Chapter 8
Empirical Evaluations with Domain Experts

KREŠIMIR MATKOVIĆ, *VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Austria*
THOMAS WISCHGOLL, *Wright State University, USA*
DAVID H. LAIDLAW, *Brown University, USA*

Abstract
Over the past thirty years, the visualization community has developed theories and models to explain visualization as a technology that augments human cognition by enabling the efficient, accurate, and timely discovery of meaningful information in data. Along the way, practitioners have also debated theories and practices for visualization evaluation: How do we generate durable, reliable evidence that a visualization is effective? Interestingly, there is still no consensus in the visualization research community how to evaluate visualization methods. The goal of this chapter is to rise awareness of still open issues in the visualization evaluation, and to discuss appropriate evaluations suitable for different visualization approaches. This includes user studies and best practices to conduct them but also other approaches for suitable evaluation of visualization. The chapter is structured as a moderated dialogue of two visualization experts.

8.1 Introduction

Over the past thirty years, the visualization community has developed theories and models to explain visualization as a technology that augments human cognition by enabling the efficient, accurate, and timely discovery of hidden information in data. Along the way, practitioners have also debated theories and practices for evaluation of visualizations: How do we generate durable, reliable evidence that visualization is effective — that visualization facilitates obtaining insight into the data in ways that are demonstrably beneficial to the user, and that it perfectly complements automatic methods in cases where problems and queries are ill-defined or hard to specify?

User studies seem as a first choice for evaluation of visualization as it is really a human-centric technique. Somewhat paradoxically, user studies are both taken for granted and controversial among visualization practitioners. On the one hand, it is

e-mail: matkovic@vrvis.at

difficult to get an application paper accepted for conference or journal publication without a "User Evaluation" section at the end of a manuscript. At the same time, user studies are often haphazardly executed and presented, leading researchers (and reviewers) to question whether user studies are helping move the field forward in any meaningful way. A proper evaluation typically is an important aspect of a visualization publication. Since visualization involves a visual interface combined with other user interface elements, it appears natural to deploy some form of user study to evaluate the system. However, this is only one type of evaluation. There can be several reasons why a user study is not appropriate for a specific visualization approach. In addition, the term "user study" is not always used correctly in the visualization community. Besides its strict definition, many visualization researcher consider almost any user experiment to be a user study. It is our goal to provide guidance toward appropriate evaluations suitable for visualization approaches. As pointed out earlier, user studies are only one way of evaluating a visualization approach.

There are guidelines on how to do proper user studies in various fields. As its name clearly implies, a user study is impossible without a user. The user, one of the fundamental players in the visualization, has to be taken with great caution when evaluating visualization research. The visualization solutions are designed for a specific user group. The target audience could be casual users on the web on one end of a spectrum or highly trained domain experts solving very specific problems on the other end. Can we evaluate visualization solutions for those two with the same methods? It hardly seems possible.

Finding the users for a user study on case visualization is, in general, simpler. Crowd sourcing mechanisms make it possible to recruit a large number of users in a relatively short time. Motivating a large number of experts who deal with a specific problem is close to impossible. First, there are not many experts that deal with very specific problem, and secondly, their time is usually too precious and they cannot afford to participate in a lengthy study. If the study has to be repeated for any reason it is also impractical as it uses up even more of the expert's valuable time. Once, we had to design an interactive exploratory visualization system geared toward systems designers for fuel injection into a Diesel engine [18]. We evaluated it with two experts with whom we collaborated on the project. It was impossible to perform a large user study on such a specific topic. However, does this still count as an evaluated solution?

Nevertheless, we should strive to evaluate our solutions. Evaluation should focus on lessons learned and take home messages for visualization researchers. This might be the most difficult part of the evaluation. We should evaluate it for the given task and users, but we should also strive to generalize it or at least to reflect on important findings from the visualization research perspective. Without such a reflection, the research is of less interest for the visualization community (it still can be of a great interest for other scientific domains and worth publishing in their journals). At the same time, there is also a certain threat in trying to generalize everything. If a solution cannot be generalized, it still can represent valuable research for the visualization community.

