Skipping the Replication Crisis in Visualization: Threats to Study Validity and How to Address Them

Robert Kosara* Steve Haroz†

Tableau Research Sorbonne Université

Abstract

Replications are rare in visualization, but if they were more common, it is not unreasonable to believe that they would show a similar rate of unreproducible results as in psychology and the social sciences. While a replication crisis in visualization would be a helpful wake-up call, examining and correcting the underlying problems in many studies is ultimately more productive.

In this paper, we survey the state of replication in visualization. We examine six threats to the validity of studies in visualization and suggest fixes for each of them. Finally, we describe possible models for publishing replications that satisfy the novelty criterion that can keep replications from being accepted.

1 Introduction

When studies in psychology get repeated, the results often fail to show what the initial study claimed to find. This calls into question whether the studies' findings were real phenomena or merely artifacts of the study design, results of questionable research practices (QRPs), or mere statistical flukes. None of these options is an attractive answer for a science, leading to this problem being dubbed the *replication crisis*.

Is visualization in a similar situation? There is no replication crisis in visualization, but that does not necessarily prove that our results are strong – we just rarely question them. The lack of a replication crisis in visualization is more likely due to the fact that there are virtually no published replications in visualization.

This is not a good or sustainable situation, however. As a young field, visualization has been able to forge ahead and not worry about the trappings of a traditional science for a while. To continue building on our knowledge about visualization, perception, reasoning with data, etc., we need to ensure that what we know is actually sound and correct [10]. Replication is a key way of doing that.

While the replication crisis in psychology is by no means over, the field is responding and proposing a number of promising solutions: preregistered studies, registered reports, and even large-scale efforts like a psychology accelerator [14] promise stronger and more reliable results. Visualization can learn from these practices even without going through a full-blown crisis.

Many studies in visualization suffer from similar issues as the ones in psychology. Visualization needs to get ahead of the game and adopt stronger methods if it wants to produce strong results and avoid an eventual crisis. Below, we outline potential problems with current study approaches and suggest solutions that lead to stronger, more scientifically sound work that stands up to scrutiny and replication.

2 The State of Replication in Visualization

While replication in visualization is generally rare, there are instances of accepted papers that replicate, or largely replicate, previous studies.

Heer and Bostock replicated some of Cleveland and McGill's experiments as a way of testing the feasibility of using crowdsourcing for perceptual experiments [7]. The goal there was not to verify the earlier results, but to test whether an online study would produce comparable results to a well-known lab study. In a similar vein, Kosara and Ziemkiewicz also replicated one of their own earlier studies using Mechanical Turk, as well as a study that was similar to Cleveland and McGill's [12].

A particularly interesting chain of replications and reanalyses was based on a paper by Rensink and Baldridge that investigated the perception of correlation in scatterplots and suggested that Weber's law was useful for modeling it [17]. Harrison et al. replicated their results and added new conditions, considerably broadening the scope into a number of additional visualization types [6]. Kay and Heer reanalyzed their data using a Bayesian framework, showed some shortcomings in their analysis, and suggested that Weber's law was not the best explanation after all [9]. Some of the authors of the initial two papers recently followed up that work by taking the studies further and breaking down possible causes for better understanding [24].

Talbot et al.'s study of line chart slopes [22] covered a larger parameter space than the original Cleveland study [1], which included the original study's space. They thus replicated that study's results as part of theirs, but also covered a larger parameter space and came to different conclusions than the original.

A partial conceptual replication of (by visualization standards) very old work and some unquestioned assumptions was done by Skau and Kosara when they questioned whether pie charts were really read by angle [11, 20]. The 1926 study by Eells [4] had many gaps and hardly had enough statistical power to remain unquestioned for so long. Some of the more recent work had looked at effectiveness but not the actual perceptual mechanism used.

Dragicevic and Jansen similarly questioned highly publicized work on the power of unrelated charts to convince people of facts, and found themselves unable to replicate those results [3]. This particular work is important because it tested a study that had attracted much attention in the press, but appears to be deeply flawed, if not outright fraudulent (some of the original authors' other work had to be retracted under the suspicion of fraud).

3 Study Validity: Threats and Remedies

There are many possible ways the validity of studies can be compromised. In this section, we divide this space into a number of categories, describe the issue, and suggest possible ways of spotting and addressing it. We adopt a threat-and-response format inspired by Munzners nested model for visualization design [15] for this.

