

# THE IMPACT OF EYENALYZER

L. Grabinger, T. Ezer, F. Hauser, J. Mottok

*OTH Regensburg (GERMANY)*

## Abstract

Empirical research poses numerous challenges for beginners. This is especially true for data analysis – a task that usually requires knowledge from two distinct areas: statistics and programming. To support prospective researchers with that task, we developed a web-based tool called *eyenalyzer*. It supports common activities in the data analysis phase of empirical studies in a way that is suitable for novices in both, statistics and programming. The present article describes a controlled experiment investigating the impact of this tool with a total of 20 participants. All of them are given a set of common data analysis tasks. Half of the participants complete the tasks using *eyenalyzer*, the other half can use anything except for *eyenalyzer*. For each task and participant, we record the time in minutes, the task score, and the perceived difficulty. The results confirm that our tool is a valuable support for novice researchers: With *eyenalyzer*, the participants are significantly faster, achieve higher scores, and perceive the tasks to be less difficult.

Keywords: empirical research, eye tracking, data analysis, tool evaluation

## 1 INTRODUCTION

Empirical research poses numerous challenges for beginners. This is especially true for data analysis – a task that usually requires knowledge from two distinct areas: statistics and programming. To support prospective researchers with data analysis, we developed a web-based tool called *eyenalyzer* – short for *eye tracking data analyzer* based on the programming language `R`. The tool was developed with eye tracking research in mind, but its application is not limited to this. It supports common activities in the data analysis phase of empirical studies (see [1]) in a way that is suitable for novices in both, statistics and programming. For a detailed introduction to the tool, please refer to [2].

This article presents a controlled experiment accessing the impact of *eyenalyzer*. Thereby, it is structured as follows: We first describe the study in detail (i.e., the design, the material, the sample, the measurements, and the analysis), then present the study results, and elaborate on their implications.

## 2 METHODS

This section details the conduction of the experiment. First, we state the design and describe the material. Afterwards, we address sample, procedure, and analysis.

### 2.1 Experimental design

For the study, we follow a *between-subjects* approach. We randomly divide our sample into *two groups*: Half of the participants are only allowed to use *eyenalyzer*, the other half are may use anything other than *eyenalyzer*, i.e., any software system, the internet or even chat bots. Communication with human third parties is prohibited for both groups. Note that, in this article, the two groups are sometimes referred to with the *eyenalyzer* logo (i.e., a lowercase e on a green-grey square) for the first group and the term *others* for the second group.

### 2.2 Instruments

**Priming:** For the study, we use two distinct primings. With the first, we ensure that all participants are familiar with the basic concepts and terms of statistics (e.g., the distinction between the terms independent and dependent variable or the construction of box plots). The second document is only provided to one of the two groups: It outlines the scope, structure, and use of our tool *eyenalyzer* for the participants assigned to working with it.