There are multiple orthogonal ways to evaluate a visualization approach, user studies being one of them. Isenberg et al. [9] provide a review of evaluation means for visualization. Some other chapters of this book also deal with evaluation [1, 28]. The following sections address different questions about user studies and evaluation in general. As the evaluation and user studies are often a hot topic of formal and informal discussions at many visualization conferences, we decided to structure this chapter as a conversation. We are fully aware that none of the extreme approaches ("user studies are a must" and "we do need user studies at all") is appropriate. We hope that the following dialogue can help in clarifying the evaluation needs of visualization research.

8.2 A Conversation about Empirical Evaluation of Visualization Approaches

Krešimir Matković:

We start this section with position statements on evaluation in visualization research from two visualization scientists Thomas Wischgoll and David Laidlaw with decades of combined visualization research experience. Thomas Wischgoll is an expert in flow visualization [27, 12], medical visualization [4], virtual reality [26], and areas of information visualization [5]; whereas David Laidlaw is an expert in multi-valued volume visualization [11], applications of visualization to science [7], virtual reality for visualization [24], visualization design [10], and visualization evaluation [13, 3, 14, 6]. Both are experts in working with domain specialists to successfully apply visualization algorithms to various disciplines. We continue the section as a moderated dialogue. So, let us start with your viewpoints on evaluation in visualization research? Do we always need it, is it an unnecessary add-on which is required by reviewers, or do you think, we should stubbornly omit it whenever possible?

Thomas Wischgoll:

A proper evaluation typically is an important aspect of a visualization publication. Since visualization involves a visual interface combined with other user interface elements, it appears natural to deploy some form of user study to evaluate the system. However, this is only one type of evaluation. There can be several reasons for why a user study is not appropriate for a specific visualization approach.

The purpose of visualization is by definition to involve humans as their target audience and to provide better insight into the data that is to be visualized. There are numerous aspects of providing better insight, however. A visualization approach could be better in terms of providing insight more quickly or in a more comprehensible way. The visualization could also be more user-friendly or intuitive, albeit

those are both fairly subjective criteria. Hence, there are several different aspects one could focus on for a user evaluation. For example, Lam et al. [15] list seven scenarios for empirical evaluations: evaluating visual data analysis and reasoning, evaluating user performance, evaluating user experience, evaluating environments and work practices, evaluating communication through visualization, evaluating visualization algorithms, and evaluating collaborative data analysis. This underscores the complexity of user evaluations as well as the different aspects a user evaluation can be used to test for with respect to a visualization approach. There is quite a number of publications in the literature to provide guidelines for evaluations. Munzner [20] suggests a nested, four-tiered model to assist in the validation of design studies. Meyer et al. [19] refined this model by adding blocks for additional flexibility to describe activities within the tiers of the original model. Focusing on design studies, Sedlmair et al. [21] propose a nine-stage model for the entire life-cycle of the visualization system and provide guidance and common pitfalls. Lam et al. [16] analyzed papers from the Information Visualization area between 2009 and 2015 to develop a framework for breaking down goals of a project to individual tasks, whereas Chen and Ebert [2] provide an ontological framework to assist in the design and evaluation of visual analytics systems.

At the same time, it is imperative to provide a proper evaluation of some sort. There are research areas outside of visualization that suffer from issues of lack of reproducibility. One reason for this is improper use of statistical methods during the evaluation or the selection of the participants which leads to misleading results. However, once a paper that applied such improper statistical methods is published, it is considered factual and researchers may not question the results despite the fact that there is a high probability that the results are invalid. This stresses the importance of properly executed user studies or any kind of evaluation for that matter. It does not serve the visualization community to publish papers with user studies or other types of evaluations that are flawed in a significant way, thereby making the findings questionable. We therefore need to find a way to make the evaluation of visualization approaches easier for researchers that ensures meaningful results. Some additional guidance may be needed, and the visualization researchers, if they want to go for a user evaluation, need to make sure they know what they want to test for and how to execute the user study properly in accordance with those goals.

David Laidlaw:

I would modify Thomas's first statement, "a proper evaluation typically is an important aspect of a visualization publication," to say that a visualization publication needs to clearly state how it extends human knowledge, successfully arguing for both the novelty and the significance of that new knowledge. Empirical evaluations can be a part of this argumentation in two ways.

