Appendix to the paper "Overcoming the Equivalent Mutant Problem: A Systematic Literature Review and a Comparative Experiment of Second Order Mutation"

Lech Madeyski, Wojciech Orzeszyna, Richard Torkar, and Mariusz Józala

Index Terms—mutation testing, equivalent mutant problem, higher order mutation, second order mutation, statistical analyses.

Appendix A
Statistical analyses

This appendix contains statistical analyses. Thanks to a personal discussion with Kitchenham the following important issues related to statistical analyses can be highlighted. First, it is worth mentioning that variables used in statistical tests ought to be independent (even for non-parametric tests), but we can not be sure that this is the case for the statistical tests used for mutant reduction analysis. The question is how to deal with the problem. In this particular situation, we could use a more stringent alpha level, because Monte Carlo simulations have suggested that when the observations in a group are correlated with one another, the nominal alpha level is no longer the Type I error rate (when a positive correlation is introduced within a group the Type I error rate increases). Moreover, in this particular case (i.e. mutant reduction) it would be perfectly fair to say that the *JudyDiffOp* algorithm performed better in terms of total number mutants than the other analysed algorithms without any statistical analyses, which are here only a kind of double-check whether real implementations behave as stated in the description of algorithms. One issue that suggests the problem might not be addressed by a higher alpha level is that the value of a random variable should not be known in advance of measuring it. But for the total number of second order mutants it appears that if one of the algorithms is used, the value for some of the other algorithms may be known. Nonetheless, presented test would be an appropriate method to test any new algorithms that use a different method to construct second order mutants (like *JudyDiffOp*). Last but not least, some of the findings can be seen as dependent but still have merit.

• *Lech Madeyski is with the Institute of Informatics, Wroclaw University of Technology, Poland. E-mail: lech.madeyski <at>pwr.wroc.pl. WWW: http://madeyski.e-informatyka.pl/*

A.1 Statistical analysis of mutants reduction

It appears that the total number of generated mutants was significantly affected by the applied mutation strategy ($\chi^2(4) = 16, p < .001$) as shown in Table 1. The difference between the p-values given by the exact and approximate methods is a cause for concern. As the sample size is small, the exact test procedures (i.e. exact probabilities of obtaining the calculated value of the test statistic or any less likely value) have to be chosen because it is unwise to rely on the χ^2 approximation [1]. The p-value based on the relevant permutation test is $p = .000 < .001$ (while the asymptotic p-value based on the χ^2 distribution is $p = .003$).

TABLE 1
Friedman test statistics for mutants reduction.

Test Statistics	
N	4
χ^2	16.000
df	4
Asymp. Sig.	.003
Exact Sig.	.000

As the overall effect from the Friedman test was significant, post hoc analyses were used to follow-up this finding. In this case, we saw two immediate ways to do non-parametric post hoc procedures: the Wilcoxon approach and the Siegel and Castellan approach [2]. Pett [3] argues that as the Wilcoxon post hoc analyses are affected by small samples, the researchers may want to use the approach by Siegel and Castellan [2].

Siegel and Castellan's procedure [2, p. 180] identifies the minimum required difference between the means of the ranks of any two conditions (designated as $CD_{F_{(\bar{R}_a - \bar{R}_b)}}$) in order for them to differ from one another at the pre-specified level of significance. The difference between the means of the ranks of the different groups are compared to z, adjusted for the number of comparisons performed, and a constant based on the total sample size and the number of conditions, k, ($k = 5$ as there are four SOM

strategies and the FOM strategy in this case) as presented in Equation 1.

$$CD_{F_{(\bar{R}_a - \bar{R}_b)}} = z_{\alpha/k(k-1)}\sqrt{\frac{k(k+1)}{6N}} \quad (1)$$

where $\alpha/k(k-1)$ comes from the fact that we are testing absolute differences and, therefore, use only upper tail probability, $\alpha/2 = .025$, divided by the number of comparisons, $k(k-1)/2 = 10$, between k strategies (conditions).

For our study $\frac{\alpha/2}{k(k-1)/2} = \alpha/k(k-1) = .05/5(5-1) = .0025$, hence $z_{\alpha/k(k-1)} = 2.81$. As a result, our critical difference is, $CD_{F_{(\bar{R}_a - \bar{R}_b)}} = 2.81\sqrt{5 * 6/6 * 4} = 3.14$.

