A Systematic Review on the Practice of Evaluating Visualization

Tobias Isenberg, *Senior Member, IEEE*, Petra Isenberg, Jian Chen, *Member, IEEE*,
Michael Sedlmair, *Member, IEEE*, and Torsten Möller, *Senior Member, IEEE*

Abstract—We present an assessment of the state and historic development of evaluation practices as reported in papers published at the IEEE Visualization conference. Our goal is to reflect on a meta-level about evaluation in our community through a systematic understanding of the characteristics and goals of presented evaluations. For this purpose we conducted a systematic review of ten years of evaluations in the published papers using and extending a coding scheme previously established by Lam et al. [2012]. The results of our review include an overview of the most common evaluation goals in the community, how they evolved over time, and how they contrast or align to those of the IEEE Information Visualization conference. In particular, we found that evaluations specific to assessing resulting images and algorithm performance are the most prevalent (with consistently 80–90% of all papers since 1997). However, especially over the last six years there is a steady increase in evaluation methods that include participants, either by evaluating their performances and subjective feedback or by evaluating their work practices and their improved analysis and reasoning capabilities using visual tools. Up to 2010, this trend in the IEEE Visualization conference was much more pronounced than in the IEEE Information Visualization conference which only showed an increasing percentage of evaluation through user performance and experience testing. Since 2011, however, also papers in IEEE Information Visualization show such an increase of evaluations of work practices and analysis as well as reasoning using visual tools. Further, we found that generally the studies reporting requirements analyses and domain-specific work practices are too informally reported which hinders cross-comparison and lowers external validity.

Index Terms—Evaluation, validation, systematic review, visualization, scientific visualization, information visualization

1 Motivation

In this paper, we report a systematic review of 581 papers from ten years of IEEE Visualization conference publications with respect to their use of evaluation. We provide a quantitative and objective report of the types of evaluations encountered in the literature. At the same time, we also qualitatively assess our observations from coding these 581 papers. Specifically, we put evaluation practices into historic perspective and assess and compare them in context to those of the larger visualization community. Our goal in pursuing this work is to get an understanding of the practices of evaluation in visualization research as a whole.

The importance of evaluation to the field of visualization has become well recognized—demonstrated by the growing body of work on how to conduct visualization evaluation and by the growing amount of research papers that incorporate some form of formal or informal evaluation. In this article we contribute to the body of work by providing a systematic assessment and understanding of the evaluation practices reflected by published peer-reviewed visualization papers that have not been subject to such a systematic assessment in the past.

Our work is based on Lam et al.'s [38] recent literature analysis, in which they identified seven evaluation scenarios in visualization research articles. Their paper is an important contribution but does not reflect on the entire visualization community. It focuses on what is known as the 'information visualization' sub-community and excludes all other visualization flavors. While Lam et al. primarily focused on identifying evaluation scenarios, our goal with this paper is different. We aim to complete the assessment for the larger visualization community by answering the question: What are evaluation practices in the 'scientific visualization' part of our community? What are similarities and differences between these sub-communities? To do so, we use and extend Lam et al.'s scenarios to systematically analyze the literature that appeared at the IEEE Visualization conference. We believe that our extended work is fundamental to understanding all subcultures in visualization and to properly sample all aspects of visualization work, not only those labeled as 'information visualization.'

By looking at the historic record, we were hoping to uncover some trends by examining how the field of visualization has been changing over the last 15 years. We were wondering whether some of the self-reflection by some of the field's leaders in the early 2000's has left its mark on our community and whether it led to more rigor in our evaluations. Likewise, our work is an opportunity to compare the IEEE Information Visualization and IEEE Visualization conferences to better understand their differences and commonalities. Our analysis of evaluation methods in visualization exposed a number of both weaknesses and strengths from which we, as a community, can learn for future work. Hence, we not only describe the current evaluation practices but also show what evaluation types are possible and how to improve their reporting in visualization papers. We thus expose exemplary papers and discuss a number of pitfalls that should be avoided.