First, an empirical evaluation can serve as a measure of the significance of a new visualization artifact, by which I mean an algorithm, interactive technique, or software system. There are numerous examples of this kind of user evaluation in

the literature as well as a number of papers that describe the process and organize examples of it. These example evaluations or studies range from small numbers of expert users sharing their opinions about a visualization artifact to quantitative performance comparisons among several artifacts that are similar enough to compare. The scope and type of evaluation and evaluators is a research design consideration, and there is no single best choice. In particularly young areas, there may not be a clearly related artifact to compare with a newly created one. In such a case, the opinions of a few domain science experts as evaluators may be sufficient to establish that a particular system or technique holds enough promise to share in a publication. But an evaluation is always stronger when there is a comparison of some kind to what has already been published. This kind of anchoring of a research result to the rest of the literature is something that our visualization field could do more consistently. There is always a most closely related artifact that has been described in a publication. In most cases, if an approach serves a need that is already being addressed, however inefficiently, it can be compared to the current inefficient approach. Such a comparison may be sufficiently self-evident that it does not even require an experiment. It does require explicit statement.

Second, an empirical study may establish new knowledge about how humans and computers interact. This kind of research is often hypothesis driven, and the user study employed serves to test the hypothesis. A hypothesis might state that certain visual cues are more easily perceived by users. Or it might state that one approach for a particular task is more efficient than another approach. What then emerges is new knowledge about whether the hypothesis is supported or falsified by experimental testing. In the best cases, the emergent knowledge can be generalized and helps to guide future research as well as the design and use of visualization artifacts.

I do not think that an empirical evaluation is a requirement for a visualization publication. However, there are a number of publications I have seen where the novelty or significance could have been established much better with an empirical evaluation. Too often those of us who are engineers create a software system or a visualization technique that is new and presume that that is enough to warrant a publication. After all, it was a lot of work! But software documentation, even of a novel piece of software, is typically insufficient to extend human knowledge in a significant way. It requires an explanation of how it is significant. Is it faster? Does it scale better? Does it more efficiently use screen real estate? Questions like these can be answered without a user study, but they do require some testing, analysis, and argumentation. And, as with a user study, in order to be demonstrably faster or more efficient, there has to be something specific to be compared to. There are other ways for something novel to be significant. Do users like it more? Does it speed their work? Does it make them more accurate? Is the experience of using it more pleasurable? If these are the ways something is significant, than some kind of empirical evaluation is likely to be essential. If the claim of significance includes "more," then the empirical evaluation likely needs to include a concrete comparison.

Some of what I have said probably sounds abstract and some perhaps even grandiose. But I believe that extending human knowledge is truly the bar for a re-

search publication. With that context, perhaps we can converge on some conclusions.

Thomas Wischgoll:

I actually agree a great deal with what David lists here. I see the evaluation as the aspect of the publication that illustrates in what way the chosen approach is an improvement over existing work. I use the term *evaluation* fairly loosely here as it could be any means of showing the benefits of the presented work. A properly executed user-study can be an effective way of accomplishing that. But there are certainly others as well as both David and I tried to hint at in our previous statements. Weber et al. [25] list 12 different ways in which an application paper can contribute to the area of visualization each of which could be shown with different empirical measures. Personally, I like quantitative and objective measures, such as the execution time of an algorithm in a very well defined test environment. On the one hand, such measures are easier to determine. But at the same time, they cannot be refuted easily either. However, given the fact that visualization algorithms are geared toward making human beings more effective at specific tasks, a user-study may be the only way to prove a certain measure.

This leads me to one of the issues with user-studies as it is of utmost importance for a user-study to be designed, set up, and executed properly to bring value to a publication. This involves the number of people to include, how to recruit participants, and the analysis of the data collected through those participants. One issue can be with domain experts as there may only be a fairly low number of experts suitable for such a user-study depending on who the visualization approach is designed for and these domain experts may not be accessible to everyone which makes reproducibility difficult. These domain experts may have a very specific mind set already based on they day to day work and that bias may be different compared to another group of domain experts.