If the difference between mean ranks (presented in Table 2) is bigger than or equal to the critical difference (3.14 in this case), then that difference is significant.

TABLE 2
Friedman test mean ranks for mutants reduction.

	Mean Rank
FOM	5.00
RandomMix	3.00
Last2First	3.00
JudyDiffOp	1.00
NeighPair	3.00

We concluded that the FOM-*JudyDiffOp* comparison ($|R_{FOM} - R_{JudyDiffOp}| = 4$) is the only one that is greater than our critical value (3.14). Hence, we can claim that the intervention (i.e. using a second order mutation strategy called *JudyDiffOp* instead of the first order mutation strategy) significantly reduces the number of mutants. This is consistent with the significance of the initial Friedman test presented in Table 1.

Usually, it is not helpful to have an effect size for a general effect tested by the Friedman test; however, effect sizes for performed comparisons are very informative [4] and can be obtained from the Wilcoxon signed-rank tests according to the following equation $r = \frac{Z}{\sqrt{N}}$ (proposed by Rosenthal [5]) in which Z is the z-score, and N is the number of observations (in each comparison we compared two strategies, each of which were measured on four projects). The effect size r in the comparisons: FOM-*JudyDiffOp*, *RandomMix-JudyDiffOp*, *Last2First-JudyDiffOp*, *NeighPair-JudyDiffOp* are all equal to .65.

However, the reported parametric effect sizes should be treated with some caution, since whatever led us to the use of non-parametric methods would also distort the parametric effect size [6]. Therefore, researchers should consider (if possible) following up statistically significant non-parametric p-values with non-parametric effect sizes, even though the major statistical software programs do not support them [6]. Vargha and Delaney's $\hat{A}_{12}$ statistics is a non-parametric effect size measure recommended by Leech and Onwuegbuzie [6], Arcuri and Briand [4] as well as Grissom and Kim [7].

Leech and Onwuegbuzie argue that $\hat{A}_{12}$ is one of the measures which are the most robust to violations of normality and heterogeneity of variance. From [8] we can get five estimated $\hat{A}_{iu} = [(R_i/n) - 1]/(l-1)$ values (where R_i is the sum of ranks, n is the number of analysed projects i.e. 4, while l is the number of matched treatments i.e. 5):
$\hat{A}_{1u} = ((20/4) - 1)/(5-1) = 1$
$\hat{A}_{2u} = ((12/4) - 1)/(5-1) = .5$
$\hat{A}_{3u} = ((12/4) - 1)/(5-1) = .5$
$\hat{A}_{4u} = ((4/4) - 1)/(5-1) = 0$
$\hat{A}_{5u} = ((12/4) - 1)/(5-1) = .5$

We can then get a value for the point estimation of the average absolute deviation from .5, $AAD = \sum_i |A_{iu} - .5|/l = (.5 + 0 + 0 + .5 + 0)/5 = .2$. $AAD + .5 = .7$ reflects a close to large level of stochastic heterogeneity, since according to the guidelines by Vargha and Delaney [8], a $\hat{A}_{12}$ statistic of .71 indicates a large effect size (see Table 1 [8]).

From [8] we can also get effect sizes in the comparisons: $\hat{A}_{FOM,JudyDiffOp} = \hat{A}_{RandomMix,JudyDiffOp} = \hat{A}_{Last2First,JudyDiffOp} = \hat{A}_{NeighPair,JudyDiffOp} = 1$ reflecting large effect sizes.

The results of the Friedman test indicate that there was a significant difference in the total number of mutants generated using the analysed mutation testing strategies i.e. the FOM and the four SOM strategies ($\chi^2 = 16.00$, $p < .001$). A post hoc analysis (Siegel and Castellan [2]), with adjustment due to 10 comparisons to accommodate increased Type 1 error, indicated that there was a significant decrease in the total number of mutants generated by means of the *JudyDiffOp* SOM strategy compared to the FOM strategy (63.5% reduction on average). No other significant pairwise differences between the analysed mutation testing strategies were obtained.

The effect sizes r and $\hat{A}_{12}$ for the comparisons (FOM-*JudyDiffOp*, *RandomMix-JudyDiffOp*, *Last2First-JudyDiffOp*, *NeighPair-JudyDiffOp*) return consistent results, .65 and 1 respectively, and hence the effect sizes are considered large according to Cohen [9], [10] and Vargha and Delaney [8], which is a substantial finding.

