VisHikers' Guide to Evaluation: Competing Considerations in Study Design

Emily Wall, Emory University, Atlanta, GA, 30322, USA
Cindy Xiong, University of Massachusetts - Amherst, Amherst, MA, 01003, USA
Yea-Seul Kim, University of Wisconsin - Madison, Madison, WI, 53706, USA

In this Viewpoint article, we describe the persistent tensions between various camps on the "right" way to conduct evaluations in visualization. Visualization as a field is the amalgamation of cognitive and perceptual sciences and computer graphics, among others. As a result, the relatively disjointed lineages in visualization understandably approach the topic of evaluation very differently. It is both a blessing and a curse to our field. It is a blessing, because the collaboration of diverse perspectives is the breeding ground of innovation. Yet it is a curse, because as a community, we have yet to resolve an appreciation for differing perspectives on the topic of evaluation. We explicate these differing expectations and conventions to appreciate the spectrum of evaluation design decisions. We describe some guiding questions that researchers may consider when designing evaluations to navigate differing readers' evaluation expectations.

Imagine that you are a visualization researcher (skip this step if you already are one). You just got the reviews back for your most recent submission. Do you dare look?

Brave fictional researcher #1 Mojo opened the web page containing the reviews. She has just submitted a paper on how people perceive pie charts. The paper contains three carefully designed studies that involve showing participants pie charts with varying design features to evaluate how quickly people can extract key statistics from them. The paper was brutally rejected. The reviews pointed out that the datasets used to generate these pie charts were too limited, and the study itself was too abstract and artificial.

Brave fictional researcher #2 Jojo also reads her reviews. Her paper introduced a technique integrated into a system, whose evaluation with in-lab participants assessed the technique's performance to promote reflection of unconscious decision-making strategies. Also brutally rejected, reviewers critiqued the lack of control in the study and the abundance of potential confounds.

Do these experiences sound familiar? As a research community, we often struggle to design experiments that balance readers' expectations. Maybe there is not much we can do about the late hours we work, but there must be something we can do about designing our studies in a way that ensures the quality of our work while at the same time meeting these differing expectations.

We assert that this tension arises as a result of *opposing lineages* in the visualization community. In particular, one must look to the diversity of fields that blend to form our field. VIS is influenced by fields such as computer graphics, semiotics, HCI, cognitive science, vision science, graphic design, and cartography, among others. Each of these fields provides varying perspectives and approaches to the challenges in visualization research,[1] and, in particular, the preferred approach to evaluation. For instance, researchers from a computation background tend to make contributions, such as models or systems, which are often evaluated with discrete metrics (e.g., accuracy) via simulation; HCI

0272-1716 © 2022 IEEE
Digital Object Identifier 10.1109/MCG.2022.3152676

researchers make contributions that are theoretical, empirical, or artifacts, but evaluation typically involves more human-centric experimentation using mixed qualitative and quantitative methods[2]; psychology researchers often make contributions in the form of empirical findings from highly controlled laboratory experiments. With this breadth and diversity in backgrounds and expectations for evaluation, it is understandable that there is often disagreement in visualization about what constitutes a well-designed evaluation.

AS A RESEARCH COMMUNITY, WE OFTEN STRUGGLE TO DESIGN EXPERIMENTS THAT BALANCE READERS' EXPECTATIONS.

We acknowledge that critique and disagreement are a natural part of research. Particularly in the VIS community, with such a broad range of backgrounds and experiences that forged our field, the spectrum of evaluation methods is likewise broad. In this Viewpoint article, we describe some of the lineages in our community and their respective traditional expectations and norms for evaluation. We wade through the collective amassment of our rejected evaluation papers to bring you lovingly distilled lessons learned. We focus on describing several decision points of evaluation design and suggest concrete guidance for ways to think through these choices. This is not a guide for how to pander to reviewers; rather, we hope that this will help guide researchers through the often conflicting goals of an evaluation.

