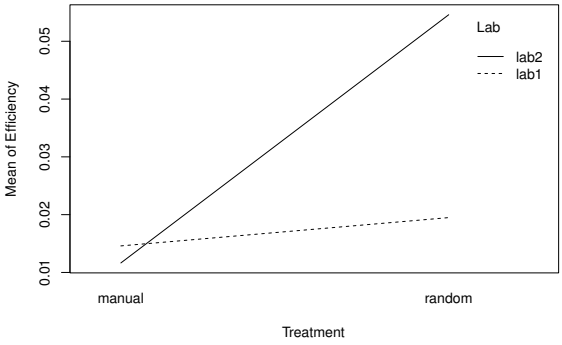
(a) Influence of Ability



(b) Influence of Experience

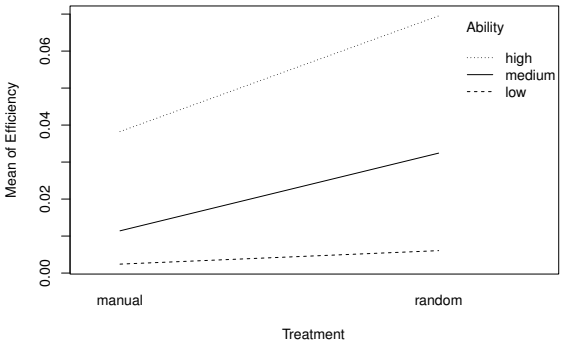


(c) Influence of System

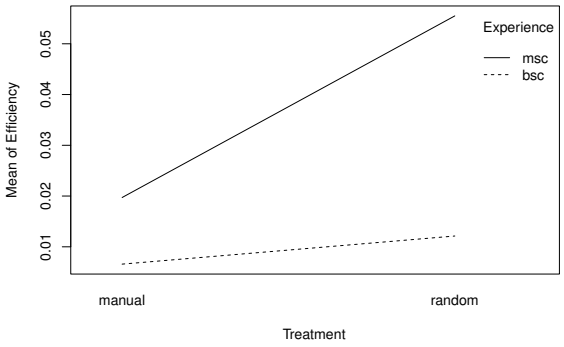


(d) Influence of Lab

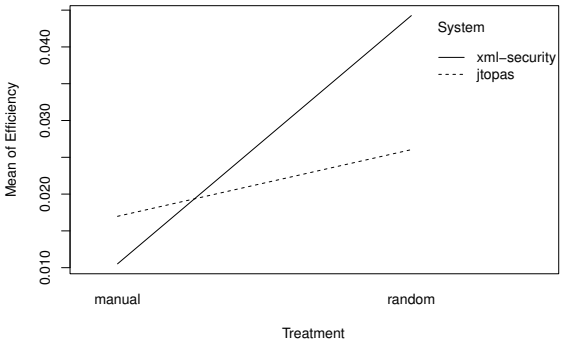
Figure 8. Interatcion plot for fixing efficiency (Experiment II).



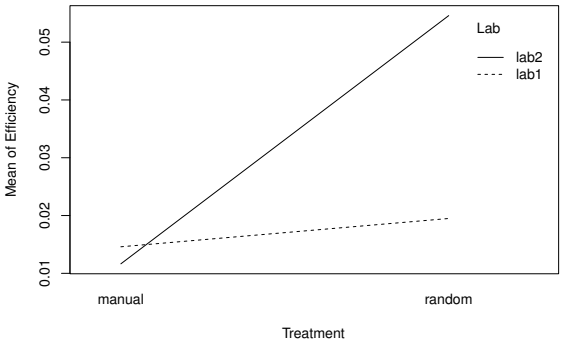
(a) Influence of Ability



(b) Influence of Experience

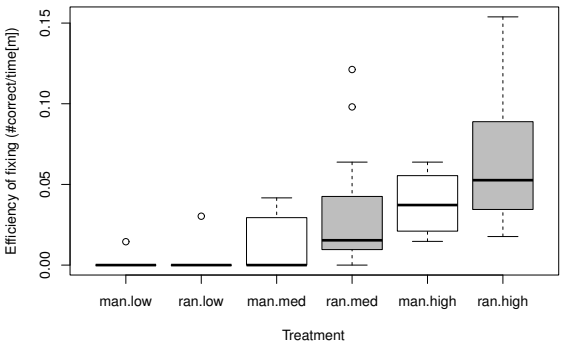


(c) Influence of System

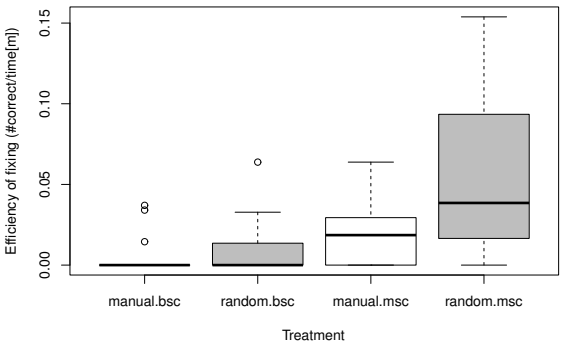


(d) Influence of Lab

Figure 9. Interatcion plot for fixing efficiency (Experiment I and II).



(a) Influence of Ability



(b) Influence of Experience

Figure 10. Boxplot of the effect of Ability and Experience on fixing efficiency.

6 Post questionnaire

question	I.median	I.p.value	II.median	II.p.value	All.median	All.p.value
Q1	0.00	0.81	0.50	0.01	0.00	0.05
Q2	2.00	0.00	1.00	0.00	1.00	0.00
Q3	0.00	0.74	0.00	0.36	0.00	0.52
Q4	1.00	0.03	0.50	0.07	1.00	0.02
Q5	0.50	0.04	0.00	0.98	0.00	0.86
Q6	0.00	0.84	0.00	0.76	0.00	0.85
Q7	-1.00	0.85	1.00	0.00	1.00	0.00
Q8	1.00	0.03	1.00	0.00	1.00	0.00

Table 37. Analysis of post quest Q1-Q8. Mann-Whitney test for the null hypothesis $median(Qx) <= 0$

question	I.median.random	I.median.manual	I.p.value	II.median.random	II.median.manual	II.p.value	All.median.random	All.median.manual	All.p.value
Q4	1	1	0.77	1.00	0.00	0.95	1.00	1	0.92
Q5	0	1	0.47	0.00	0.00	0.21	0.00	0	0.58
Q6	0	0	0.55	0.00	0.00	0.36	0.00	0	0.24
Q7	-1	0	0.12	1.00	1.00	0.89	1.00	1	0.53
Q8	1	1	0.48	1.00	1.00	0.98	1.00	1	0.72
Q9	1	0	0.27	1.00	1.00	0.82	1.00	0	0.46
Q10	0	1	0.14	-1.00	-1.00	0.95	-1.00	-1	0.59
Q11	-1	-1	0.48	1.50	1.00	0.89	0.00	1	0.92
Q12				0.00	0.00	1.00	0.00	0	1.00
Q13				-1.00	-1.00	0.81	-1.00	-1	0.81
Q14				1.00	1.00	0.95	1.00	1	0.95
Q15				0.00	0.00	0.36	0.00	0	0.36
Q16				1.00	0.00	0.27	1.00	0	0.27
Q17				0.00	-1.00	0.67	0.00	-1	0.67

Table 38. Analysis of post quest Q4-Q17. Mann-Whitey test for the null hypothesis $median(Qx_{random}) = median(Qx_{manual})$