Exploring Reproducibility in Visualization

Jean-Daniel Fekete, Juliana Freire

Department: Visualization Viewpoints
Editor: Theresa-Marie Rhyne, theresamarierhyne@gmail.com

Exploring Reproducibility in Visualization

Jean-Daniel Fekete
Université Paris-Saclay, CNRS, Inria, LRI, France

Juliana Freire
Tandon School of Engineering, New York University

***Abstract*—The American National Academies of Sciences, Engineering, and Medicine (NASEM) has recently released the report "Reproducibility and Replicability in Science". The report has prompted discussions within many disciplines about the extent of the current adoption of reproducibility and replicability, the challenges involved in publishing reproducible results as well as strategies for improving it. We organized a panel at the IEEE VIS conference 2019 to start a discussion on the reproducibility challenges faced by the visualization community and how those challenges might be addressed. In this viewpoint, we summarize key findings of the NASEM report, the panel discussion, and outline a set of recommendations for the visualization community.**

Reproducibility and Replicability

Visualization is having a substantial impact in science, industry, and in everyday life. In our data-driven world, visualizations are key to obtaining and communicating insights, and they are increasingly used to guide decisions, as highlighted, for example, by the plethora of visualizations used to convey the status and impact of the COVID-19 pandemic.

Unfortunately, the practice of reproducibility and replicability (R&R) has not been widely adopted by visualization researchers. This is in contrast to other sub-areas of computer science, where these issues have been widely discussed and initiatives have been established, for example, several conferences and journals have instituted reproducibility review for their papers [1]–[3].

In light of the recent American National Academies of Sciences, Engineering, and Medicine (NASEM) report on "Reproducibility and Replicability in Science" [4], we organized a panel at IEEE VIS 2019 to start a discussion to both better understand the unique challenges for visualization research and discuss steps the community can take to improve the adoption of R&R best practices. In this viewpoint, as context, we give a brief overview of findings and recommendations in the NASEM report. We then discuss the challenges involved in reproducing and replicating visualization research by examining different types of papers. We conclude with recommendations for the visualization community.

Reproducibility and Replicability in Science

Science aims to reveal the structure and behavior of physical and natural phenomena through observation and experiments. It advances through the discovery of new knowledge, which often involves the confirmation or extension of prior scientific results. Repeated findings of consistent results tend to confirm the validity of a given scientific conclusion, while repeated failures raise doubts about the conclusion. Revisiting and reusing past results — or as Newton once said, "standing on the shoulders of giants" — is thus the standard paradigm of all sciences. This requires that scientific results be accompanied by all details required to repeat them.

In recent years, concerns over reproducibility and replicability (R&R) have been expressed in both scientific and popular media [5][1]. In response to these concerns, in the House Science Committee American Innovation and Competitiveness Act 2017, the US Congress directed the National Science Foundation to engage with the NASEM to assess "research and data reproducibility and replicability issues in interdisciplinary research" and make "recommendations for improving rigor and transparency in scientific research".

The NASEM produced the consensus report "Reproducibility and Replication in Science" [4], of which one of the authors was a committee member. The report defines the terms *reproducibility* and *replicability* as applied to scientific and engineering research, assesses and ascertains the extent of issues of reproducibility and replicability, considers the impacts to the health of science as an enterprise and the public's perception of science, and provides findings and recommendations for improving rigor and transparency in scientific research. We discuss some of these aspects as they relate to visualization research.

Definitions: Reproducibility and Replicability

The terms reproducibility and replicability have been used in arbitrary and sometimes contradictory ways in different scientific disciplines. The NASEM report defines reproducibility to mean computational reproducibility: obtaining consistent computational results using the same input data, computational steps, methods, and conditions of analysis. Replicability means obtaining consistent measurements or results using new data, methods, and/or conditions in a study aimed at the same or similar scientific questions. As concrete examples, one expects that results described in papers that propose new algorithms or computational systems should be reproducible, while it should be possible to replicate the results in papers that describe user studies.

Transparency and the Assessment of Reproducibility and Replicability

Transparency is crucial to reproduce or replicate a result. While data and computation have transformed many scientific disciplines and led to important discoveries, this revolution has yet to be widely reflected in how results are published:

[1] https://www.economist.com/briefing/2013/10/18/trouble-at-the-lab

publications often fail to include the necessary information to reproduce their results. There are several ongoing efforts that incentivize or mandate authors to make their work transparent. For example, Nature has a detailed list of reporting requirements for papers they publish. Several Association for Computer Machinery (ACM) conferences and journals have instituted formal processes to evaluate the reproducibility of their publications [6]. While different venues have adopted different guidelines, they often require authors to submit data, code, and information about the computational environment, so that reviewers are able to re-run the experiments.