Since user studies by definition involve human beings, they are susceptible to different types of biases. This makes the selection of the user pool for the study all the more important. A good user study typically documents very well the selection criteria and processes that were used as well as the entire procedure used to conduct the user study itself. This then aids in improving the reproducibility and should enable the reader to at the very least better understand of what was tested.

Ultimately, I would like to reiterate the importance of a properly executed user-study. If the user study was poorly executed it does not only provide little to no value to the field. But to make matters worse, it actually provides misleading data. For example, it could suggest that an approach works better than it actually does and the user-study then provides a false sense of confidence in the approach. There are other research areas outside of visualization that suffer from exactly that problem and in the end the trustworthiness in that research area suffers from it.

David Laidlaw:

I think that we are in agreement that any experiment should be properly designed. That is easy to say, but quite difficult to do. The design of an experiment can be as intricate and complex as the design of a software system or other artifact. Thomas mentioned several design elements. One is the number of participants. There is no universally proper number of participants – in some cases, a single participant is sufficient for an experiment to be well designed.

I think that our field could benefit from two changes. The first change is to better educate ourselves in how to design experiments. It is hard to know how to do everything, especially in a field like our where we need to communicate across multiple disciplines. But if we are going to use experiments as a core element of our research, we need to know how to design them. We do not need to be the best experimental designers in the world, but we do need to know the basics.

The second change is to better appreciate the good parts of an imperfectly designed experiment. As with any creative artifact, a viewer (or reviewer) can always find ways in which it can be changed or improved. With an algorithm or even a software system, small changes are typically easily implemented. With an experimental design, small changes to the experimental procedures mean doing all of the experiment again. Reviewers often do not seem to weigh the cost of suggested changes against the marginal benefits. I think that most researchers who submit experimental work would like their work to be judged on its merits, not critiqued for re-execution.

Krešimir Matković:

I have noticed that Thomas used the term "user studies" while David used the term "empirical evaluation". Both of you have reached an agreement that "user studies" are not essential for a research paper involving domain experts. I am wondering what is the place of other empirical evaluation methods, such as surveys, discussion groups, think aloud, user testimony, observation diaries, and so on. Should an application paper be published in top journals in the field without any empirical evaluation?

Thomas Wischgoll:

This obviously depends on the application paper. In some cases a user study or other form of empirical evaluation can be very helpful in terms of evaluating the proposed application technique. The list of empirical evaluations is certainly quite extensive which increases the likelihood of one of them being an appropriate evaluation technique. Which one to choose depends on various factors, such as the target audience and what solution the technique is trying to solve. If the targeted audience is fairly small, a user testimony may be more feasible than discussion groups, for example.

However, I do believe that it should be possible to get an application paper published in a top journal without an user-based evaluation. If the authors can make the case for their method to be more effective for a particular application in some way that would be a valid evaluation that shows the usefulness of the approach. For example, the approach could utilize some optimization that makes it perform faster leading to a more effective use of the user's time. I do, however, recognize the fact that lines get blurry fairly fast when we talk about the user's time. If the time saving comes from a more efficient user interface in some way, a more thorough empirical evaluation would be warranted. On the other hand, if the increase in efficiency solely stems from algorithmic improvements, other forms of evaluations can be sufficient.

David Laidlaw:

I used "empirical evaluation" because it is in the title of this part of the book. I consider the choice of which type of empirical evaluation to be a central part of the experimental design. I stand by my assertion that empirical evaluation can be a part of a visualization paper, but I do not think that it must be. If the novelty and significance of the work is compelling without an empirical evaluation, then a paper does not need the evaluation. Some examples of visualization papers with no empirical evaluation and over 500 citations are: Force-Directed Edge Bundling for Graph Visualization, by Danny Holten and Jarke J. van Wijk [8], Marching Cubes, by Lorensen and Cline [17], and The Application Visualization System: A Computational Environment for Scientific Visualization by Upson et al., including me [23].

