You Can't Publish Replication Studies (and How to Anyways)

Position Paper

Ghulam Jilani Quadri*
University of South Florida

Paul Rosen†
University of South Florida

ABSTRACT

Reproducibility has been increasingly encouraged by communities of science in order to validate experimental conclusions, and replication studies represent a significant opportunity to vision scientists wishing contribute new perceptual models, methods, or insights to the visualization community. Unfortunately, the notion of replication of previous studies does not lend itself to how we communicate research findings. Simple put, studies that re-conduct and confirm earlier results do not hold any novelty, a key element to the modern research publication system. Nevertheless, savvy researchers have discovered ways to produce replication studies by embedding them into other sufficiently novel studies. In this position paper, we define three methods—re-evaluation, expansion, and specialization—for embedding a replication study into a novel published work. Within this context, we provide a non-exhaustive case study on replications of Cleveland and McGill's seminal work on graphical perception. As it turns out, numerous replication studies have been carried out based on that work, which have both confirmed prior findings and shined new light on our understanding of human perception. Finally, we discuss how publishing a true replication study should be avoided, while providing suggestions for how vision scientists and others can still use replication studies as a vehicle to producing visualization research publications.

Keywords: Replication, empirical studies, graphical perception, vision science.

1 INTRODUCTION

A replication study is a re-evaluation, re-confirmation, or extension of an original study. These studies can be performed under similar experimental conditions to the original studies to validate conclusions or under varying experimental conditions to gain more knowledge [13,29], and they provide an important concrete baseline, which is useful to improving cross-study validity [24]. The journal *Science* labeled replication as the "scientific gold standard" [16] and in others' words "in replication, the private chimera becomes the communal fact" [1].

Unfortunately, the positivist notion of reproducibility does not offer the novelty of qualitative research itself [38]. In a literature review of 891 papers by Hornbaek et al., they found only 3% of papers in Human-Computer Interaction (HCI) attempted replication [13], and at a recent BELIV workshop multiple researchers pointed out that it is rare to find replication work in information visualization research [22,37].

Efforts have been made to encourage researchers to improve reproducibility and conduct more replication studies. For example, it is increasingly common for authors to voluntarily (and sometimes mandatorily) provide their experimental data and other supplemental materials to accommodate future replication. Since 2012, the EuroRVVV[1] Workshop on Reproducibility, Verification, and Validation in Visualization has been promoting and supporting replication in visualization. The RepliCHI[2] workshop, organized as a part of the ACM CHI conference, promotes a shift towards favoring replication in the HCI research community [43]. Finally, the BELIV[3] workshop at IEEE VIS conference is also focused on replication and reproducibility [22,37,41].

A highly debated usability evaluation paper in ACM CHI highlighted the lack of replication studies in HCI [8]. It argued that *reviewers do not value replication work*, and therefore papers are not published, despite the intrinsic value in their work. The reluctance from the reviewers and publication venues to consider replication studies as a valuable contribution has lead to few replication studies being published [8, 13, 43]. Most of the replications studies successfully published are of controversial findings or of highly-cited works [3,4,6,8,13,17,33]. Instead of attempting replication studies, HCI and information visualization researchers have been pushed towards novel and organic findings [13,27,43]. Nevertheless, savvy researchers have been successful at publishing replication studies in less conventional ways, in particular, by embedding them into other sufficiently novel studies. Many of these studies have been at the intersection of vision science and visualization.

Recent workshops and conferences on vision science and visualization have established the potential for methodologies, experimental techniques, and user studies on perceptual judgements from vision science to benefit visualization. However, cross-field contribution is challenging. For vision scientists wanting to interact with the visualization community, *replication* can be a viable and relatively low-risk area where they can contribute.

