Why Ask Why? Considering Motivation in Visualization Evaluation

Position Paper *

Michael Gleicher
Department of Computer Sciences
University of Wisconsin - Madison
gleicher@cs.wisc.edu

ABSTRACT

My position is that improving evaluation for visualization requires more than developing more sophisticated evaluation methods. It also requires improving the efficacy of evaluations, which involves issues such as how evaluations are applied, reported, and assessed. Considering the motivations for evaluation in visualization offers a way to explore these issues, but it requires us to develop a vocabulary for discussion. This paper proposes some initial terminology for discussing the motivations of evaluation. Specifically, the scales of *actionability* and *persuasiveness* can provide a framework for understanding the motivations of evaluation, and how these relate to the interests of various stakeholders in visualizations. It can help keep issues such as audience, reporting and assessment in focus as evaluation expands to new methods.

Keywords

Visualization, Evaluation

1. INTRODUCTION

Visualization serves many rich, complex, and open-ended goals. Its challenges range from helping a scientist make discoveries using a massive amount of data to aiding a mass audience in understanding a complex topic and helping an analyst detect anomalies in a complex network. Evaluation can play many important roles in helping visualization serve its goals. For example, it can help assess how visualizations (and the systems that create them) serve user needs, and can be used by developers to inform the design and implementation of their visualizations. The premise of this paper is that considering the motivations of evaluation can serve as a useful tool in assessing and improving evaluations.

*This paper is a revised version my original position paper, based on the extensive feedback I received on the original.

BELIV 2012, October 14–15, 2012, Seattle, WA, USA

Evaluation for visualization is challenging for a number of reasons. For example, evaluation serves a variety of stakeholders and purposes, and often "what really matters" is a complex and/or long-term outcome that depends on many factors. In response to these challenges, the evaluation methods used in visualization are becoming more diverse and sophisticated. For example, sophisticated empirical designs offer to quantify "what really matters" by measuring high-level outcomes (such as "insights" [7] or "learning" [2]), crowdsourced experiments offer to allow exploration of large parameter spaces, and biometric measurements (eye-tracking, functional near-infrared spectroscopy[8], skin capacitance, ...) offer to quantify viewer response. However, the increasing diversity, complexity and frequency of evaluation raises new questions in how to insure their efficacy.

Effective evaluation requires more than just having sophisticated methods: it requires understanding how these methods serve the motivations, and how they are communicated and interpreted by their audiences (e.g. potential users, other researchers, reviewers, ...). As evaluations become more complex, understanding and assessing them becomes complicated. Therefore, as we introduce new evaluation methods, we must also consider how the use of these methods will be evaluated.

The assessment of evaluation must be be made relative to the motivations of the evaluation, as these high-level goals and their stakeholders are diverse. However, we lack a vocabulary for discussing the motivations for evaluation. As a first step towards a broader consideration of the efficacy of evaluations I propose some vocabulary that provides an initial framework for discussion. The core of the framework is the idea that evaluations can be assessed on their *persuasiveness* and their *actionability* to their audience. This framework allows for considering the range of evaluation motivations, the stakeholders in evaluations, issues in reporting evaluations, and the need for translating empirical results into useful form.

2. MOTIVATIONS FOR EVALUATION

There are many motivations for evaluation, and certain evaluation techniques may serve some motivations better than others. Therefore, I believe it is important that we correctly match motivations to evaluations, which includes gaining a better understanding of the motivations for evaluation, as well as making sure technique development considers a wide range of needs.

The motivation for evaluation (the "why") is subtly differ-

ent than the specific goal of the evaluation (the "what," as in "what we might learn"). For example, a specific goal of an evaluation may be to learn how a system performs as the data size scales. Motivations (or high-level goals) for this specific goal might be to learn where the bottleneck in the system is in order to improve the scalability, to convince a potential user that the system will work on their data, or to persuade a reviewer that the system is better than last year's publication. Munzner's nested model [6] provides a framework for considering the types of specific goals ("what"), and how we can choose the correct type of evaluation (or validation) depending on what we want to learn. There are a similar range of motivations (why do we want to learn that thing), and I believe there is a similar range of appropriateness relationships between the what and why.

