Differing Results: Debugging with Intelligence via Probabilistic Inference

Summary: The authors developed an approach to include probabilistic information during the debugging process. Specifically, they describe the probability of how correct variables and statements are. They compare the approach to a standard debugger [**?**].

- Independent variable: Debugging approach (2 levels, operationalized as the prototype of the authors and a standard debugger (pdb))
- Tasks: Participants should solve a debugging tasks in 4 different programs (lcsubstr, quadratic, fibonaci, mergesort)
- Dependent variables: Task completion time [metric scale]
- Null hypothesis:
  - There is no significant performance difference between the two groups
- Results:
  - The authors rejected the null hypothesis, such that their prototype reduced the debugging time for each task.

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

Person/Task	lcsubstr		quadratic		fibonaci		mergesort	
	pdb	prototype	pdb	prototype	pdb	prototype	pdb	prototype
N1/Y1	11.56	6.37	33.1	16.32	15.03	14.55	10.54	12.5
N2/Y2	30.06	8.2	10.14	11.22	27	16.11	18.13	12.23
N3/Y3	10.3	5.29	21.5	10.58	13.31	11.81	15.04	14.32
N4/Y4	5.8	5.5	19.31	17.76	17.24	18.33	18.33	13.9
N5/Y5	2.85	6.96	10.45	14.4	12.45	14.78	20.78	15.61
N6/Y6	21.11	4.89	12.8	10.33	25.69	15.6	20.5	13.8
N7/Y7	23.11	7.36	27.45	14	34.46	13.28	30.5	11.44
N8/Y8	12.71	8.11	20.45	15.4	18.51	18.44	14.33	14.49
Mean	14.69	6.59	19.40	13.75	20.46	15.36	18.52	13.54
p (original)	0.006*		0.037*		0.049*		0.014*	
t (df)	2.576 (7)		2.176 (7)		2.760 (7)		2.170 (7)	
p (reanalysis)	0.018		0.033		0.061		0.033	
t (df)/p (aggregated)	t = 4.489, df = 31, p ¡ 0.001							

*These p values cannot be replicated (see row p (reanalysis)). The original p values remain significant after FDR correction, but our computed p values do not indicate a significant difference for any task.

We can replicate the results of the authors that adding probabilistic inference to describe the probability of how correct variables and statements are can make debugging faster. However, we found several inconsistencies in the analysis of the authors: First, there was no correction for multiple comparison, which, however, would not change the results. Second, and more importantly, we cannot replicate the p values with the data provided in the paper. We reran the one-sided paired Wilcoxon test with R[1], and we come to different p values (provided in Table 1). With an FDR correction, a significant difference only for the first task remains, which is the task with the highest speed-up. When applying the full pipeline (i.e., check for normality, a t test (data for all tasks are normally distributed), and the FDR correction), no significant difference remains. This highlights once again the importance of (a) applying the correct analysis pipeline and (b) describe it in detail, as it reveals that the debugger might not in general be more efficient for debugging than the standard debugger.

[1] Unfortunately, the authors did not specify the details, so we used the test that comes closest to the authors' p values. Given the hypothesis of the authors, we should have actually used a two-sided version.