Table 1: Task completion time and correctness for static and maintenance tasks. All values are computed by us.

Kind of annotation	Time static	Correct/incorrect static	Time maintenance	Correct/incorrect maintenance
Color	352.86	26 / 18	657.13	76 / 12
Ifdef	554.12	19 / 23	515.99	62 / 22
t test/χ^2	3.93	1.14	−2.49	3.52
p value	0.000	0.285	0.012	0.061

Differing Results: Do Background Colors Improve Program Comprehension in the #ifdef Hell?

We have conducted a study to evaluate the effect of background colors on maintenance in configurable software [**?**]. We shortly summarize the important aspects of the study[1]:

- Independent variable: Variability annotations for configurable software (2 levels, operationalized with background colors and `#ifdefs`)
- Tasks: 6 program comprehension tasks in two categories (plus a warm-up task that was not analyzed)
  - Feature location, referred to as static (2 tasks)
  - Bug location, referred to as maintenance (4 tasks)
- Dependent variables:
  1. Task completion time [metric scale]
  2. Correctness of a provided solution (correct, incorrect) [nominal scale]
- Null hypotheses:

$H_0 static$: The kind of annotation does not affect task completion time for static tasks

$H_0 maintenance$: The kind of annotation does not affect task completion time for maintenance tasks

$H_0 correctness$: The kind of annotation does not affect correctness

- Results:

$H_0 static$: Significant difference for the static tasks regarding task completion time

$H_0 maintenance$: Significant difference for one of the maintenance tasks regarding task completion time

$H_0 correctness$: No significant difference regarding correctness

In Table 1, we summarize the results of the reanalysis. As aggregation function, we applied the mean for the response times and the sum for correctness. We aggregated the 2 static tasks and the 2 maintenance tasks, as both constitute different categories of task with our analysis. We could replicate that background colors lead to shorter task completion times for static tasks ($H_0 static$), and that the kind of annotation does not affect the correctness for any of the tasks ($H_0 correctness$). However, we reject $H_0 maintenance$, indicating an effect of the kind of annotation on the completion time for maintenance tasks. Looking at the direction of the effect, participants with background colors are significantly slower than participants with `#ifdefs`. However, this conclusion would be incorrect, because one particular maintenance task was substantially different than the others, so it skewed the results. In this specific task, participants had to work with a class that was entirely annotated with a red background color, causing visual fatigue and slowing down participants. Thus, with a blind aggregation, we would arrive at the incorrect conclusion that background colors slow down participants for maintenance tasks. We pick up on the discussion of choosing a suitable level of aggregation in Section **??**.

[1] Link to data: `https://www.infosun.fim.uni-passau.de/se/janet/colors/index.php`