BACKGROUND

Evaluation in VIS has been a hotly contested topic for quite some time. Critiques often stem from the community's dissatisfaction with insufficient evaluation techniques as well as lack of clarity or efficacy in applying evaluation techniques. For instance, researchers often express concerns with methods, such as measuring the accuracy and time for users to perform benchmark tasks with a visualization, which does not provide researchers or developers insights into the benefits of a visualization or visualization tool nor actionable items to improve them. This reflects a growing need in the visualization community to consider evaluation techniques across all stages of development to generate research-driven evidence demonstrating the benefits of visualization.[3]

Emerging Work

In 2006, the BELIV workshop emerged as a respected venue for novel evaluation methodologies.[4] The initiative has inspired an abundance of progress in evaluation methodologies. For example, Shneiderman and Plaisant proposed a method called multidimensional in-depth long-term case studies (MILCS), which assesses visualization tools based on observations, interviews, surveys, and an expert user's likelihood to achieve their goals with the tool over an extended period of time to obtain multiple perspective on the tool's effectiveness.[5]

Stasko[6] proposed a framework describing the value of visualization, which contained components describing how a visualization can provide time savings and insights, convey data, and inspire user confidence in data. Inspired by this framework, Wall *et al.*[7] created a heuristic-based methodology that enables evaluators to identify the strengths and weaknesses of a visualization by quantitatively rating ~~heuristics~~ along several dimensions. Researchers have also referenced evaluation taxonomies from education and humanities literature, such as evaluating a visualization using learning outcomes and Bloom's taxonomy.[8]

The initiative has also motivated researchers to create guides on how to conduct evaluation studies. For example, Elliott *et al.*[9] introduced a lexicon of experimental design for empirical user studies, applying methodologies from human visual perception studies to evaluate visualizations, describing novel experimental paradigms, and dependent measures specific to the visualization community. Sedlmair *et al.*[10] provided a guide for conducting design studies over nine stages (learn, winnow, cast, discover, design, implement, deploy, reflect, and write) and discussed potential pitfalls.

Other researchers categorized these evaluation methods based on their high-level purposes in answering a research question[11] and mapped out how these purposes connect to the appropriate broader research contribution.[2] For example, Munzner[12] proposed a nested model that guides visualization researchers to select the appropriate evaluation approach in four levels of visualization design and validation: characterizing the task and data, abstracting the characterization into operations, designing visual encoding and interactions, and creating algorithms to execute the techniques efficiently.

~~But~~ ~~d~~espite these efforts, the debate on how a visualization should be evaluated rages on. Researchers and practitioners debate the criteria for measuring the value of a visualization,[6] expressing concerns on the reproducibility of evaluation studies,[13] and continue to share new expectations for visualization evaluations. Also, quietly in the background, lurks the same debate in the form of a long discussion between paper reviewers and paper chairs.

Contributions

Recently, the visualization research community's flagship conference venue, IEEE VIS, has gone through a remodel where, instead of separating submissions into three subconferences (VAST, InfoVis, and SciVis), submissions are now distributed into the following six areas: theoretical and empirical; applications; systems and rendering; representations and interaction; data transformations; and analytics and decisions. These areas may be roughly conceived of as contribution types. Alternatively, Wobbrock and Kientz[2] describe seven contribution types in HCI research, including empirical (new observations), artifact (new tools), methodological (new practices), theoretical (new concept or model), dataset (new corpus), survey (new reflection on a collection of past work), and opinion (new perspective).

What makes visualization papers (or generally speaking, any interdisciplinary academic research) especially strong and unique contributions to the scientific community is that one visualization paper often touches on multiple areas and brings multiple forms of contribution. For example, a visualization paper submitted to the theoretical and empirical track at IEEE VIS could, in addition to its empirical contributions, introduce a novel research method, a new dataset, and a comprehensive background section that synthesized a large amount of past work to be considered a survey contribution. A paper submitted to the analytics and decisions area might contribute to the visualization community a new data analytic system (an artifact), along with an empirical observation that reveals new theoretical insights.

There are many ways to describe possible contribution types, but here we will focus on the following three: factor, system, and technique, which often have different evaluation expectations. A *factor* contribution usually tells the story of how one design element impacts the visualization and its interpretation. For example, Ceja *et al.*[14] demonstrated that the aspect ratio of a visualization can influence how accurately people perceive data. A *system* contribution showcases a novel tool to help people build visualizations or analyze data, such as Gratzl *et al.*'s work,[15] which supports visual exploration of rank data. For systems, evaluations are needed in order to make claims about how effective the system is. Finally, a *technique* contribution points to one specific component (often in a system) and demonstrates that manipulation of that technique can impact how people make sense of visualizations. For example, Wall *et al.*[16] demonstrated an approach to displaying user interaction history that may increase awareness of cognitive or societal biases that drive behavior and decisions in data analytics.