In prior work, a dichotomy of motivations has been used (c.f. [1]). The dichotomy comes from educational assessment which makes a distinction between formative and summative assessment. The former refers to assessment done during the process of the learning experience, and the latter refers to assessment done afterwards (e.g. a final exam, or standardized testing in a school). The connotation is that summative assessment usually is done simply for the sake of assessment, and therefore is often used a a pejorative term: it provides little value to the learning process. In HCI and design, the terms have different connotations. They occur in different phases of the design process, implying different methods. Others (c.f. [3]) have discussed the ramifications of stressing summative evaluations in research.

Defining the terms based on when the evaluations occur in the development cycle can be problematic for visualization. For example, iteration means that a summative assessment may inform the next iteration, while an early and informal assessment may help achieve buy-in from potential users. Altering the terms slightly makes them more focused on motivation:

Informative motivations for assessment are to improve the quality of the system (or inform future designs) and, therefore are part of the design process.

Summative motivations for assessment are to understand the qualities of the system.

Any particular evaluation may serve both summative and informative motivations. For example, a performance evaluation may seek to inform further design (such as optimization), or more summative motivations (to convince a potential user that the tool will work in practice). While the informative/summative spectrum may serve in describing the range of motivations, it does not offer direct guidance in how to assess the range of possible evaluations.

As a different framework and basic vocabulary for considering stakeholders and motivations in evaluation, I suggest considering two categories of motivations of (good) evaluations: they can guide and persuade. Any evaluation does both to some degree:

Actionability is a measure of how well an evaluation informs the audience to do some thing. An evaluation may provide specific guidance, or may be non-prescriptive.

Persuasiveness is a measure of how well an evaluation convinces the audience to believe some thing. An evaluation may be a convincing argument or specious.

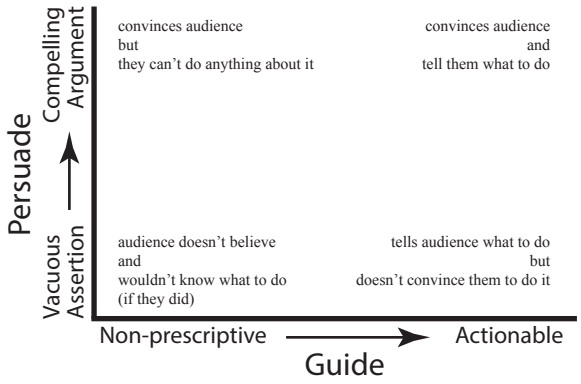

Figure 1: Actionability and persuasiveness provide two independent scales for assessing a visualization.

Actionability and persuasiveness provide two independent scales for assessing a visualization (see Figure 1). Any visualization may do well at one or the other or both.

Actionability and persuasiveness can be considered for whatever an evaluation is trying to show (the "thing") and whomever it is trying to show to (the "audience"). Both are goals in all evaluations, although their relative importance varies. Also, the meaning of the two qualities depends on the context (the what and who). Consider an evaluation of two bioinformatics visualization tools that shows that researchers are more productive with one than another (for example, using an insight-based methodology [7]). For a biology lab director, this evaluation is actionable: it advises them to which tool to buy. However, to a visualization researcher it may be hard to use this evaluation to inform the design of other tools. At the same time, the statistical methodology of the study may be very convincing to a visualization researcher versed in them, but less persuasive to others. In contrast, a pundit proclaiming that one should use one color scheme over another may provide actionable advise to a visualization researcher developing a tool, but not to a user selecting one. However, this researcher may prefer empirical evidence to be persuaded to follow the advice, whereas others may be more accepting of something based on the authority and credentials of the source.

Actionability and persuasiveness are just two of many possible scales on which to assess visualization evaluations. Other scales include generality (the size of the audience, or how often actions suggested will apply) and usefulness (how much of a difference taking the action will make). These scales seem independent (e.g. it is possible to be persuasive and actionable, but not general or useful).

3. IMPROVING EVALUATIONS

Actionability and persuasiveness are positive properties in evaluation. There are situations where one or the other may not be necessary, for example, something far-looking might be inspiring even if there is no action to take from it. But generally, evaluations should seek actionability and persuasiveness.

Improving the actionability of evaluations is important, although there is little general guidance on how to do so. The work of Meyer et al [5] offers a promising start by dis-

cussing how evaluations can lead to *guidelines*, introducing a general way to look at how evaluations can be actionable. Improving the actionability of evaluations, especially to the audience of researchers and developers, is an important problem to be addressed.

Achieving persuasive evaluations involves a range of factors. Part of persuasiveness involves the choice of goals and methods: designing good experiments that measure the correct things. However, assessment and reporting are also an essential part of persuasion. If the audience does not understand an evaluation, or believe in its correctness, it is unlikely to persuade.