For ACM, publications that pass the reproducibility evaluation receive badges that are associated with the paper in the ACM Digital Library [6]. The Computer Graphics community has created the Graphics Replicability Stamp Initiative (GRSI[2]). An independent group of volunteers evaluates the reproducibility of papers accepted to a set of eligible journals, including IEEE TVCG and EG Computer Graphics Forum. As of today, there are few visualization articles with a stamp (6 for TVCG, 4 for CGF), and most of these are about computer graphics.

It is often straightforward to assess reproducibility of a computational result, but there can be challenges, for example, when experiments that involve sensitive data that cannot be shared or that require special hardware that is not widely available (we give some examples in the next section). Nonetheless, we note that even when it is not possible to attain full reproducibility, transparency can support partial R&R, which can be sufficient to support the associated research claims. Assessing the replicability of a study, on the other hand, can be costly and complicated. For example, it may require the re-implementation of the methods described in a paper or re-doing a user study with a different cohort. Even if a study was rigorously conducted and transparently reported, it may fail to replicate. A failure to replicate may be due to the inherent variability in the system under study or the inability to control complex variables.

While the NASEM definitions capture the general notions for R&R, the PRIMAD model [7] provides a more flexible way to express different levels of R&R. It defines a set of variables that are associated with an experiment: **P**latform (the computational environment), **R**esearch objectives (the main goal of the experiment), **I**mplementation (source code and binaries), **M**ethods and algorithms used to achieve the research goals, **A**ctors, **D**ata (input and intermediate data). These variables are used to describe which aspects of the experiment can be changed while still attaining reproducible results. In addition, there are other dimensions that can be considered for qualifying the level of reproducibility of experiment, including coverage, i.e., how much of the experiment can be reproduced, and longevity, which relates to the ability to reproduce experiments (long) after they were created [8].

Recommendations

All of the recommendations in the report have the goal of increasing transparency, specificity, completeness, and accuracy of the way in which science is conducted. We have identified two recommendations in the report that are particularly relevant to visualization research. We transcribe them below ipsis verbis:

RECOMMENDATION 4-1: To help ensure the reproducibility of computational results, researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results to enable other researchers to repeat the analysis, unless such information is restricted by non-public data policies. That information should include the data, study methods, and computational environment:

- *the input data used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are non-deterministic and cannot be reproduced in principle;*
- *a detailed description of the study methods (ideally in executable form) together with its computational steps and associated parameters; and*
- *information about the computational environment where the study was originally executed, such as operating system, hardware archi-*

[2] http://www.replicabilitystamp.org

tecture, and library dependencies (which are relationships described in and managed by a software dependency manager tool to mitigate problems that occur when installed software packages have dependencies on specific versions of other software packages).

RECOMMENDATION 5-3: Journals should disclose their policies relevant to achieving reproducibility and replicability. Moreover, the strength of the claims made in a journal article should reflect the reproducibility and replicability standards to which an article is held, with stronger claims reserved for higher expected levels of reproducibility and replicability.

Furthermore, we encourage journals and conferences to:

- *Set and achieve desired standards of reproducibility and replicability, and to make this one of their priorities. For example, journals could decide which level they wish to achieve for each Transparency and Openness Promotion (TOP) guideline and work towards that goal.*
- *Adopt policies to reduce the likelihood of avoidable or reprehensible sources of non-replicability. For example, they may consider implementing incentives and/or requirements for research materials transparency, design and analysis plan transparency, enhanced review of statistical methods, study and/or analysis plan pre-registration, and replication studies.*
- *Require that all research reports include a thoughtful discussion of the uncertainty in measurements and conclusions, and make this a review criterion.*

Reproducibility and Replication in Light of the IEEE VIS Paper Types

Improving reproducibility and replicability (R&R) in visualization research is a worthy goal but what does it mean concretely? The question is not new and there have already been several workshops dedicated to reproducibility in visualization, in particular, the "EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization" Series, that started in 2013[3].

[3]https://diglib.eg.org/handle/10.2312/980

Should all the research articles follow the same methods, or are there a few variations? Are some types of research immune to reproducibility issues? In this section, we discuss this question from the coarse grain of the five types of papers used to classify IEEE VIS submissions. This classification is familiar to researchers and practitioners who already submitted at the VIS conference. These types are:

1) Technique & Algorithm
2) System
3) Application & Design Study
4) Empirical Study
5) Theory & Model

Although the actual paper types may not be well aligned with the methods to achieve reproducibility, they are well understood by the community so we use them as a basis for the discussion. Since a given R&R method can be applicable to aspects of different data types, to avoid repetition, we introduce the methods in the first section where they are applicable, and refer to them later if needed for other types of articles.