> **Finding**: The second order mutation strategy called *JudyDiffOp* significantly reduced the total number of generated mutants in comparison with the first order mutation. The size of the effect was large and in favour of *JudyDiffOp*.

The magnitude of the observed effect is an indicator of practical importance of *JudyDiffOp* second order mutation technique, which, in turn, comes from the fact that the major computational cost of mutation testing is incurred when running mutants against test cases. Therefore, a way to reduce the cost of mutation testing is to reduce the number of mutants we generate [11].

A.2 Statistical analysis for equivalent mutant reduction

The cross tabulation (Table 3) contains the number of cases that fall into each combination of categories *IsEquivalent* and *IsSOM*.

TABLE 3
IsSOM * *IsEquivalent* Crosstabulation.

			IsEquivalent		
			False	True	Total
IsSOM	False	Count	131	69	200
		Expected Count	159.0	41.0	200.0
	True	Count	664	136	800
		Expected Count	636.0	164.0	800.0
Total		Count	795	205	1000

We observed a lower ratio of equivalent to non-equivalent mutants when the SOM strategy was used. A plausible explanation might be that the number of equivalent mutants is generally lower than the number of non-equivalent mutants. Hence, almost every equivalent mutant will probably be combined with a non-equivalent mutant. According to Polo et al. [12, Table I], such a combination will always produce one second-order non-equivalent mutant.

Pearson's χ^2 test examines the association between two categorical variables (in this case the type of mutation strategy, FOM vs. SOM, and whether the generated mutant was equivalent or not). The assumption for χ^2 is that all expected frequencies should be greater than 5. It should be clear from Table 3 that the smallest expected count is 41 for equivalent mutants (*IsEquivalent=True*) generated by means of FOM (*IsSOM=False*). This value exceeds 5 and so the assumption for χ^2 that all expected frequencies should be greater than 5 has been met. The value of the χ^2 statistic (given in Table 4) is 30.066 and this value is significant ($p < .001$), indicating that the type of mutation strategy had a significant effect on whether a mutant would be equivalent.

TABLE 4
χ^2 test to examine the association between two categorical variables *IsEquivalent* and *IsSOM*.

	Value	**df**	**Exact Sig. (2-sided)**	**Point Probability**
Pearson Chi-Square	30.066	1	.000	
Continuity Correction(a)	29.002	1		
Likelihood Ratio	27.376	1	.000	
Fisher's Exact Test			.000	
Linear-by-Linear Association	30.036	1	.000	.000
N of Valid Cases	1000			

A useful measure of effect size for categorical data is the odds ratio since, according to Rosenthal [13], it is unaffected by the proportions in each cell. The odds ratio is the odds of non-equivalent mutants generated by means of SOM divided by the odds of non-equivalent mutants obtained by means of FOM:

$$odds_{non-equiv.\ in\ SOM} = \frac{\#\ of\ non-equiv.\ mutants\ in\ SOM}{\#\ of\ equiv.\ mutants\ in\ SOM} = \frac{664}{136} = 4.88$$

$$odds_{non-equiv.\ in\ FOM} = \frac{\#\ of\ non-equiv.\ mutants\ in\ FOM}{\#\ of\ equiv.\ mutants\ in\ FOM} = \frac{131}{69} = 1.90$$

$$odds\ ratio = \frac{odds_{non-equiv.\ in\ SOM}}{odds_{non-equiv.\ in\ FOM}} = 2.57$$

There was a significant association between the type of mutation strategy (i.e. first order vs. second order mutation) and whether a mutant would be equivalent ($\chi^2(1) = 30.066$, $p < .001$). This seems to be based on the fact that, due to the odds ratio, the odds of non-equivalent mutants were 2.57 times higher if they were generated by the second order mutation strategy rather than the first order mutation strategy. It is considered a medium effect size, according to Rosenthal [13], which is a substantial finding. It is even more so because handling the equivalent mutants forms an undecidable problem [14], [15].

Finding: The second order mutation significantly reduced the number of equivalent mutants in comparison to the first order mutation. The size of the effect was medium.

