Re-Evaluating the Efficiency of Physical Visualizations: A Simple Multiverse Analysis

PIERRE DRAGICEVIC, Inria
YVONNE JANSEN, CNRS Sorbonne Université

A previous study has shown that moving 3D data visualizations to the physical world can improve users' efficiency at information retrieval tasks. Here, we re-analyze a subset of the experimental data using a multiverse analysis approach. Results from this multiverse analysis are presented as explorable explanations, and can be interactively explored in this paper. The study's findings appear to be robust to choices in statistical analysis.

Additional Key Words and Phrases: Physical visualization, Multiverse analysis

ACM Reference Format:
Pierre Dragicevic and Yvonne Jansen. 2018. Re-Evaluating the Efficiency of Physical Visualizations: A Simple Multiverse Analysis. In . ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Whereas traditional visualizations map data to pixels or ink, physical visualizations (or "data physicalizations") map data to physical form. While physical visualizations are compelling as an art form, it is unclear whether they can help users carry out actual information visualization tasks.

Five years ago, a study was published comparing physical to on-screen visualizations in their ability to support basic information retrieval tasks [5]. Interactive 2D representations were clearly the fastest, but a gain in speed was observed when transitioning from on-screen to physical 3D representations. Overall, the study suggested that features unique to physical objects – such as their ability to be directly touched – can facilitate information retrieval.

These results however only hold for the particular data analysis that was conducted. A group of statisticians and methodologists recently argued that results from a single analysis can be unreliable [9]. They recommended researchers to conduct instead *multiverse analyses*, i.e., to perform all reasonable analyses of their data and report a summary of their outcomes. While Steegen et al. show how to summarize all outcomes using p-values, here we use an interactive approach based on Bret Victor's concept of *explorable explanation* [10].

2 EXPERIMENT

The study consisted of two experiments. In the first experiment, participants were presented with 3D bar charts showing country indicator data, and were asked simple questions about the data. The 3D bar charts were presented both on a screen and in physical form (see Figure 1). The on-screen bar chart could be rotated in all directions with the mouse. Both a regular and a stereoscopic display were tested. An interactive 2D bar chart was also used as a control condition.

Manuscript submitted to ACM

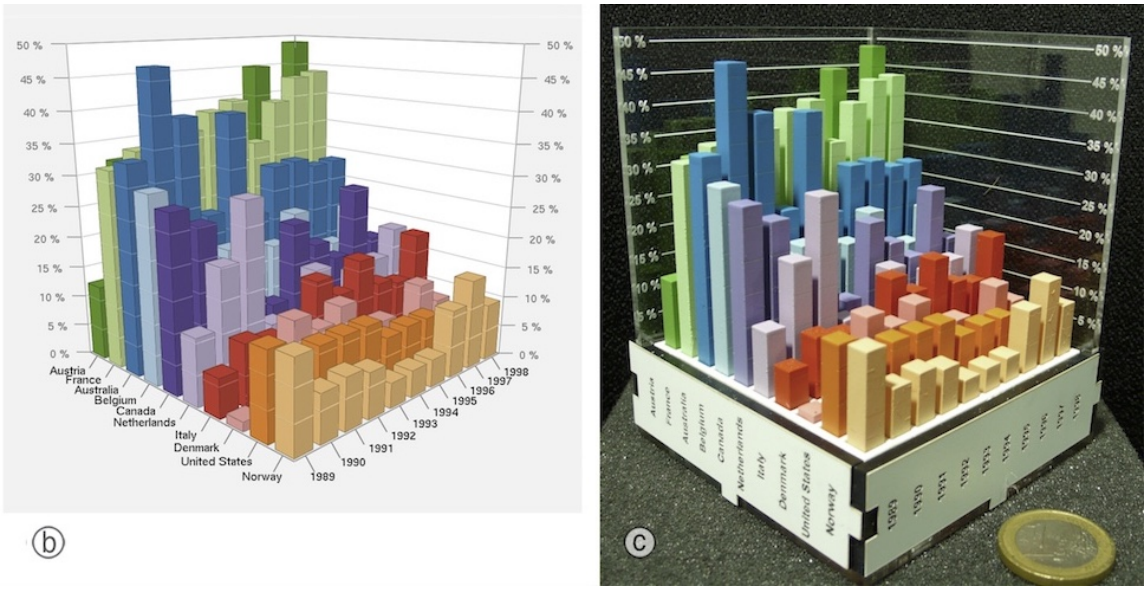

Fig. 1. 3D bar chart, on-screen and physical.

$$\textit{mouse} \xrightarrow[\textit{rotation}]{+\textit{direct}} \textit{prop} \xrightarrow[\textit{realism}]{+\textit{visual}} \textit{no touch} \xrightarrow{+\textit{touch}} \textit{touch}$$

Fig. 2. Effects of interest.

Accuracy was high across all conditions, but average completion time was lower with physical 3D bar charts than with on-screen 3D bar charts.