Krešimir Matković:

Somehow along the lines of both of you, I remember the capstone talk of Jarke van Wijk at IEEE VIS 2013. He said, "Develop new methods/interface/software that are so awesome, cool, impressive, compelling, fascinating, and exciting that reviewers, colleagues, users are totally convinced just by looking at your work and some examples." Should we advise our younger colleagues (and the whole community) to focus on research which does not need evaluation? Can we say that a need for evaluation indicates a lack of awesomeness, coolness, impressiveness in our research? Furter, Smith and Pell in their famous paper in the medical domain [22] argue that randomized control trials are sometimes simply not needed and, still, considered a must in medical research. They illustrate their point on a fictitious case of controlled trials on parachute usage in prevention of death and major trauma related to gravitational challenge. Are you aware of the cases when a visualization paper has been rejected due to a missing study in spite of obvious benefits of the proposed method?

Thomas Wischgoll:

If it is a ground-breaking new technique that is proposed and can stand on its own, I do think that it can be a valuable contribution even without a formal evaluation. If you look at the history, even for VIS, there are a number of publications that fall into that category, most of them probably earlier in history than later.

But this is where it gets tricky: The authors would still have to provide some indication as to why this method is ground-breaking in some aspect. If the benefits can be easily described for the reader to follow, then one may get by without a more formal evaluation. But if not then some form of evaluation in the form of a user study or some other metrics would be warranted.

Looking at the historical context of some of the more successful VIS papers, it used to be relatively common to describe a novel method based on a sample use case or application. The application was then used as some form of evaluation to showcase the utility of the method. The approach then would be picked up by other researchers and extended to different applications who then may include additional evaluation. Over time, this can build a very thorough use case analysis of a visualization algorithm and thus provide great additions to the state-of-the-art in its entirety.

Part two of your question refers to some of the points I was trying to make earlier. But I do believe that it is worthwhile stressing some of those aspects more thoroughly. I think most of us know of recent cases of papers being rejected due to some reviewers considering the evaluation insufficient. In the medical field, randomized control studies seem to be the gold standard for some types of research. However, there are a lot of issues starting with the question as to how one would pick a truly randomized group of people that at the same time reflects the average composition of the population. Sometimes, the size of the group is used as a measure to guarantee this. Other times, statistical methods are applied to reduce the fact that the randomized group was not as reflective of the population as desired. For example, the effect of smoking is sometimes eliminated statistically for that reason. In that case, it would be important to disclose the exact methods used to perform that elimination step in my view. The current debate about reproducibility particularly in the medical domain supports this need. What this example shows is that there are a lot of reasons for why a randomized user study may not show what the authors say they do if the user study was not carefully planned and all the steps taken described clearly in the publication. This is why I would always prefer an evaluation based on some quantitative metric. However, this is not feasible in many cases in the area of visualization since after all it involves visual interpretation by the user.

David Laidlaw:

The papers I mentioned at the end of my last answer are examples of highly-cited publications without empirical evaluations. I do not know if they are “awesome, cool, impressive, compelling, fascinating, and exciting,” but reviewers and citing authors at least found them compelling enough to accept and cite. That suggest

that they were judged novel and significant. Empirical evaluations are not always needed.

As far as part two of your question, there are certainly examples of both false positives and false negatives in our review process. I have seen manuscripts that should have been accepted be rejected because a reviewer insisted that a user study was missing or flawed. I have also seen papers be accepted without sufficient evidence of significance. As a field, we can always strive to improve, and improvements in reviewing would be welcome. One major challenge here is that judging design is difficult. Each design must be judged on its merits in the context of all related work. And if that were not challenging enough, the related work is constantly growing and changing, so the evaluation criteria are, too. I do not think that there is a simple answer here; we need to keep discussing and growing as a community. And we need to avoid making rigid rules.

Krešimir Matković:

Someone could argue that peer-reviews represents a sufficient evaluation for (some of) the visualization research. Suppose that there is an awesome method and if reviewers like it we can consider it evaluated. Could you briefly comment on it, is it also a form of evaluation? Further, I think a commercial success of an innovative visualization method which has not been formally evaluated by means of a user study could also be considered as an evaluation. Could you, please, briefly comment on these thoughts.

David Laidlaw:

Evidence of the significance of research work can take many forms. We have agreed that empirical evaluations can be part of that evidence. Sometimes statements of self-evidence can, too. Other types of evidence might include download counts for software, estimates of installed base or number of users, publication of results created using a visualization artifact, commercial success, or awards. Peer review is the way to evaluate whether a visualization artifact is sufficiently novel and significant, but what is presented should be evidence supporting a positive evaluation. The peer reviewers evaluate that evidence. What that evidence comprises is a part of the research design process.