Technique & Algorithm

According to the IEEE VIS site: [the] "technique should ideally be of general application rather than being restricted to a single task or single source of data, and the exposition should be focused on what the technique does, how it does it, the tasks and datasets for which this new method is appropriate, and what the computational and other costs are. Evaluation is likely to strengthen technique papers."

The articles of this type match perfectly the needs for reproducibility if they provide an evaluation. There are already accepted methods to facilitate the reproducibility of algorithms and techniques, assuming they are assessed quantitatively. Technique papers used to be accepted without quantitative evaluation in the past but this is becoming less common nowadays, as the reviewers' standards have raised in this respect.

R&R Method 1 (Algorithmic Reproducibility):

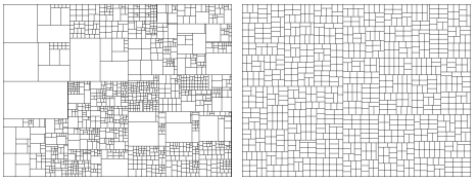
For most articles on algorithms, reproducibility means that an external practitioner should be able to obtain, from the authors, the software

to reproduce the results reported in an article, including figures and tables. Reproducibility entails running the program or algorithms with the right parameters on the benchmark or example datasets (using a compatible operating system and hardware), and obtaining results that are identical or consistent with the reported results, i.e., the differences should not challenge or invalidate the conclusions of the article. For algorithms, the typical measures to reproduce are related to execution time, memory usage, and usually some quality measures. In computer graphics, the measures can also be frame-per-second, or the images generated, or quality measures on the images. In algorithms for visualization, the measures are similar to computer graphics except that the quality measures can be visualization specific or tied to perceptual features.

The use of benchmark data further improves the usefulness of the results and avoids the suspicion that the authors have cherry-picked datasets that behave particularly well with the algorithm. For replication, the results of the algorithm are not enough, some objective function should be specified to test the quality of the replication.

R&R Method 2 (Technique R&R):
R&R of visualization or interaction techniques is less standard in computer science and borrows methods from psychology and human-computer interaction (HCI). The problem has been studied in the HCI and visualization communities to some extent [9], [10] and in the proceedings of the "Eurovis Workshop on Reproducibility, Verification, and Validation in Visualization".

Techniques are typically assessed by measuring time and errors, although many more measures exist as attested by the BELIV workshop series (see https://beliv-workshop.github.io/). Compared to algorithms, simply comparing raw results is not enough to assess replication because humans are involved as users of the techniques, and humans exhibit individual differences. Evaluations of techniques, just like algorithms, involve datasets and results, but they additionally require tasks operationalized as a set of lower-level actions, and statistical methods to measure accurately the time and errors, including experimental design, data cleaning, and data analyses.



A popular example of algorithm regarding R&R is the Treemap technique for visualizing a rooted tree as a recursive containment of boxes, made popular by Ben Shneiderman. The original "slice and dice" algorithm is trivial to implement but exhibits artifacts. Bruls et al. [1] have improved the method to obtain boxes that are more square. They have published an algorithm that implements a heuristics solution since the optimal one would be too complex. This Squarified Treemap algorithm has been replicated and is now implemented in all the standard information visualization systems and toolkits. Yet, the main idea for squarifying the boxes composing a treemap is its objective function to optimize for replication and improvement. A follow-up article by Bederson et al. [2] studies other Treemap algorithms and compare them to the results of the Squarified Treemap based on the original objective function. Bederson et al. provided the source code of their algorithms as well as the test program that has been used to generate the figures of their article. It turns out that almost 20 years after the publication, the source code (in Java) still compiles and runs on some platforms.

