Evaluation of the Comprehensiveness of Bar Charts with and without Stacking Functionality using Eye-Tracking

Falk Böschen
Kiel University
fboe@informatik.uni-kiel.de

Benjamin Strobel
IPN - Leibniz Institute for Science and Mathematics Education
strobel@ipn.uni-kiel.de

Steffen Goos
Kiel University
stg@informatik.uni-kiel.de

Christoph Liebers
Kiel University
clie@informatik.uni-kiel.de

Bastian Rathje
Kiel University
bar@informatik.uni-kiel.de

Ansgar Scherp
Kiel University and
ZBW - Leibniz Information Centre for Economics
a.scherp@zbw.eu

ABSTRACT

Bar charts are widely used to visualize core results of experiments in research papers or display statistics in news, media, and other reports. However, visualizations like bar charts are mostly manually designed, static presentations of data without the option of adaption to a user's needs. But so far, it is unknown whether interactivity improves the understanding of charts. In this work, we compare static with dynamic bar charts, which offer an interactive stacking option. We assess the efficiency, effectiveness, and satisfaction when answering questions regarding the content of a bar chart. An eye-tracker is used to measure the efficiency. We have conducted a between group experiment with 38 participants. While one group had to solve the aggregation tasks using stackable, i. e., interactive bar charts, the other group was limited to static visualizations. Even though new interactive features require familiarization, we found that the stacking feature significantly helps completing the tasks with respect to efficiency, effectiveness, and satisfaction for bar charts of varying complexity.

Keywords

Interactive Bar Charts; Stacking; User Study; Eye-Tracking

1. INTRODUCTION

Bar charts are commonly used to visualize data in research papers or printed material. With few exceptions, they are static and convey a predefined message. However, the aggregation of multiple values is often necessary for common comparison and aggregation tasks (e. g., to sum up the costs over certain period of time or to compare two periods). In these tasks, it would be helpful to provide the users a possibility to alter the chart by mouse interaction. In our experiment, we introduce the option to interactively aggregate bars as stacks to help users in tasks that require aggregation of data. We use an eye-tracker to measure the efficiency, which is reflected in the number of saccades between areas of interests (e. g., bars and labels). In general, saccades are fast movements of the eye between two fixations. In our analysis, we are specifically interested in eye movements between areas of interests on the bar chart.

We hypothesize that solving aggregation tasks is faster, more precise, and more satisfying when using interactive bar charts than using static charts, regardless of data complexity. Thus, we assess different data complexities in our experiment to examine if interactive bar charts are always an improvement. To test our hypothesis, we have conducted a between-group experiment with 38 participants. Each participant solved the same 10 tasks in random order. We analyze the eye movements as well as the completion time to assess the efficiency. We designed the tasks in such a way that the answer to be provided by the participants is entered as numerical value. Thus, we can compute the effectiveness as the relative deviation of the entered value from the exact value. After the experiment, the participants filled a questionnaire which is used to gain insight about the user satisfaction and validate the experimental setup.

The results of our experiment support our hypothesis. All observed differences are significant, except for the effectiveness in tasks with complex charts. In the following, we briefly present the related research. Subsequently, we describe the apparatus, procedure, tasks, and participants. We present the results and interpret them, before we conclude.

2. RELATED WORK

Most similar to our research is the work by Abell et al. [1], who also investigated some interaction feature on stacked bar charts. In their work, the users could adapt which data row (i. e., bar) of the stacked bar chart is plotted on the x-axis. This allows to interactively align a user selected data row with the base line of the chart to enable easier comparison. Abell et al. conducted a small experiment with 10 participants and were able to identify improvements for three out of five task types (e. g., find min/max, compare sums,

CHIIR '17, March 7–11, 2017, Oslo, Norway.