In summary, the contributions of our paper are threefold. First, we objectively report the current evaluation practices in the visualization community. This is a quantitative report, focusing on the works in the IEEE Visualization conference, complementing the work done by Lam et al. [38]. Second, we give a historical overview of the use of evaluation in the visualization community as reported in the IEEE Information Visualization and IEEE Visualization conferences and put evaluation practices into perspective. This is a qualitative assessment and provides a historical perspective by comparing current and past evaluation practices. And, third, we provide information for researchers conducting evaluation by assisting them to identify, justify, and refine evaluation approaches as well as helping them to recognize and avoid pitfalls that can be learned from previous research.

2 Fundamentals and Related Work

There are two traditions of evaluation that the visualization community draws from—evaluation in the sciences (both social and natural) and evaluation in design. On the one hand, scientists try to understand the world and seek a representative model, often a mathematical model (e. g., Newton's law or Fitts' law), while designers and engineers introduce a tool and henceforth seek to alter the world in which they live and

- *Tobias Isenberg is with INRIA, France. E-mail: tobias.isenberg@inria.fr.*
- *Petra Isenberg is with INRIA, France. E-mail: petra.isenberg@inria.fr.*
- *Jian Chen is with the University of Maryland, Baltimore County, USA. E-mail: jichen@umbc.edu.*
- *Michael Sedlmair is with the University of Vienna, Austria. E-mail: michael.sedlmair@univie.ac.at.*
- *Torsten Möller is with the University of Vienna, Austria. E-mail: torsten.moeller@univie.ac.at.*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

with which they interact. Science is concerned with model validation and reproducibility. In addition, in the computational sciences, the mathematical model is turned into a computer algorithm. This invokes challenges of verifying the algorithm based on the mathematical model. For designers, the focus is what is called a 'user,' putting the emphasis on the 'human-in-the-loop.' Hence, aspects of tool functionality, usability, and aesthetics are of concern.

2.1 Validation, Verification, and Reproducibility

In computational science, *validation* refers to the process of ensuring the correctness of a conceptual or mathematical model with the salient aspects of reality [4]. In contrast, *verification* refers to the process of determining the accuracy of an algorithmic implementation with respect to the mathematical model. It is important to point out the dilemma in science that theories and models cannot be validated, only invalidated. Hence, the process of validation and verification tends to be a difficult one and of empirical nature. It is thus common to test one's algorithms and models on a number of well-chosen test cases. In the words of Karl Popper, one of the prominent philosophers of science: "So long as a theory withstands severe tests and is not superseded by another theory in the course of scientific progress, we may say that it has 'proved its mettle' " [54].

Simply computing an (absolute or relative) error measure between a known, highly accurate solution and a current algorithmic output tends to be a standard in code verification and is common practice in visualization research. However, Etiene et al. [17] recently have pointed out that asymptotic error measures can be more powerful in finding problems in an implementation.

Reproducibility of experiments is essential during the validation process. Experiments are often the basis for our conceptual models of the real world, but come with imprecision attached. Being able to quantify this error and reproduce the measurements greatly increases the confidence in the model and theories—which is important in the social and natural sciences alike. Based on this notion a movement in the computational sciences has been established known as *reproducible research* [15]. It advocates the publication of data and source code together with the paper in order to improve independent validation of the proposed models and thereby increasing the trust in them and to accelerate scientific progress. To address these issues in our community (see also, e. g., [16, 21, 31, 67]), recently the EuroRVVV workshop[1] has been created to discuss, in particular, problems of reproducibility, verification, and validation in visualization research.

2.2 Human-In-The-Loop

The Human-Computer Interaction (HCI) community has focused on understanding the human-centered-design process specific to computational tools. Hence, while the functionality (effectiveness) of a tool is of primary concern, usability (efficiency) and aesthetics (affect) play an important role as well [61, 63]. Just like for the scientific method, there is no possibility to fool-proof a tool. Hence, in both cases—science and engineering—evaluation is based on empirical methods.