In this position paper, we highlight how researchers have integrated replications of prior studies into new studies by including them with cutting-edge contributions. In this way, these new works confirm the prior studies and introduce a novel idea or methodology. In particular, we found three common methods of integration: (1) *re-evaluating* under different demographics and/or participant environments; (2) *expanding* upon a study's conclusions by new experiments that elucidate additional information or deepen understanding; or (3) *specializing* the knowledge to a specific domain by elaborating on experimental conclusions. To demonstrate these three methods at the intersection of vision science and visualization, we highlight a number of replicated works of Cleveland and McGill's graphical perception paper [4], whose objectives are similar to the original ranking of quantitative judgment effectiveness of graphical encodings. We review their innovative contributions and contribution to replication-based validation. Finally, we discuss our perspective on how vision scientists can use this information for producing replication works in the current publication environment.

2 BACKGROUND

The visualization field is "an empire built on sand" with a weak foundation and in need of replication to strengthen many of the assumptions used by the field [21]. In a recent study, Kosara and

*e-mail: ghulamjilani@mail.usf.edu
†e-mail: prosen@usf.edu

[1] https://eagereyes.org/blog/2018/eurorvvv-call-for-papers
[2] https://replichi.org
[3] https://beliv-workshop.github.io

Haroz pointed out the *replication crisis*—very few replication studies are attempted in visualization, let alone published [22]. They examined six threats to the validity of studies in visualization and provided suggestions for replications, like outlining study design flaws and understanding or re-running misinterpreted results. Their suggestions help to minimize the threats to producing scientifically sound work.

Sukumar et al. provided guidelines for experimental design to encourage researchers to conduct more replication studies in information visualization [37]. They provided a list of possible experimenter biases that can occur related to devising hypotheses, independent and dependent variables, tasks, experimental procedures, sampling, experimenter behavior, and experimental setting, and they focused their discussion on designing and running sound experiments.

Hornbaek et al. developed a prescriptive definition for replication studies [13]. They stated that a replication must name and reference the original work, and they must state how their work confirms or extends the prior study. They further distinguished the replicated work with three categories—*strict*, *partial*, and *conceptual*—based on their literature review. A similar type of distinguished category can be found in Kosara and Haroz work—*reanalysis*, *direct replication*, *conceptual replication* [22]. These characterizations vary by the amount of originality from previous work that has been kept intact. *Strict replication* uses the same variables with the intent to reproduce the exact same study. Strict replications usually only replicate and confirm the original findings, which is what draws reviewers reluctance. *Partial replication* modifies the original study for testing within a different environment or with different participants. *Conceptual replication* studies investigate the same study but with different metrics, settings, or judgment criteria.

We distinguish our contribution from previous studies [13, 22, 37] by focusing on a taxonomy that does not focus on the quantity of overlap or similarity of study design in replication studies, but instead it focuses on the *types of novel contributions* that are associated with the replication studies.

3 Novelty and Replication

Hornbaek et al.'s work [13], along with the Kosara and Haroz work [22], essentially considered the level of similarity to the original study when classifying replication. As a complementary measure to that, we consider classifying replication studies by the *objective* of the novel contributions of the study. By understanding the types of novel contributions that blend with replication studies, other researchers, like those from vision science, can mimic the contribution styles to produce their own replication studies.

3.1 Taxonomy of Replication

Our evaluation of prior work shows that the vast majority of replicated work in information visualization falls within one of the three following categories.

3.1.1 Re-evaluate an Experiment's Objective

As practices change, software and hardware advance, and new techniques or information become available, some research studies whose conclusions were once considered solid require re-evaluation using these new contexts. Re-evaluation studies confirm the findings of an original study with different environment setting, while attempting to reproduce the objectives as closely as possible [13]. These replications help to establish if results from the prior study can be repeated to increase confidence in its validity. For example, Kosara and Ziemkiewicz replicated [23] their own earlier studies [2, 36] on pie charts using Amazon Mechanical Turk in order to re-evaluate whether an online environment produced the same results. Additional examples of this type of replication study are [11, 18].

