Examining Design-Centric Test Participants in Graphical Perception Experiments

G. Guo[1], B. Dy[2], N. Ibrahim[2], S.C. Joyce[2] and A. Poorthuis[2]

[1]Georgia Institute of Technology, USA
[2]Singapore University of Technology and Design, Singapore

Abstract

In this paper, we replicate a foundational study in graphical perception, and compare our findings from using design-centric participants with that of previous studies. We also assess the visual accuracy of two groups, students and professionals, both with design backgrounds, to identify the potential effects of participants' backgrounds on their ability to accurately read charts. Our findings demonstrate that results for reading accuracy for different chart types of previous empirical studies [CM84, HB10] are applicable to participants of design backgrounds. We also demonstrate that besides significant differences in response time, there are no significant differences in reading accuracy between the student and professional groups in our study. This indicates that, despite bias in research participants for visualization research, previous conclusions about graphical perception are likely applicable across different populations and possibly work fields.

CCS Concepts

• ***Human-centered computing*** → *Visualization; Empirical studies in visualization;*

1. Introduction

Visualizations play a key role in many decision making contexts. However, there is relatively little empirical research on exactly how visualizations are read and used to generate actionable insights in decision-making processes. Despite overall attention to this interaction between visualization and decision making [Bre94, CHGF09, Den75, Fla71, RWJ75, Tuf97] especially in cartography, empirical and experimental visualization research has for various reasons focused on the assessment of relatively narrow, specific tasks that are tested with specific, convenient respondent populations.

For example, previous research that looks at human graphical perception for use in data visualization, starting with the classic experiment by Cleveland & McGill (1984) [CM84] and studies that follow the former closely [HB10, TSA14] as well as other research [KH18, SK16] are generally limited in their choice of research participants. Test subjects are often drawn from the researchers' university students or other 'convenient' participant populations like Amazon Mechanical Turk ('MTurk') [Ama]. This issue has been well-documented and critiqued, especially in the field of psychology, mainly under the label 'WEIRD' [HHN10]. In short: people from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) backgrounds may not be representative of all people. More specifically in the context of this research, undergraduate students or previous work on Mechanical Turk may not be representative of design professionals (i.e. those working in fields such as architecture, urban design, product design) and thus, design decision-makers. This is a potential issue if we design visualization systems for decision making for specific domains based on recommendations derived from empirical research on such groups.

From this, we make the following primary contributions:

- We successfully replicate earlier empirical work on the graphical perception of different visual encodings, consistent with theoretical predictions on which visualization techniques are most effective [Mac86].
- We compare the performance of specific participant populations to better understand whether and how frequent participant populations in empirical research (i.e. students and MTurk workers) may introduce specific sampling bias. We show that students and professionals perform similarly in terms of reading accuracy but that students perform significantly faster in graphical perception tasks.

2. Research Goals

Our primary goal is to assess the potential effect of different participant backgrounds on graphical perception (specifically student vis-a-vis design professionals). We hypothesize that the background of a research participant affects the reading accuracy of a visualization, and that in particular, spatially trained design professionals may be better at reading certain charts, i.e. those that rely on area and proportion to communicate data, than others.

DOI: 10.2312/evs.20201046

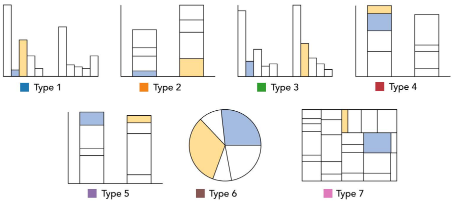
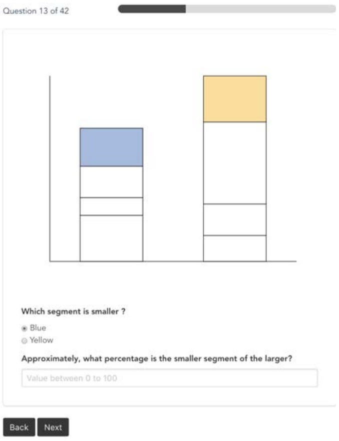


Figure 1: *Chart types used in the experiment.*



Question 13 of 42

Which segment is smaller ?

● Blue
○ Yellow

Approximately, what percentage is the smaller segment of the larger?