ISBN 978-1-4503-4677-1/17/03. . . $15.00

DOI: http://dx.doi.org/10.1145/3020165.3022147

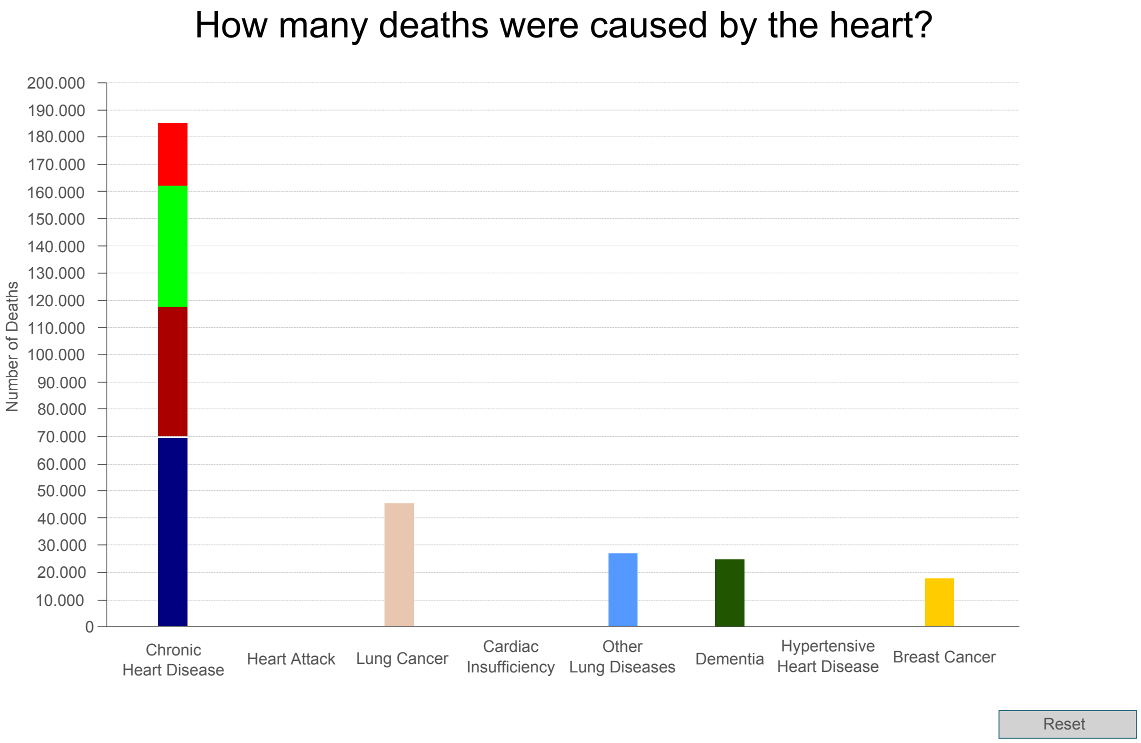

Figure 1: A chart where the bars no. 2, 4, and 7 are stacked onto bar no. 1.

identify trends) when using the interactive chart. The improvement was measured by assessing the number of correct, incorrect, and not-given answers (option "I don't know") per scenario. Furthermore, they investigated the number of interactions in the interactive case as well as the time needed by the participants until they provided the answer. However, they did not perform any significance tests on their results, perhaps due to the small size of 10 participants in their experiment, and they did not use an eye-tracker in their evaluation. Burch et al. [2] investigated the users' behavior by exploring different layouts for hierarchical structures like tree maps. They compared three types: traditional, orthogonal, and radial node-link diagrams. Eye-tracking was used to detect the visual exploration behaviors. They found out that traditional and orthogonal layouts have a significantly better performance. Kim and Lombardino [5] analyzed the influence of the complexity of bar charts on user's comprehension compared with plain text using eye-tracking data. They discovered that providing answers to complex questions using charts is significantly more efficient than text. However, when the complexity of the chart increases, this advantage becomes less apparent. Unlike our work, their experiment did not contain any interactive features.

3. EXPERIMENT

We describe the apparatus, tasks, procedure, and measures. Finally, we report descriptive statistics about our participants.