3.1.2 Expand an Experiment's Objective

Due to the efforts required for performing human studies, the scope of studies is often kept small, leaving the need to expand conclusions with followup studies. Replication studies can be expanded beyond the objectives and conclusions of the prior studies by conducting themselves under different experimental conditions to make a novel contribution. By experimenting under different conditions, these replication studies serve as alternative means to validate prior results or as a means to generalize results [10, 15, 20]. For example, Rensink and Baldridge investigated the perception of correlation in scatterplots and suggested that Weber's law was useful for modeling it [32]. A replication study by Harrison et al. supplemented new conditions, in order to broaden the scope into a number of additional visualization types [10] (which was itself extended further through re-evaluation type paper [18]). Some additional examples of these types of replication studies are [19, 28, 31].

3.1.3 Specialize an Experiment's Objective

Studies often result in generalizable conclusions that need to be studied in new or more specific contexts than those of the original study. By taking the original study as the fundamental base, these works consider if and how much of the knowledge acquired in the original study is transferable to a new, different, or more specific domain. Performing these replication studies of different contexts under similar conditions with respect to setting, experiment environments, or metrics, aims to establish the validity within that specialized domain. Rensink and Baldridge's [32] study on modelling correlation perception was also replicated by Yang et al. [44] in order to further investigate visual features in modelling perceptual processes. The objective of finding perceived correlation in a scatterplot is synonymous with perceiving its visual features and quite unrelated to one's statistical training. The results of this replication study showed how visual features provide a baseline for model-approaches in visualization evaluation and design. Additional examples of this type of replication study can be found in [12, 39].

4 Case Study

With a focus on perceptual work in information visualization, we selected the seminal and highly cited work by Cleveland and McGill on graphical perception [4] to highlight examples of replication studies. Graphical perception is an anchor of visualization research [42] with the potential to improve the efficiency and effectiveness of the automatic representation of data [25]. We have observed the vast influence of this paper in last two decades of information visualization research. According to Google Scholar, this work has been cite more than 1585 times as of July 1, 2019, and aspects of their study have been replicated many times.

4.1 Graphical Perception

Research has demonstrated viewer's perceptual judgment significantly influences effective visualization design. The representa-

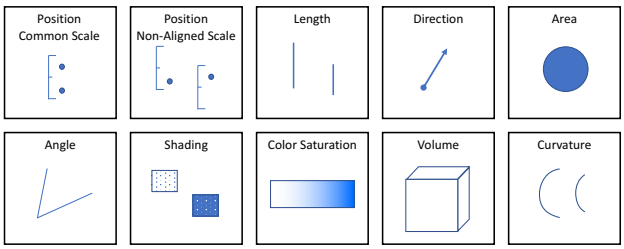

Figure 1: Reproduction of the Cleveland and McGill's graphical encoding channels [4].

Table 1: Selected publications that are representative replication studies of Cleveland and McGill [4].

Publication	Venue	Year	Hornbaek et al.'s Category [13]	Replication Type	Types of Graphical Perception (from Fig. 1)
Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design [11]	ACM CHI	2010	Conceptual	**Re-evaluate**	Length, Position & Angle
Evaluating Interactive Graphical Encodings for Data Visualization [34]	IEEE TVCG	2018	Conceptual	**Re-evaluate**	12 encodings (interactive encodings)
Graphical Perception of Continuous Quantitative Maps: The Effect of Spatial Frequency and Color Map Design [31]	ACM CHI	2018	Conceptual	**Re-evaluate**	Color
The Impact of Social Information on Visual Judgments [14]	ACM CHI	2011	Partial	**Expand**	Length, Position & Area
Influencing Visual Judgment through Affective Priming [9]	ACM CHI	2013	Conceptual	**Expand**	Length, Position & Angle
Learning Perceptual Kernels for Visualization Design [5]	IEEE TVCG	2014	Partial	**Expand**	Shape, Color & Size
Sizing the Horizon: the Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations [12]	ACM CHI	2009	Partial	**Specialize**	Chart Height & Layering
Four Experiments on the Perception of Bar Charts [40]	IEEE TVCG	2014	Partial	**Specialize**	Length & Position (for bar charts)

tion of data in a visualization is encoded with specific elements on the display, also known as the graphical encodings, include *position*, *length*, *angle*, *area*, *volume*, *shading*, *direction*, *curvature*, and *color* [30]. Fig. 1 represents these 10 elementary *perceptual tasks*[4] from Cleveland and McGill's work that people use to extract quantitative information from graphs.