Value between 0 to 100

Figure 2: *A sample question from our web platform*

We do this through a replication of Cleveland & McGill's and Heer & Bostock's studies with the two aforementioned populations. We replicate both studies' tests on proportion estimation across different spatial encodings (position, length, angle, area). We also assess whether there is any significant difference between student and professional participants' performance in reading visualizations, and to which chart types these are applicable.

3. Experiment Design

Seven chart types consisting of stacked and grouped bar charts, pie charts and treemaps were tested in this experiment. Types 1-6 correspond to the first six types used in Heer & Bostock's proportional judgement experiment and Type 7 corresponds to the treemap used in their area judgement experiment (see Figure 1). All charts were generated using the same set of values as well, mimicking Cleveland & McGill's original position-length experiment.

The student group was made up of students currently undertaking Architecture and Engineering courses at a Design and Technology school in Singapore. The design professionals were recruited through design networks, such as the National Design Council of Singapore (NDC), and included professionals from different design professions who are active in industry and academia, for example in architecture, urban design, product design and the like.

All participants submitted their responses using a web platform developed for this purpose through specific invitation links generated for each group (students and professionals). This allowed us to track the status of different groups individually. Each participant went through an explanatory introduction and two practice questions before attempting the actual experiment. Participants were advised not to spend too much time on each question and instead to try to make a quick, intuitive visual judgment without using any precise measuring techniques.

Each participant was asked to respond to 42 questions/stimuli in random order. Each question consisted of a unique chart belonging to one of the seven chart types tested. Each chart had two coloured segments, one blue and the other yellow. Participants were first asked to indicate which segment was smaller by choosing the colour of the segment and were then asked to judge the percentage the smaller segment was of the larger segment (see Figure 2). Participants had to answer both parts before continuing to the next question. The time spent on each question and the device details were recorded along with the participants' responses.

We adopt the same measure of accuracy (log absolute error) as Cleveland & McGill [CM84] and Heer & Bostock [HB10]:

$$\log_2(|judged\ percent - true\ percent|) + \frac{1}{8} \tag{1}$$

Our measure for response time is the natural logarithm of the difference in seconds between a participant's entry and exit of a question:

$$\log_e(|exit\ time - enter\ time|) \tag{2}$$

To test between-group differences, we used ANOVA and followed up with Tukey post-hoc tests. We used a Q-Q plot to determine the validity of the F-test, and found that our time and accuracy results follow the normal trend closely. The data collected in this study can be made available upon request.

4. Results and Discussion

There were a total of 123 respondents, of which 87 were students and 36 were professionals. The age of students ranged from 19 to 32, with a mean age of 23. The age of professionals ranged from 21 to 63, with a mean age of 35.

Only the fully completed stimuli/questions (4602) were included in further analysis. From these completed questions, 298 (6.48%) were excluded from accuracy analysis as these participants chose not to provide their occupation or listed their profession as 'Other'. We also looked at the time taken for each response. The online and unpaid nature of our questionnaire meant that compared to the MTurk study [HB10] and the controlled environment of the [CM84], there was a significantly higher chance of participants taking longer than was reasonable to answer questions. Hence, we omitted another 404 outlier responses that did not fall within the interquartile range. This led to a total of 702 (15.3%) of responses being excluded from the analysis of time taken.

4.1. Replication of proportional judgement experiments

The log absolute errors measured in this experiment are, on average, slightly higher than in Heer & Bostock's paper (see Figure 3). This is likely due to their exclusion of response that differed from the true difference by more than 40%. No such exclusion was performed here.

We found a significant effect of chart type on response accuracy ($F(6,4297) = 68.868, p < 0.05$). A further Tukey post-hoc analysis

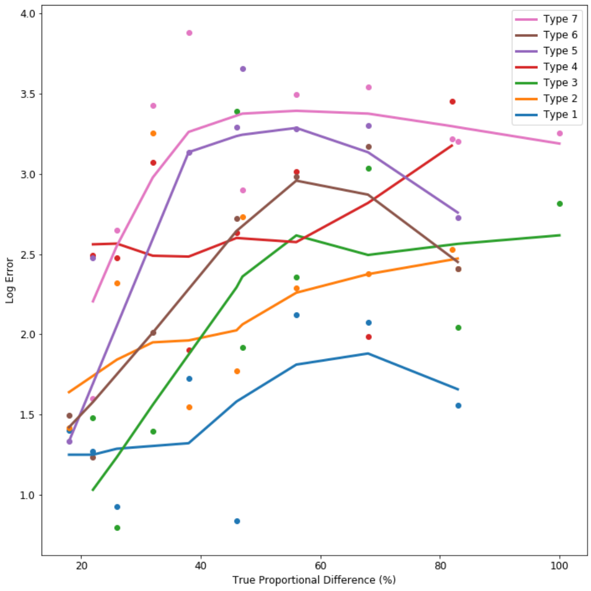


Figure 3: *Midmeans of log absolute errors against true percentages for each proportional judgment type. Superimposed curves were computed with lowess smoothing. The log absolute errors are, on average, slightly higher than in [HB10].*

