Chapter 12
Open Challenges in Empirical Visualization Research

Caroline Ziemkiewicz, Min Chen, David H. Laidlaw, Bernhard Preim and Daniel Weiskopf

Abstract In recent years, empirical studies have increasingly been seen as a core part of visualization research, and user evaluations have proliferated. It is broadly understood that new techniques and applications must be formally validated in order to be seen as meaningful contributions. However, these efforts continue to face the numerous challenges involved in validating complex software techniques that exist in a wide variety of use contexts. The authors, who represent perspectives from across visualization research and applications, discuss the leading challenges that must be addressed for empirical research to have the greatest possible impact on visualization in the years to come. These include challenges in developing research questions and hypotheses, designing effective experiments and qualitative methods, and executing studies in specialized domains. We discuss those challenges that have not yet been solved and possible approaches to addressing them. This chapter provides an informal survey and proposes a road map for moving forward to a more cohesive and grounded use of empirical studies in visualization research.

C. Ziemkiewicz (✉)
Forrester Research, Inc, Cambridge, MA, USA
e-mail: cziemkiewicz@forrester.com

M. Chen
University of Oxford, Oxford, UK

D. H. Laidlaw
Brown University, Providence, RI, USA

B. Preim
Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany

D. Weiskopf
University of Stuttgart, Stuttgart, Germany

M. Chen et al. (eds.), *Foundations of Data Visualization*,
https://doi.org/10.1007/978-3-030-34444-3_12

12.1 Introduction

The visualization field has long had a complex relationship with empirical validation. In the 2000's, as it was quickly growing from a niche graphics subfield into a major research area of its own, there was a proliferation of reports and panels on the major unsolved challenges in visualization. A common theme in these challenges was a need to reliably prove the value of visualization. For example, Keim et al. [16] noted the issue of user acceptability; if domain users did not see how visualization could help them, they would not adopt it, and so there was no way to test whether visualization helped them. In a report directed at national funding agencies in the USA, Johnson et al. [14] cite similar challenges in demonstrating value and involving domain scientists in research. This example suggests some of the practical context behind this push for validation. As visualization researchers sought support for their work, it was necessary to find objective metrics that could show the value of their methods.

With this backdrop, evaluation in visualization has historically focused on user studies that either measure the effectiveness of visualization versus a traditional method or the relative effectiveness of two or more visualization techniques. However, while our overall approach to empirical research has remained much the same, the context around it has changed dramatically. Visualization has been adopted widely in commercial and government settings. As "big data" became a household term, the value of visualization came to be broadly understood as an efficient interface between people and information. Empirical visualization research has yielded general guidelines that are familiar to commercial designers outside the research community.

In this new context, it may be necessary to revisit the role of empirical research in visualization. In a world where visualization is assumed to have value, demonstrating that a visualization is usable may no longer be sufficient validation. In a moment like this, it is worthwhile to look back at what challenges have been addressed and which remain open. In 2004, Plaisant [22] identified what were then the major challenges in visualization evaluation. In some cases, the visualization community has made substantial progress in these challenges: for example, building task taxonomies [2, 5, 25], adding to the variety of evaluation approaches [12, 21, 27], and using contests to develop benchmark problems and datasets [23].

However, there are other challenges named in 2004 that remain unsolved today. Even as visualization grows more popular, the core problem of motivating domain users to buy into research continues to be a limitation. As a community, while we have developed more techniques for evaluation, we have not consistently established best practices for either research methodology or experimental stimulus design. Researchers continue to face challenges in controlling the experimental parameters in study design. It is possible that many of these issues could be addressed by a greater understanding of related psychological fields, but incorporating that understanding is a nontrivial exercise. Each of these challenges faces unique obstacles, but there is promising work that points to possible ways of addressing them.

12.2 Challenge 1: Motivating Domain Experts

12.2.1 Current Challenges

The difficulty of substantially involving domain experts in research has perhaps been the most cited challenge in visualization evaluation, and remains as relevant today as ever. Empirical research strongly benefits from realistic assessments of current technology as well as from realistic evaluations of research prototypes. In the early stages of development, the observation of experts solving real problems is essential to understand workflows, processes, constraints, and non-routine factors that would not be detected with questionnaires or interviews carried out at places distant from the working environment.