1. M. Bruls, K. Huizing, and J. J. v. Wijk, "Squarified Treemaps," in *Eurographics / IEEE VGTC Symposium on Visualization*, W. de Leeuw and R. van Liere, Eds. The Eurographics Association, 2000. [Online]. Available: http://doi.org/10.2312/VisSym/VisSym00/033-042
2. B. B. Bederson, B. Shneiderman, and M. Wattenberg, "Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies," *ACM Trans. Graph.*, vol. 21, no. 4, p. 833–854, Oct. 2002. [Online]. Available: https://doi.org/10.1145/571647.571649

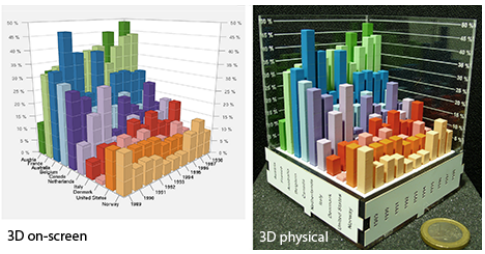
All of these should be described and explained in the article, and ideally provided as scripts for replication. Although social scientists could see this operation as reproducibility since everything is controlled except the human variability, which is dealt with using statistics, according to the NASEM report, reproducibility means computational reproducibility which cannot be achieved due to the variability of humans.

Replication involves the tasks and claims of improvements (the objective function), such as "technique X is significantly faster than technique Y for task T on datasets $\{D_1, D_2, \ldots, D_n\}$". To account for human variability, the results (time, errors) should be gathered by repeating the experiments multiple times and validating statistically that the time is significantly shorter on the multiple results and the errors are significantly lower. Some algorithms should also be assessed with similar methods, for example when they use stochastic methods that cannot be reproduced deterministically.

To summarize, R&R for research on techniques & algorithms is well understood and consists of transparently providing information about the technique to replicate or algorithm to reproduce, in source code preferably, the data used to validate it, the detailed experimental protocol suitable for reproducibility, the code for analyzing the results of the experiments, and all the information required to reproduce the figures provided in the paper. For replicability, an objective function should be specified to assess the quality of conformity of the results compared to a baseline. Conducting research on algorithms or techniques with the goal of being reproducible and/or replicable takes practice but is well documented and possible for most research projects. However, it should be planned ahead since R&R problems can be difficult to solve when they are discovered later (e.g., asking for IRB approval for disclosing user data).

System

According to the IEEE VIS website: "The system that is described is both novel and important, and has been implemented. Here, the focus should be on the design decisions, the implications for software/hardware structure, and comparison with



A particularly challenging example of a replicable technique is "Evaluating the Efficiency of Physical Visualizations" [1]. It compares physical to on-screen visualizations for 3D bar charts and shows that physical visualizations can improve users' efficiency at information retrieval tasks. It relies on a user study to compare the techniques across multiple datasets and tasks. It points to a site containing the material necessary for replicability: www.aviz.fr/phys. The experimental material is made available, with the instructions to create the physical visualizations using a laser cutter, some electronic circuits, datasets, tasks/questions asked to the participants, results of the participants, scripts to perform the statistical analysis, and the graphs published in the final article. This example is an extreme case where reproducibility involves building physical objects and electronic systems to capture physical motions. Most of the techniques reported in the visualization conferences are much simpler to replicate.

1. Y. Jansen, P. Dragicevic, and J.-D. Fekete, "Evaluating the efficiency of physical visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: ACM, 2013, p. 2593–2602. [Online]. Available: http://doi.org/10.1145/2470654.2481359

other systems. The comparison includes a specific discussion of how the described system differs from and is, in some significant respects, superior to those systems."

R&R Method 3 (System R&R):
Articles describing systems cannot use the same methods to assess their R&R as for the techniques & algorithms because a system is typically more complex than a set of techniques and algorithms. Testing all of the capabilities of a system in a controlled fashion would lead to a combinatorial explosion of experiments.

Instead, systems are assessed for targeted tasks, targeted users, and report targeted measures, sometimes compared to other systems. In some cases, these assessments can be formally defined and re-applied in the future to reproduce the findings in the article. An important distinction concerns machine-based assessments vs. human-based assessments. Articles assessing systems based only on machine time and resources are well understood in terms of R&R; the database and HPC domains are used to publish these kinds of articles and have developed solid methods to do so.

When humans are involved in the assessment, the problem is more complicated and not well streamlined yet. For example, a system targeted at domain experts can only be tested by these experts, and they may be hard to find. Still, experiments with humans can be performed with multiple tasks and datasets, just like techniques, and to enable replicability and comparison between systems, the tasks and datasets should be provided. The different contests organized at the IEEE VIS conference (InfoVis Contest, VAST Challenge, and SciVis Contest) were meant to compare solutions at the system level to common problems and tasks, and foster reproducibility sometimes, but more often replicability. When introducing a new system, it is fair to assess it using a standard benchmark when available, or by replicating a previous study, in addition to other tasks and datasets if needed.