3.1 Statistical Fluke

Threat: A study can lead to a statistically significant finding by accident. The common cutoff of alpha=0.05 still allows for a 5% false-positive rate (or 1 in 20).

A single study's ability to explain a phenomenon or demonstrate a phenomenon or effect depends on the underlying noise and the number of observations. But even studies with lots of observations can be driven by chance. The more observations are collected (via a

*e-mail: rkosara@tableau.com

†e-mail: steve.haroz@gmail.com

large study or several smaller ones) and collectively analyzed, the more reliable the findings are. Visualization tends to treat every single study as proof of the effect under study, while older and more established sciences like physics work differently: replications are routine and are required for phenomena to be accepted. This is good scientific practice and also statistically necessary – after all, a 5% false positive rate means that one out of every 20 studies in visualization (potentially several each year!) reports on an effect that does not exist.

The remedy for this kind of problem is direct replication, i.e., running the exact same study again. The results of these studies need to be published whether they agree with the initial one or not, otherwise they lead to the *file-drawer problem*: studies whose results do not reach the significance threshold are not published [21]. This is problematic because it makes it unlikely for erroneous results to get corrected.

Another way to address this issue are meta-analyses. These are run across similar studies and are common in the life sciences and medicine. To be able to conduct these, papers need to report more data, however. Ideally, all the data analyzed for the reported results should be available, and research methodology needs to be described in complete, minute detail, to be sure that the studies being analyzed are comparable. At the very least, more detailed results than the usual F-statistics need to be reported, however.

3.2 Questionable Research Practices

Threat: Statistical analysis of study results allows significant leeway that can lead to false positives.

Visualization is a great tool for data exploration, and many researchers enjoy exploring data. Unfortunately, the data collected in studies is not the right place for this kind of analysis. It leads to what have been called *researcher degrees of freedom* [19] and *the garden of forking paths* [5]: statistics that are shaped by decisions made after the fact, and that invariably "help" the analysis get to a statistically significant result.

A related problem is that even when the commonly-accepted 0.05 cutoff is reached, p-values between 0.1 and 0.5 are actually much less likely than ones below 0.1 when the effect is in fact present [18]. "Just significant" p-values are therefore suspect, even if not necessarily of any nefarious data manipulation, but an indicator that a replication of the study is needed to increase confidence in its results.

A remedy for researcher degrees of freedom is preregistration [16]. The study procedure, as well as the analysis, are described in sufficient detail and deposited in a repository that is timestamped and cannot be manipulated later (such as a registration on the Open Science Framework[1]). Once the study is run, the analysis must then follow the procedure and justify any deviations from it.

Besides preregistration, a formal education in research methods, experience, and conscientiousness regarding the many questionable practices [23] can help reduce (but not eliminate) the chance of exponentially expanding the false positive rate.

3.3 Analysis Problems

Threat: The data analysis is flawed through the application of the wrong statistics, incorrect comparisons, etc.

This is perhaps the most mundane reason results can be flawed, but also one of the more difficult ones to detect. While it is often possible to spot multiple-comparison errors and a few others, many more are difficult to impossible to find. When comparing means, it is important to make sure the statistical package or software actually computes means and doesn't take the means as its input (this can be spotted in the text when the degrees of freedom in the reported F-statistics are too high, which often leads to absurdly small p values). Similarly, groups need to be filtered correctly to be compared, t-tests, χ^2 tests, ANOVAs, etc. need to be applied correctly, etc.

While some of these problems can be spotted in the manuscript, the only real way to ensure correct analysis is to publish all study data *and* analysis scripts, code, etc. This lets others examine the process and not only spot problems, but reanalyze the data and make meaningful corrections. Over the last few years, there has been a slowly emerging trend of publishing study data, though it is by no means a given. Analysis code is often not included, if only because authors feel it is "messy" – similar to the reluctance in publishing source code.

Publishing analysis code, even if messy, has the huge advantage that it lets others examine what was done and re-run the code on the same data. It also protects the authors from others making mistakes (or different assumptions) when reanalyzing the data and claiming to have found deviating results.

3.4 Study Design Flaws

Threat: Poor study design can lead to misleading or inconclusive results.

Study design is a more challenging task than often acknowledged: a good study needs to control a huge number of possible factors in order to focus on just the few that are under investigation. The experiment also needs to actually address the question, which is not always a given. In trying to keep the parameter space from exploding, it is easy to lose track of how the experiment actually relates to the initial question.