Similarly, empirical evaluations of a prototype strongly benefit from a high degree of realism. If they are carried out at the workplace of the domain experts and serve to solve real tasks, domain experts are fully motivated. In contrast, if artificial or archival data are used, the motivation is lower. Additionally, many visualization methods are aimed at niche user groups with advanced training, such as scientists, medical professionals, and analysts. The tasks these users engage in are frequently complex and involve learning, problem solving, or decision making. However, many empirical studies use simplified low-level tasks, such as basic perceptual tasks, searching for data, navigation, or routine activity. Evaluations in the developer's laboratory using such abstracted low-level tasks are much simpler to carry out, but often the results have at best a very indirect relation to the true activities of users.

Apart from designing abstracted studies outside of the domain context, another common approach to this problem is to involve domain experts briefly at key points in the process. For example, a researcher might develop a tool, then have a domain expert to evaluate it using an interview or other form of qualitative feedback. While this can be a way to work around the domain expert's schedule, it asks the user to evaluate something for which they have no prior context. A pair of surveys of evaluation methods used in visualization papers argues for a systematic lack of process evaluation methods such as requirements gathering and analysis of user workflows [13, 18]. Without this key context, tools are likely to be disconnected from the user's work context, and the value of their feedback may be limited.

12.2.2 Possible Approaches

One of the reasons for the systematic lack of process evaluation methods is that they are difficult to publish except as part of a lengthy design study. An approach to address this problem may be to create venues for such papers, for example, by designing a workshop around them or by introducing a new paper type. Another possibility would be to investigate methods that combine controlled and uncontrolled empirical methods; for example, contextual inquiry, and observational studies in a laboratory environment [18]. Ultimately, as Sedlmair et al. [26] point out, adoption of a system in the field is a problem to approach at the organizational level, not on the level of individual end users. Visualization researchers who wish to motivate domain experts must learn to observe and integrate with the experts' environment and work context.

12.3 Challenge 2: Systematic Lack of Research Methodology Skills

12.3.1 Current Challenges

A major factor that limits the effectiveness of empirical research in visualization is that visualization researchers, especially those from a computer science background, are not guaranteed to be trained in basic human subjects research methodology. In the field of psychology, from which visualization researchers often borrow approaches, there is no shortage of researchers who have been trained to design and conduct controlled empirical studies and formal qualitative research. In the field of visualization, the number of researchers who have had direct experience in designing and conducting empirical studies is significantly smaller. Computer science education does not prioritize these skills, as evidenced by the fact that user-centered design and research methods are not included as part of the core computer science curriculum [15]. This lack of skilled resources means that it is impossible to conduct large numbers of high-quality studies on any given topic, leaving many core research questions unanswered.

This skill deficit is reflected in visualization research in a number of ways. One of the most widespread is a lack of detailed and consistent statistical reporting of empirical results [6, 7, 13]. Researchers who present studies without using the appropriate statistical tests, making corrections for multiple comparisons, or reporting effect sizes not only limit the impact of their own work but make it difficult or impossible to produce meta-analyses and surveys. Moreover, it is still not uncommon to see papers with evaluations that consist only of unstructured feedback from a small number of experts. Contributing to this problem is a broad lack of knowledge about qualitative methods that leads to confusion between qualitative research and informal feedback-gathering [13].

This problem affects all of visualization but can be especially difficult in scientific visualization (SciVis), where researchers are less likely to come from a human–computer interaction (HCI) background. SciVis research often requires specialized algorithmic knowledge, and the social context of computer science education frequently puts distance between these "hard" algorithmic skills and the "soft" skills of user research. SciVis researchers face the additional challenge of balancing collaboration. Information visualization researchers dealing with generic or broadly understood data may forego domain collaborators in favor of psychologists or HCI experts, but SciVis researchers almost always need to collaborate with experts from a scientific domain. Coordinating multiple collaborations, especially among in-demand experts, carries significant risks. As a result, teams including SciVis researchers, domain experts, and empirical research specialists remain relatively rare.

12.3.2 Possible Approaches

While visualization researchers understand the value of collaboration with experts in empirical methods, this does not always translate into active participation in such collaborations. Providing specific funding incentives has the potential to push these partnerships forward. As an example, cooperation between visual analytics and data analysis in Germany was initiated by a national research priority program on Scalable Visual Analytics which encouraged collaborations between both fields and between funded projects [17]. To address the skills gap within the community, one possibility is to revise standards for computer science curricula to include user-centered research as a core topic [15]. A more immediate action could be to compile a community portal to collect resources on empirical methods, similar to efforts such as The Fluid Project [1] but tailored to the specific needs of visualization researchers.