Several system articles published at IEEE VIS also come with their source code on a public repository such as github.com, an important step in R&R, allowing the system to be tested or used by others. Some of these articles also describe controlled studies but very few provide all the material required to reproduce or replicate them. In particular, installing a non-trivial system is rarely straightforward and requires a lot of detailed instructions and tuning parameters that are not published in the scientific articles for good reasons but cannot be recovered if they are not archived properly and made available in a public repository. In the past, it has been common practice in research to ask the authors for this material, and very often, the authors provide it in a form more or less easy to reuse and more or less complete. In our experience, the answer is quite often positive and deserves acknowledgment. Having a structured process for authors to submit the necessary artifacts as reproducibility evaluation initiatives do, would encourage authors to make these available at publication time.

Even when all the material for reproducing a study is available, there are limitations to the reproducibility of systems, just like with techniques. Systems become obsolete, or rely on obsolete libraries, or rely on special hardware. This problem is well documented in other domains and special frameworks are meant to allow the archival or systems for a long time, using virtual machines such as Docker [11] and automatic dependency tracking, such as ReproZip [12]. For hardware, the problem can be more difficult, and for commercial black-box hardware, there is no easy way to reproduce results after the company manufacturing it is out of business or the product has disappeared. Yet, scientific articles describing systems for special hardware are usually trying to generalize or abstract the features of this hardware, leading the way to replicability.

For modern systems based on web technologies, reproducibility is a problem due to the fast evolution of these technologies and the use of distributed web-based services that also evolve quickly and cannot be saved for later reuse. Most interactive systems of the early 21st century are designed to run on a web browser, using a combination of JavaScript — which evolves quickly and not always in a backward compatible fashion, libraries that are also evolving quickly and are fetched from the web using "Content Delivery Networks" (CDN) instead of local files, with risks of seeing the CDN disappear or stop serving an old version of a library. Finally, many

features that used to be delivered as libraries in traditional languages and operating systems are now implemented as remote services, creating extra dependencies with code that cannot be archived. Web-based systems are a nightmare for reproducibility, yet they become the norm for interactive systems. Designing a web-based system for reproducibility is still possible but requires a lot of care, and probably giving up on some features provided by online services that offer no guarantee of sustainability.

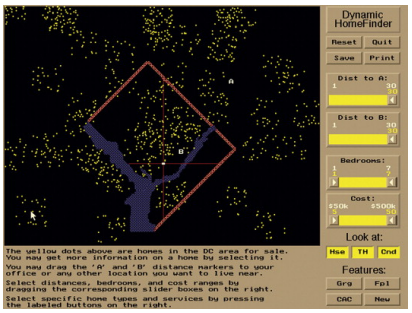
Application & Design Study

According to the IEEE VIS website: "These papers typically include an encapsulated description of a problem domain and the questions to be resolved by visualization, then describe the application of visualization to the task, any novel techniques developed, and how the visualization solution answered the questions posed. The results of the study, including insights generated in the application domain and visualization knowledge generated through the research process, should be clearly conveyed."

Applications and design studies cannot be reproduced but rather replicated in the context of the tasks they were designed to support.

R&R Method 4 (Application R&R):
Substantial efforts have been devoted towards trying to replicate visualization applications, in particular, with the different contests conducted by IEEE VIS. The VAST Challenge is organized every year to challenge visual analytics systems to solve 3 or 4 challenges [10]. Therefore, several applications can compete and disclose how well they performed in some of the challenges. Before that, the InfoVis Contest [9] was also providing datasets and tasks focused on applications, and the SciVis community is also organizing contests to address specific applications every year. However, there will never be a benchmark or a contest for all the possible applications, therefore more work should be done to improve the replication methods. They currently rely on qualitative methods that are difficult to replicate. The contest results are assessed by a jury.

Replication is usually not hampered by time, but comparison to the baseline can be difficult if the system supporting the application cannot be run after a certain time — this is related to the longevity aspect of reproducibility. The same solutions as the ones used to make sure a system will run in the future can be used for applications.



An interesting historical example of a reproducible application study is the *Dynamic HomeFinder* [1]. It showcases the first dynamic queries UI for visualization. It was meant to help to rent or to buy an apartment or a house with the support of a map and multidimensional queries. The program was written for Windows95 and distributed in executable form. It is still usable on most current systems; they emulate Windows95 quite well. The demo is very useful to showcase the design of dynamic query techniques that have been described in 1992, well-cited, but are still rarely well implemented nowadays.

1. C. Williamson and B. Shneiderman, "The dynamic homefinder: Evaluating dynamic queries in a real-estate information exploration system," in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '92. New York, NY, USA: ACM, 1992, p. 338–346. [Online]. Available: http://doi.org/10.1145/133160.133216

For design studies, replication is much harder because the design of systems evolves and introduce new standards and conventions; old designs are usually very noticeable and perceived as obsolete. Yet, with emulators, many old systems along with their designs can still be run and used as baselines for comparing interaction of visualization designs on various tasks and datasets.