Understanding the role of perception in the choice of graphical encodings is critical to visualization designers. Based on 10 common graph types—distribution function plots, bar charts, pie charts, divided bar charts, statistical map, curve-difference charts, Cartesian graphs, triple scatterplots, volume charts, and juxtaposed Cartesian graphs—Cleveland and McGill ranked these perceptual tasks by the accuracy of quantitative information extraction. Mackinlay produced one of the earliest comprehensive rankings of graphical encodings by data type, as shown in Fig. 2 [25]. The ranking has been further validated and elucidated through followup (replication) studies [5, 7, 11, 12, 26, 31, 34, 35, 39].

4.2 Example Replication Studies

We surveyed major visualization publication venues (IEEE TVCG, ACM CHI, EG EuroVis, and IEEE PacificVis) for papers that cited and then replicated aspects of Cleveland and McGill's work. Our non-comprehensive analysis is primarily limited to the eight papers listed in Table 1. For each, we provide the objective of the replication study, followed by a high-level description of the novelty added to the replication, which contributed to the visualization research.

4.2.1 Re-evaluate an Experiment's Objective

Heer and Bostock's graphical perception study replicated ranking on effectiveness of visual encodings, such as length, position, and angle, using an alternative crowdsourced subject pool on Amazon Mechanical Turk (AMT) [11]. Further, they extended the study on additional encodings, such as circular area (e.g., bubble charts) and rectangular area (e.g., treemaps). The ranking order that resulted from their studies were similar to Cleveland and McGill's rankings. The main novelty of this paper was not to validate the findings of Cleveland and McGill, but to test the viability of online user study like crowdsourcing. The results showed that AMT could serve as a viable user study platform for visualization research.

Another replication of Cleveland and McGill was carried out by Saket et al. on 12 graphical encodings to study their effectiveness in terms of task completion time and accuracy, when using them for interaction [34]. The objective of replication was same as that of original study but re-evaluated on *interactive* graphical encodings. Their ranking followed and confirmed the findings of Cleveland and McGill's, except for a significant difference between length and angle in terms of accuracy.

4.2.2 Expand an Experiment's Objective

Hullman et al. [14] and Harrison et al. [9] replicated the study of Cleveland and McGill but ranked their quantitative judgement effectiveness on the basis of social information[5] and affective priming[6], respectively. The AMT-based perceptual studies demonstrated social information and affective priming can significantly influence user's visual judgment [9, 14]. The findings on social information can be applied to collaborative visualization systems to produce more accurate results on individual interpretations in a social context, and the findings on affective priming showed that it can influence accuracy in common graphical perception tasks.

The concept of graphical perception has extended to various branches of visualization—maps, color, visual properties, etc. Another replication of Cleveland and McGill, towards color mapping on continuous maps, found that spatial frequency significantly impacts the effectiveness of color encodings [31]. The granular level of novelty to this work is based on conceptual replication of previous work and applying it to a new domain of continuous maps.

A partial replication of Cleveland and McGill studied how visual encodings, such as color, shape and size, affect a user's way of

[4]The term *perceptual task* originates from the concept that viewer performs a mental task to extract quantitative values represented on graphs [4].

[5]Social information represents the creation and processing of information from multiple features by a group.

[6]Affective priming is the impact of emotional biases, whose study involves manipulating valence and/or arousal via emotional stimuli.

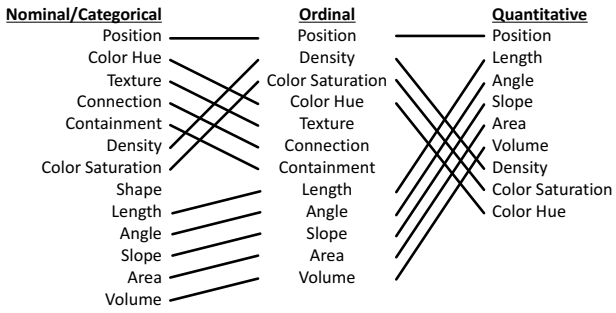

Figure 2: Reproduction of Mackinlay graphical encoding rankings [25].