A.3 Statistical analysis of the number of live mutants

The number of not killed mutants was significantly affected by the applied mutation strategy ($\chi^2(4) = 14.20$, $p < .001$) as shown in Table 5.

TABLE 5
Friedman test statistics for the number of live mutants.

Test Statistics	
N	4
χ^2	14.200
df	4
Asymp. Sig.	.007
Exact Sig.	.000

As the effect was significant, a post hoc analysis (as suggested by Siegel and Castellan [2]) was used to follow-up on this finding.

For our study, the critical difference was $CD_{F_{(\bar{R}_a - \bar{R}_b)}} = 3.14$. Hence, according to the mean ranks presented in Table 6, we concluded that the FOM-*JudyDiffOp* comparison ($|R_{FOM} - R_{JudyDiffOp}| = 3.75$) is the only one that is greater than our critical value of 3.14. Hence, we can claim that the intervention (i.e. using the second order mutation strategy called *JudyDiffOp* instead of the first order mutation) significantly reduces the number of not killed mutants. This is consistent with the significance of the Friedman test as presented in Table 5.

The effect sizes r in the comparisons (FOM-*JudyDiffOp*, *Last2First-JudyDiffOp*, *NeighPair-JudyDiffOp*) are equal to .65 and hence considered large according to Cohen [9], [10], which is considered to be a substantial finding. The effect size r in the comparison (*RandomMix-JudyDiffOp*) is equal to .39 and, therefore, considered large according to [9], which is a substantial finding as well.

TABLE 6
Friedman test mean ranks for the number of live mutants.

	Mean Rank
FOM	5.00
RandomMix	3.00
Last2First	3.00
JudyDiffOp	1.00
NeighPair	3.00

We also calculated a non-parametric effect size measure as proposed by Vargha and Delaney. From [8] we can get five estimated $\hat{A}_{iu}$ values as follows: $\hat{A}_{1u} = 1$, $\hat{A}_{2u} = .5$, $\hat{A}_{3u} = .5$, $\hat{A}_{4u} = 0$, $\hat{A}_{5u} = .5$

Using these values we can then get a point estimation of $AAD = \sum_i |A_{iu} - .5|/l = 1/5 = .2$. $AAD + .5 = .7$ reflects a close to large level of stochastic heterogeneity [8].

From [8] we can also get effect sizes in the comparisons $\hat{A}_{FOM,JudyDiffOp} = \hat{A}_{Last2First,JudyDiffOp} = \hat{A}_{NeighPair,JudyDiffOp} = 4/4 = 1$ which reflects large effect sizes. $\hat{A}_{RandomMix,JudyDiffOp} = 3/4 = .75$, which also reflects a large effect size.

Finding: The second order mutation strategy, called *JudyDiffOp*, significantly reduced the number of not killed mutants in comparison with the first order mutation. The size of the effect was large.

The magnitude of the observed effect is an indicator of the practical importance which, in turn, comes from the extremely high cost of manual classification of not killed mutants (as equivalent or non-equivalent).

A.4 Statistical analysis of the time of mutation testing process

The number of seconds spent on testing mutants was significantly affected by the applied mutation strategy ($\chi^2(4) = 13.600$, $p = .001$), as shown in Table 7.

TABLE 7
Friedman test statistics for the time of mutation testing.

Test Statistics	
N	4
χ^2	13.600
df	4
Asymp. Sig.	.009
Exact Sig.	.001

According to the mean ranks shown in Table 8, we can conclude that the FOM-*JudyDiffOp* comparison ($|R_{FOM} - R_{JudyDiffOp}| = 4.00$) is the only one that is greater than our critical value of 3.14. Hence, we can conclude that the intervention (i.e. using the SOM strategy called *JudyDiffOp* instead of FOM) significantly reduced the time (measured in the number of seconds) spent on testing mutants. This is also consistent with the results presented in Table 7.

TABLE 8
Friedman test mean ranks for the time of mutation testing.

	Mean Rank
FOM	5.00
RandomMix	3.00
Last2First	3.50
JudyDiffOp	1.00
NeighPair	2.50

Effect size analysis revealed that r equaled .65 and is hence considered large in the comparisons (FOM-*JudyDiffOp*, *RandomMix*-*JudyDiffOp*, *Last2First*-*JudyDiffOp*, *NeighPair*-*JudyDiffOp*).

