A Guide to Designing Experiments to Test Statistical Graphics

Review

The paper introduces an overview of existing methods and procedures of testing for statistical graphics, covering topics such as tools that capture participant responses (including their interactions with the statistical graphic), screening experiments to narrow or quantify the parameter choices in the simulation model, and sharing their experiences (and mistakes) in running graphical experiments. Authors outline several factors that warrant attention such as participants' cognitive load, the design of practice demonstrations, and the demographics and individual effects of the participants. Finally, the importance of pilot testing is emphasized, along with a discussion of methods for analyzing experimental data.

Overall, we think the paper makes valuable contributions to designing experiments for testing statistical graphics, however, the writing requires significant improvement in rigor, greater attention to detail would be welcomed and more practical guidance is desired. The latter point seems critical if this paper is supposed to be a guide.

There are issues with nomenclature, tone and clarity. Below is not an exhaustive list.

- The title refers to “statistical graphics”, however in text authors use “charts”. In fact on Page 1, line 27: Authors write “charts and data visualization” but why the distinction?
- The authors call the visualization lineup protocol as “statistical lineups” but is this really a good name for it? It has not been called as such before and I don’t think it is a good idea to give unintuitive names (it will confuse future research in the area). If you must give it a name, I would prefer “visualization lineups” but it should really be defined, e.g. “In this paper, visualization lineups refer to …".
- The tone is sometimes casual, e.g. Page 8, Line 33: “system is very, very good”, Page 13, Line 5: “a combination of density and rug geoms”. These sound colloquial

and it is not of an academic convention to write things like “very, very” and “rug geoms”.

- Citation style issues. E.g., page 8, Line 3 “Lu et al (Lu e al, 2022)” should be just “Lu et al (2022)”.
- Being specific, e.g. Page 1 line 43: be explicit that “test charts _with human evaluation_”.
- I would appreciate a more informative caption for Figure 1.

As this is a guide, should it not be accessible to a general statistical audience? The authors seem to assume prior knowledge in the area. E.g.

- The description of the lineup itself and its role in visual testing is somewhat abstract. Figure 5 might need to be introduced earlier for better clarity. Similarly, no explanation to how the see-value is calculated – if this is expected knowledge, could authors give recommendation of what literature readers should read to know the appropriate background, otherwise I think it is useful to make this guide self-contained and describe it briefly.
- Page 1 line 46: the concept of internal and external validity is not explained. This might be common knowledge in psychometrics or psychology but I don’t think it is in statistics.
- Is there a guideline or reference for “think-aloud” processes? It’s a concept that not a lot of statisticians would be familiar with so it would be nice to know more about this and the benefits of this approach.
- Should you assume your audience knows what a “PHP” (page 11, line 18) is? Same goes for Qualtric surveys, Amazon Mechanical Turk, Prolific and Reddit. These have no reference/link and little to no explanation. Given technology platforms can evolve rapidly, perhaps authors would like to give more context so that their paper can be understood by future readers. E.g. hypothetically suppose if Reddit disappears today then researchers in 20 years may not know what Reddit is.
- What is IRB? Page 10, Line 43: “make the argument to IRB that our research is exempt”.
- Also NORC panel is not defined.
- Rasch models on Page 12, Line 39 not defined nor cited.

Guidance on the following practical aspects seem important but not discussed:

- how many participants should be recruited?
- what number of plots should be shown to participants in a lineup?

- mention of ethic approval? A reminder that people do need to have ethics approval lined up and linked in their study to use results for publication?
- can authors comment on financial costs related to participant recruitment?
- any guide to ensure representative sample in participant recruitment? Do authors adopt particular method or strategies?
- what about filtering out obviously non-serious participants?
- Authors mention usage of Amazon Mechanical Turk, Prolific, Reddit and social media. Could they provide advice on which online recruitment platform to use?
- any advice for the design of interface that participants use?
- Any examples of qualitative questions asked?

The paper is titled “A Guide to Designing Experiments to Test Statistical Graphics” but perhaps it is better to rename this as “A Guide to Designing Online Experiments to Test Statistical Graphics” as the experiments seem to be only for online experiments?

The key reference is missing: Heer and Bostock (2010) Crowdsourcing graphical perception: using mechanical turk to assess visualization design.

Minor technical issues

- Page 1 line 36: grammar of graphics technically doesn’t “classify” charts but just is a means to specify one.
- Page 2 line 46: “target plot with signal” — but there can be a target plot with no signal right?
- “automatically transcribed using text-to-speech functions of large models” — why LLM particularly? What about the approaches that pre-dates LLM?
- Page 6, line 40: what is "(shah?)” supposed to be?
- The metadata doesn’t match paper title. The paper title is "A Guide to Designing Experiments to Test Statistical Graphics” but the metadata title is "Designing Statistical Graphics Experiments”.