Table 1: Task completion times per task and participant. Highlighted cells indicate analysis we replicated of the authors (dark gray) or of the aggregated analysis (light gray).

Group	Participant	Task1	Task2	Task3
Microbat	P1	5.7	8.1	10
	P2	10	9.8	7.8
	P3	9.7	4.2	7
	P4	12.1	9.5	10.5
	P5	20.4	7.3	13.5
	P6	16	11.4	6.5
	P7	12.2	10.7	11.2
	P8	33.2	22.9	12.6
Whyline	P9	15.5	18	12.1
	P10	10.2	25.5	25.4
	P11	25.5	10.1	19.5
	P12	36.2	32.7	25.1
	P13	35.2	35.1	35.3
	P14	42.3	34.8	13
	P15	27.2	47.4	22.5
	P16	48.6	39.5	43.4
Microbat	Average	14.9	10.5	9.9
Whyline		30.1	30.4	24.5
	W/t(df)	-2.746	-2.94 (14)	-3.790
	p	0.0158	0.003	0.002
Microbat	Average		11.76	
Whyline			28.34	
	W		-4.815	
	p		< 0.001	

We cannot replicate the exact p values.

Same Results: Feedback-Based Debugging

Summary: The authors present their approach to integrate light-weight feedback in the debugging process, and based on this, automatically recommend suspicious execution traces. They implement their approach in the tool Microbat, and compare it to the Whyline tool. They found that participants who used Microbat are significantly faster in debugging [**?**].

- Independent variable: Tool (2 levels, operationalized with Microbat and Whyline)
- Tasks: 3 debugging tasks on three different programs
- Dependent variable: Task completion time (referred to as performance) [metric scale]
- Null hypothesis:
  - No difference in performance between the two groups for none of the tasks
- Results:
  - The authors reject all three null hypotheses, such that Microbat reduces the task completion time for all tasks.

In the reanalysis, we aggregated the data over all three tasks, and we come to the same conclusion as the authors: The Wilcoxon test showed that over all tasks, the participants who used Microbat were significantly faster than the participants who used Whyline. We used the Wilcoxon test (not the t test), since the Shapiro-Wilk test revealed a deviation from normality. The authors did not correct the p level for multiple comparison, but an FDR correction does not lead to different conclusions.