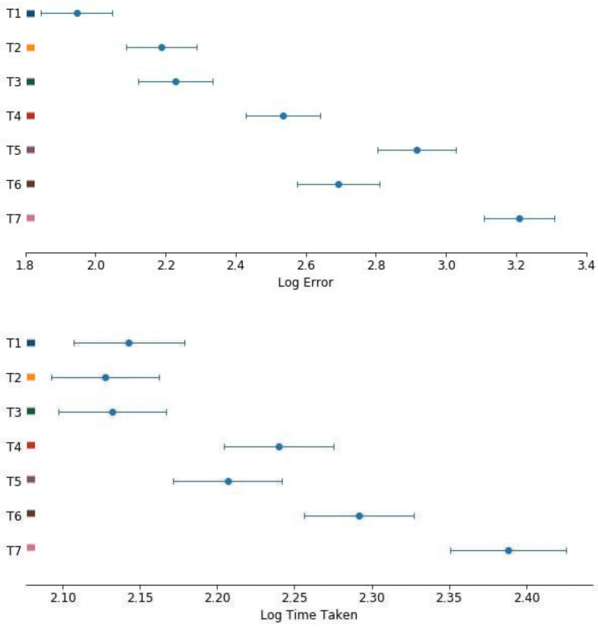


Figure 4: *95% confidence intervals of log errors (top), time taken (bottom) by chart type. Relative performance of each chart (log errors) is comparable to results by [CM84], and [HB10].*

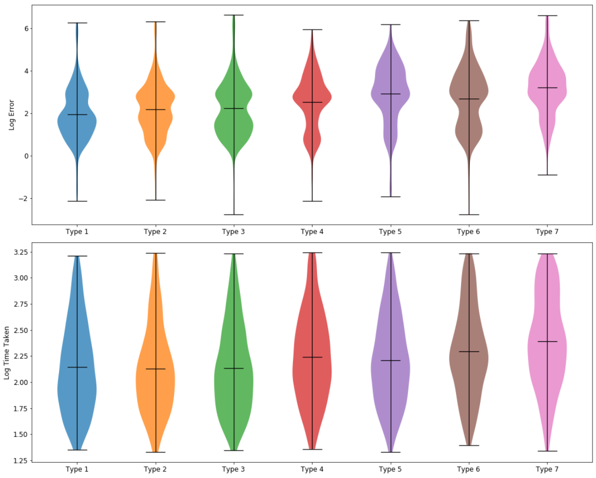


Figure 5: *Top: Violin plot of log error by chart type. Bottom: Violin plot of log time taken by chart type. These results have a similar order of accuracy between the chart types for both log error and log response time as with [CM84] and [HB10].*

found that with the exception of Type 2 (stacked bar) and Type 3 (dodged bar) charts, which had similar performance, all chart types were significantly different from each other in terms of participant accuracy. This result is in line with previous results from Cleveland and McGill [CM84], and Heer & Bostock [HB10]. They found a similar order of accuracy between the chart types, with grouped bar charts as the best performing chart types and tree maps as the worst. In our experiment, Type 1 has the lowest error (1.95), while Type 7 has the highest error (3.21).

In addition to differences in accuracy between chart types, we also found a significant difference in response time ($F(6,3949) = 28.144, p < 0.05$), as seen in Figure 4. Due to strongly left-skewed data for response time, a natural log transform was performed on the measurements before further analysis. A Tukey post-hoc test found that response time was not significantly different for chart Types 1, 2, 3, and 5, while chart Types 4, 5, 6 were also similar. Type 2 has the fastest average response time (9.2 seconds), while Type 7 is 30% slower (12.0 seconds).

Additionally, a correlation analysis between log error and time taken for each chart type reveals no or very weak relationships between log error and time taken for all charts. Most of these correlations are not statistically significant ($p > 0.05$). This suggests that taking more time to give a response does not significantly alter the accuracy of the judgment, regardless of chart type.