Thomas Wischgoll:

I completely agree with David. The paper needs to provide some guidance to the reviewers as well as to how to judge the quality of the results. That is the purpose the evaluation serves. A paper can provide some other form of evidence to provide a feel for the quality of the results. But the peer-review process is not really a replacement

as the reviewers do not have access to the full software, for example, they have to solely rely on what is presented in the paper (and potentially additional material, such as a video).

Krešimir Matković:

We have mentioned various forms of evaluation so far. Case studies, user studies, design studies, surveys, etc. It seems that many visualization researcher are not trained in performing studies. Moreover, they do not know which type of study to choose, and what is the difference between some of them. As we do not have a common visualization curriculum at universities, is there a way that all visualization researchers have the same understanding of what evaluation means, in particular user studies? Should we try to find new means to promote importance of evaluation to all, or do you think the most researchers already know it, or do you, maybe, think that awesome research does not need such an evaluation anyhow?

Thomas Wischgoll:

User studies and several other forms of evaluation are probably something the typical visualization researcher is not trained on all that much. This is especially true for the typical scientific visualization researcher. I think in the information visualization community, user studies are significantly more common as the focus and target audience is often times broader, i.e. methods are designed for a larger group of people in information visualization whereas scientific visualization applications sometimes only have a very limited number of users. But of course there are exceptions to that statement as well. But to your point, it seems to me that due to the fact that in the information visualization realm user studies are more common, researchers there probably have a little more experience in that area. However, I would assume that many people were not rigorously trained on that area. So some guidance could be helpful even though there are a number of visualization researchers who attempted to provide such guidance in several publications throughout the recent past.

In several of our studies in the past, albeit not all of them directly related to visualization, we included researchers from psychology and human factors engineering to ensure that the studies follow the necessary scientific rigour on the evaluation part and make sure that the conclusions drawn from the collected data are accurate. In both of those areas, user studies of some form are fairly common and drawing from the experience of those researchers can be very helpful. There is also a lot of existing research on the perceptual side that is relevant to the area of visualization that one can directly tap into and avoid repeating the same type of research or at least use it as a baseline for a formal study.

But to answer your question more directly, I think that most visualization researchers are aware of the need for evaluation and to some extent of the different forms of evaluations, including user studies. However, the degree of what is required

to ensure that the results form such a user study are valid may not be as known to everyone in the community. Some of the other chapters in this book try to give an answer to some of those aspects but there probably is a lot more that could be provided. After all, this is a fairly big topic with lots of different avenues that one can take. And not all of them are valid or appropriate for a given visualization approach.

8.3 Concluding Remarks

The empirical evaluation of the visualization research is far from trivial. As we described above, there are many facets that should be taken care of. We have to select an elevation method, then find appropriate users, correctly execute the evaluation and present results properly. We agreed that we do need evaluation, and, at the same time, it is clear that it is possible to have a valuable and innovative visualization paper without a user study. Finding the proper evaluation method might be tricky.

As not all members of the visualization community have a proper training in performing studies, many visualization researchers and reviewers often colloquially referred to a controlled laboratory experiment as a user study, while many others consider any empirical study involving the **actual** users is a user study. These two definitions not only are quite different, but also have a limited amount of overlapping. A controlled laboratory experiment is typically conducted in a university environment and its participants commonly include a good number of students and university staff. In many applications, a visualization tool or system is designed for a specific group of users who have better knowledge about the data to be visualized and the tasks to be performed. Although it is possible to design a controlled laboratory experiment with domain experts as participants, this approach is not commonly used because (i) it is difficult to design a set of stimuli for complex scenarios, (ii) the variation of users' expertise typically becomes a confounding effect, and (iii) the users may find performing tasks in a controlled setting time-consuming or somehow patronizing. Unlike controlled laboratory experiments or semi-controlled crowd sourcing studies, evaluation with domain experts is difficult to reproduce since others cannot easily replicate the same real-world settings, application-specific tasks, and domain experts with similar knowledge. Nevertheless, the lack of reproducibility is a naturally-occurring feature rather than a shortcoming.