The five estimated $\hat{A}_{iu}$ values can be obtained as follows: $\hat{A}_{1u} = 1$, $\hat{A}_{2u} = .5$, $\hat{A}_{3u} = .625$, $\hat{A}_{4u} = 0$, $\hat{A}_{5u} = .375$ and we can then get a value for the point estimate of $AAD = \sum_i |A_{iu} - .5|/l = 1.25/5 = .25$. $AAD + .5 = .75$ reflects a close to large level of stochastic heterogeneity [8].

The effect sizes in the comparisons $\hat{A}_{FOM,JudyDiffOp} = \hat{A}_{RandomMix,JudyDiffOp} = \hat{A}_{Last2First,JudyDiffOp} = \hat{A}_{NeighPair,JudyDiffOp} = 1$ reflect large effect sizes [8].

Finding: The second order mutation strategy called *JudyDiffOp* significantly reduced the mutation testing time in comparison with first order mutation. The size of the effect was large.

A.5 Statistical analysis of manual mutants' classification time

On the basis of a large sample of 1,000 (200 FOM and 800 SOM) manually classified mutations (e.g. Schuler and Zeller [16] used a sample of 140 manually classified first order mutations) we are able to evaluate whether the time spent on manual mutant classification was significantly affected by the applied mutation strategy (i.e. FOM and the different SOM strategies) by means of the independent t-test to compare two means (a test for normality revealed that a t-test was suitable to use). We assumed two levels (i.e. FOM and SOM) in the independent variable (i.e. order of mutation). Table 9 presents the descriptive statistics (mean value, standard deviation, and standard error) for FOM and SOM.

TABLE 9
Descriptive statistics of mutants classification time.

Mutation strategy	N	Mean	Std. Dev.	Std. Err.
FOM	200	736.74	331.89	23.47
SOM	800	576.18	296.87	10.50
Total	1000	608.29	310.74	9.83

Levene's test, presented in Table 10, indicates that: *i*) the assumption of homogeneity of variance has not been violated ($p = .30 > .05$) and; *ii*) that using SOM instead of FOM significantly affected the time needed for the manual classification of mutants as equivalent or non-equivalent ($t(998) = 6.68$, $p < .001$).

TABLE 10
Mutants classification time (independent samples test).

	Levene's test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. 2-tailed	Mean Diff.	Std. Err. Diff.	95% CI of the Difference Lower	Upper
Equal variances assumed	1.087	.297	6.677	998	.000	160.559	24.047	113.370	207.748

TABLE 11
Descriptive statistics for mutants classification time for each technique.

Mutation strategy	N	Mean	Std. Dev.	Std. Err.	95% CI for Mean Lower Bound	Upper Bound	Min.	Max.
FOM	200	736.74	331.89	23.47	690.46	783.01	125.00	1600.00
RandomMix	200	574.07	285.60	20.20	534.24	613.89	30.00	1210.00
Last2First	200	574.25	310.47	21.95	530.96	617.54	55.00	1560.00
JudyDiffOp	200	549.34	299.54	21.18	507.57	591.11	10.00	1155.00
NeighPair	200	607.05	290.66	20.55	566.52	647.58	15.00	1300.00
Total	1000	608.29	310.74	9.83	589.01	627.57	10.00	1600.00

TABLE 12
Mutants classification time (test of homogeneity of variances).

Levene Statistic	df1	df2	Sig.
.439	4	995	.781

TABLE 13
Mutants classification time (ANOVA test).

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4460966	4	1115242	12.061	.000
Within Groups	92002141	995	92464		
Total	96463107	999			

We can then use the following equation to calculate effect size:

$$r_{\text{FOM vs. SOM}} = \sqrt{\frac{t^2}{t^2 + df}} = .21 \tag{2}$$

The effect size r in the comparison (SOM vs. FOM) is equal to .21 and, hence, considered medium according to Cohen [9], [10].

Finding: The second-order mutation strategy significantly reduced the time needed for the manual classification of mutants as equivalent or non-equivalent in comparison with the first-order mutation. The size of the effect was medium.

A more detailed analysis may be conducted by means of a one-way analysis of variance (ANOVA) where we assume five levels (i.e. FOM, *RandomMix*, *Last2First*, *JudyDiffOp*, *NeighPair*) of the independent variable (mutation strategy applied).