Confounds, variables that are not controlled or counterballanced but still vary between conditions, are a common occurrence in visualization studies. They can influence both the results and the possible explanations, but are rarely appropriately considered. Similarly, functional dependencies between variables can reduce the effective parameter space and make an effect appear that is really just the result of a direct, and usually known, relationship that has no bearing on the actual question.

Keeping the possible combinations of parameters under control is also a common problem, and it can lead to experiments that do not completely cover their parameter space, which then leads to wrong conclusions. An example of this from visualization is banking to 45°: Cleveland's original study [1] found that the ideal mean slope of two lines (for most accurate slope comparison) was 45°. This was later shown to be an artifact of the study design, which didn't test the full range of possible angles and slope ratios [22].

There are two main ways to discover this type of problem: experience and conceptual replications. Given the inexact science that is designing experiments, there is no procedure for doing this. Experience helps spot common mistakes, as does meticulous documentation (which enables reviewers and later readers to find problems).

Conceptual replication is the stronger method of the two. Instead of repeating the same, possibly flawed, experiment, it consists of designing a new experiment to test the same underlying effect or phenomenon. If a different experiment finds the same effect, it is much more likely to be real. Physics and other 'hard' sciences demand conceptual replication before they will accept the results of a new study, especially one that produces surprising or counterintuitive results.

3.5 Overgeneralized Conclusions

Threat: The results of the study are overinterpreted or overgeneralized beyond what the experimental results support.

While studies are often designed to be extremely well controlled and test a very narrow effect, the goal is to show something much broader. The perception of size in different shapes, say, may differ in a certain way – but only for the shapes tested. Any other shape might behave differently, even if the study carefully tested different

[1] https://osf.io

classes of shapes. It is easy to make overly broad statements about one's findings that are not fully supported by the data.

Reviewers often push back against broad statements like this, and for good reason. The broad statements can be phrased as questions: given that we found this to be true for these specific cases, perhaps there is a more general effect here? It behooves authors to be precise and specific in order to stay within the realm of a scientific paper. At the same time, it is a valid desire to produce research that is useful and applicable, which is not possible when it has to be couched in disclaimers and narrow restrictions.

This kind of problem is relatively easy to spot by comparing the claims with the study design and noting where the claims extend beyond what was tested. In the study, the breadth of possible claims can be increased by broadening the scope of conditions studied, making sure the study populations are sufficiently diverse (e.g., was this only tested with male subjects, or subjects of college age, etc.?).

Conceptual replications also play a crucial role: the same authors or others can repeat a similar study on a slightly different set of conditions, a different population, etc. If they find the effect as well, a broader formulation is more acceptable and more likely to be correct.

3.6 Misinterpreted Results

Threat: The claimed mechanism is not actually the correct or only explanation for the observations from the study.

An example of this is the angle component of Cleveland and McGill's graphical perception paper [2]. They showed their participants pie charts and assumed that the visual cue used was angle. This was recently shown to not be the only explanation (area and arc are also possible), and in fact the least likely one [11, 20].

Detecting these issues is possible through careful scrutiny of the methods, analysis, and conclusions in the paper, as well as through simple hunches: perhaps the explanation in the paper feels wrong, or a reader has a different possible explanation in mind.

Misinterpreted results are not necessarily a fault in the original research. In fact, science largely progresses because accepted explanations are found to be wanting, or additional evidence and ideas call for the reexamination of the existing knowledge [13]. The transition from Newtonian physics to Einstein's relativity theory is a well-known example of such a transition, but similar ones happen on a much smaller scale all the time.

One mechanism for this sort of transition is the comment paper or comments associated with a published paper. While this is virtually unheard of in visualization, it is common in statistics and other fields to invite comments on papers about to be published. Those are then published together with the paper and can offer additional ideas or propose alternative interpretations of results. They can serve as valuable starting points for further research. Even after publication, journals in many fields (including visualization, at least TVCG has a *comments paper* category) accept short comments as valid contributions.

4 Types of Replication

There are different kinds of replications of studies, from pure data reanalysis to repeating the exact same study, to designing an entirely new study to investigate the same underlying effect. Each one addresses a different threat described in the previous section, and each provides different new information.

4.1 Reanalysis

Perhaps the simplest kind of replication is to reanalyze the data gathered from the study. This was done to great effect in the example of the Weber's law papers described in Secion 2. Reanalysis can serve different purposes: ensure the soundness of the mathematics and statistics, and test possible alternative explanations.