In our opinion, we should strive to better understanding of the evaluation in the visualization community. There is a vibrant subgroup of the community which organizes BELIV workshop at the IEEE VIS conference. As we do not have a common visualization curriculum across all university we recommend to include evaluation topic in teaching of visualization whenever possible. The current state of user studies in visualization often seems like something that has to be done in order for a paper to get accepted instead of contributing to the merit of the paper. This is definitely wrong. We should all learn when to use a user study and when some other means of evaluation is more appropriate. An unsuitable user studies creates more

harm then good to a paper. A proper evaluation enriches the paper and helps in its acceptance, for sure.

Our take home message is, unless your work is really "so awesome, cool, impressive, compelling, fascinating, and exciting that reviewers, colleagues, users are totally convinced just by looking at your work and some examples," consider finding a proper means of evaluation. It will certainly make the paper better. And, even if you write only awesome and fascinating papers, do your homework in study of evaluation methods. We need brilliant minds in evaluation as well. This is probably the only way to ensure an exciting future for visualization research. We did a lot in the last 30 years; we should ensure there will be another fruitful 30 years.

Acknowledgements

The authors would like to thank Laura McNamara for numerous discussions. Her input was very valuable and it helped improve the chapter considerably.

References

1. Abdul-Rahman, A., Chen, M., Laidlaw, D.H.: A survey of variables used in empirical studies for visualization. In: M. Chen, H. Hauser, P. Rheingans, G. Scheuermann (eds.) Foundations of Data Visualization. Springer (2019)
2. Chen, M., Ebert, D.S.: An ontological framework for supporting the design and evaluation of visual analytics systems. Computer Graphics Forum **38**(3), 131–144 (2019). DOI 10.1111/cgf.13677. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13677
3. Demiralp, C., Jackson, C., Karelitz, D., Zhang, S., Laidlaw, D.H.: Cave and fishtank virtual-reality displays: A qualitative and quantitative comparison. IEEE Transactions on Visualization and Computer Graphics **12**(3), 323–330 (2006)
4. Gillmann, C., Wischgoll, T., Hamann, B., Hagen, H.: Accurate and reliable extraction of surfaces from image data using a multi-dimensional uncertainty model. Graphical Models **99**, 13 – 21 (2018). DOI https://doi.org/10.1016/j.gmod.2018.07.004. URL http://www.sciencedirect.com/science/article/pii/S1524070318300365
5. Glendenning, K., Wischgoll, T., Harris, J., Vickery, R., Blaha, L.: Parameter space visualization for large-scale datasets using parallel coordinate plots. Journal of Imaging Science and Technology **60**(1), 10,406–1–10,406–8 (2016)
6. Gomez, S.R., Guo, H., Ziemkiewicz, C., Laidlaw, D.H.: An insight- and task-based methodology for evaluating spatiotemporal visual analytics. In: Proceedings of IEEE VAST (2014)
7. Guo, H., Gomez, S.R., Ziemkiewicz, C., Laidlaw, D.H.: A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. In: Proceedings of IEEE VAST (2015)
8. Holten, D., Van Wijk, J.J.: Force-Directed Edge Bundling for Graph Visualization. Computer Graphics Forum **28**(3), 983–990 (2009). DOI 10.1111/j.1467-8659.2009.01450.x
9. Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., Möller, T.: A Systematic Review on the Practice of Evaluating Visualization. IEEE Transactions on Visualization and Computer Graphics **19**(12), 2818–2827 (2013). DOI 10.1109/TVCG.2013.126