interpreting data [5]. A perceptual kernel[7] is estimated for set of perceptual stimuli, based on size, shape, color, and combinations, to assess the effectiveness of visual representations from reported results of crowdsourced experiments. They compared six stimuli using different set of judgment types—Likert rating among pairs, ordinal triplet comparisons, and manual spatial arrangements—to existing perceptual kernels and demonstrated how kernels can be applied to automate visualization design decisions. The novel contribution from this replication work is fixed on similarity judgement using perceptual kernels, extending the concept of the prior work to a particular domain.

4.2.3 Specialize an Experiment's Objective

Talbot et al.'s [40] study on bar charts provides insights on comparing adjacent, separated, aligned, and non-aligned bars in types of bar charts. This replication used the foundational work of Cleveland and McGill to focus on perceptual tasks related to bar charts only. Distractors (i.e., intermediate bars between bars being compared) affect the comparison between types of bar charts, and inconsistent placement of marking dots in the original study affect user accuracy on perceptual tasks.

In similar fashion, the specialized domain replication work of Heer et al. [12] extended the concept of graphical encoding effect of user judgement to the specialized area of chart height and layering on time and accuracy of a value comparison task. Their findings on estimation accuracy across charts identified transition points in smaller charts, where accuracy and estimation time decreases with size.

For all of the studies mentioned, the additional novelty of the paper helped it stand out beyond just the replication—the studies verified previous results, but they also covered a larger parameter space and/or came to different conclusions than the original study.

5 Discussion

The overwhelming value of reproducibility is undeniable—from the ability to have third party verification of claims and conclusions, to the development of new insights expanding upon old conclusions, to understanding changes in user expectations and technologies with respect to prior study findings. While general effort toward improved reproducibility have had some success (slow progress but trending in the right direction), replication studies have remained somewhat in the shadows. Further attention needs to be paid to this particular area of reproducibility. One natural avenue is for vision scientists interested in applying their expertise to the area of visualization to contribute replication studies that either re-evaluate, expand, and/or specialize previous studies.

5.1 You Can't Publish (Strict) Replication Studies

As long as novelty remains a necessary contribution that reviewers in the visualization community acknowledge, the simple fact is that it will be incredibly difficult to publish strict replication studies. By their nature, the contribution of strict replication studies is not novelty, thus the reluctance from the reviewers to acknowledge any value. Though some may exist, we did not find any strict replication studies of Cleveland and McGill during our literature survey. *Our opinion is that it remains in the best interests of researchers, in vision science or otherwise, to avoid trying to publish any strict replication study.*

In many ways, the struggles of replication parallel those of application papers in visualization. When application papers are discussed in the halls of the IEEE VIS conference, everyone agrees on their value and wishes there was more acceptance for them. As soon as those individuals review an application paper, a stamp of 'limited novelty' (the review equivalent of the 'kiss of death') is applied and the paper is promptly rejected. It is only by wrapping the application in novelty that these papers are ever accepted in the main conference tracks. A variety of attempts have been made to correct for this issue, most have been failures. The introduction of *Application Spotlights* this year is the latest attempt, whose success has yet to be determined. The point is, *those looking to promote reproducibility and replication studies should look to the history of application papers for some insights as to what approaches are likely (and unlikely) to succeed moving forward.*

5.2 How to Publish Replication Studies Anyways

In this paper, we argue that *the best way for individual researchers to publish replication studies is to distinguish the work with the help of added novelty*. Considering partial and conceptual style replications bridges the gap between original work and the innovations required for publications. For those new innovations, we demonstrated how prior works used *re-evaluation*, *extension*, and *specialization* to help frame their novel contribution around the replication study, enabling the replication to provide value to the overall contribution of the paper.