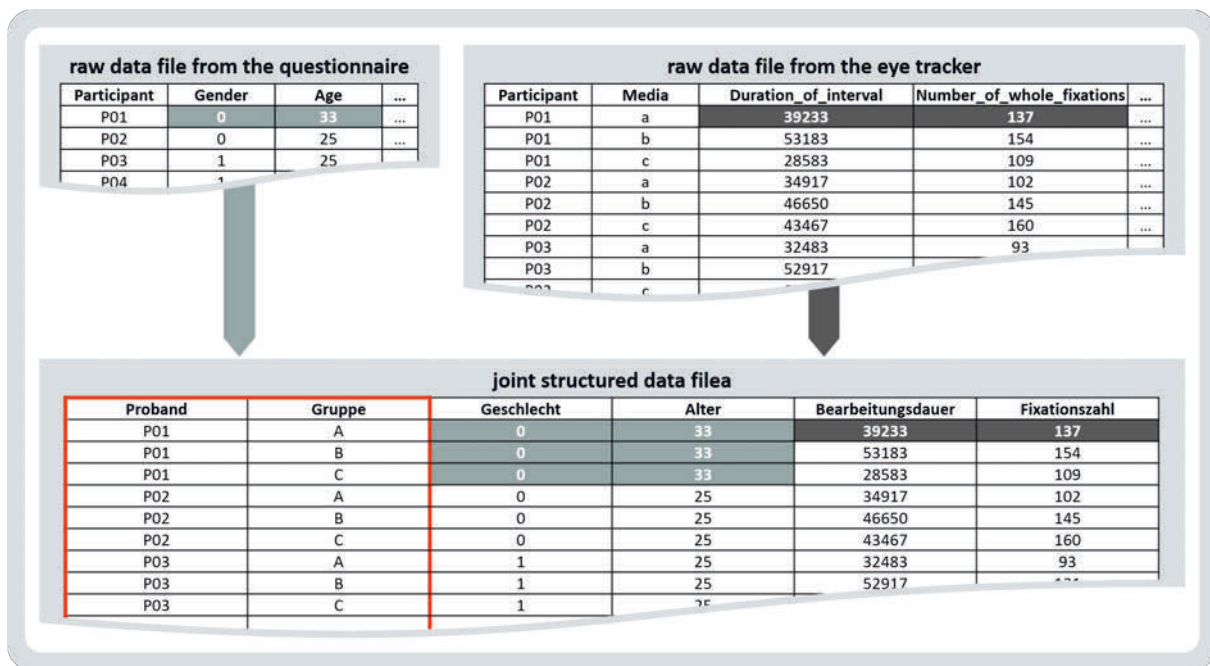


Figure 1. Task of building the joint structured data file.

**Tasks:** During the study, the participants encounter a total of six data analysis tasks – one for each application area or action page of eyenalyzer. All tasks are embedded in a hypothetical study setting that can be summarized as follows: 29 participants each complete three tasks (i.e., A, B, and C); An eye tracker is used to record the time and the eye tracking metric number of fixations for each task and each participant, while a questionnaire gives age and gender of each participant. This setting is based on an earlier study of the authors [3, 4]; The data used is adapted from these works.

The first task (T1) is to create a joint structured data file<sup>1</sup> from the two raw data files of eye tracker and questionnaire; The desired file structure is given by the setting (i.e., within-subjects with 29 participants and three values of the independent variable). The task is outlined in Fig. 1; The columns relevant for the structure are framed in red while the data transfer is indicated by highlighting. As the study was conducted in German, the column names in the target file are also in German; They can be translated as follows: *Proband* = participant, *Gruppe* = group, *Geschlecht* = gender, *Alter* = age, *Bearbeitungsdauer* = time, *Fixationszahl* = fixation count. As second task (T2), the joint structured data file is to be adapted: The values 0 or 1 in the gender column are to be replaced with the German words for female or male, respectively, and all rows belonging to participants older than 50 years are to be removed. The remaining tasks use the adapted version of the structured data file. In the third task (T3), one needs to calculate the means and standard deviations of the fixation counts for each group as well as the frequency with which each combination of gender and age occurs (i.e., the respective number of participants). The fourth task (T4) is to render two plots: A box plot of fixation count by group (on the x-axis) and gender (on the legend) as well as a scatter plot of fixation count over time. In task five (T5), two hypothesis tests for group differences are to be carried out – both times the fixation count is to be compared, once between the groups and once between genders while filtered for group A. The last task (T6) is to get the correlation between fixation counts and time, again for group A only.

**Data files:** The participants are provided with four data files; One each with the raw data from questionnaire (i.e., questionnaire.xlsx) and eye tracker (i.e., eyetracker.tsv), one with the joint structured data file following T1 (i.e., dataRaw.csv), and one version of the latter adapted following T2 (i.e., data.csv). Note that the latter two files do not correspond to the actual results of tasks T1 or T2. To avoid subsequent errors, the participants do not work on the self-created files, but with given data files; Yet, to prevent cheating, we give not the actual resulting data file, but a substitute, in which in each case the data values for two participants are swapped: For dataRaw.csv we exchange the values from group B for participants P25 and P28 while for data.csv it is participants P12 and P14 from group A.

<sup>1</sup> Following [2], a structured data file is "a tabular file with one row per combination of participant and independent variable and one additional column per dependent variable".