10. Keefe, D., Acevedo, D., Moscovich, T., Laidlaw, D.H., LaViola, J.: Cavepainting: A fully immersive 3D artistic medium and interactive experience. In: Proceedings of ACM Symposium on Interactive 3D Graphics, pp. 85–93 (2001)
11. Kirby, M., Marmanis, H., Laidlaw, D.H.: Visualizing multivalued data from 2D incompressible flows using concepts from painting. In: Proceedings of IEEE Visualization 1999, pp. 333–340 (1999)
12. Koehler, C., Wischgoll, T., Dong, H., Gaston, Z.: Vortex visualization in ultra low reynolds number insect flight. IEEE Transactions on Visualization and Computer Graphics **17**(12), 2071–2079 (2011). DOI 10.1109/TVCG.2011.260
13. Kosara, R., Healey, C.G., Interrante, V., Laidlaw, D.H., Ware, C.: User studies: Why, how, and when. Computer Graphics and Applications **23**(4), 20–25 (2003)
14. Laidlaw, D.H., Kirby, M., Jackson, C., Davidson, J.S., Miller, T., DaSilva, M., Warren, W., Tarr, M.: Comparing 2D vector field visualization methods: A user study. In IEEE Transactions on Visualization and Computer Graphics **11**(1), 59–70 (2005)
15. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical Studies in Information Visualization: Seven Scenarios. IEEE Transactions on Visualization and Computer Graphics **18**(9), 1520–1536 (2012). DOI 10.1109/TVCG.2011.279
16. Lam, H., Tory, M., Munzner, T.: Bridging from goals to tasks with design study analysis reports. IEEE Transactions on Visualization and Computer Graphics **24**(1), 435–445 (2018). DOI 10.1109/TVCG.2017.2744319
17. Lorensen, W.E., Cline, H.E.: Marching Cubes: A High Resolution 3D Surface Construction Algorithm. SIGGRAPH Comput. Graph. **21**(4), 163–169 (1987). DOI 10.1145/37402.37422
18. Matkovic, K., Gracanin, D., Jelovic, M., Hauser, H.: Interactive visual steering - rapid visual prototyping of a common rail injection system. IEEE Transactions on Visualization and Computer Graphics **14**(6), 1699–1706 (2008). DOI 10.1109/TVCG.2008.145
19. Meyer, M., Sedlmair, M., Munzner, T.: The four-level nested model revisited: Blocks and guidelines. In: Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization, BELIV '12, pp. 11:1–11:6. ACM, New York, NY, USA (2012). DOI 10.1145/2442576.2442587. URL http://doi.acm.org/10.1145/2442576.2442587
20. Munzner, T.: A nested model for visualization design and validation. IEEE Transactions on Visualization and Computer Graphics **15**(6), 921–928 (2009). DOI 10.1109/TVCG.2009.111
21. Sedlmair, M., Meyer, M., Munzner, T.: Design study methodology: Reflections from the trenches and the stacks. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2431–2440 (2012). DOI 10.1109/TVCG.2012.213
22. Smith, G.C.S., Pell, J.P.: Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ **327**(7429), 1459–1461 (2003). DOI 10.1136/bmj.327.7429.1459. URL https://www.bmj.com/content/327/7429/1459
23. Upson, C., Faulhaber, T.A., Kamins, D., Laidlaw, D., Schlegel, D., Vroom, J., Gurwitz, R., van Dam, A.: The Application Visualization System: a Computational Environment for Scientific Visualization. IEEE Computer Graphics and Applications **9**(4), 30–42 (1989). DOI 10.1109/38.31462
24. van Dam, A., Laidlaw, D.H., Simpson, R.M.: Experiments in immersive virtual reality for scientific visualization. Computers and Graphics **26**(4), 535–555 (2002)
25. Weber, G.H., Carpendale, S., Ebert, D., Fisher, B., Hagen, H., Shneiderman, B., Ynnerman, A.: Apply or die: On the role and assessment of application papers in visualization. IEEE Computer Graphics and Applications **37**(3), 96–104 (2017). DOI 10.1109/MCG.2017.51
26. Wischgoll, T., Glines, M., Whitlock, T., Guthrie, B.R., Mowrey, C.M., Parikh, P.J., Flach, J.: Display infrastructure for virtual environments (dive). Journal of Imaging Science and Technology **61**(6), 60,406–1–60,406–11 (2017)
27. Wischgoll, T., Scheuermann, G.: Detection and visualization of closed streamlines in planar flows. IEEE Transactions on Visualization and Computer Graphics **7**(2), 165–172 (2001). DOI 10.1109/2945.928168

28. Ziemkiewicz, C., Chen, M., Laidlaw, D., Preim, B., Weiskopf, D.: Open challenges in empirical visualization research. In: M. Chen, H. Hauser, P. Rheingans, G. Scheuermann (eds.) Foundations of Data Visualization. Springer (2019)