While the general practices in HCI are also applicable to our field, visualization research has several unique properties that have led researchers to reflect on how to best study and evaluate visualization tools. Therefore, with BELIV[2] a dedicated workshop series has been established on this topic, similar to EuroRVVV.

Carpendale [9] provides an excellent overview of different empirical evaluation approaches and strategies as they can be applied to visualization research. In particular, she describes quantitative and qualitative evaluation methods, highlighting the advantages and challenges for both. Important differences between a quantitative, controlled approach and a more qualitative study technique that aims at measuring insight using open-ended protocols have also been argued well by North [50]. While not specifically focused on visualization, a further important paper on finding the right study approach is Greenberg and Buxton's "usability evaluation considered harmful (some of the time)" [23], in which they discuss pitfalls of over-focusing on usability studies.

Munzner's Nested Model [47] identifies four levels of visualization design—problem characterization, data and task abstraction, visual encoding/interaction design, and algorithm design—and provides guidance on valid evaluation methods for these different design levels. Tory and Möller [65] specifically discuss the role of human factors in visualization and advocate various methods applied in user-centered design processes. Based on such methods, Sedlmair et al. [59] provide a design study methodology that guides the selection of such evaluation methods in problem-driven and collaborative visualization projects.

2.3 Novel Evaluation Methods

While visualization research borrows heavily from other disciplines, several researchers have either developed new methods or discussed in detail how certain approaches need to be extended for visualization evaluation. In particular, empiric evaluation and the consideration of human factors are actively being discussed [3, 9, 10, 11, 19, 36, 46, 53, 65, 68]. Here we highlight some select methods discussed in the past.

North's insight-based method mentioned above [50] is one prominent example of a novel method for visualization evaluation. Others have reflected on the use of a critical inspection as a form of evaluating visualizations. This inspection can either be done by the authors of an article themselves or by reporting feedback from external experts. This "critical thinking about visualization" [34] has to be neutral and be backed up by facts as Kosara points out. In addition, Tory and Möller [66] argue that domain expert feedback can be a viable complement to controlled studies, both for heuristic evaluation of usability as well as for understanding the support of high-level cognitive tasks. However, not only the judgment of domain experts but also that of visual experts such as artists, graphic designers, or illustrators can be useful as has been shown in a few cases [1, 27, 29, 30]—in particular since 'critique' as a technique originated from teaching in the visual arts [35]. In visualization, it can be used in combination with techniques such as sketching and ideation when developing new visualization techniques [26]. For evaluating the impact of visualization tools on practices of real users, specific forms of case studies have been suggested [62]. Even gameplay as a form of human computing can be used to involve a wide audience in evaluating visualizations [2]. Van Wijk [69, 70] employed an economic model to assess the "value of visualization" based on effectiveness and efficiency. He explains the success or its lack with this model for a number of example visualization techniques and tools. Van Wijk's model, however, is based on the correct estimation of costs and benefits for a tool, which is often difficult to obtain in practice.

2.4 The Practice of Evaluation in Visualization Research

While the cited researchers have reflected on methodological approaches for evaluation, others have looked systematically at evaluation in visualization (e. g., [11, 52, 65]). Most influential for our work is the meta-study on evaluation goals by Lam et al. [38]. They examined 850 papers published at the IEEE Information Visualization symposium/ conference (InfoVis), in Palgrave's Journal of Information Visualization (IV), at the IEEE Symposium on Visual Analytics in Science and Technology (VAST), as well as at the Eurographics Symposium/Conference on Visualization (EuroVis)[3] in the years 1995–2010. Based on their analysis, Lam et al. identified seven scenarios that delineate empirical evaluation goals and visualization questions prevalent in the examined visualization publications. They found that, in the chosen sample of papers, the use of evaluation in visualization papers is steadily increasing but that the types of evaluation most frequently used are those that examine people's performances, users experience, and objective measures of algorithm quality and performance. However, because Lam et al.'s systematic review of the use of evaluation in visualization

[1] EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization; see http://www.eurorvvv.org/ .