While we have to assume that authors are meticulous in their work, mistakes happen, statistics can be misunderstood and misapplied, participants and data points may be removed a little bit too generously, etc. Reanalysis can spot these mistakes and judgment calls and can point them out. Visualization does not have a history of corrections and withdrawn papers, but both are common practice in many other fields to correct the inevitable mistakes and correct the record when problems are found.

The more exciting use of reanalysis is to test the potential for alternative hypotheses. This may be done as a sort of pilot or feasibility study for another experiment, or simply to test a hunch. Both are valuable because they bring more eyes to the data and help broaden the possible interpretations that are being considered – thus moving science forward.

4.2 Direct Replication

Simply repeating the exact same experiment might appear pointless, but it is critical from a statistical point of view and has other important advantages. A single study only represents a single sample, which may show a significant effect by chance (which at 5% is not even that low). Even if all conditions were exactly the same, a pure, exact replication is valuable. If it "succeeds" (shows the same effect), it makes the original study more believable and the studied/claimed effect more likely to be real; if it "fails" (does not show the effect), it raises interesting questions about the actual existence of the effect and demands more replications (it can also serve as the impetus for a conceptual replication to rule out the study design as a confounding factor).

Of course, no replication is ever exact, since it happens at a different time, with different participants, (hopefully small) deviations from the study protocol, etc. A successful replication thus usually means that the effect is robust against a variety of factors. It also usually broadens the diversity of the participant pool, thus increasing trust in the effect being universal (and not limited to a certain gender, age group, education level, etc.).

4.3 Conceptual Replication

Both of the above replications are focused on the experiment as initially run, rather than the underlying effect. The effect is arguably much more important than any experiment – after all, the experiment is not an end in itself, it is a means to test or find an effect.

A conceptual replication therefore consists of designing and running a new experiment that aims to test for the same effect or phenomenon. An effect that is detected by two or more different experiments is much more robust and likely to really exist. Conceptual replication is the modus operandi in fields like physics, where phenomena like gravitational waves, or the existence of a new particle, need to be shown not only by different labs conducting very similar experiments, but also a variety of different experiments that all test the same underlying theory.

4.4 Registered Reports

A registered report is not necessarily a replication, but it is closely related (and registered reports are a good mechanism for replications). It consists of two phases: first, the study design and analysis are specified, and any pilot studies are run. The resulting paper, which at that point does not contain the results of the actual study, is submitted for review and accepted or declined based on the methodology alone. Once the paper is accepted, the study is run, the data are analyzed and the results written up according to the accepted methodology, and the final manuscript is only reviewed for adherence to the originally-described methods, for explanation of any deviations, and for any interptation of the results.

Registered reports do not suffer from the file drawer problem, since the paper gets published whether or not it finds the expected effect. Since they are reviewed based on their methodology alone, that needs to be described in sufficient detail for reviewers to be able to understand and believe that such a study would yield an interesting result. That requires not just clear and well-motivated methodological choices but also a sufficient understanding of the underlying theory to make clear predictions. If such a study does not find the expected effect, it raises interesting questions about that theory, and if it is a replication it does the same for the study it replicates. Similarly, when the expected effect is found, confidence in it is higher given the preregistered nature of the study.

5 Ways to Publish Replications in Visualization

The threats to study validity itemized above are not just of academic interest, we are aware of examples for all of them, even though we refrained from citing many specific examples (especially recent ones). Many studies in visualization are flawed to varying degrees, and we believe that many would not hold up to replication.

What can be done to improve research methods in visualization? We propose a few possible ways below.

5.1 Replication and Novelty: Build-Upon Studies

Publishing replications is extraordinarily difficult in visualization because they are not considered novel. Both of separately experienced this as reviewers, having to push hard to get the very rare replications accepted. How can we make replications acceptable in a field that demands novelty above all else?

While we propose more structural changes to the field below, there is a simpler way: using replications as starting points to build on. This kind of paper replicates an existing study as a pilot or first step. Depending on the results of this, further studies are then conducted or new methods developed that are novel. This way, the paper has a novel component even if the replication is not considered novel at all.

An existing example of this type of paper is Heer and Bostock's crowdsourcing paper [7], which validated crowdsourcing by means of a replication, but then also added new studies about area perception.