Here we only re-analyze the second experiment, whose goal was to better understand why physical visualizations appear to be superior. The experiment involved an "enhanced" version of the on-screen 3D chart and an "impoverished" version of the physical 3D chart. The enhanced on-screen chart was rotated using a 3D-tracked tangible prop instead of a mouse. The impoverished physical chart consisted of the same physical object but participants were instructed not to use their fingers for marking. There were 4 conditions:

- *physical touch*: physical 3D bar charts where touch was explicitly encouraged in the instructions.
- *physical no touch*: same charts as above except subjects were told not to use their fingers to mark points of interest (labels and bars).
- *virtual prop*: on-screen 3D bar charts with a tangible prop for controlling 3D rotation.
- *virtual mouse*: same charts as above, but 3D rotation was mouse-controlled.

These manipulations were meant to answer three questions: *1)* how important is direct touch in the physical condition? *2)* how important is rotation by direct manipulation? *3)* how important is visual realism? Visual realism referred to the higher perceptual richness of physical objects compared to on-screen objects, especially concerning depth cues. Figure 2 summarizes the three effects of interest.

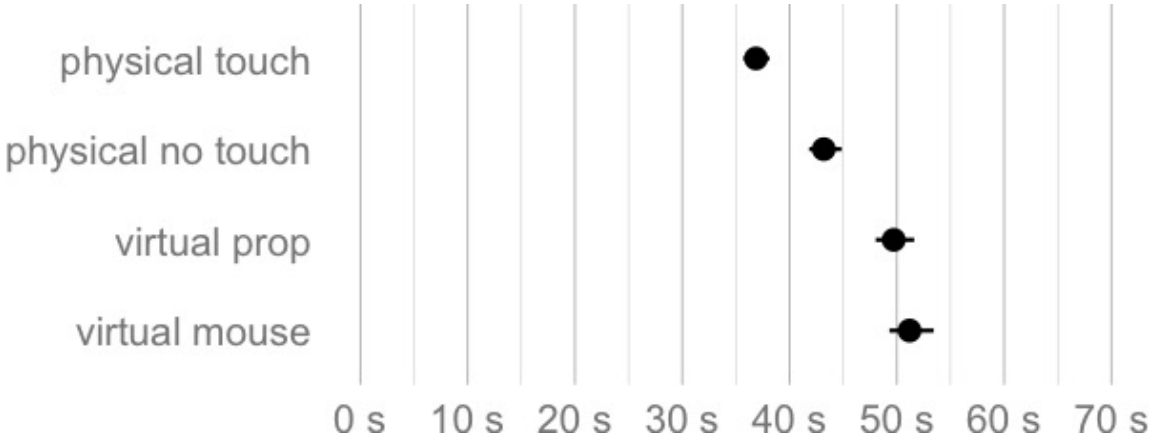

Fig. 3. Average task completion time (arithmetic mean) per condition. Error bars are 50% t-based CIs.

Sixteen participants were recruited, all of whom saw the four conditions in counterbalanced order. For more details about the experiment, please refer to [5].

3 RESULTS

Like the original paper we use an estimation approach, meaning that we report and interpret all results based on (unstandardized) effect sizes and their interval estimates [4]. We explain how to translate the results into statistical significance language to provide a point of reference, but we warn the reader against the pitfalls of dichotomous interpretations [1].

We focus our analysis on task completion times, reported in Figure 3 and Figure 4. Dots indicate sample means, while error bars are 50% confidence intervals computed on untransformed data using the BCa bootstrap method.

Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the population mean 50% of the time across replications, under some assumptions [7]. In practice, if we assume we have very little prior knowledge about population means, each interval can be informally interpreted as a range of plausible values for the population mean, with the midpoint being more likely than the endpoints [3].

Figure 3 shows the mean completion time for each condition. At first sight, *physical touch* appears to be faster than the other conditions. However, since condition is a within-subject factor, it is preferable to examine within-subject differences [3], shown in Figure 4.

Figure 4 shows the pairwise differences between mean completion times. A value lower than 0 (i.e., on the left side of the dark line) means the condition on the left is faster than the condition on the right. The confidence intervals are Bonferroni-corrected. Since the individual confidence level is 50%, an interval that does not contain the value 0 indicates a statistically significant difference at the =.17 level. The probability of getting at least one such interval if all 3 population means were zero (i.e., the familywise error rate) is =.42. Likewise, the simultaneous confidence level is 58%, meaning that if we replicate our experiment over and over, we should expect the 3 confidence intervals to capture all 3 population means 58% of the time.