This categorization complements the nested model proposed by Munzner[12] by focusing on the end-product of visualization research, the evaluation of which can consist of any combination of the four levels from Munzner's work.[12]

Challenges

While the multifaceted contributions of visualization papers can lead to significant innovation, they also introduce many challenges in the paper review process. To evaluate a paper, the primary reviewer needs to gather a group of reviewers with diverse backgrounds and ~~expertize~~ to ensure a holistic evaluation of the paper's contributions.

This is especially beneficial for papers that touch on multiple areas and make multiple different contributions. However, one prominent issue often arises: reviewers from different areas might judge the paper in terms of its contribution in the one area they are familiar with, without considering the other forms of contribution the paper brings. This can lead reviewers to find the work underwhelming.

As a result, it becomes increasingly difficult for one paper to reconcile the differing expectations from a group of authors and a group of reviewers with diverse experiences. The authors may feel pressured to make artificial additions or omissions to please the reviewers. For example, the long-running academic cliché laments that reviewer number two makes unreasonable demands, such as asking authors to conduct a full-fledged controlled study, in stark contrast to other reviewers who would prefer to see an ecologically valid study!

There are occasions where it is unnecessary or impossible to run a perfectly balanced experiment or user study that covers all possible confounds and simultaneously maintains ecological validity. In fact, many academics argue that it is a detriment to theoretical advancement to attempt to maximize external validity in a given experiment.[17] We need to collectively acknowledge that no perfect experiment exists, and one paper typically cannot solve an entire problem space. Papers ought to be judged on the experiments conducted and contributions made, rather than the ones they did not.

THINKING ABOUT STUDY DESIGN

We structure our discussion of evaluation guidance around the three common types of contributions

RESEARCH DESIGN RESOURCES

There are a variety of articles on research methodologies in computing and psychology that we found helpful in building this guide to evaluation design. Listed below are a few of our favorites.

R. Elio, J. Hoover, I. Nikolaidis, M. Salavatipour, L. Stewart, and K. Wong, "About computing science research methodology," 2011.

L. Berkowitz and E. Donnerstein, "External validity is more than skin deep: Some answers to criticisms of laboratory experiments.," *Amer. Psychologist*, vol. 37, no. 3, p. 245, 1982.

E. Wall, M. Agnihotri, L. Matzen, K. Divis, M. Haass, A. Endert, and J. Stasko, "A heuristic approach to value-driven evaluation of visualizations," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 491–500, 2018.

B. Shneiderman and C. Plaisant, "Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies," in *Proc. AVI Workshop BEyond Time Errors: Novel Eval. Methods Inf. Vis.*, 2006, pp. 1– 7.

A. Burns, C. Xiong, S. Franconeri, A. Cairo, and N. Mahyar, "How to evaluate data visualizations across different levels of understanding," *IEEE Workshop Eval. Beyond-Methodol. Approaches Vis.*, pp. 19–28, 2020.

M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2431–2440, 2012.

described earlier in the "Background" section: factor, system, and technique contributions. How a user study is designed by researchers and evaluated by reviewers should depend on the type of contribution, the paper claims to make. We identify a guide of several components to help researchers design their studies and likewise help reviewers evaluate these studies.

Formulating Research Questions

Research begins by defining research questions and corresponding claims researchers hope to address. The research question should be directly connected to the type of contributions, the paper sets out to make. Munzner's[12] nested model for visualization design and validation emphasized the importance of asking the right research question, because addressing "the wrong problem" threatens the validity of every step downstream in the research process. In this section, we provide guidance on choosing an evaluation that aligns with the specific research questions for a feature, system, and technique.

Example: Let's imagine you are a visualization researcher specializing in cognitive bias in visual analytics. You want to conduct a study that focuses on bias mitigation. Inspired by previous work demonstrating that having a user externalize their prior belief through drawing can increase data recall,[18] you come up with the idea that people will be less susceptible to cognitive biases in visual analytics if they can compare their mental representation of the relationship between variables to the actual relationship between variables. There are three ways you can approach your work: from a factor, system, or technique perspective.

