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

Position Paper

Robert Kosara*
Tableau Research

Steve Haroz†
Sorbonne Université

ABSTRACT

Replications are rare in visualization research, but if they were more common, it is not unreasonable to believe that they would show a similar rate of unreproducible results as in the psychological and social sciences. While a replication crisis in visualization research 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 research. We examine six threats to the validity of studies in visualization and suggest ways to address 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, many results fail to show what the initial study claimed to find. This inconsistency calls into question whether the original studies' findings were real phenomena or merely artifacts of the study design, results of questionable research practices (QRPs), or statistical flukes. None of these options is an attractive answer for a science, leading to this problem being dubbed the *replication crisis*.

Is visualization research 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 very few replications in visualization are attempted, let alone published.

This situation raises questions about the validity of the field's results. As a young field that largely grew from computer science and design, visualization research has prioritized applicability and aesthetics over falsifiability. To build on our scientific understanding about how we perceive, reason with, and remember visual information, we must put increased value on the soundness and correctness of conclusions [14]. Replication is one component of improving the validity of the field's conclusions.

While the replication crisis in psychology is by no means over, the field is responding and implementing a number of promising solutions: preregistration, registered reports, requiring open data and study materials [7], and even large-scale efforts like a psychology accelerator [18] 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 to the ones in psychology. 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.

*e-mail: rkosara@tableau.com
†e-mail: replication@steveharoz.com

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 [11]. The goal 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 [16].

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 [21]. Harrison et al. replicated their results and added new conditions, considerably broadening the scope into a number of additional visualization types [10]. 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 [13]. Some of the authors of the initial two papers recently followed up on that work by taking the studies further and breaking down possible visual properties that correlate with subject responses [28].

Talbot et al.'s study of line chart slopes [26] 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.

An experiment by Gramazio, Schloss, and Laidlaw [6] included a subset of the conditions from one of the visual search experiments by Haroz and Whitney [9]. When trying to find a uniquely colored item in a display, the original study found an interaction effect on performance between the number of colors and the layout of the items. Gramazio et al. used different colors and layouts but very closely replicated the nominal results of the original experiment. They also extended the original by manipulating the number of items and found limitations on the original conclusion.

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 [15, 24]. 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 they were 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 (other work by one of the original authors has been 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 Munzner's model for visualization design [19].

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 an effect depends on the measurement noise and the number of observations. But collecting more observations (via a large study or several smaller ones) only decreases the likelihood of noise driving a result; it does not eliminate the possibility. More observations allow for more noise to be accounted for and increase the reliability of the findings. 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.

Any type of replication can reduce the likelihood of a statistical fluke. Direct replication – running the exact same study again – can be particularly effective at reducing this problem by providing more observations. 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 [25]. If negative replications are not published or have a higher bar for publication, erroneous results are unlikely to be corrected.

If multiple similar studies are run, and not all find the same results, meta-analyses can aggregate the evidence. They are common in the life sciences and medicine, where effect sizes tend to be unaffected by minor variations in experiment design. However, to be able to conduct meta-analyses, papers need to consistently report effect sizes, including confidence intervals. Ideally, all the data analyzed for the reported results should be available, and experiment methods need 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 p-values need to be reported.

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 has been called *researcher degrees of freedom* [23] or *the garden of forking paths* [5]: statistics that are shaped by decisions made after observing the data, 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 [22]. "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.

Visualization research is often especially susceptible to motivated reasoning, which may increase the likelihood that researchers perform questionable research practices (QRPs). When researchers test a design that they created against some existing alternative, the paper is unlikely to be published if "their design" performs worse than a common alternative. They may therefore try adding more subjects, excluding subjects based on new criteria, changing the statistical test, or even dropping experiments until they get the results they want. These QRPs, motivated reasoning, and the file-drawer problem may explain why almost all publications in visualization find that the authors' novel design is somehow better than existing techniques.

A remedy for researcher degrees of freedom is preregistration [20]. 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 (e.g., the Open Science Framework[1]). Once the study is run, the analysis must then follow the procedure or justify any deviations from it.

Besides preregistration, a formal education in research methods, experience, and conscientiousness regarding the many questionable practices [27] 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.