4.2. Differences between students and professionals

Now that we have established that our results generally replicate earlier studies, we turn towards analyzing the difference between the student and professional participant groups. If all stimuli (regardless of chart type) are taken together, we find no significant effect of occupation on response accuracy ($F(1,4302) = 0.592, p < 0.4415$) (see Figure 6). Even when we analyze the difference in accuracy between students and professionals for each of the seven chart types, none of the differences are statistically significant. This is a potential indication that the relationship between visual encoding and reading accuracy discussed in previous work is not sensitive to sampling bias present in many empirical, experimental designs. While this result runs counter to our original hypothesis (i.e. the background of a research participant affects the reading accuracy of

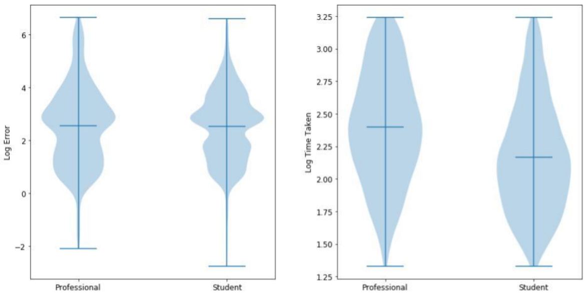


Figure 6: *Left: Violin plot of log error by profession. Right: Violin plot of log time taken by profession. Students and professionals have similar accuracy, but students are significantly faster.*

a visualization), other sampling biases not captured by our current study design may affect the relationship between visual encoding and reading accuracy.

However, we do find a significant effect of occupation on response time taken ($F(1,3954) = 202.675, p < 0.01$). On average, a student takes 8.7 seconds to respond, while professionals take 11.0 seconds. We are not entirely sure what may be the underlying cause behind this difference. It could be that students are indeed faster in reading and assessing visualizations but it could also be that students felt more rushed to complete the experiment; or that they are more comfortable with digital surveys; or generally more trained in test taking. Part of this could also be confounded by the student population being younger, as we did find a significant positive effect of age on response time ($r = 0.447; p < 0.01$).

4.3. Additional findings

Apart from these main findings, we also find that the observed errors have a relationship to the actual, 'true' difference between the two segments that the respondents are asked to compare (see Figure 7). In general, the mean error gets lower if the difference between the two segments is smaller, implying that as the difference between the areas of two segments increases, participants find it more difficult to accurately estimate said difference. Interestingly, we also find that the error variance is noticeably higher for both very small and very large differences.

Finally, no correlation is found between the time taken by participants and their accuracy on a given chart type. Although the choice of chart affects both the response time and accuracy, *within* each chart type there is no or very weak correlation between response time and accuracy. In other words, 'slower' chart types are generally less accurate but on an individual level taking more time to read a chart does not lead to higher accuracy.

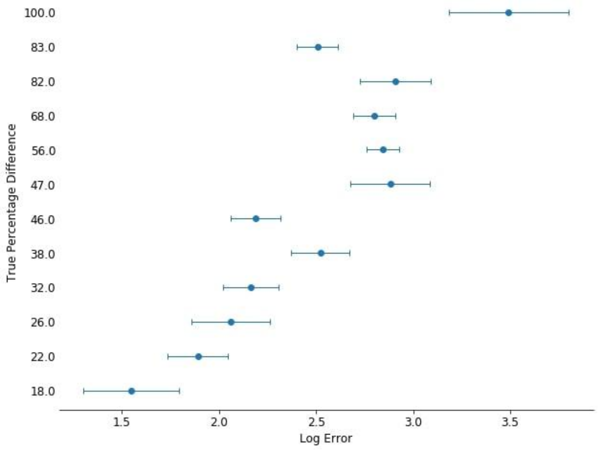


Figure 7: *95% confidence intervals of log errors against true differences between chart segments. As the difference between the two segments to be compared increases, participants find it harder to accurately estimate the difference.*

5. Conclusion

As the previous section illustrates, earlier empirical study and results on the graphical perception of different visual encodings was successfully replicated in the current study. In this replication, we tested on two different groups of participants to evaluate the effect of a participant's background on graphical perception. Our impetus for this research design was the idea that student populations (commonly used as research participants in visualization research) may not be representative of the larger population, or in our case for decision-makers in design fields specifically. Furthermore, there was interest that spatially trained design professionals may be better at reading certain charts (i.e. treemaps) than others.