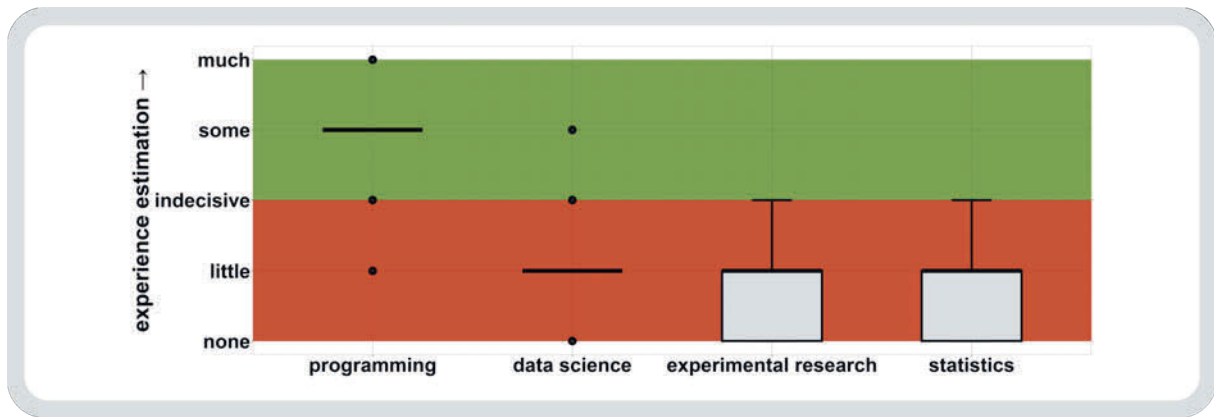


Figure 2. Prior experience of participants.

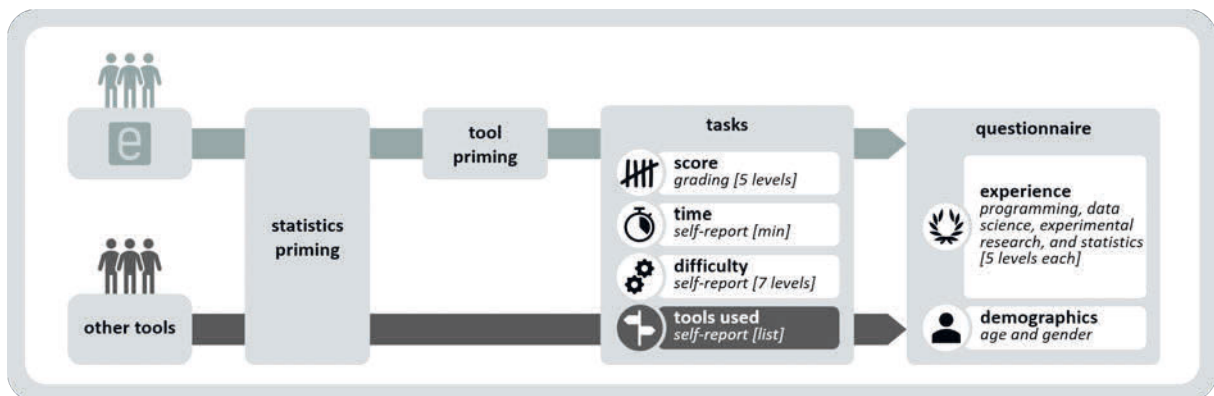


Figure 3. Experimental procedure.

**Questionnaire:** We gather data on participant characteristics (i.e., prior experience and demographics) by an online questionnaire realized in **LimeSurvey** version 3.13.2+180709. For experience, we use a 5-point Likert scale (i.e., *none*, *little*, *indecisive*, *some*, *much*).

## 2.3 Sample

A total of 20 participants took part in this study; Only two of the participants were female, one in each of the two groups. All participants are engineering undergraduates, hence mostly in their early twenties (i.e.,  $M = 22.25$   $SD = 1.68$ ) and equipped with quite good programming experience but practically no prior knowledge of data science, experimental research, or statistics (see Fig. 2). Both age and Prior experience hardly differ between the two groups – appropriate hypothesis tests (following section 2.5) are not significant.

## 2.4 Data collection

Prior to the actual start of the study, all participants sign an informed consent form and receive the material (i.e., the priming relevant to their group, instructions on the tasks, and the necessary data files). Within the next two hours, they proceed in any order; Yet, they were advised to start with the priming. For each task, the participants provide the working time (i.e., start and end time) and rate the difficulty on a 7-point Likert scale (i.e., *very difficult*, *difficult*, *rather difficult*, *indecisive*, *rather easy*, *easy*, *very easy*). Participants who do not work with *eyenalyzer*, but with other tools, are also asked to specify the tools they have used (e.g., Microsoft Excel). Once they either finished all tasks or the time was up (i.e., 120 minutes after receiving the material), they complete the online questionnaire and submit their task results (i.e., `data1.csv` for T1, `data2.csv` for T2, `boxplot.png` as well as `scatterplot.png` for T4, and the filled instruction file for the remaining tasks). The above explanations are summarized in Fig. 3.