3.1 Apparatus

The experiment software ran on a notebook with Microsoft Windows 10 to which the Tobii eye-tracker was connected via USB 3.0. The eye-tracker runs with a frequency of 60 hertz. The monitor is a 24" screen with a resolution of 1920x1200 pixels. We used two identical eye-trackers in parallel, distributed among two adjacent rooms with comparable environmental conditions.

3.2 Tasks

We defined 10 aggregation tasks which consist of a bar chart and a question that the participant has to answer by looking at the chart. Each task either requires to report the sum of multiple bars or the difference between two sums of bars. We decided to use real world data and not artificially created data to avoid a potential bias induced by artificial tasks. The data/charts on which each task is based are taken from different German websites that provide statistical data (e. g., Destatis[1]). Five of the tasks (A1) present simple bar charts, while the other five tasks (A2) show complex bar charts. The simple bar charts have two dimensions and eight bars. The complex bar charts have a third dimension, which is defined by additional data rows, and 15 bars. All questions require a numerical value as answer, e. g., the number of deaths caused by a heart illness (see Figure 1).

3.3 Procedure

Before the experiment started, the procedure of the experiment was explained to the participants and they signed an informed consent form. Each participant was pseudo-randomly assigned to either the experimental group (interactive) or control group (static) to assure balanced groups. Afterwards the eye-tracker was calibrated for each partici-

[1] https://www.destatis.de/EN/Homepage.htmls

pant. Then a training task followed so that the participants of both groups could get familiar with the type of questions in the experiment and how to answer them. In addition, participants in the experimental group had time to get familiar with the interactive stacking feature.

Each task consists of three screens. First, the task is displayed. After looking at a synchronization point (red dot) at the top of the screen for two seconds the next screen is loaded. The synchronization point unifies the starting point on the bar chart screen for all participants and thus avoids the so-called center bias, observed in earlier eye-tracking experiments [7]. The second screen shows the bar chart and the participants have to analyze the chart to find the answer to the question. Participants in the experimental group are able to stack bars via drag-and-drop as depicted in Figure 1. The figure shows a user-modified bar chart where the second, fourth, and seventh bar are stacked onto the first bar. Once the answer is found, the participants must press the space bar to access the third screen, where they enter their answer. It is important to note that there are no tool-tips displaying the exact values when hovering over a bar. Thus, the participants have to estimate the value(s) by looking at the charts. After completing a training task, the main experiment started. Each participant completed the ten tasks (5×A1 and 5×A2), one at a time, in random order. Subsequently, the participants filled a questionnaire to collect feedback regarding their perception of the experiment.

3.4 Measures

We measured the effectiveness by calculating the deviation of the answers, provided by the participants, from the ground truth. Since all answers are numerical, but of different scale, we compute the deviation as standard percentage.

We computed two measures for the efficiency: First, we computed the task completion time which is the time the participant spent to solve a task. The second efficiency measure is computed using eye-tracking information, which can be more accurate because it is possible to detect when participants were distracted from the screen. Furthermore, it allows for an in depth analysis of the steps each participant took to solve a task. We decided to assess the eye-tracking information on an area-of-interest (AOI) based approach. Each bar, as well as the axes and labels of the coordinate systems and the question, had its own rectangular AOI. Thus, in the interactive setting we had to keep track of the AOIs that change their coordinates due to the stacking of the bars. The detected fixations on the AOIs where entered in a transition matrix [4] to calculate how efficient a participant was at solving a task. A participant was more efficient if he had less transitions (eye movements between AOIs) per task.

Information about the satisfaction of the participants was collected with a questionnaire based on IsoMetrics [3]. We defined three blocks of five questions. The blocks assess the satisfaction with the visualizations, the tasks, and the interactivity. The questions regarding the interactivity are only answered by participants of the interactive group. All questions are based on an ordinal 5-Point-Likert scale from strongly disagree (––) to strongly agree (++). We encoded the scale from 1 (––) to 5 (++) to evaluate the results. In addition, the participants provided information about their age, job, gender, corrective lenses or glasses, and some general feedback.