12.4 Challenge 3: Data Collection and Generation

12.4.1 Current Challenges

Although visualization researchers have made considerable progress in recent years in developing formal taxonomies and models of evaluation tasks, there has been less emphasis on developing repeatable approaches to data generation. In a field where the nature of the data can considerably change the effectiveness of the method being tested, unrealistic data is a serious threat to ecological validity. Examples include data at a scale much smaller than would be encountered in real tasks, data that lacks the errors and inconsistencies common to real datasets, and data with strong statistical patterns that might not normally be present. While benchmark datasets are useful for comparison, they often do not capture these real-world data challenges.

At the same time, real-world datasets can be difficult to collect and use for a variety of reasons, such as privacy, size, protection of proprietary information, or legal restrictions on dissemination [26]. A common approach in such situations are to build sanitized datasets by removing or perturbing sensitive information. However, security research has shown that even sophisticated privacy-preserving data mining methods can be vulnerable to re-identification, especially in cases where multiple data sources can be combined [20]. Even in cases where real data can be used as-is, it can be difficult to generalize evaluation results from a single dataset. Finding multiple datasets that represent a realistic range of conditions only compounds the problem.

Generative data models can be an effective approach to this problem, but they require careful design to avoid biases [24]. A generative data model can be used to automate the generation of multiple datasets with desired properties, which can address the issue of testing against multiple valid datasets to support generalization. There are a number of significant technical challenges associated with such models; many involve complex simulations, and as most models are developed for one-off cases, standardized techniques and replication are rare. Moreover, interactions between a generative model and a visualization technique can be difficult to predict. There is no guarantee that a model that produces data with desirable characteristics will still have the same characteristics after being processed as part of a given visualization algorithm.

12.4.2 Possible Approaches

The type of formalization that has been applied to tasks and visual representations in recent years has helped to produce more rigorous and controlled experiments. However, the way we describe data is still most often in the terms used by Jacques Bertin fifty years ago [3]. More specialized typologies of data that take into account contemporary concerns such as scale, heterogeneity, and uncertainty could go a long way toward defining a design space in which datasets used in experiments can vary. Generative models have the potential to address many problems in data collection, but the field will advance more quickly if designers of generative models adopt open practices and make models available for replication and benchmarking. As more such models are made available, it will be possible to identify best practices and guidelines for further development [24].

12.5 Challenge 4: Experimental Design Space and Tradeoffs

12.5.1 Current Challenges

At the core of many of the challenges in visualization evaluation is that it involves the combination of two highly complex systems: the human user and the data visualization system. In such a situation, the number of experimental variables that must be controlled can quickly become unmanageable. The skills deficit discussed in Sect. 12.3 compounds this problem, as there is a lack of institutional knowledge about how to balance tradeoffs and control variables in experimental design. This leads to a number of issues affecting the ecological validity of experiments as well as the ability of other researchers to evaluate and make use of experimental results.

One common problem is when the assumption that a system should be evaluated in one experiment leads to overstuffed design. In some ways, this problem has been exacerbated by the increased push for no system to go unevaluated. While this is an admirable goal, in practice, treating evaluation as a box that must be checked often leads to user studies that either lack a clear hypothesis or attempt to test too many hypotheses at once. Such user studies often suffer from a mismatch in validation method to type of contribution; for example, a paper whose primary contribution is a novel visual encoding does not necessarily require a task-based evaluation, provided the authors make no claims about improving performance on that specific task [19]. Nonetheless, user studies remain common in such situations, often using ad hoc tasks that have not been rigorously designed.

Knowledge of appropriate design space tradeoffs also affects the quality of reviewing. Lack of familiarity with empirical methods is one issue, but partial familiarity can cause its own share of problems. A reviewer with knowledge of only one method may apply the rigor metrics of that method to an unrelated one, leading to inappropriate evaluations [6]. For example, a researcher who uses qualitative methods may receive criticism for not including statistical analyses suited for quantitative methods. A better understanding of the experimental design space, and an acknowledgement that no one study can cover it exhaustively, remains elusive.