## 2.5 Data analysis

**Grading:** While the time, the degree of difficulty, the tools used, and the participant characteristics are specified by the participants themselves, the score is determined by grading. For each task and participant, we assign one of *not completed*, *wrong*, *borderline*, *acceptable*, or *correct*. The answers are *correct* if they fully comply with the assignment<sup>2</sup>, *wrong* if there are content errors<sup>3</sup>, *acceptable* if there are minor discrepancies of the same kind<sup>4</sup>, and *borderline* if there are several of the latter or, for inferential statistics, correct method but incorrect values. If a task consists of multiple sub-tasks (i.e., tasks T3, T4 and T5), those sub-tasks are graded individually before the average of both values is calculated and rounded down to the smaller whole number. Rounding down ensures that, for example, *correct* and *acceptable* still results in *acceptable*, while *correct* and *wrong* results in *borderline*.

**Hypotheses:** With this study, we assess whether the use of eyenalyzer affects the performance on typical data analysis tasks of eye tracking studies. In terms of hypotheses, this is:

**H1** With eyenalyzer, data analysis tasks are completed with better results than with other tools.

**H2** With eyenalyzer, data analysis tasks are completed in less time than with other tools.

**H3** With eyenalyzer, data analysis tasks are perceived to be less difficult than with other tools.

**Testing:** We aim to compare data between two unpaired groups, which means we need an *unpaired t-test* or its non-parametric alternative, the *Mann-Whitney-U test*. The latter is chosen if assumption of normality is violated (i.e., if the *Shapiro-Wilk test* is significant for at least one of the groups). If otherwise, the normality assumption is met, we rely on *Levene's test* for deciding between *Student's* and *Welch's t-test* – we choose the former, if Levene's test is not significant and homogeneity of variances can therefore be assumed. For the Mann-Whitney-U tests, we use the *z-approximation* in the presence of ties, and the *continuity correction* for strongly differing group sample sizes after NA-removal. If we have multiple data points (i.e., when comparing score, time, or difficulty for all tasks combined), we first compute the *median* of these data points<sup>5</sup>. For all hypothesis tests, we use a significance level of  $\alpha=0.05$ . For a detailed description of the test procedures applied please refer to [5].

## 3 RESULTS

In this section, we present the results of comparing score, time, and difficulty between participants working with and without eyenalyzer. Each quantity is visualized separately in Fig. 4 to 6; The hypothesis test results following section 2.5<sup>6</sup> are shown combined in Table 1. To calculate the tests, the scores and difficulty estimates are transformed into numbers (i.e., correct = 4 and very easy = 7). To ease interpretation, the results of the group comparisons are color-coded: Green for significance with at least moderate effect and red vice versa. Note that in the table, some cells have the value *not applicable*. This is because, for one, Levene's test is not required for non-normally distributed data. For the other, T5 was only completed by two participants of the *others*-group; The Shapiro-Wilk test is only applicable for at least three samples in a group.

Note that the ten participants that did not work with eyenalyzer mostly relied on **Microsoft Excel** (i.e., eight participants for T1, seven for T2, seven for T3, five for T4, two for T5, and two for T6). Three participants reported the use of **ChatGPT**, one for tasks T1 to T4, two others for T5, and one of the latter for T6. Two participants checked **Stack Overflow**, one for tasks T2 and T3, the other one for all tasks but T4. **Python** was mentioned by two participants, once for task T1 to T3 and once for T2 and T3 only. Apart of that, individual participants reported **DATAstab** (T4), **ExtendOffice** (T1 to T3), **MATLAB** (T3), **StatistikGuru** (T6), or a not further specified tsv-to-csv-converter (T1).

<sup>2</sup>For tabular data, row and column sorting is not taken into account.

<sup>3</sup>This includes, for example, incorrect mapping in T1, relative instead of absolute frequencies in T3, missing grouping for box plots in T4, or inappropriate hypothesis tests in T5.