3.5 Participant Statistics

In total, we had 38 participants in our experiment (9 female). The participants were equally distributed over the two conditions of the between group design (i.e. interactive group and static control group). In terms of visual aids, 12 participants wore glasses and 4 had contact lenses. The average age of the participants of the control group is about 24.84 years and the average age of the participants of the experimental group is about 26.16 years. The participants were recruited in the area of Kiel University with most of them being students (33).

4. RESULTS AND DISCUSSION

Effectiveness: Table 1 shows the average percentage answer deviation (M) and standard deviation (SD) per task for both groups. The answers to the simple tasks in the experimental group deviate on average only about 1.66% from the correct value, while the answers in the control group deviate about 6.67%. Regarding the answers to the complex tasks, the answers of the experimental group differ about 34.51% from the correct value while the answers of the control group differ about 90.62%. Thus, we can see a difference between simple and complex tasks and experimental and control group.

Table 1: Average deviation of the participants' answers per task as percentages

	Interactive		Static	
Task	**M**	**SD**	**M**	**SD**
A1.1	0.79	0.77	8.50	18.63
A1.2	0.13	0.56	3.98	12.73
A1.3	0.23	0.11	3.24	8.18
A1.4	6.39	20.84	9.11	16.10
A1.5	0.74	1.86	8.53	16.53
A1	1.66	9.47	6.67	14.79
A2.1	14.47	17.31	96.05	215.10
A2.2	17.63	48.63	48.71	91.45
A2.3	61.51	116.80	63.16	96.93
A2.4	71.05	80.09	218.86	365.39
A2.5	7.89	25.07	26.32	38.62
A2	34.51	71.91	90.62	206.85

In order to evaluate the effectiveness, significance tests have been conducted. First, we checked for normal distribution using Shapiro-Wilk, which showed that the data is not normally distributed (p<.001). Thus, we used the non-parametric one-sided Wilcoxon Signed Rank Test to see whether there is a significant difference in answer deviation between the test and control group for A1 and A2. We choose $\alpha = .05$. The tasks using simple bar charts showed a significant difference (W=3,456, p<.003) and the tasks using complex bar charts differed significantly as well (W=3,722, p<.02). We also assessed the significance over both task complexities together and the results also differed significantly (W=14,984, p<0.002).

Efficiency: We calculate the efficiency based on the task duration as well as eye-tracking information. Table 2 shows the average task duration for the test and control group on A1, the tasks with simple bar charts, and A2, the tasks with complex bar charts.

Table 2: Average efficiency measure for the test and control group split up by level of complexity (A1 vs. A2) and experiment conditions (interactive vs. static)

		Task Duration		Transitions	
Subset	**n**	**M**	**SD**	**M**	**SD**
Int-A1	95	22.12	9.05	13.31	6.65
Stat-A1	95	39.68	25.52	17.20	10.07
Int-A2	95	43.63	16.43	41.76	19.06
Stat-A2	95	58.60	26.74	53.22	26.64

We can see that the average task duration for the tasks with simple bar charts in the experimental group (M=22.12, SD=9.05) is lower than the average in the control group (M=39.68, SD=25.52). For the tasks with complex bar charts, the average of the experimental group (M=44.63, SD=16.43) is also lower than the average of the control group (M=58.6, SD=26.74). We first conducted a Shapiro-Wilk test, which resulted in detecting a non-normal distribution (p<.001). Thus, we used a one-sided Wilcoxon Signed Rank Test to compute the significance of the difference. The results for the tasks with simple bar charts (W=2,409, p<.001) as well as the results for the tasks with complex bar charts (W=2,890, p<.001) are significant. We conclude that the experimental group is significantly more efficient than the control group based on the task duration.