Empirical Study

According to the IEEE VIS website: "An empirical study paper explores the usage of visualization by people, and presents a study, either qualitative or quantitative, of visualization techniques or systems."

Articles of this category are evolving year after year towards open science, pre-registration, transparency, and reproducibility. Practically, the methods used are the same as for Techniques; the domains of psychology, HCI, and the natural sciences have been working hard to achieve more replicability in their publications. Psychology, just like many natural sciences, has suffered a replication crisis in the last decade, challenging landmark articles and the career of well-known researchers. Therefore, the new generation of experimental psychologists and natural science researchers is strongly promoting the use of R&R for new publications and is active in developing tools to improve the best practices. One of the authors is managing a laboratory where many empirical studies are performed every year and can attest that the road towards open science is hard and takes time and skills to achieve. Yet, it does improve science and avoids many methodological pitfalls that have polluted experimental sciences in the past. It seems that the peer-pressure, from the ongoing practices and some vocal supporters of replicability [13], is sufficient for this category of articles to adopt open science methods.

Theory & Model

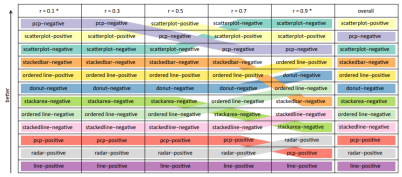
According to the IEEE VIS website: "These papers do not require implementation, but contribute by illuminating how visualization techniques complement and exploit properties of human vision and cognition, as well as how researchers conduct effective and rigorous visualization studies."

Articles from this type include articles framing a mathematical theory, conceptual models, taxonomies, and ontologies. R&R do not always make sense for these kinds of articles, except when the theories or models are described using equations. Then, according to the standard epistemology of science, a mathematical theory can be falsified or demonstrated, and a natural science theory can only be falsified or remain plausible. In the humanities, theories and models

Just like interaction techniques, quantitative empirical studies can be replicated using methods borrowed from HCI and psychology. A few VIS empirical studies have been reproduced and sometimes improved, for example, the seminal experiment by Cleveland and McGill on the ranking of visual variables [1]. The original study dating from 1984 was not very precise in its exact setup, but the intents were clear and the tasks to support the experiments were specified well enough to be replicable. It has been replicated and extended by Jock Mackinlay [2] but he did not confirm his results with empirical studies, explaining: "Although this extension was developed using existing psychophysical results and various analyses of the different perceptual tasks, it has not been empirically verified". In 2010, Heer and Bostock have replicated the study using crowdsourcing [3], and confirmed the results. The source material used to perform the study is specified in the article on a web page that does not exist anymore so, technically, the study is now reproducible but practically, some hunting is necessary to find the data.

---

1. W. S. Cleveland and R. McGill, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984. [Online]. Available: http://www.jstor.org/stable/2288400
2. J. Mackinlay, "Automating the Design of Graphical Presentations of Relational Information," *ACM Trans. Graph.*, vol. 5, no. 2, p. 110–141, Apr. 1986. [Online]. Available: https://doi.org/10.1145/22949.22950
3. J. Heer and M. Bostock, "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, p. 203–212. [Online]. Available: https://doi.org/10.1145/1753326.1753357



For example, a nice success story concerns the Weber Law (which is a model), well-known in psychophysics. It describes the relation between the actual change in a physical stimulus and the perceived change. The law has been successfully used by Harrison et al. [1] to model the perception of correlation with different visualizations. The material for the study has been provided by the authors in a public repository and reused by Kay & Heer [2] for an alternative and insightful data analysis.

1. L. Harrison, F. Yang, S. Franconeri, and R. Chang, "Ranking Visualizations of Correlation Using Weber's Law," *Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1943–1952, 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2014.2346979
2. M. Kay and J. Heer, "Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation," *Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 469–478, 2016. [Online]. Available: http://dx.doi.or/10.1109/TVCG.2015.2467671

are less formalized but are sometimes described with equations too that could be verified in the context of an epistemological system.

Still, a theory and model in natural sciences, supported by equations, can be challenged with data, and reproducing the validation is important, in particular, to compare the accuracy of different models according to measures of the phenomenon they model. Natural science theories can be challenged and compared to competing theories. In visualization and interaction, we have few theories but many models, and they can be tested in a way similar to algorithms. The theory and model articles that are replicable use the same methods as algorithms, techniques, and systems.