<sup>4</sup>This includes, for example, adapted column names in T1, the wrong amount of decimal places, or typos as well as copy-paste-errors such as copying frequencies twice in T3 or missing the lower sign for the p-value in T6.

<sup>5</sup>Note that the *all*-values in the figures are also calculated in this way.

<sup>6</sup>We use Student's t-test for time in T1, Welch's t-test for time in T3, and Mann-Whitney-U tests for the others. For all Mann-Whitney-U tests, we have ties and therefore apply z-approximation. For time in T4, T5, and T6 we have strongly varying samples sizes due to uncompleted tasks (i.e., five measurements in *others* for T4, two measurements in *others* for T5, and three measurements in *others* for T6 with ten measurements each in *eyenalyzer*); Hence, we use continuity correction there.

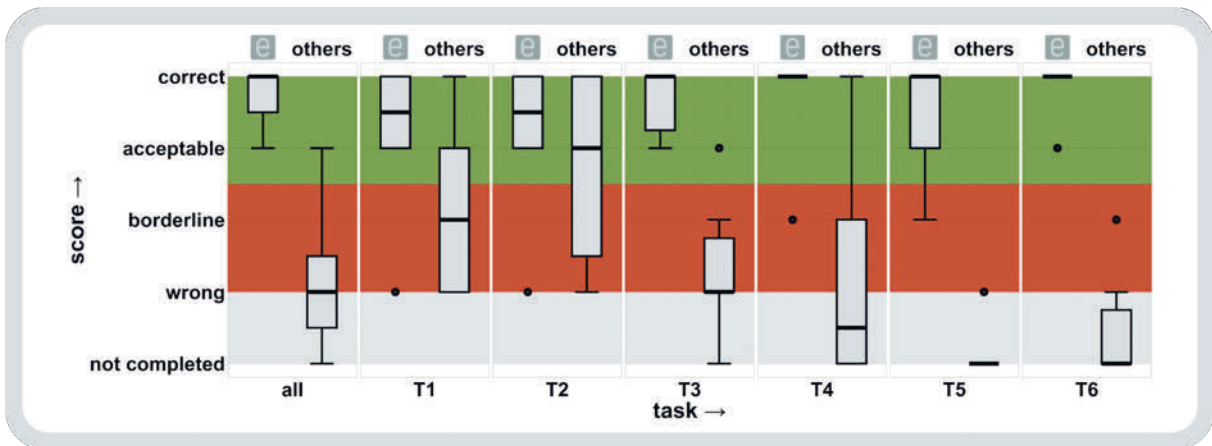


Figure 4. Score per task between groups.

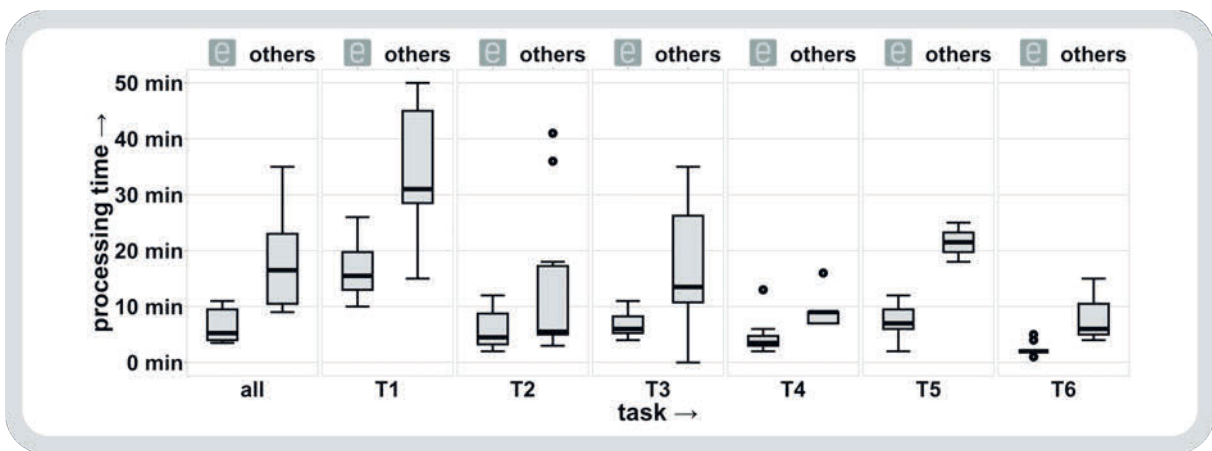


Figure 5. Time per task between groups.