5.2 Paper Categories and Reviewing Policies

A simple change that would allow replications to be published would be the introduction of new paper categories and associated reviewing policies. Given the crucial novelty question, this would have to include some education of reviewers as well, who might still balk at accepting papers they do not consider to be novel.

Guidelines would have to be clear about what kinds of replications they accept (perhaps only preregistered ones) and what the criteria for acceptance should be. While it has been suggested that any replication should involve the original authors [8], we believe that to be counterproductive as a general rule. Instead, the authors of the original work should be invited to comment on the replication paper.

We believe that replications can, at the very least, serve a purpose similar to literature surveys: give graduate students exposure to research methods and aid in their training. Given the importance of publishing to obtain a computer science degree, being able to publish replications is the only way they can ever become part grad students' training.

5.3 Journal of Visualization Experiments and Methods

Even with the build-upon model and policies, publishing pure replications will remain challenging. As the field grows and matures, it needs more and more specialized publication venues. One of them could be a journal specializing in experiment design, novel methods, and replications. The latter would include registered reports, which address the file drawer problem.

Similar to the policies suggestion above, such a journal would have to be very clear in its criteria for different categories of papers to make the different types of paper acceptable: pure methodology, novel study design, pure replication, reanalysis, registered report, etc.

6 Conclusions

Improving research methods and establishing replications as a viable type of publication in visualization will require effort from the entire field. This change cannot be made purely top-down (via policies, etc.) or bottom-up (via stubborn submission of replications). Authors, reviewers, papers chairs, steering committee members all need to help to make this happen. We believe not only that this is a worthwhile effort, but that it is crucial to increasing the strength of the field and protecting it from a full-scale replication crisis.

References

[1] W. S. Cleveland. A Model for Studying Display Methods of Statistical Graphics. *Journal of Computational and Graphical Statistics*, 2(4):323–343, Dec. 1993.

[2] W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[3] P. Dragicevic and Y. Jansen. Blinded with Science or Informed by Charts? A Replication Study. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):781–790, 2018.

[4] W. C. Eells. The Relative Merits of Circles and Bars for Representing Component Parts. *Journal of the American Statistical Association*, 21(154):119–132, 1926.

[5] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Technical report.

[6] L. Harrison, F. Yang, S. L. Franconeri, and R. Chang. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(12):1077–2626, 2014.

[7] J. Heer and M. Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings CHI*, pp. 203–212, 2010.

[8] D. Kahneman. A New Etiquette for Replication. *Social Psychology*, 45(4):310–311, 2014.

[9] M. Kay and J. Heer. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):1077–2626, 2016.

[10] R. Kosara. An Empire Built On Sand: Reexamining What We Think We Know About Visualization. In *Proceedings BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, 2016.

[11] R. Kosara and D. Skau. Judgment Error in Pie Chart Variations. In *Short Paper Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization EuroVis*, pp. 91–95. The Eurographics Association, 2016.

[12] R. Kosara and C. Ziemkiewicz. Do Mechanical Turks Dream of Square Pie Charts? In *Proceedings BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, pp. 373–382, 2010.

[13] T. S. Kuhn. ***The Structure of Scientific Revolutions***. University of Chicago Press, 1962.

[14] H. Moshontz. The Psychological Science Accelerator: Advancing Psychology through a Distributed Collaborative Network. Technical report, PsyArXiv, 2018.

[15] T. Munzner. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[16] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, Mar. 2018.

[17] R. A. Rensink and G. Baldridge. The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010.

[18] U. Schimmack. The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17(4):551–566, 2012.

[19] J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-Positive Psychology. *Psychological Science*, 22(11):1359–1366, 2011.

[20] D. Skau and R. Kosara. Arcs, Angles, or Areas: Individual Data Encodings in Pie and Donut Charts. *Computer Graphics Forum*, 35(3):121–130, 2016.

[21] T. D. Sterling. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance–Or Vice Versa. *Journal of the American Statistical Association*, 54(285):30–34, 1959.

[22] J. Talbot, J. Gerth, and P. Hanrahan. An Empirical Model of Slope Ratio Comparisons. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2613–2620, 2012.

[23] J. M. Wicherts, C. L. S. Veldkamp, H. E. M. Augusteijn, M. Bakker, R. C. M. van Aert, and M. A. L. M. van Assen. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, 7(e124):108, Nov. 2016.

[24] F. Yang, L. Harrison, R. Rensink, S. L. Franconeri, and R. Chang. Correlation Judgment and Visualization Features: A Comparative Study. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, to appear, 2018.