Table 1: Number of correctly identified defects and gaze area of what participants looked at. All values are computed by us.

	Correctness			Gaze behavior		
Error category	Correct	Incorrect	Timeout	Source	Error	Navigation
Semantic	31	98	35	70.67	19.67	9.33
Dependency	66	10	34	76.5	14.5	8.5
Type mismatch	104	25	1	66.5	24	10
Syntax	50	2	25	65	20	15
Other	70	28	12	73	16	11
χ^2 / F		198.16			4.624	
p value		0.000			1	

Same Results: Do Developers Read Compiler Error Messages?

In the study by Barik and others, developers' gaze behavior was observed with eye tracking [**?**]. The authors evaluated how different categories of errors affect the gaze behavior of participants. We shortly summarize the results[1]:

- Independent variable: Error categories (5 levels, derived from frequent errors of a large set of builds)
- Tasks: 10 tasks to identify a defect based on a compiler error message
- Dependent variables:
  1. Correctness, i.e., whether a provided solution was correct or incorrect [nominal scale]
  2. Gaze behavior, i.e., amount of time participants' gaze was on a certain area (source code, error, navigation) of the IDE used for the study [metric scale]
- Research questions:

  RQ corr.: How effective and efficient are developers at resolving error messages for different categories of errors?

  RQ gaze: Do developers read error messages?
- Results: (mostly verbal description)

  RQ corr.: Solution for correctness is skewed; significantly different solution times for correct and incorrect answers

  RQ gaze: Participants' gaze is most of the time on the source-code area (65 % to 80 %), then the error area (13 % to 25 %)

In Table 1, we summarize the results of the reanalysis. As aggregation function, we applied the sum for correctness and the mean for the gaze behavior. We aggregated all tasks of each category according to Table 1 of the original paper. Then, we were able to conduct a more nuanced analysis. For correctness, the category of error affected the number of correct solutions (*RQ correctness*). This was especially apparent for the semantic category (many incorrect responses) and the type-mismatch category (many correct responses). With a task-wise analysis, this influence of the different categories of error messages did not become clear, but it revealed only that different tasks affected correctness. With our aggregation, we found that semantic error messages may be problematic, but error messages regarding type errors do not seem to pose much of a challenge. Furthermore, we can state that the category of error does not affect the gaze behavior (*RQ gaze*), such that participants spent a large amount of time on the source code. In contrast to the original study, we can narrow the amount of time from 65 % to 80 % down to 65 % to 76.5 %, and from 13 % to 25 % down to 16 % to 24 %, giving a more specific estimation of participants' gaze behavior. Thus, with the aggregation, a more nuanced analysis of the effect of different kinds of errors on programmer behavior was possible.

[1] Link to data: `http://static.barik.net/barik/gazerbeams/`