The importance of persuasiveness raises issues when the evaluation methods are unfamiliar and/or complex: how do we report the evaluations (as evaluators) and how do we assess those evaluations (e.g. as the audience)? In cases where the methodology is familiar, standards of practice and reporting have emerged. We know what aspects of the experimental design to report, what types of statistical tests are typically applied[1]. The standards for experimental reporting tell us not only what information to report, but what we might expect for it (e.g. what kinds of confidence levels we consider as compelling evidence). Part of the attraction of traditional "time and errors" evaluations are that we have standards for reporting and assessing them. But for newer, or more visualization-specific questions, we have no pre-existing basis. How practical does a model task need to be? How realistic does artificial data need to be to serve in an insight quantification? Can the process of developing new methods include consideration of how the results should be assessed and reported?

Novel and sophisticated evaluation methods provide a trade-off. New methods can offer improved persuasiveness and/or actionability. However, the unfamiliarity and complexity of a sophisticated new method may make its results harder to interpret and translate to practice. Developers of evaluation methods can consider how the applications of their methods may guide and persuade in order to help others use them effectively. An expanded arsenal of methods provides more choices to select one that is appropriate to the goals and audience of an evaluation task. Understanding the motivations of evaluation can help in making such choices.

Personal Statement

I am a Professor of Computer Sciences at the University of Wisconsin, where I am founder and co-director of the Visual Computing Group. I am interested in understanding how we can use our knowledge of perception and artistic traditions in order to create better tools to help people communicate and comprehend. My research and education efforts including not only visualization, but also computer graphics, computer animation, image and video processing, human-robot interaction, human computer interaction, and virtual reality. My work in visualization has included collaborations with several domains including literary scholarship, educational science, virology, genetics, structural biology, and microscopy. I teach courses on visualization, animation, computer games, and graphics.

While I consider myself a relative newcomer to the field of visualization, I have a longer history in other areas where similar issues arise. The complexity of evaluation in Visualization has strong parallels in fields such as animation, image and video processing, and virtual reality. My visualization work has included a variety of evaluation methods, ranging from ad hoc (and, sometimes, admittedly cursory) observations of deployments with domain experts to highly controlled crowd-sourced perceptual studies. Our visualization projects (and publications) have considered a wide range of aspects to evaluate ranging from the applicability of a system to addressing difficult domain challenges to the accuracy, robustness, and scalability of our techniques. The evaluations have run the gamut from formal to ad hoc, and from theoretical to empirical.

[1] Although, the standards of practice may be questioned[4].

Acknowledgments

Miriah Meyer and Michael Sedlmair helped reshape my thoughts on this topic. Our conversations lead to the persuade and guide space. My thoughts on evaluation and empiricism have been shaped through conversations with many people. Much of my experience with doing evaluation has been vicariously through my students.

My work on visualization is supported by NSF awards IIS-0946598, CMMI - 0941013, and IIS-1162037, and work in specific domains are supported by NIH award R01 AU974787 and a Mellon Foundation grant. My work on HCI and animation, which has certainly influenced the thought here, is supported by NSF awards IIS-0941013 and IIS-1208632.

4. REFERENCES

[1] K. Andrews. Evaluation Comes In Many Guises. In *BELIV '08 Workshop*, 2008.

[2] R. Chang, C. Ziemkiewicz, R. Pyzh, J. Kielman, and W. Ribarsky. Learning-based evaluation of visual analytic systems. In *Proceedings of the 3rd BELIV'10 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization - BELIV '10*, pages 29–34, New York, New York, USA, Apr. 2010. ACM Press.

[3] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, page 111, New York, New York, USA, Apr. 2008. ACM Press.

[4] M. Kaptein and J. Robertson. Rethinking statistical analysis methods for CHI. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 1105, New York, New York, USA, May 2012. ACM Press.

[5] M. Meyer, M. Sedlmair, and T. Munzner. The Four-Level Nested Model Revisited: Blocks and Guidelines, 2012. In this proceedings.

[6] T. Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–8, Jan. 2009.

[7] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, May 2006.

[8] E. Peck, E. Solovey, S. Su, R. Jacob, and R. Chang. Near to the brain: Functional near-infrared spectroscopy as a lightweight brain imaging technique for visualization. In *IEEE Conference on Information Visualization (InfoVis) Posters*, 2011.