Paper Type	Method
Algorithm & Technique	Algorithm repr. methods HCI/Psy repl. methods
System	HPC or DB repr. methods HCI repl. methods
Application & Design Study	Repl. methods using benchmarks HCI repl. methods using representative tasks & datasets
Empirical Study	HCI/Psy repl. methods
Theory & Model	Diverse methods w.r.t. concrete theory or model

Table 1. R&R methods by IEEE VIS Paper Type

DISCUSSION

The benefits of R&R are well explained in the NASEM report, but it does not come for free, there is a cost related to learning the right methods, summarized in Table 1, and using the right tools to ensure reproducibility. There is no silver bullet to address these questions and the easiest process to follow it is to start early and iterate to develop the necessary know-how. In a lab, it usually starts by asking new students to develop their research in a reproducible manner and by cross-testing the different projects for reproducibility, solving the issues as they come in light of the literature on tools for R&R [4, Chap. 6].

For human-based experiments, raw results of each subject should be reported for replicability. This might become an issue regarding the IRB approvals since data disclosure could expose sensitive information about the subjects. This point is exacerbated by the RGPD regulations in the EU and similar regulations in other parts of the world. Still, several research laboratories have been able to obtain IRB approvals to disclose raw results anonymously and with a strict minimum amount of associated information to avoid identifying the subjects. Sharing these IRB applications and recommendations would help the community, both to understand the threats of gathering and disclosing some kinds of measures, and to provide solutions to maintaining anonymity while still disclosing essential information.

While many perceive attaining R&R difficult and time consuming, there are already many tools (e.g., ReproZip [12], Docker [11], Jupyter) and infrastructure—including repositories (zenodo.org, osf.org)—that make it easier to publish transparent, R&R results. It is worth noting that there are possible limitations, re-

garding humans, hardware, and software, that can hamper reproducibility: the need for special skills, obsolete hardware, and obsolete software. In particular, special hardware is an obstacle to reproducibility. Visualization is particularly rich in special hardware, from HPC to display technologies like VR, AR, wall-sized displays, and physical visualizations.

Interactive systems are also considered an issue for replication and visualization relies deeply on interaction. More research and methods should be developed to reconcile interaction and replication since both are essential: data exploration requires interaction and Science needs replication.

Finally, we should note that 1) reproducibility does not necessarily imply correctness—an incorrect result can be reproduced, and 2) the inability to (completely) reproduce or replicate a result also does not imply the result is incorrect. However, it is only through R&R and transparency that other researchers can confirm and check the correctness of the computations, attempt to replicate the experiment and understand the full context of how to interpret the results.

CONCLUSION

In this viewpoint, we discussed some of the findings and recommendations in the NASEM report of Reproducibility and Replicability in Science [4]. We also discussed challenges involved in attaining reproducibility and replicability for different types of visualization papers.

There are already visualization researchers that follow best practices for openness and include code and data with their papers. There are also notable examples of impactful contributions that have become widely adopted as open-source tools. But too few visualization articles are reproducible or replicable. We mentioned a few, the ten articles with the Graphics Replicability Stamp, and much more exist, but they are still the exception rather than the rule.

The community should do better. We can start by implementing some steps that have already been adopted in other sub-areas of computer science. Conference organizers and journal editors can establish a process for authors to submit artifacts for reproducing or replicating their results, institute R&R evaluation (e.g., like the Graphics Replicability Stamp Initiative), and provide incentives to authors (e.g., the ACM SIGMOD Reproducible Paper Award[4]) and to reviewers (e.g., the reproducibility report co-authored by reviewers and authors in the Information Systems Journal[5]).

Researchers can plan for R&R and adopt R&R best practices in their day-to-day work. There are already many tools and infrastructure that help with this. However, researchers may not be aware of these. Therefore it is important to educate the community on the 'whys' and 'hows' of reproducibility. Different approaches can be used to disseminate this information, from tutorials at conferences to the publication of reports that detail how reproducibility was attained for different papers/experiments and that can serve as a guide for others.[6]

While we focused on visualization research and publications, R&R is essential for visualization in the wild. Visualization has become an integral part of the data science pipeline, and as decisions are made based on data and results of analyses that include visualizations. Therefore, our recommendations are also applicable to visualization practitioners. In fact, by addressing the R&R challenges in visualization, the research community has the opportunity to lead the way in solutions that can be widely used in the visualization practices.