Analysis problems are perhaps the most mundane reason results can be flawed, but can also be one of the more difficult problems to detect. While it is often possible to spot multiple-comparison errors, many other issues are difficult or impossible to find. For example, when analyzing data from within-subject experiments with many repeated conditions, it is important to ensure that the statistical method's assumptions about observation independence are followed. More specifically, an ANOVA does not account for multiple observations from the same condition by one subject. This problem can artificially inflate the effect size and shrink the p-values, but it can easily be spotted in the text when the degrees of freedom of the F-statistics are too high. Similarly, groups need to be filtered correctly to be compared, t-tests, χ^2 tests, ANOVAs, or Bayesian analyses need to be applied correctly.

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 counterbalanced 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

[1] `https://osf.io`

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 [26].

There are two main ways to discover this type of problem: experience and conceptual replications. Just like with programming, experiment design experience helps spot common mistakes, as does meticulous documentation (which enables reviewers and later readers to find problems).

Conceptual replication is also critical. Instead of repeating the same, possibly flawed, experiment, a new experiment can 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 overgeneralized beyond the experiment's findings.

Visualization papers often conflate two research approaches which have very different goals: (1) a user study that aims to provide applicable and actionable comparison, and (2) a set of experiments which seek generalizable understanding.

Many visualization user studies test "real-world designs" by comparing fully functional designs that are different in many ways. This approach makes sense when improving user performance for a specific context is the goal, as a user study can provide easily interpretable and directly applicable results. For example, the producer of a commercial security visualization package or a stock analysis platform aims to compare designs rather than understand why one is better than another. Unfortunately from a research standpoint, the lack of careful isolation of variables in user studies makes it difficult if not impossible to determine whether the results generalize to any scenario without identical designs and tasks. Without an explanation for why an effect occurs, there is rarely an indication of what and how much can change while maintaining the benefits of a particular design. User studies are applicable and useful but rarely generalizable.

The goal of scientific experiments, however, is to explain some phenomenon and understand and predict how it may manifest in other circumstances. Accomplishing this generalizability requires experiments that carefully isolate variables and control all changes between conditions. Furthermore they require clear hypotheses to interpret results. However, the near infinite number of combinations of variables needed to understanding how people use a fully functional visualization cannot be reasonably isolated and tested in an experiment. Therefore, experiments are often limited in complexity and number of variables. Though they are generalizable, they may not always be directly applicable by more application-oriented researchers.

Both complex user studies and carefully controlled experiments have merits and limitations. However, care should be taken when inferring generalizable results from a user study and when assuming that scientific experiments can predict applied scenarios. Use of the wrong approach can be detected by comparing the complexity or variable isolation in the methods with the claims of generalizability or applicability.

Nevertheless, user studies and controlled experiments can serve as avenues for conceptual replications of each other. If the results of a simple controlled experiment are replicated in an applied scenario, the underlying mechanism is robust to confounds and moderators. Likewise, if an effect found in a user study can be isolated in a carefully controlled experiment, researchers can better attempt to understand why it occurs and how it would generalize.

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 conclusion was recently shown to not be the only explanation (area and arc are also possible), and in fact the least likely one [15, 24].

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 [17]. 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, since 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

Requirements: original data and original analysis scripts

Perhaps the simplest kind of replication is to reanalyze the data gathered from the study to reproduce the results. This approach was taken 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 models.

While we generally assume that authors are meticulous in their work, mistakes happen, results can be transcribed incorrectly, 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. While retraction is rare in other fields, most have established procedures for correcting 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

Requirements: original experiment code and original analysis scripts

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, especially sampling error. It also usually broadens the diversity of the participant pool, thus increasing trust in the effect being universal (and not limited to a specific population).

4.3 Conceptual Replication

Requirements: well documented methods and analyses in the original paper

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 under different circumstances. 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.

5 Registered Reports

A registered report is an approach to peer review that separates review of the research methods from the results. It consists of two phases: first, the study design, analysis, and a priori interpretations of possible results 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 methods alone. Once the paper is accepted, the study is run; the data are analyzed; and the results are written up according to the accepted methods. Studies with iterative experiments can undergo iterative rounds of review and data collection. The final manuscript is only reviewed for adherence to the originally-described methods, for explanation of any deviations, and for any interpretation of the results.