Figure 4 provides good evidence that *i) physical touch* is faster on average than *physical no touch*, and that *ii) physical no touch* is faster than *virtual prop*. This suggests that both visual realism (e.g., rich depth cues) and physical touch can

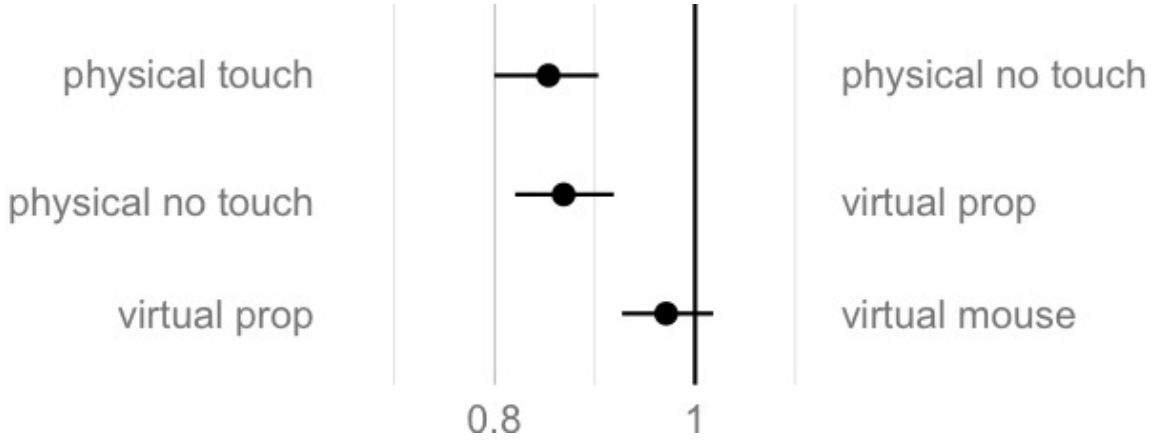

Fig. 4. Differences between mean completion times (arithmetic means) between conditions. Error bars are Bonferroni-corrected BCa bootstrap CIs.

facilitate basic information retrieval. Importantly, these two properties are unique to physical objects and are hard to faithfully reproduce in virtual setups. In contrast, we could not find evidence that physical object rotation (as opposed to mouse-operated rotation) provides a performance advantage for information retrieval.

4 DISCUSSION AND CONCLUSION

Our findings for experiment 2 are in line with the previously published study [5]. In the present article, the default analysis options reflect the choices made in the previously published analysis – thus, the figures are by default identical. On top of this, we consider alternative choices in statistical analysis and presentation, which together yield a total 56 unique analyses and results. The conclusions are largely robust to these choices. Results are less clean with untransformed data, likely because abnormally high completion times are given as much weight as other observations. The use of a log transformation addresses this issue without the need for outlier removal [8].

Meanwhile, the use of bootstrap CIs makes the results slightly stronger, although this is likely because bootstrap CIs are slightly too liberal for small sample sizes [6].

We did not re-analyze experiment 1 to keep this article simple. Since experiment 1 used four conditions and the reported analysis included a figure with seven comparisons [5], it is possible that some of the effects become much less conclusive after correcting for multiplicity. Multiplicity correction is however a contested practice [2], thus it is generally best to consider both corrected and uncorrected interval estimates.

The goal of this article was to illustrate how the ideas of *multiverse analysis* [9] and of *explorable explanation* [10] can be combined to produce more transparent and more compelling statistical reports. We only provided a few analysis options, and many more options could have been included. In addition, our choice of analysis options was highly personal and subjective. Steegen et al. have argued that multiverse analyses are necessarily incomplete and subjective, but are nonetheless way more transparent than conventional analyses where no information is provided about the robustness or fragility of researchers' findings [9].

REFERENCES

[1] Valentin Amrhein, Fränzi Korner-Nievergelt, and Tobias Roth. 2017. The earth is flat (p> 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ* 5 (2017), e3544.

[2] T Baguley. 2012. Serious Stats: A Guide to Advanced Statistics for the Behavioral Science.

[3] Geoff Cumming. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis.* Routledge.

[4] Pierre Dragicevic. 2016. *Fair Statistical Communication in HCI.* Springer International Publishing, Cham, 291–330.

[5] Yvonne Jansen, Pierre Dragicevic, and Jean-Daniel Fekete. 2013. Evaluating the Efficiency of Physical Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13).* ACM, New York, NY, USA, 2593–2602.

[6] Kris N Kirby and Daniel Gerlanc. 2013. BootES: an R package for bootstrap confidence intervals on effect sizes. *Behavior research methods* 45, 4 (12 2013), 905–927.

[7] Richard D Morey, Rink Hoekstra, Jeffrey N Rouder, Michael D Lee, and Eric-Jan Wagenmakers. 2016. The fallacy of placing confidence in confidence intervals. http://dx.doi.org/10.3758/s13423-015-0947-8. *Psychonomic bulletin review* 23, 1 (2 2016), 103–123.

[8] Jeff Sauro and James R Lewis. 2010. Average task times in usability tests: what to report?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, Atlanta, Georgia, USA, 2347–2350.

[9] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing Transparency Through a Multiverse Analysis. http://dx.doi.org/10.1177/1745691616658637. *Perspectives on psychological science: a journal of the Association for Psychological Science* 11, 5 (9 2016), 702–712.

[10] Bret Victor. 2011. Explorable explanations. *Bret Victor* 10 (2011).