For vision researchers new to visualization, the concept of replication can serve as a bridge into the field. Replication enables the researcher to become deeply familiar with a specific visualization topic, while contributing to the field. Taking the replication as a base, *re-evaluating*, *expanding*, or *specializing* enables them to contribute added novelty to the field. In this way, the familiarization process (i.e., the replication study) has both personal and community value. For example, re-evaluation work can be used as a foundation for a researcher introducing themselves to visualization community, or an expansion can be used to evaluate conclusions under different conditions, not previously considered. However, we believe specializing represents the best opportunity. Vision scientist can leverage their in deep knowledge of human perception in efforts to replicate and specialize prior visualization studies, thereby bringing innovative ideas to the visualization community. In similar fashion, findings from the vision science community can be replicated and specialized into the context of visualization. In the end, visualization replication studies by vision scientists represent a win-win. First, it engages both communities in a dialog that advances knowledge in both communities. Second, we all benefit when the experimental conclusions we rely on are re-validated and further elaborated upon.

Replication does not necessarily need new experimentation. Already completed studies can be utilized to generate new state-of-the-art work. Researchers can make use of available experimental data, study designs, and comparable results from original studies in order to formulate new high-level research questions. A *re-evaluation* of an earlier study can reduce the possibility that conclusions are the result of a statistical fluke, flawed analysis, or a flaw in the study design [22]. New perspective on the analysis of that data can be used to *expand* or *specialize* the domain of the work.

Finally, it is important to recognize that many non-replication studies over the years could have corroborated the conclusions of earlier studies by performing small replication studies of key findings. We are not necessarily advocating that all or even the majority of prior studies should be replicated. However, *reviewers could encourage more replication by allocating "bonus points" in reviews containing some form of replication.* We are already seeing this with reproducibility in general. In a personal communication with one InfoVis paper chair, they noted a measurably higher average score for papers that included their data in the submission.

Acknowledgments

We would like to thank our reviewers for their valuable feedback on our work. This project is supported in part by National Science Foundation (IIS-1513616 and IIS-1845204).

[7] Perceptual kernel is a distance matrix derived from aggregate perceptual judgments. It contains pairwise perceptual dissimilarity values for a specific set of perceptual stimuli.

REFERENCES

[1] C. Bazerman. Book Review of "Changing Order: Replication and Induction in Scientific Practice" by H. M. Collins. *Philosophy of the Social Sciences/Philosophie des Sciences Sociales*, 19(1):115, 1989.

[2] E. Bertini, N. Elmqvist, and T. Wischgoll. Judgment Error in Pie Chart Variations. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*, pp. 91–95, 2016.

[3] R. Bronstein. Publication Politics, Experimenter Bias and the Replication Process in Social Science Research. *Journal of Social Behavior and Personality*, 5(4):71, 1990.

[4] W. 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.

[5] Ç. Demiralp, M. Bernstein, and J. Heer. Learning Perceptual Kernels for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1933–1942, 2014.

[6] 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, 2017.

[7] C. Gramazio, K. Schloss, and D. Laidlaw. The Relation Between Visualization Size, Grouping, and User Performance. *IEEE Transactions on Visualization and Computer Graphics*, 2014.

[8] S. Greenberg and B. Buxton. Usability Evaluation Considered Harmful (Some of the Time). In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 111–120, 2008.

[9] L. Harrison, D. Skau, S. Franconeri, A. Lu, and R. Chang. Influencing Visual Judgment through Affective Priming. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 2949–2958, 2013.

[10] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014.

[11] J. Heer and M. Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *ACM SIGCHI: on Human Factors in Computing Systems*, pp. 203–212, 2010.

[12] J. Heer, N. Kong, and M. Agrawala. Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2009.

[13] K. Hornbæk, S. Sander, J. A. Bargas-Avila, and J. Grue Simonsen. Is Once Enough?: On the Extent and Content of Replications in Human-Computer Interaction. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 3523–3532, 2014.

[14] J. Hullman, E. Adar, and P. Shah. The Impact of Social Information on Visual Judgments. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1461–1470, 2011.

[15] M. Jakobsen and K. Hornbæk. Interactive Visualizations on Large and Small Displays: The Interrelation of Display Size, Information Space, and Scale. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2336–2345, 2013.

[16] B. Jasny, G. Chin, L. Chong, and S. Vignieri. *Again, and Again, and Again...* American Association for the Advancement of Science, 2011.

