Table 1: Reanalysis of the study on tool support for code clones. All values are computed by us.

Approach	Completion Time			Test Case Failures		
	Low	Medium	High	Low	Medium	High
				Mean/Sum	Mean/Sum	Mean/Sum
CCDemon	77.94	338.94	244.38	0 /0	0.69/11	1.31//21
MCIDiff	164.38	361.56	341.88	0.06/1	1.38/22	0.5/8
W	5.5	71	23	0	30	23
p	0.001	0.90	0.038	1	0.820	0.143
t/χ^2	-4.473		-2.115	1	3.667	5.828
df	15		15	1	1	1
p	<0.001		0.051	> 0.05	>0.05	<0.05
ANOVA	Effect of Group: F: 3.663 p: 0.0588					
	Effect of Complexity: F: 14.648 p < 0.001					
	Complexity * Group: F: 0.421 p: 0.6579					

Differing Results: Clone-Based and Interactive Recommendation for Modifying Pasted Code

Summary: The authors developed an approach (CCDemon) to automatically recommend whether and where pasted code should be edited. In a study, they compared their approach to a state-of-the-art baseline, MCIDiff.

- Independent variable: Tool support for working with code clones (2 levels, opertionalized with MCIDiff and CCDemon)
- Tasks: 6 programming tasks in 3 levels of complexity (low, medium, high)
- Dependent variables:
  1. Task completion time [metric scale]
  2. Number of failed tests [nominal scale]
- Null hypotheses:
  1. Tool support does not affect task completion time
  2. Tool support does not affect the number of failed tests
- Results:
  1. Significant difference in completion time for three of the tasks
  2. No significant difference in failed tests for any of the task

In our reanalysis, we summarized the tasks by complexity as indicated in Table 6 of the paper, so Tasks 1 and 2 are summarized to low complexity, Tasks 3 and 4 to medium complexity, and Tasks 5 and 6 to high complexity. In Table 1, we summarize the results of our reanalysis.

For low and high complexity, we find a significant difference in terms of response time in favor of CCDemon. Regarding failed test cases, we could confirm the results of the authors that there is no difference between the two approaches. Furthermore, we found inconsistencies in the analysis of the authors. First, the authors applied a Wilcoxon test, which is not the optimal choice neither for response times nor for failed test cases. For response times, they are interval scaled and for low and high complexity, normally distributed, and variance homogeneity also holds, so that, at least, for low and high complexity, a t test would have been the better choice, because it has more statistical power. For completeness, we have added a t test to Table 1, and the significant difference for the high-complexity task barely holds. Regarding test cases, since these are frequency data, a χ^2 test would have been the more fitting choice. We also added this to the table, and we observe a significant difference for high complexity tasks, but in favor of the control group (i.e., the control group produces fewer test case failures). Thus, we would conclude that the approach of the authors leads to more errors.

Another reason for our deviating in results is that the authors did not correct for multiple testing [?]. Applying an FDR correction to the task-wise analysis of the authors would make all significant differences vanish [?]. For our reanalysis on the aggregated data, however, one difference would still remain. With the FDR correction, the difference for the low complexity task still holds, showing that the approach of the authors makes participants significantly faster for low complexity tasks. Thus, in this case, aggregating the data for this task would make a stronger statement in favor of the approach of the authors.

Going on step further, one could even conduct a two-way ANOVA with complexity as second factor. In this case, the approach would have no significant effect, and instead, the effect of task complexity would be the only significant effect. So, our reanalysis shows how important it is to choose the correct analysis method. We can either show that the tool of the authors has a positive or no effect (results of the Wilcoxon tests and t tests), that the approach has a negative effect (χ^2 test of test case failures), or that the complexity of the tasks seems to be the determining factor in the response time differences (ANOVA). The best conclusion that we can draw here is that the data are inconclusive and that we would need further studies to provide a more definitive answer.