Our findings, however, did not show a significant difference between students and design professionals in terms of reading accuracy, nor did we see any difference between design trained participants in this test and the non-spatially trained participants in the prior Cleveland & McGill (1984) and Heer & Bostock (2010) experiments. This implies a similar parity of capability between students and design professionals, and holds for both overall accuracy, and for the differences in accuracy between different chart types. Beyond accuracy, we do observe a significant difference in how much time students and design professionals take to complete chart reading tasks, with students being significantly faster.

In summary, the theoretical principles from visualization theory and recommendations based on previous empirical studies on the accuracy of different visual encodings could potentially apply beyond the often-used populations of students and MTurk workers. However, as indicated by the differences in response time, specific sub-groups, such as design professionals and other decision-makers, may indeed read visualizations in slightly different, nuanced ways. This has implications on the development of more complex visualizations, such as those for use in design visualization systems oriented towards decision making, and requires additional follow-up research.

6. Acknowledgements

The material reported in this document is supported by the SUTD-MIT International Design Centre (IDC). Any findings, conclusions, recommendations, or opinions expressed in this document are those of the author(s) and do not necessary reflect the views of the IDC.

References

[Ama] AMAZON: Amazon Mechanical Turk. https://www.mturk.com/. Accessed: 2019–06–13. 1

[Bre94] BREWER C. A.: Chapter 7 – Color use guidelines for mapping and visualization. In *Visualization in Modern Cartography*, Maceachern A. M., Taylor D. R. F., (Eds.), vol. 2 of *Modern Cartography Series*. Academic Press, 1994, pp. 123–147. doi:https://doi.org/10.1016/B978--0--08--042415--6.50014--4. 1

[CHGF09] COLTEKIN A., HEIL B., GARLANDINI S., FABRIKANT S.: Evaluating the effectiveness of interactive map interface designs: a case study integrating usability metrics with eye–movement analysis. *Cartography and Geographic Information Science 36*, 1 (2009), 5–17. URL: https://doi.org/10.1559/152304009787340197, doi:10.1559/152304009787340197. 1

[CM84] CLEVELAND W. S., MCGILL R.: Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association 79*, 387 (sep 1984), 531–554. doi:10.2307/2288400. 1, 2, 3

[Den75] DENT B. D.: Communication aspects of value–by–area cartograms. *The American Cartographer 2*, 2 (1975), 154–168. doi:10.1559/152304075784313278. 1

[Fla71] FLANNERY J. J.: The relative effectiveness of some common graduated point symbols in the presentation of quantitative data. *Cartographica: The International Journal for Geographic Information and Geovisualization 8* (dec 1971), 96–109. doi:10.3138/J647--1776--745H--3667. 1

[HB10] HEER J., BOSTOCK M.: Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), CHI '10, pp. 203–212. URL: http://doi.acm.org/10.1145/1753326.1753357, doi:10.1145/1753326.1753357. 1, 2, 3

[HHN10] HENRICH J., HEINE S. J., NORENZAYAN A.: Most people are not WEIRD. *Nature 466*, 7302 (2010), 29. doi:10.1038/466029a. 1

[KH18] KIM Y., HEER J.: Assessing effects of task and data distribution on the effectiveness of visual encodings. *Computer Graphics Forum 37*, 3 (Jul 2018), 157–167. doi:10.1111/cgf.13409. 1

[Mac86] MACKINLAY J.: Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics 5*, 2 (Apr. 1986), 110—-141. doi:10.1145/22949.22950. 1

[RWJ75] ROTH R., WOODRUFF A., JOHNSON Z.: Value–by–alpha maps: An alternative technique to the cartogram. *The Cartographic Journal 47*, 2 (1975), 130–140. doi:10.1179/000870409X12488753453372. 1

[SK16] SKAU D., KOSARA R.: Arcs, angles, or areas: Individual data encodings in pie and donut charts. *Computer Graphics Forum 35*, 3 (7 2016), 121–130. doi:10.1111f/cgf.12888. 1

[TSA14] TALBOT J., SETLUR V., ANAND A.: Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 2152–2160. doi:10.1109/TVCG.2014.2346320. 1

[Tuf97] TUFTE E. R.: *Visual and statistical thinking: Displays of evidence for decision making*, 1st ed. Graphics Press, Chesire, Conneticut, 1997. 1