If you want to make a *factor* contribution and determine possible associations or causal relations between factors, it is probably a good idea to conduct a highly controlled user study where the only difference between the conditions tested is that factor. For the case study described, a good research question might be "how does comparing a mental representation of the relationship between two variables to the actual relationship influence one's interpretation of data?"

If you want to make a *system* contribution, your study should test whether your system improves an existing visualization workflow, based on valid user behaviors and intentions.[6,12] The comparison to be made here should be between the outcome from when people use your system and the outcome from when people do not. In the case study described, you will probably want to design and build a visualization system that can support bias mitigation. A good research question might be "will people be less biased when they analyze data using my system?" Notice that the research question does not specifically talk about the effect of any specific design elements in your system, as it is up to you to how you want to operationalize these elements.[19]

TABLE 1. Description of three contribution types that we will focus on.

Contribution type	Description	Example RQ
Factor	Usually tells the story of how one design element impacts the visualization and its interpretation	How does comparing a mental representation of the relationship between two variables to the actual relationship influence one's interpretation of data?
System	Showcases a novel tool to help people build visualizations or analyze data	Will people be less biased when they analyze data using my system?
Technique	Points to one specific physical component (often in a system) and demonstrates that a manipulation of that technique can impact how people make sense of visualizations	Is the process of comparing mental representations to actual data mitigating cognitive bias in data analysis with my system?

Your system might include a novel technique that allows users to compare their mental representations of a data relationship, and you are sure that is the key to mitigate biases. It may be very tempting to add in your research question that this feature in the system mitigates biases. However, that turns your paper's contribution into a *technique*, rather than a *system*, and it will need to be evaluated differently, because a system is a complex collections of multiple techniques. Perhaps there is one technique in the system that is the key driver to mitigate bias, or perhaps the system pushes people to do analytic tasks in a certain order, and that order is what truly mitigates biases. If you additionally want to make claims about *why* your system works, you need to ask additional research questions at the technique level.

> *DUE TO OPPOSING LINEAGES, THERE MAY BE A TENSION AMONG READERS' EXPECTATIONS FOR WHAT THE RESEARCH QUESTIONS OUGHT TO BE THAT YOU ADDRESS.*

If you want to make contributions at a *technique* level, you should think about how your technique can make a visualization system "better." The comparison you want to make in your study should be between the outcome for when people use a system with the target technique and the outcome for when people use the same system without the target technique, where other potential confounds are isolated. Notice how the system is kept constant in the comparison in this research question. This is because you need to justify the capability of *the technique itself* to combat the threat to the validity.[12] In the running example on bias mitigation, you might want to make a claim of why your system works to mitigate bias. A good question might be "is the process of comparing mental representations to actual data mitigating cognitive bias in data analysis with my system?" These contribution types are tabulated in Table 1.

Due to opposing lineages, there may be a tension among readers' expectations for what the research questions ought to be that you address. It is possible then, or perhaps encouraged, to ask multiple research questions from different perspectives in your paper. A research question can be exploratory (which aims to navigate problem spaces to formulate hypotheses, often formulated prior to seeing data) or confirmatory (which aims to test a preexisting hypothesis the researchers have, often formulated after seeing data). It is critical to make sure, as mentioned in Munzner's work[12] that the research question always matches the output. The incongruency between the two tends to be a common source of critique from readers. It is on you to appropriately scope the research question and to motivate the problem space in the introduction and throughout ~~this~~ article to communicate why the exact questions contribute to visualization design and systems.

Designing Conditions

Once you have formulated your research question, you should have a sense of what type of comparison you want to make in your study to validate your hypotheses and test the capabilities of your technique or system. This means coming up with the right testing conditions to answer your research questions.[19]

Example: Let's continue with the bias mitigation example. Let's say your focus is at a *factor*-level and your research question is "how does comparing a mental representation of the relationship between two variables to the actual relationship influence one's interpretation of data?" The key comparison here is how people interpret data in two scenarios: when they are able to compare a mental representation of the relationship to the actual data, and when they are not able to (also known as the control condition). Your study design should cover at least these two situations.

If your research focuses on evaluating your *system*, let's say your research question is "will people be less biased when they analyze data using my system?" You should include conditions that help you make the comparison between people's performance using your system versus another system or no system.