[17] K. Jones, P. Derby, and E. Schmidlin. An Investigation of the Prevalence of Replication Research in Human Factors. *Human Factors*, 52(5):586–595, 2010.

[18] M. Kay and J. Heer. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):469–478, 2015.

[19] S.-H. Kim, Z. Dong, H. Xian, B. Upatising, and J.-S. Yi. Does an Eye Tracker Tell the Truth About Visualizations?: Findings While Investigating Visualizations for Decision Making. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2421–2430, 2012.

[20] N. Kong, J. Heer, and M. Agrawala. Perceptual Guidelines for Creating Rectangular Treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):990–998, 2010.

[21] R. Kosara. An Empire Built on Sand: Reexamining What We Think We Know About Visualization. In *Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 162–168, 2016.

[22] R. Kosara and S. Haroz. Skipping the Replication Crisis in Visualization: Threats to Study Validity and How to Address Them: Position Paper. In *IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, pp. 102–107, 2018.

[23] R. Kosara and C. Ziemkiewicz. Do Mechanical Turks Dream of Square Pie Charts? In *Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 63–70, 2010.

[24] C. Lallemand, V. Koenig, and G. Gronier. Replicating an International Survey on User Experience: Challenges, Successes and Limitations. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2013.

[25] J. Mackinlay. Automating the Design of Graphical Presentations of Relational Information. *ACM Transactions On Graphics (TOG)*, 5(2):110–141, 1986.

[26] L. Neumann, K. Matkovic, and W. Purgathofer. Perception Based Color Image Difference. *Computer Graphics Forum*, 17(3):233–241, 1998.

[27] W. Newman. A Preliminary Analysis of the Products of HCI Research, Using Pro Forma Abstracts. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, vol. 94, pp. 278–284, 1994.

[28] A. Ottley, E. M. Peck, L. Harrison, D. Afergan, C. Ziemkiewicz, H. Taylor, P. Han, and R. Chang. Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):529–538, 2015.

[29] R. Peng. Reproducible Research and Biostatistics. *Biostatistics*, 10(3):405–408, 2009.

[30] S. Pinker. A Theory of Graph Comprehension. *Artificial Intelligence and the Future of Testing*, pp. 73–126, 1990.

[31] K. Reda, P. Nalawade, and K. Ansah-Koi. Graphical Perception of Continuous Quantitative Maps: The Effects of Spatial Frequency and Colormap Design. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, p. 272, 2018.

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

[33] R. Rosenthal. Replication in Behavioral Research. *Journal of Social Behavior and Personality*, 5(4):1, 1990.

[34] B. Saket, A. Srinivasan, E. D. Ragan, and A. Endert. Evaluating Interactive Graphical Encodings for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(3):1316–1330, 2018.

[35] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A Taxonomy of Visual Cluster Separation Factors. *Computer Graphics Forum*, 31(3pt4):1335–1344, 2012.

[36] 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.

[37] P. Sukumar and R. Metoyer. Towards Designing Unbiased Replication Studies in Information Visualization. In *IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, pp. 93–101, 2018.

[38] P. T. Sukumar and R. Metoyer. Replication and Transparency of Qualitative Research from a Constructivist Perspective. *OSF Preprints*, 2019.

[39] D. Szafir. Modeling Color Difference for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, 2018.

[40] J. Talbot, V. Setlur, and A. Anand. Four Experiments on the Perception of Bar Charts. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2152–2160, 2014.

[41] A. C. Valdez, A. K. Schaar, J. R. Hildebrandt, and M. Ziefle. Requirements for Reproducibility of Research in Situational and Spatio-Temporal Visualization: Position Paper. In *IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, pp. 53–59, 2018.

[42] C. Ware. *Information Visualization: Perception for Design*. Elsevier, 2012.

[43] M. Wilson, E. Chi, S. Reeves, and D. Coyle. RepliCHI: The Workshop II. In *ACM SIGCHI Conference on Human Factors in Computing Systems (Extended Abstracts)*, pp. 33–36, 2014.

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