The assessment of the efficiency using eye-tracking information shows similar results (see Table 2). Here, the average for the experimental group is lower than the average of the control group and for both levels of task complexity. Due to the fact that the data is again not following a normal distribution, a Wilcoxon Signed Ranks Test was used to determine the significance. The results are significant for the tasks with simple bar charts (W=3,540, p<.005) as well as for the tasks with complex bar charts (W=3,314, p<.001). Thus, we can conclude that the interactive approach is more efficient, regardless of the task complexity.

Satisfaction: Finally, we analyzed the satisfaction based on the answer given by the participants in the questionnaire at the end of the experiment. The answers differed on average only marginally between the experimental group and the control group. The only major difference was in the perception of the task clarity, where the experimental group gave an average rating of about 2.7, while the control groups' rating was on average 3.5. The expected differences with respect to task complexity and problems when solving the tasks could not be observed. However, the opportunity to interact with the bars in terms of stacking was highly rated (M=4.84) and the interaction feature was assessed as intuitive (M=3.79). Furthermore, the participants of the experimental group stated that the tasks would have been too difficult to solve without the interaction (M=3.79).

Discussion: The interactive stacking feature tends to increase the efficiency and effectiveness in task performance regardless of the task complexity. From a cognitive perspective, we assume that the stacking features facilitates cognitive processes that are more difficult and error-prone in static displays such as mental movement of bar elements to estimate the stacked height or the extraction of multiple values and subsequent computation of sums [6]. After stacking multiple bars, participants can rely on perceptual processes to determine the total height rather than depending on computations that are limited by the working memory capacity of the human mind.

Our results further confirm the findings of Kim and Lombardino (see Related Work section) that the efficiency decreases with increasing complexity. This can be observed for the static as well as the interactive condition. The results regarding effectiveness of task A2.5 suggest that the definition of our task complexity should be discussed. Complexity depends on many factors like number of dimensions, label length, or number of digits. We only considered the first for our complexity definition. However, the other two factors seem to be important for the required memory load as well. Thus, they should be investigated in future experiments. Finally, a more detailed assessment of the recorded eye-movements could provide more insights on how the improvement in effectiveness and efficiency was achieved.

5. CONCLUSION

We have reported the results of our eye-tracking experiment with 38 participants to evaluate whether the possibility to modify a bar chart improves its comprehensibility. Our results support our hypothesis that the use of the stacking feature significantly improves the efficiency and effectiveness, independent of the complexity of the bar chart. We observed that the use of the stacking feature is perceived as a benefit, although the task complexity perception did not differ from the one reported by the control group.

Acknowledgement. This research was co-financed by the EU H2020 project MOVING (http://www.moving-project.eu/) under contract no 693092. We thank all participants of our experiment for their input and time.

6. REFERENCES

[1] W. Abell, C. Churcher, and J. Lee. An evaluation of interactive stacked bar charts. *IJCAT*, 34(4):285–290, 2009.

[2] M. Burch, N. Konevtsova, J. Heinrich, M. Hoeferlin, and D. Weiskopf. Evaluation of traditional, orthogonal, and radial tree diagrams by an eye tracking study. *IEEE TVCG*, 17(12):2440–2448, 2011.

[3] G. Gediga, K. Hamborg, and I. Düntsch. The isometrics usability inventory: An operationalization of ISO 9241-10 supporting summative and formative evaluation of software systems. *Behaviour & IT*, 18(3):151–164, 1999.

[4] J. H. Goldberg and X. P. Kotval. Computer interface evaluation using eye movements: methods and constructs. In *International Journal of Industrial Ergonomics*, volume 24, pages 631–645, 1999.

[5] S. Kim and L. J. Lombardino. Comparing graphs and text: Effects of complexity and task. *Journal of Eye Movement Research*, 8(3), 2015.

[6] S. Pinker. A theory of graph comprehension. In R. O. Freedle, editor, *Artificial Intelligence and the Future of Testing*, chapter 4, pages 73–126. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, England, 1990.

[7] T. Walber, A. Scherp, and S. Staab. Benefiting from users' gaze: selection of image regions from eye tracking information for provided tags. *Multimedia Tools Appl.*, 71(1):363–390, 2014.