12.5.2 Possible Approaches

In psychology and related disciplines, it is common to publish a series of related studies in a single publication, which each experiment building on the knowledge gained in the previous one. Such a structure allows researchers to produce more tightly controlled individual study designs while still approaching a larger research question. While linked studies of this type are sometimes seen in visualization perception research [4, 11, 29], it may also be a useful method for technique or system evaluation. In this model, user evaluations may even be published separately from the system itself, which in many cases may require more limited validation methods. In order to improve the control of variables in study design, one possibility is to publish and promote evaluation checklists, a method that has been used effectively in other domains [10].

12.6 Challenge 5: Engagement with Relevant Psychology Fields

12.6.1 Current Challenges

As in other human-centered computer science disciplines, understanding visualization depends heavily on understanding the people who use it. Psychology is a key component of any empirical research in visualization. Yet explicit engagement with psychology research remains infrequent outside of a few specific areas, such as research on color scale design [28]. This can lead to findings that are divorced from important context. An experiment on how well a user can remember information in a particular visual representation must take into account the expected performance of visual working memory in general; a field study observing adoption of a system cannot be generalized without a working knowledge of how quickly new technology is usually adopted in workplaces. Perhaps the most pervasive example of such issues is the widespread assumption that the effectiveness of a visual representation can be generalized between users without taking into account natural variation in spatial ability and other cognitive factors [30].

This lack of engagement with psychology also causes issues when it leads to ignorance of challenges in psychological research that affect visualization researchers as well. The difficulty of integrating knowledge gained in increasingly specialized subfields was named “The Grand Challenge” of psychology by Axel Cleeremans of Université libre de Bruxelles in 2010 [8]. Clearly this is a concern for visualization as well, as community discussion at the 2018 IEEE VIS Conference centered around the problem of unifying the diverging fields of scientific visualization, information visualization, and visual analytics. Visualization researchers are also just beginning to take notice of the replication crisis in psychology [9], but have yet to adopt the reforms made by psychologists in its wake.

These challenges themselves can create pitfalls for outside researchers looking to make use of psychological findings or methods. The complexity of psychology’s many disparate fields, and lack of communication between these fields [8], can obscure important connections and make it difficult to know where to start looking for answers. Visualization researchers very often know that psychology is important to their work, but without clear goals and an understanding of the research space, it is rare for sustained productive conversation to happen between the two disciplines.

12.6.2 Possible Approaches

In order to engage more fully with psychology research, it may be necessary to modernize our research practices to meet the changes made by psychologists in recent years. For example, adopting open science practices, especially sharing data and code (where possible), would be a positive step for the visualization field on its own. But it could also foster collaboration by making materials and tools available to psychology researchers themselves. In some cases, these experts may have visualization needs that our community is unaware of, and a greater degree of communication may help reveal them. This can be a challenging process, as publication cultures and research goals will vary across fields. Work to identify common ground and mutual goals will be a necessary first step. We can also learn from psychology now by addressing some of the known issues that affect both fields; for example, submitting research reports ahead of performing experiments in order to reduce positive effect bias.

12.7 Conclusion and Next Steps

In this chapter, we have discussed five key challenges in empirical visualization research in detail and proposed possible approaches to addressing them. By doing so, we hope to build on the successes of the past in developing a research agenda for the future. It is vital to note the areas in which we have made progress as well as those where challenges remain. Empirical studies in visualization have advanced in many ways over the past decade, as has visualization itself. But even as the value of visualization becomes more broadly accepted, the current evaluation paradigm more often than not focuses on testing whether a visualization is generally effective or not. By addressing these challenges, we hope to make space for research that goes beyond this paradigm to answer more specific, contextualized, and meaningful questions that drive the future of visualization research.

References

1. Administrator, F.P.: Fluid project wiki. https://wiki.fluidproject.org
2. Amar, R., Eagan, J., Stasko, J.: Low-level components of analytic activity in information visualization. In: IEEE Symposium on Information Visualization, 2005. INFOVIS 2005, pp. 111–117. IEEE (2005)
3. Bertin, J., Berg, W.J., Wainer, H.: Semiology of Graphics: Diagrams, Networks, Maps. University of Wisconsin Press, Madison (1983)
4. Bezerianos, A., Isenberg, P.: Perception of visual variables on tiled wall-sized displays for information visualization applications. IEEE Trans. Vis. Comput. Graph. **18**(12), 2516–2525 (2012)
5. Brehmer, M., Munzner, T.: A multi-level typology of abstract visualization tasks. IEEE Trans. Vis. Comput. Graph. **19**(12), 2376–2385 (2013)
6. Carpendale, S.: Evaluating information visualizations. Information Visualization, pp. 19–45. Springer, Berlin (2008)
7. Chen, C., Yu, Y.: Empirical studies of information visualization: a meta-analysis. Int. J. Hum.-Comput. Stud. **53**(5), 851–866 (2000)
8. Cleeremans, A.: The grand challenge for psychology. APS Observer **23**(8) (2010)
9. Collaboration, O.S., et al.: Estimating the reproducibility of psychological science. Science **349**(6251), aac4716 (2015)
10. Crisan, A., Elliott, M.: How to evaluate an evaluation study? comparing and contrasting practices in vis with those of other disciplines. In: Proceedings of the 2018 Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization. IEEE (2018)