Let's say your contribution is at the *technique*-level, and your paper asks "is the process of comparing mental representations to actual data mitigating cognitive bias in data analysis with my system?" You should minimally test people's performance in your system with or without this technique by keeping everything else constant so you know exactly to what extent this technique has an effect on user performance.

> *THE MOST COMMON TENSION AMONG REVIEWERS WITH RESPECT TO EXPERIMENTAL CONDITIONS TYPICALLY LIES IN AN ISOLATION OF CONFOUNDING VARIABLES.*

The most common tension among reviewers with respect to experimental conditions typically lies in an isolation of confounding variables. To summarize, a system-level question where the conditions are "using system" and "not using system" can only warrant system-level conclusions, such as "we demonstrate that our system can help people perform better than no system." In this scenario, you cannot make technique-level claims and say "we demonstrate that our system can help people perform better because of this technique" because you did not design study conditions to specifically test the effect of the technique, nor can you make factor claims about the mechanism of mental comparison toward mitigating bias since confounding variables were not isolated.

Internal and External Validity

In designing the experiment, you may wish to also ensure both internal and external validity.[19] This forms another source of tension for authors to manage. One common complaint from reviewers for factor-based investigations is that these controlled studies lack external validity, meaning how well the outcome of a study can be expected to apply to other settings in general. On the other hand, reviewers also often complain that system-based studies lack internal validity, which focus~~es~~ on eliminating alternative explanations for a finding. So to address these common issues, we discuss a few study designs to provide an idea how to enhance both external and internal validity. Note, however, that it may not always be desirable to balance both internal and external validity. For instance, theoretical advancements may be slowed by overindexing on external validity.

Example: For a *factor*-level investigation, say the question is "how does comparing a mental representation of the relationship between two variables to the actual relationship influence one's interpretation of data~~.~~" You probably want to operationalize all the relevant factors mentioned in your research question~~,~~ and create sets of conditions to test the effect of each factor so you can pinpoint the factor(s) driving your effect.

For example, what is a mental representation of a relationship? Does this mean thinking about it? Drawing it out? Verbally describing ~~via~~ a sentence? What kind of variables do you want to focus on? Continuous variables? Discrete variables? If you want to come up with generalizable results, you might want to test all types of mental representations on both continuous and discrete variables. Then, your conditions should cover all permutations of these two factors ({thinking, drawing, verbalizing} x {discrete, continuous}) to yield six conditions.

But that is not all, you can keep asking yourself to further operationalize other factors in your research question: what is the actual relationship people should be comparing their mental representation with? Is this a visualization made from the underlying data? Perhaps a verbal description of a key insight? How would you measure "influence"? What about interpretation of data? Now you realize that this design space will expand exponentially as the number of permutations and combinations grows by adding more conditions to make your results more generalizable, but the amount of resources is finite. You cannot possibly run a well-powered study with 147 conditions and cram your findings in a nine-page paper. So now what?

We recommend you start by listing the entire design space of the experiment. Suppose you want to design an experiment with three experimental variables (A, B, and C). First, you should list all the levels within each variable to exhaust the possibilities. For example, you could identify three levels for each variable (A1, A2, A3, B1, B2, B3, C1, C2, C3). Ideally, you can test all conjugated conditions, totaling 27 conditions. However, there are several reasons why testing all the conditions may not be necessary. For example, prior work demonstrated that the combination of A1, B1, and C1 does not improve the task performance. You can eliminate the condition to save some time. You can also consult existing theories to narrow down the condition space.

YOU CANNOT POSSIBLY RUN A WELL-POWERED STUDY WITH 147 CONDITIONS AND CRAM YOUR FINDINGS IN A NINE-PAGE PAPER. SO NOW WHAT?

Explicitly listing all the variables, levels, and the combined conditions is a necessary step toward thinking about the entire space first to ensure internal validity. Researchers can then narrow down the space based on prior work and *communicate this process in their article*. This will enable reviewers and readers to follow the reasoning behind why a limited set of conditions have been tested and the rationale for those choices. Based on this, you can circle back to your research questions and re-scope it to match your study conditions (e.g., if you only tested the effect on continuous variables but not discrete variables, you may scope the research question down to explicitly focus on continuous variables). In a similar way, for *system*-level investigations, the goal is to demonstrate that your system actually works. To ensure internal validity, you want to make sure the only difference between your two conditions is whether your system is being used to complete the task or not. That means external variables like the task being tested, the participants' level of expertise in the task domain, among other things, should be kept constant.