Table 11 shows descriptive statistics (mean value, standard deviation, and standard error) for each mutation testing strategy.

One of the assumptions of ANOVA is that the variances within experimental conditions are similar. Levene's test, which tests the null hypothesis that the variances of the groups are the same, indicated (as presented in Table 12), that the assumption of homogeneity of variance had not been violated ($F(4, 995) = .44$, $p = .78 > .05$). Table 13 presents a summary of the ANOVA.

The results show that using different mutation testing strategies significantly affected the time needed for the manual classification of mutants as equivalent or non-equivalent ($F(4, 995) = 12.06$, $p < .001$).

Since we have no specific hypotheses about the effect mutation testing strategies might have on mutants classification time, we can look at some post hoc tests to compare all mutation testing strategies with each other. The Holm-Bonferroni correction (Table 14) was used in order to control the Type I error rate.

All of the comparisons between the FOM strategy and higher order mutation strategies (*RandomMix*, *Last2First*, *JudyDiffOp*, *NeighPair*) appeared to be highly significant.

We then calculated the effect size on the basis of two measures of variance, the between-group effect SS_M and the total amount of variance in the data SS_T (values taken from Table 13):

$$r = \sqrt{\frac{SS_M}{SS_T}} = \sqrt{\frac{4460966.066}{96463107.056}} = .22 \tag{3}$$

This constitutes a medium effect; however, it is even more interesting to have effect sizes for the comparisons between FOM and different SOM strategies. Hence, we used the following equation to calculate effect sizes for comparisons (the required t-statistic are presented in Table 15.):

$$r_{contrast} = \sqrt{\frac{t^2}{t^2 + df}} \tag{4}$$

TABLE 14
Mutants classification time (post hoc Holm-Bonferroni correction).

(I) Mut. Strategy	(J) Mut. Strategy	(I-J) Mean Difference	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
FOM	*RandomMix*	162.670(*)	30.408	.000	77.123	248.217
	Last2First	162.485(*)	30.409	.000	76.938	248.032
	JudyDiffOp	187.395(*)	30.408	.000	101.848	272.942
	NeighPair	129.685(*)	30.408	.000	44.138	215.232
RandomMix	FOM	-162.670(*)	30.408	.000	-248.217	-77.123
	Last2First	-.185	30.408	1.000	-85.732	85.362
	JudyDiffOp	24.725	30.408	1.000	-60.822	110.272
	NeighPair	-32.985	30.408	1.000	-118.532	52.562
Last2First	FOM	-162.485(*)	30.408	.000	-248.032	-76.938
	RandomMix	.185	30.408	1.000	-85.362	85.732
	JudyDiffOp	24.910	30.408	1.000	-60.637	110.457
	NeighPair	-32.800	30.408	1.000	-118.347	52.747
JudyDiffOp	FOM	-187.395(*)	30.408	.000	-272.942	-101.848
	RandomMix	-24.725	30.408	1.000	-110.272	60.822
	Last2First	-24.910	30.408	1.000	-110.457	60.637
	NeighPair	-57.710	30.408	.580	-143.257	27.837
NeighPair	FOM	-129.685(*)	30.408	.000	-215.232	-44.138
	RandomMix	32.985	30.408	1.000	-52.562	118.532
	Last2First	32.800	30.408	1.000	-52.747	118.347
	JudyDiffOp	57.710	30.408	.580	-27.837	143.257

* The mean difference is significant at the .05 level.

TABLE 15
Mutants classification time (comparison results).

Simple Contrast		
RandomMix vs. FOM	Contrast Estimate	-162.670
	Hypothesized Value	0
	Diff.(Estimate-Hypothesized)	-162.670
	Std. Error	30.408
	t	-5.350
	Sig.	.000
95% CI for Difference	Lower Bound	-222.341
	Upper Bound	-102.999
Last2First vs. FOM	Contrast Estimate	-162.485
	Hypothesized Value	0
	Diff.(Estimate-Hypothesized)	-162.485
	Std. Error	30.408
	t	-5.343
	Sig.	.000
95% CI for Difference	Lower Bound	-222.156
	Upper Bound	-102.814
JudyDiffOp vs. FOM	Contrast Estimate	-187.395
	Hypothesized Value	0
	Diff.(Estimate-Hypothesized)	-187.395
	Std. Error	30.408
	t	-6.163
	Sig.	.000
95% CI for Difference	Lower Bound	-247.066
	Upper Bound	-127.724
NeighPair vs. FOM	Contrast Estimate	-129.685
	Hypothesized Value	0
	Diff.(Estimate-Hypothesized)	-129.685
	Std. Error	30.408
	t	-4.265
	Sig.	.000
95% CI for Difference	Lower Bound	-189.356
	Upper Bound	-70.014