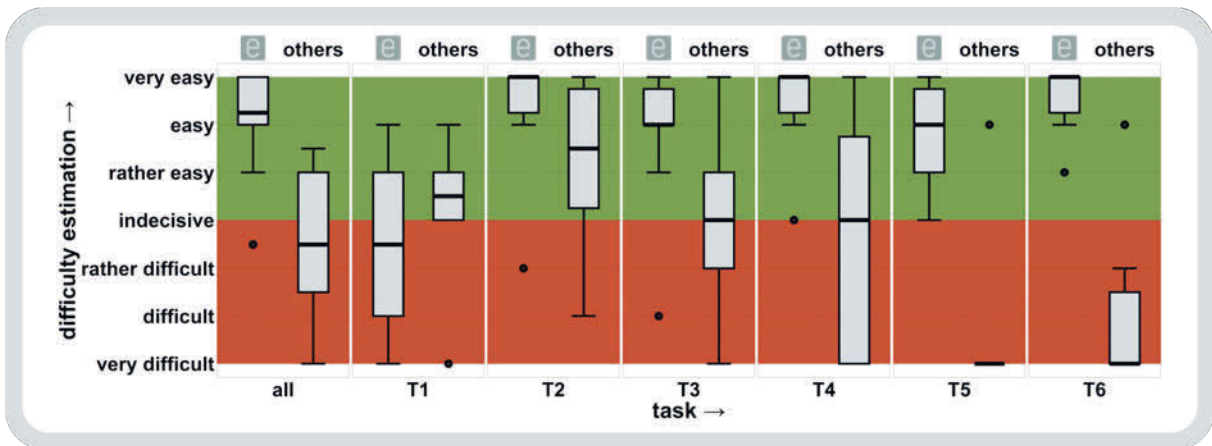


Figure 6. Difficulty per task between groups.

## 4 DISCUSSION

Our findings partially support our hypotheses. The differences in score, time, and difficulty estimation are significant for all tasks combined as well as for the statistical tasks (i.e., T3 to T6). However, the situation is different for T1 and T2. For T2, differences are visible for all three quantities, but do not reach significance. In T1, the participants need significantly less time and achieve significantly higher scores, but tend to rate the task as more difficult. This could be because they are confronted with a process that is unfamiliar to them (i.e., building a structured data file).

Table 1. Differences in performance metrics between groups (cells are marked green for significant differences with at least a moderate effect and red vice versa).

Quantity	Task	Normality (E)	Normality (others)	Homogeneity of variances	Group comparison
score	T1	W = 0.73, p = .002	W = 0.84, p = .046	not applicable	z = 75.00, p = .048, r = .44
	T2	W = 0.73, p = .002	W = 0.77, p = .007	not applicable	z = 59.50, p = .439, r = .17
	T3	W = 0.59, p < .001	W = 0.84, p = .045	not applicable	z = 98.50, p < .001, r = .85
	T4	W = 0.37, p < .001	W = 0.78, p = .007	not applicable	z = 88.00, p = .001, r = .71
	T5	W = 0.72, p = .001	W = 0.51, p < .001	not applicable	z = 100.00, p < .001, r = .89
	T6	W = 0.51, p < .001	W = 0.65, p < .001	not applicable	z = 100.00, p < .001, r = .89
	all	W = 0.65, p < .001	W = 0.93, p = .418	not applicable	z = 99.50, p < .001, r = .86
time	T1	W = 0.94, p = .513	W = 0.89, p = .170	L(1, 18) = 3.59, p = .074	t(18) = 4.07, p = .001, g = 1.74
	T2	W = 0.84, p = .046	W = 0.75, p = .003	not applicable	z = 26.50, p = .071, r = .40
	T3	W = 0.88, p = .123	W = 0.89, p = .164	L(1, 18) = 5.47, p = .031	t(9.78) = 2.91, p = .016, g = 1.25
	T4	W = 0.69, p = .001	W = 0.76, p = .035	not applicable	z = 4.00, p = .011, r = .67
	T5	W = 0.96, p = .786	not applicable	not applicable	z = 0.00, p = .039, r = .63
	T6	W = 0.76, p = .005	W = 0.88, p = .328	not applicable	z = 1.50, p = .020, r = .67
	all	W = 0.84, p = .041	W = 0.91, p = .274	not applicable	z = 7.00, p = .001, r = .73
difficulty	T1	W = 0.90, p = .229	W = 0.79, p = .012	not applicable	z = 37.50, p = .329, r = .22
	T2	W = 0.59, p < .001	W = 0.86, p = .076	not applicable	z = 69.50, p = .114, r = .35
	T3	W = 0.71, p = .001	W = 0.98, p = .982	not applicable	z = 79.50, p = .023, r = .51
	T4	W = 0.60, p < .001	W = 0.83, p = .036	not applicable	z = 87.50, p = .003, r = .66
	T5	W = 0.89, p = .191	W = 0.51, p < .001	not applicable	z = 89.00, p = .002, r = .69
	T6	W = 0.63, p < .001	W = 0.62, p < .001	not applicable	z = 95.00, p < .001, r = .80
	all	W = 0.82, p = .024	W = 0.95, p = .712	not applicable	z = 92.50, p = .001, r = .72