If you also want to make *technique*-level claims, then you need to ensure that the only thing that differs between your conditions is whether that specific technique exists or not. You should not compare a system with the targeted technique to a system without it, unless the two systems are identical to each other. Otherwise it violates internal validity; since there could be other differences in the system that make people perform better/worse, in addition to the technique of interest.

Choosing a Task

Now that you have your research questions formulated and your study conditions scoped, it is time to think about what tasks you want users to complete for your study. The internal and external validity as well as the claims that you want to make should be considered in choosing a task in your study.

NOW THAT YOU HAVE YOUR RESEARCH QUESTIONS FORMULATED AND YOUR STUDY CONDITIONS SCOPED, IT IS TIME TO THINK ABOUT WHAT TASKS YOU WANT USERS TO COMPLETE FOR YOUR STUDY.

For a *factor*-level contribution, it is important to abstract your task so that you can create an isolated environment to pinpoint the effect of your manipulation, such as how Amar *et al.*[20] and Brehmer and Munzner[21] have abstracted analytic tasks for researchers to use to evaluate visualizations. However, tension can arise when the task is too abstracted since the study can lose generalizability to real-world settings. In these situations, the task may feel artificial to users, and the decisions they make in completing the task may no longer be good approximations of their actions in the real world. In these cases, an abundance of clarity in communicating the choices and tradeoffs can preempt many readers' concerns.

Example: In the bias-mitigation scenario, let's say we need some task that can capture a user's prior belief of a relationship, so we can see how that belief changes as biases are introduced or mitigated. Since you cannot randomly assign people to suddenly believe in one thing or another, if you want full control over people's prior beliefs, you likely need to make up an entirely artificial scenario and prime people with beliefs (e.g., the likelihood of a plant growing on an alien planet). But this will make the task seem contrived, and participants might not take it seriously or respond in ways that would reflect their behaviors with real long-held beliefs.

Alternatively, you can take the organic approach and select scenarios where people's true belief can be easily measured and predicted and recruit people that hold a specific belief. This will likely resemble real-world scenarios more. However, beliefs that are easily measured and predicted are often associated with strong emotions, such as political orientation. For such strong or emotionally charged beliefs, you might not be able to observe changes in belief related to biases. Although realistic, these emotional attachments may influence your results.

For a *system-* or *technique*-level contribution, abstracting a task to be short and simple can isolate confounding variables and make the specific task outcome more easily measured and controlled. However, systems are rarely designed for very simple tasks (e.g., Wall *et al.*'s work[16]). This makes evaluating performance via a simple task artificial and less useful for real-world usage scenarios. On the other hand, using real-world tasks usually means factors that may not be of research interest also play a part in the user workflow. This makes the outcome noisier to measure and the effect of a system or a technique enhancing performance on one specific task more difficult to isolate from the influence of other factors or steps in the workflow. For example, let's say the system we designed mitigates bias by helping people see correlations in data more accurately. The abstracted control task may be to have users extract correlations from visualizations. In this case, the user views a visualization in the system and estimates a correlation value. These correlation estimates are compared to estimates made when the users view the same visualization in a different system. But reading correlation values from scatterplots is rarely the ultimate goal in a real-world system that supports a data analysis workflow. Just because people can more accurately read correlations from one system over another does not mean they are going to analyze the data, think about the data, or present the data in a less biased way. So alternatively, you may want to design a task that more closely resembles the real world.

Due to this tradeoff and potential tensions that can result between an abstracted, fully controlled task and a complex, real-world task, we recommend researchers to *consider planning for multiple studies*. It is possible to start with a more controlled setting in the first study to detect and quantify an effect~~,~~ and then move to a more realistic setting in subsequent experiments where you examine whether the effect generalizes. This tradeoff in task choice is also a form of tradeoff of research contributions.

Picking a Dataset

The next consideration is with which dataset to design your visualization task. To ensure internal validity, you may want to generate your own datasets or look for datasets with specific characteristics. This way, you will have more control over what the visualization looks like and what type of analytic tasks users can perform. Some characteristics of a dataset that are manipulable might include the distribution (e.g., normal, uniform), how many discrete versus continuous variables there are, the number of abnormal/outlier points, the size, etc.