11. Heer, J., Bostock, M.: Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 203–212. ACM (2010)
12. Isenberg, P., Zuk, T., Collins, C., Carpendale, S.: Grounded evaluation of information visualizations. In: Proceedings of the 2008 Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization, p. 6. ACM (2008)
13. Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., Möller, T.: A systematic review on the practice of evaluating visualization. IEEE Trans. Vis. Comput. Graph. **19**(12), 2818–2827 (2013)
14. Johnson, C., Moorhead, R., Munzner, T., Pfister, H., Rheingans, P., Yoo, T.S.: NIH-NSF visualization research challenges report. Institute of Electrical and Electronics Engineers (2005)
15. Joint Task Force on Computing Curricula, A.f.C.M.A., Society, I.C.: Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science, p. 999133. ACM, New York (2013)
16. Keim, D.A., Mansmann, F., Schneidewind, J., Ziegler, H.: Challenges in visual data analysis. In: 10th International Conference on Information Visualization, IV 2006, pp. 9–16. IEEE (2006)
17. Konstanz, U.: Scalable visual analytics: Interactive visual analysis systems of complex information spaces. http://www.visualanalytics.de/node/2
18. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical studies in information visualization: seven scenarios. IEEE Trans. Vis. Comput. Graph. **18**(9), 1520–1536 (2012)
19. Munzner, T.: A nested process model for visualization design and validation. IEEE Trans. Vis. Comput. Graph. **15**(6), 921–928 (2009)
20. Narayanan, A., Shmatikov, V.: Myths and fallacies of personally identifiable information. Commun. ACM **53**(6), 24–26 (2010)
21. North, C.: Toward measuring visualization insight. IEEE Comput. Graph. Appl. **26**(3), 6–9 (2006)
22. Plaisant, C.: The challenge of information visualization evaluation. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 109–116. ACM (2004)
23. Plaisant, C., Fekete, J.D., Grinstein, G.: Promoting insight-based evaluation of visualizations: from contest to benchmark repository. IEEE Trans. Vis. Comput. Graph. **14**(1), 120–134 (2008)
24. Schulz, C., Nocaj, A., El-Assady, M., Frey, S., Hlawatsch, M., Hund, M., Karch, G., Netzel, R., Schätzle, C., Butt, M., et al.: Generative data models for validation and evaluation of visualization techniques. In: Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization, pp. 112–124. ACM (2016)
25. Schulz, H.J., Nocke, T., Heitzler, M., Schumann, H.: A design space of visualization tasks. IEEE Trans. Vis. Comput. Graph. **19**(12), 2366–2375 (2013)
26. Sedlmair, M., Isenberg, P., Baur, D., Butz, A.: Information visualization evaluation in large companies: Challenges, experiences and recommendations. Inf. Vis. **10**(3), 248–266 (2011)
27. Shneiderman, B., Plaisant, C.: Strategies for evaluating information visualization tools: multidimensional in-depth long-term case studies. In: Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization, pp. 1–7. ACM (2006)
28. Silva, S., Santos, B.S., Madeira, J.: Using color in visualization: a survey. Comput. Graph. **35**(2), 320–333 (2011)
29. Tory, M., Kirkpatrick, A.E., Atkins, M.S., Moller, T.: Visualization task performance with 2d, 3d, and combination displays. IEEE Trans. Vis. Comput. Graph. **12**(1), 2–13 (2006)
30. Ziemkiewicz, C., Ottley, A., Crouser, R.J., Chauncey, K., Su, S.L., Chang, R.: Understanding visualization by understanding individual users. IEEE Comput. Graph. Appl. **32**(6), 88–94 (2012)