## 5 CONCLUSION

The previous pages described the design, conduction, and results of a controlled experiment accessing the impact of a data analysis tool called eyenalyzer. To conclude this article, we briefly highlight the contribution, address limitations and outline future work in this area.

### 5.1 Contribution

With this article and the controlled experiment presented in it, we demonstrate the effect that eyenalyzer has on performance in common data analysis tasks. We show that eyenalyzer leads to faster, better and easier work. In other words, we prove the impact of eyenalyzer and thus the need for the tool, in the words of Trey Smith [6]:

*"If your presence doesn't make an impact,  
your absence won't make a difference."*

This article also highlights the capabilities of eyenalyzer: All analyses presented in Fig. 4 to 6 and Table 1 can basically be performed with eyenalyzer. Note that we did not use eyenalyzer for the figures, as we wanted to color the background to ease interpretation and add an in-place legend at the top; We also calculated the group comparison of time for T5 by hand, as eyenalyzer does not calculate group comparisons for less than three samples per group, as this does not allow for Shapiro-Wilk tests and thus a well-founded analysis.

### 5.2 Limitations

It is important to bear in mind that the sample of the controlled experiment presented is relatively small (i.e., ten participants in each of the groups) and homogeneous (i.e., all engineering undergraduates, only two of whom are female). The former means that we may be accessing individual effects; Yet, we still obtain significant results. The latter can also be invalidated, as such sample composition reflects the actualities of software engineering, the main research area of the authors' laboratory.

### 5.3 Future work

With regard to the limitations, we provide replication material at **Zenodo** and invite others to replicate the experiment. As a further part of future work, we want to test eyenalyzer in real-world use rather than in an artificial environment and investigate its usability and learnability in detail.

## ACKNOWLEDGEMENTS

The paper is supported by the 'German Federal Ministry of Education and Research' (BMBF) within the funding project HASKI (FKZ: 16DHBKI035).

## REFERENCES

- [1] L. Grabinger, F. Hauser, C. Wolff, and J. Mottok, "On eye tracking in software engineering," *SN COMPUT. SCI.*, vol. 5, pp. 1–20, 7 2024.
- [2] L. Grabinger and J. Mottok, "Statistical analysis of eye movement data for beginners," in *Proceedings of Mensch und Computer 2024*, MuC '24, (New York, NY, USA), ACM, 2024.
- [3] L. Grabinger, F. Hauser, and J. Mottok, "Assessing the presentation of causal graphs and an application of gestalt principles with eye tracking," in *Proceedings of the 29th IEEE International Conference on Software Analysis, Evolution and Reengineering*, SANER '22, (New York, NY, USA), pp. 1267–1274, IEEE, 2022.
- [4] L. Grabinger, F. Hauser, and J. Mottok, "On the perception of graph layouts," *J. Softw.*, vol. 36, pp. 1–18, 5 2024.
- [5] L. Grabinger and J. Mottok, "On selecting hypothesis tests for group differences," in *Proceedings of the 17th annual International Conference of Education, Research and Innovation (ICERI '24)*, (Valencia, Spain), pp. 1–11, IATED, 2024.
- [6] Goodreads, "Quote by Trey Smith." Retrieved from <https://www.goodreads.com/quotes/1136019-if-your-presence-doesn-t-make-an-impact-your-absence-won-t>, 2024.