Example: Continuing with the bias mitigation example, you want to see if people can become less biased in their data analysis with your intervention. Ideally you should test the effectiveness of your intervention with several different datasets with differing characteristics to see how much your results can generalize before making claims. For example, it is possible that your intervention can only mitigate biases when the dataset is normally distributed~~, but~~ does not work when the dataset has more than 15% outliers.

However, there are too many possible characteristics of a dataset for it to be possible to control for everything. Identifying every manipulable characteristic of a dataset and creating a separate condition for each will likely consume too many resources~~,~~ and dilute the focus of the research question. In addition, researcher-generated datasets may not resemble real-world datasets, reducing the external validity of the study as the study results may not generalize to real-world settings. These competing considerations can be another source of tension. A reader from a psychology background may expect these variables to be controlled for, while a reader from an HCI background may expect to see a realistic dataset.

We offer the following two considerations to help researchers who wish to balance these competing concerns in user studies: 1) start from a real-world dataset and manipulate the characteristics to gain control (e.g., by adding or removing columns or rows, altering the distribution of a variable, etc.), or 2) list the assumptions of a realistic data generating process and simulate the data that also meets the control characteristics you need. These options give you the flexibility to have both realism in the user's perception of the data while maintaining control in the methods of analysis; but if you cannot have both, you need to choose one and stick to it.

From a Reviewer's Perspective

The flip-side of this guidance for researchers likewise applies to reviewers. Reviewers should *assess the work according to how relevant the claims are to visualizations and whether the evidence supports the claims*. For papers that claim a factor-level contribution, assess how well the factor was isolated in influencing a phenomenon. For papers that claim a system-level contribution, assess whether the study is designed in such a way that it can capture differences in alternative systems toward helping people achieve their goals under the same conditions. For a paper that claims a technique-level contribution, reviewers should assess that the same system functions significantly differently with and without the technique. Furthermore, we encourage that reviewers *assess the contributions of the actual research that was done*, giving benefit of the doubt and, where appropriate, opportunity for authors to respond or revise when the language or framing of that research lacks precision (within reason).

CLOSING THOUGHTS

In an ideal world, we would be able to satisfy all internal and external validity goals in our evaluations. Realistically, however, we do not have infinite resources to realize all these oft-competing constraints. In this Viewpoint, we have described some practical guidelines to help VisHikers navigate the galaxy of evaluation. While these guidelines will hopefully serve as a reasonable starting point in designing an evaluation and communicating those study design choices with precision, there are a number of other considerations we have barely touched on. We want to emphasize that there are other elements to consider in your study in addition to the experimental conditions.

> *IMPORTANTLY, HOWEVER, THE PERFECT STUDY DOES NOT EXIST.*

For example, while it may be tempting to identify multiple conditions to test in one study, unless you *conduct proper power analysis* to ensure you have the appropriate power to test your hypotheses, you risk collecting noisy measurements and observing unrepresentative effects. Similarly, you must *be careful not to overstate the contribution*. If you only compared people's performance using ~~their~~ system versus another state-of-the-art system using task A and B, then claim that "our system performs better at tasks A and B than the state-of-the-art system," rather than "our system performs better than this other system," or "our system is the best." Researchers must also consider how to carefully formulate their hypotheses, how to appropriately measure the phenomena of interest, how to design a realistic data-generating process, and hopefully balance all of this within the context of a study that has some practical significance.

Importantly, however, the perfect study does not exist. There will always be tradeoffs that need to be weighed and managed. We hope that this guidance will help re-enforce a critical and comprehensive lens for researchers to consider their evaluation designs.

ACKNOWLEDGMENTS

The authors would like to thank Jessica Hullman and John Stasko for their invaluable feedback.