Reference category is FOM

Hence, r in the comparisons are as follows:

$r_{RandomMix\ vs.\ FOM} = .167$

$r_{Last2First\ vs.\ FOM} = .167$

$r_{JudyDiffOp\ vs.\ FOM} = .192$

$r_{NeighPair\ vs.\ FOM} = .134$

The aforementioned effect sizes are considered small but close to medium (especially in the case of $r_{JudyDiffOp\ vs.\ FOM}$) according to Cohen [9], [10].

Finding: Each of the second-order mutation strategies (i.e. *JudyDiffOp*, *RandomMix*, *Last2First*, *NeighPair*) significantly reduced the time needed for the manual classification of mutants as equivalent or non-equivalent in comparison with the first-order mutation. The size of the effects were considered small to medium.

References

[1] R. Mundry and J. Fisher, "Use of statistical programs for nonparametric tests of small samples often leads to incorrect P values: examples from Animal Behaviour," *Animal Behaviour*, vol. 56, pp. 256–259, 1998.

[2] S. Siegel and J. Castellan, *Nonparametric statistics for the behavioral sciences*, 2nd ed. New York: McGraw-Hill, 1988.

[3] M. A. Pett, *Nonparametric Statistics in Health Care Research: Statistics for Small Samples and Unusual Distributions*. Sage Publications, Inc, 1997.

[4] A. Arcuri and L. Briand, "A practical guide for using statistical tests to assess randomized algorithms in software engineering," in *ACM/IEEE International Conference on Software Engineering (ICSE)*. IEEE, 2011, pp. 1–10.

[5] R. Rosenthal, *Meta-analytic Procedures for Social Research*, 2nd ed. SAGE Publications, 1991.

[6] N. L. Leech and A. J. Onwuegbuzie, "A call for greater use of nonparametric statistics," Annual Meeting of the Mid-South Educational Research Association (Chattanooga, TN, November 6-8, 2002, Tech. Rep., 2002.

[7] R. J. Grissom and J. J. Kim, *Effect Sizes for Research: A Broad Practical Approach*. Indianapolis, Indiana, USA: Psychology Press, 2005.

[8] A. Vargha and H. D. Delaney, "A critique and improvement of the CL common language effect size statistics of McGraw and Wong," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 2, pp. 101–132, 2000.

[9] J. W. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, New York, USA: Lawrence Erlbaum Associates, 1988.
[10] J. W. Cohen, "A power primer," *Psychological Bulletin*, vol. 112, pp. 155–159, 1992.
[11] J. Offutt, A. Lee, G. Rothermel, R. H. Untch, and C. Zapf, "An experimental determination of sufficient mutant operators," *ACM Transactions on Software Engineering Methodology*, vol. 5, no. 2, pp. 99–118, April 1996.
[12] M. Polo, M. Piattini, and I. García-Rodríguez, "Decreasing the cost of mutation testing with second-order mutants," *Softw. Test. Verif. Reliab.*, vol. 19, pp. 111–131, June 2009.
[13] J. A. Rosenthal, "Qualitative descriptors of strength of association and effect size," *Journal of Social Service Research*, vol. 21, no. 4, pp. 37–59, 1996.
[14] T. A. Budd and D. Angluin, "Two notions of correctness and their relation to testing," *Acta Informatica*, vol. 18, pp. 31–45, 1982.
[15] A. J. Offutt and J. Pan, "Detecting equivalent mutants and the feasible path problem," in *Proc. Eleventh Annual Conf. 'Systems Integrity Computer Assurance COMPASS '96 Software Safety. Process Security'*, 1996, pp. 224–236.
[16] D. Schuler and A. Zeller, "Covering and uncovering equivalent mutants," *Software Testing, Verification and Reliability*, pp. n/a–n/a, 2012. [Online]. Available: http://dx.doi.org/10.1002/stvr.1473