Registered reports do not suffer from the file drawer problem, since the paper is published whether or not its results are expected or surprising. Being independent of results makes them a useful publication approach for large scale or expensive replications. Both positive replications that increase the certainty of an effect and negative replications that contradict a reported effect all increase our knowledge, and the chances of publication should not be influenced by the possible controversy of the results.

Registered reports are a relatively new approach to review and publication. More information about them is available at `https://cos.io/rr/`

6 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.

6.1 Replication and Novelty: Build-Upon Studies

Publishing replications is extraordinarily difficult in visualization because they are not considered novel. Both authors have separately experienced this as reviewers, having to push hard to get the very rare submitted 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. A replication of even a well-established effect can serve as sanity check to demonstrate that the methods, measurements, and experiment code work as expected. The replication can also serve as a control condition for further studies or a point of comparison for developing new methods of testing the known phenomenon. Such a paper would have a novel component even if the replication is not considered novel at all.

An example of this type of paper is Heer and Bostock's crowdsourcing paper [11], which validated crowdsourcing by means of a replication and then also added new studies about area perception.

6.2 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 category would require some education of reviewers who might still balk at accepting papers they do not consider to be novel. Furthermore, 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 [12], 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.

6.3 Facilitating Replication with Open Practices

In order to replicate or reanalyze previous work, all code, materials, and measured data used to support a paper's conclusions should be accessible [8]. Sometimes, the methods text offers sufficient information to allow for replication, but there are often minor details and parameters that are only available in the code. Hiding or otherwise preventing access to experiment code can prevent accurate replication and can make the original authors appear to be nefariously concealing information. Researchers performing replications should also not need to beg and plead with the original authors to gain access to materials needed to replicate a study. These materials should be made available as a matter of course.

6.4 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 for a conceptual replication, exact replication, reanalysis, registered report, etc.

6.5 Journal of Visualization Science, Empiricism, and Methods

A problem with the current publishing model in visualization research is that the various categories of papers are indistinguishable once published. Papers that have an art and design focus with no empirical component are published alongside multi-experiment papers with little discussion of aesthetic concerns. These different types of papers go through reviews that have completely different standards, yet that process is not indicated on the published paper. When reading a published visualization paper, a person does not know what standard of review it went through.

It may be beneficial to separate all empirical and scientific research in visualization into its own journal. Such an approach would allow the scientific community in visualization to establish review and publication standards independently of the engineering and artistic communities. More importantly, readers would know that every paper in the journal went through a peer review that required evidence for every claim. This journal could also define more specific guidelines of when replications warrant peer-review, standards for registered reports and preregistrations, and criteria for data and material sharing.

Such a separation should not be seen as a disparagement of engineering or design. Instead, it would be a clear acknowledgement that the different communities within visualization research have different metrics for what constitutes "good work." However, establishing different standards and possibly publication venues does not need to divide the field. As many researchers pursue multiple disciplines, and many people appreciate keeping up with other subfields, maintaining a joined conference would continue to allow multiple communities to share their recent and in-progress work with each other.

7 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] C. C. Gramazio, K. B. Schloss, and D. H. Laidlaw. The relation between visualization size, grouping, and user performance. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1953–1962, Dec. 2014. doi: 10.1109/TVCG.2014.2346983

[7] T. E. Hardwicke, M. B. Mathur, K. MacDonald, G. Nilsonne, G. C. Banks, M. C. Kidwell, A. Hofelich Mohr, E. Clayton, E. J. Yoon, M. Henry Tessler, R. L. Lenne, S. Altman, B. Long, and M. C. Frank. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, 5(8):180448, Aug. 2018. doi: 10.1098/rsos.180448

[8] S. Haroz. Open practices in visualization research, August 2018. doi: 10.31219/osf.io/8ag3w

[9] S. Haroz and D. Whitney. How Capacity Limits of Attention Influence Information Visualization Effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2402–2410, Dec. 2012. Citation Key: Haroz2012. doi: 10.1109/TVCG.2012.233

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

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

[20] 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.

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

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.