REFERENCES

1. R. M. Kirby and M. Meyer, "Visualization collaborations: What works and why," *IEEE Comput. Graphics Appl.*, vol. 33, no. 6, pp. 82–88, Nov./Dec. 2013.
2. J. O. Wobbrock and J. A. Kientz, "Research contributions in human-computer interaction," *Interactions*, vol. 23, no. 3, pp. 38–44, 2016.
3. C. Plaisant, "The challenge of information visualization evaluation," in *Proc. Work. Conf. Adv. Vis. Interfaces*, 2004, pp. 109–116.
4. B. Lee, C. Plaisant, C. Parr, J. Fekete, and N. Henry, "Task taxonomy for graph visualization," in *Proc. AVI Workshop Beyond Time Errors: Novel Eval. Methods Inf. Vis.*, 2006, pp. 1–5.
5. B. Shneiderman and C. Plaisant, "Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies," in *Proc. AVI Workshop Beyond Time Errors: Novel Eval. Methods Inf. Vis.*, 2006, pp. 1–7.
6. J. Stasko, "Value-driven evaluation of visualizations," in *Proc. 5th Workshop Beyond Time Errors: Novel Eval. Methods Vis.*, 2014, pp. 46–53.
7. E. Wall *et al.*, "A heuristic approach to value-driven evaluation of visualizations," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 491–500, Jan. 2019.
8. A. Burns, C. Xiong, S. Franconeri, A. Cairo, and N. Mahyar, "How to evaluate data visualizations across different levels of understanding," in *Proc. IEEE Workshop Eval. Beyond-Methodological Approaches Vis.*, 2020, pp. 19–28.
9. M. A. Elliott, C. Nothelfer, C. Xiong, and D. A. Szafir, "A design space of vision science methods for visualization research," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1117–1127, Feb. 2021.
10. M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012.
11. N. Elmqvist and J. S. Yi, "Patterns for visualization evaluation," *Inf. Vis.*, vol. 14, no. 3, pp. 250–269, 2015.
12. T. Munzner, "A nested model for visualization design and validation," *IEEE Trans. Vis. Comput. Graphics*, vol. 15, no. 6, pp. 921–928, Nov./Dec. 2009.

13. J.-D. Fekete and J. Freire, “Exploring reproducibility in visualization,” *IEEE Comput. Graphics Appl.*, vol. 40, no. 5, pp. 108–119, Sep./Oct. 2020.
14. C. R. Ceja, C. M. McColeman, C. Xiong, and S. L. Franconeri, “Truth or square: Aspect ratio biases recall of position encodings,” *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1054–1062, Feb. 2021.
15. S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, “LineUp: Visual analysis of multi-attribute rankings,” *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2277–2286, Dec. 2013.
16. E. Wall, A. Narechania, A. Coscia, J. Paden, and A. Endert, “Left, right, and gender: Exploring interaction traces to mitigate human biases,” *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 966–975, Jan. 2022.
17. B. J. Calder, L. W. Phillips, and A. M. Tybout, “The concept of external validity,” *J. Consum. Res.*, vol. 9, no. 3, pp. 240–244, 1982.
18. Y.-S. Kim, K. Reinecke, and J. Hullman, “Explaining the gap: Visualizing one’s predictions improves recall and comprehension of data,” in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 1375–1386.
19. P. C. Cozby and S. Bates, *Methods in Behavioral Research*. Mountain View, CA, USA: Mayfield Pub., 1985.
20. R. Amar, J. Eagan, and J. Stasko, “Low-level components of analytic activity in information visualization,” in *Proc. IEEE Symp. Inf. Vis.*, 2005, pp. 111–117.
21. M. Brehmer and T. Munzner, “A multi-level typology of abstract visualization tasks,” *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2376–2385, Dec. 2013.

EMILY WALL is an Assistant Professor with Emory University, Atlanta, GA, USA. Her research focuses on decision-making with data and visualizations, particularly as applied to problems of societal concern. She received the Ph.D. degree in computer science from the Georgia Institute of Technology, Atlanta. She is the corresponding author of this article. Contact her at emily.wall@emory.edu.

CINDY XIONG is an Assistant Professor with the University of Massachusetts Amherst, Amherst, MA, USA. Her research interests include perception, cognition, and data visualization, and she investigates how humans perceive, interpret, and make decisions from visualized data. She received the Ph.D. degree in psychology from Northwestern University, Evanston, IL, USA. Contact her at cindy.xiong@cs.umass.edu.

YEA-SEUL KIM is currently an Assistant Professor with the University of Wisconsin-Madison, Madison, WI, USA. Her research focuses on developing tools and algorithms to help people with varying abilities interact with data and visualizations. She received the Ph.D. degree in information science from the University of Washington, Seattle, WA, USA. Contact her at yeaseul.kim@cs.wisc.edu.

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com