Differing Results: An Empirical Study on the Impact of Static Typing on Software Maintainability

Summary: The paper describes an experiment to evaluate how static and dynamic type systems affect the maintainability of software [**?**].

- Independent variables:
  1. Type system (2 levels, operationalized with Java (static) and Groovy (dynamic))
  2. Tasks (9 levels, operationalized with 9 programming tasks of 3 different categories: Class-identification task (CIT, 5 tasks), type-error fixing task (TEFT, 2 tasks), semantic-error fixing task (SEFT, 2 tasks))
- Dependent variable: Task completion time [metric scale]
- Null hypotheses:

$H0_{CIT}$: Static type systems have no influence on development time for the task category *class identification*

$H0_{TEFT}$: Static type systems have no influence on development time task category *type-error fixing*

$H0_{SEFT}$: Static type systems have no influence on the debugging time for the task category *semantic-error fixing*

- Results:

$H0_{CIT}$: Difference in favor of static type system for all but one class-identification task

$H0_{TEFT}$: Difference in favor of static type system for all type-error fixing task

$H0_{SEFT}$: No difference for the semantic-error fixing task

Although the authors defined different task categories, they did not do a per-category analysis, but a per-task analysis (and analysis of all tasks combined). This allowed them to disentangle the effect of specific tasks, but at the cost of statistical power. The authors did not adjust the p level for multiple comparisons; with an FDR correction, we did not conclude a significant difference for CIT2 (p value of 0.033). Next, we aggregated the data per task category, and then compared whether there is a difference within subjects between Round 1 and Round 2 regarding the development time, and also ran an ANOVA. We compare the data and significance of the original analysis and our aggregated analysis in Table 1. Regarding the within-subject comparison, we come to different conclusions than the authors. First, we do not replicate the rejection of $H0_{CIT}$, since the difference is significant for only one of the two groups in favor of the static type system (i.e., the group that started with the dynamic type system). For the group that started with the static type system, the difference is not significant. Different than the authors, who assume a favor for the static type system because of the difference for one group, we find this too liberal regarding the null hypothesis, such that we do not reject $H0_{CIT}$. For the remaining two hypotheses, we come to the same conclusion. Regarding the ANOVA, we come to the same conclusion of the authors, such that we have a significant main effect of the task category and a significant interaction effect of task category and type system in both rounds. We also found the significant main effect of the type system in the second round. To summarize, the aggregated reanalysis itself did not let us draw different conclusions than the authors. The difference regarding rejecting $H0_{CIT}$ is caused by a different interpretation of when the null hypothesis can be rejected or accepted.

Table 1: Average task completion times. The lower part of the table contains the aggregated times per task category. Highlighted cells indicate analysis we replicated of the authors (dark gray) or of the aggregated analysis (light gray).

		CIT1	CIT2	CIT3	CIT4	CIT5	TEFT1	TEFT2	SEFT1	SEFT2
Round 1	Dynamic	822.41	616.35	1073.71	992.18	1004.24	757.88	788.00	1022.35	538.59
	Static	762.06	1194.81	1069.88	1105.88	837.69	303.13	168.25	1613.69	794.44
Round 2	Dynamic	812.44	725.81	1297.94	1062.31	1226.88	1108.88	914.31	591.63	311.94
	Static	321.65	391.35	570.47	564.82	552.00	172.53	126.82	638.82	237.35
Within-Group Comparison (Dynamic first)		V: 153, **p: <0.001**	V: 148, **p: <0.001**	V: 147, **p: <0.001**	V: 149, **p: <0.001**	V: 153, **p: <0.001**	V: 153, **p: <0.001**	V: 153, **p: <0.001**	V: 123, **p: 0.027**	V: 123, **p: 0.027**
Within-Group Comparison (Static first)		V: 70, p: 0.94	V: 109, p: 0.033	V: 44, p:0.231	V: 76, p: 0.706	V: 11, **p: 0.002**	V: 2, **p: <0.001**	V: 1, **p: <0.001**	V: 132, **p: <0.001**	V: 125, **p: 0.002**
Round 1	ANOVA	Type system: F: 0.089, p: 0.768, $\eta^2 = 0.0005$ Task: F: 11.402, p: <0.001, $\eta^2 = 0.190$ Type system x Task: F: 6.851, p: <0.001, $\eta^2 = 0.114$								
Round 2	ANOVA	Type system: F: 42.54, p: <0.001, $\eta^2 = 0.219$ Task: F: 9.91, p: <0.001, $\eta^2 = 0.130$ Type system x Task: F: 6.45, p: <0.001, $\eta^2 = 0.085$								
Round 1	Dynamic			901.78			772.94		780.47	
	Static			994.06			235.69		1204.06	
Round 2	Dynamic			1025.08			1011.59		451.78	
	Static			480.06			149.68		438.09	
Round 1	ANOVA	Type system: F: 0.007, p: 0.934, $\eta^2 = 0.00006$ Task: F: 26.28, p: <0.001, $\eta^2 = 0.234$ Type system x Task: F: 22.66, p: <0.001, $\eta^2 = 0.202$								
Round 2	ANOVA	Type system: F: 27.83, p: <0.001, $\eta^2 = 0.282$ Task: F: 13.57, p: <0.001, $\eta^2 = 0.076$ Type system x Task: F: 27.45, p: <0.001, $\eta^2 = 0.154$								
Within-Group Comparison (Dynamic first)				V: 153, **p: <0.001**			V: 153, **p: <0.001**		V: 134, **p: 0.004**	
Within-Group Comparison (Static first)				V: 57, p: 0.597			V: 0, **p: <0.001**		V: 135, **p: <0.001**	

We could not replicate the exact values of the authors, but they do not deviate considerably.