The standards for reporting research results have evolved and striving for R&R will undoubtedly become standard, in the opinion of the authors. With data science relying on complex methods at every step of the analysis pipeline, including the visualizations, the burden of R&R should be on the authors and not on the readers to avoid suspicion and time wasted trying to guess details that are irrelevant for the science but essential for the technicality of the work.

Having a wider adoption of R&R will enable others to revisit, reuse, and extend visualization research more easily. This has the potential to accelerate progress in the area as well as magnify its impact. There are also direct benefits

[4] https://sigmod.org/sigmod-awards/sigmod-most-reproducible-paper-award

[5] https://www.elsevier.com/journals/information-systems/0306-4379/guide-for-authors

[6] See for example https://www.practicereproducibleresearch.org which describes 31 use cases of reproducible research workflows, written by academic researchers.

to authors: besides being able to reproduce and extend their own work, recent studies indicate that reproducibility increases impact, visibility, and research quality [14], [15].

While our discussion is by no means exhaustive, we hope it will help spur a dialogue in the community about how to address these and other challenges.

ACKNOWLEDGMENT

We thank Tina Winters (NASEM) for her help in framing and organizing our panel at IEEE Vis. We thank the other panel members, Carlos Scheidegger and Steve Haroz, as well as the audience for a lively discussion. We would also like to thank David Koop and Cláudio Silva for their input on the panel topics. We also thank Maria Beatriz Silva for creating the front cover image.

REFERENCES

1. M. A. Heroux, "Editorial: ACM TOMS Replicated Computational Results Initiative," *ACM Trans. Math. Softw.*, vol. 41, no. 3, Jun. 2015. [Online]. Available: https://doi.org/10.1145/2743015
2. P. Bonnet, S. Manegold, M. Bjørling, W. Cao, J. Gonzalez, J. A. Granados, N. Hall, S. Idreos, M. Ivanova, R. Johnson, D. Koop, T. Kraska, R. Müller, D. Olteanu, P. Papotti, C. Reilly, D. Tsirogiannis, C. Yu, J. Freire, and D. E. Shasha, "Repeatability and workability evaluation of SIGMOD 2011," *SIGMOD Rec.*, vol. 40, no. 2, pp. 45–48, 2011. [Online]. Available: https://doi.org/10.1145/2034863.2034873
3. J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, and H. Larochelle, "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)," 2020.
4. National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science*. The National Academies Press, 2019. [Online]. Available: https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science
5. C. G. Begley and J. P. Ioannidis, "Reproducibility in science: improving the standard for basic and preclinical research," *Circulation research*, vol. 116, no. 1, pp. 116–126, 2015.
6. R. F. Boisvert, "Incentivizing reproducibility," *Communications of the ACM*, vol. 59, no. 10, p. 5, 2016. [Online]. Available: https://doi.org/10.1145/43965.43966
7. J. Freire, N. Fuhr, and A. Rauber, "Reproducibility of Data-Oriented Experiments in e-Science (Dagstuhl Seminar 16041)," *Dagstuhl Reports*, vol. 6, no. 1, pp. 108–159, 2016. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2016/5817
8. J. Freire, P. Bonnet, and D. Shasha, "Computational Reproducibility: State-of-the-Art, Challenges, and Database Research Opportunities," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '12. ACM, 2012, pp. 593–596.
9. C. Plaisant, J.-D. Fekete, and G. Grinstein, "Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository," *Transactions on Visualization and Computer Graphics*, vol. 14, no. 1, pp. 120–134, 2008.
10. J. Scholtz, M. A. Whiting, C. Plaisant, and G. Grinstein, "A Reflection on Seven Years of the VAST Challenge," in *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, ser. BELIV '12. New York, NY, USA: ACM, 2012.
11. D. Merkel, "Docker: Lightweight linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, Mar. 2014.
12. F. Chirigati, R. Rampin, D. Shasha, and J. Freire, "ReproZip: Computational Reproducibility With Ease," in *Proceedings of the 2016 International Conference on Management of Data*, ser. SIGMOD '16. ACM, 2016, pp. 2085–2088.
13. R. Kosara and S. Haroz, "Skipping the replication crisis in visualization: Threats to study validity and how to address them : Position paper," in *2018 IEEE Evaluation and Beyond—Methodological Approaches for Visualization (BELIV)*, Oct 2018, pp. 102–107.
14. C. G. Begley and L. M. Ellis, "Drug development: Raise standards for preclinical cancer research," *Nature*, vol. 483, no. 7391, pp. 531–533, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1038/483531a
15. P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible Research in Signal Processing," *Signal Processing Magazine, IEEE*, vol. 26, no. 3, pp. 37–47, 2009.