[2] Beyond Time and Errors: Novel Evaluation Methods for (Information) Visualization; see http://www.beliv.org/ .

[3] Of the papers published at the EuroVis conferences, (following their reviewers' request) Lam et al. [38] excluded those papers they (Lam et al.) classified as "pure SciVis papers [. . .] based on visualization type: (e. g., pure volume, molecular, fibre-bundle, or flow visualization)."

is restricted to what they consider to be 'information' visualization work, it only provides a part of the whole picture. In our own work we use Lam et al.'s set of seven scenarios to analyze the use of evaluation in the research published at the IEEE Visualization conference and compare our results to those by Lam et al. Their rigorous methodological coding approach and the resulting descriptive scenarios give us a straight-forward way to build and extend upon their work. Furthermore, it allows us to compare different practices and trends in the visualization sub-communities of 'scientific' and 'information' visualization, and to draw conclusion based upon these findings.

3 Approach and Methodology

In order to get a systematic overview of the state of evaluation in visualization we conducted a rigorous qualitative literature review. Qualitative literature reviews are a standard technique in many areas of science to objectively report on current knowledge or practices on a topic of interest [22]. As a comprehensive overview, a literature review can help to place a topic or practices into perspective. We approached our literature review as discussed in the following sections.

3.1 Choice of Literature

To get a comprehensive overview of the use of evaluation in visualization as a whole we assessed the respective practices in the IEEE Visualization conference (now IEEE Scientific Visualization) and compared it later to the previous assessment by Lam et al. [38] of the IEEE Information Visualization conference. This approach allowed us to inspect a good cross-section of topics, approaches, and solutions common to the visualization community. Out of the past 23 years of the IEEE Visualization conference, we chose to code the past seven years (2012–2006) as well as 2003, 2000, and 1997. Coding the past seven years allowed us to reflect on current practices, while coding the earlier years allowed us to put results into historical perspective.

3.2 Choice of Codes

We based our coding scheme on the seven scenarios presented by Lam et al. [38]. Each scenario was assigned as a code. In the process of coding, we extended the initial list by one code (**QRI**). We also decided to rename Lam et al.'s **VA** (Visualization Algorithms) code into **AP** (Algorithm Performance) to more accurately reflect our findings. Based on these changes, we used the following list of codes:

UWP *Understanding Environments and Work Practices:* This code includes evaluations that derive an understanding of the work, analysis, or information processing practices by a given group of people with or without software use. Common examples are evaluations with experts to understand their data analysis needs and requirements for developing a visualization.

VDAR *Visual Data Analysis and Reasoning:* This code includes evaluations that assess how a visualization tool supports analysis and reasoning about data and helps to derive relevant knowledge in a given domain. Example evaluations include those that study experts using a tool on their data and analyzing how they can solve domain-specific questions with a new tool.

CTV *Evaluating Communication Through Visualization:* This code includes evaluations that assess the communicative value of a visualization or visual representation in regards to goals such as teaching/learning, idea presentation, or casual use. For example, a study that assesses how well a visualization can communicate medical information to a patient would fall into this category.

CDA *Evaluating Collaborative Data Analysis:* Evaluations in this group try to understand to what extent a visualization tool supports collaborative data analysis by groups of people.

While the previous scenarios focused on the process of data analysis, the remainder focuses on understanding visualizations or visualization systems and algorithms:

UP *User Performance:* Evaluations in this category objectively measure how specific features affect the performance of people with a system. Controlled experiments using time and error are typical example methods in this category.

UE *User Experience:* This code includes evaluations that elicit subjective feedback and opinions on a visualization (tool). Interviews and Likert-scale questionnaires are common methods to do so.

AP *Algorithm Performance:* Evaluations in this category quantitatively study the performance or quality of visualization algorithms. The most common examples include measurements of rendering speed or memory performance. This scenario was originally called VA (Visualization Algorithms) in Lam et al.'s [38] paper.

QRI *Qualitative Result Inspection:* Evaluations in this category are evaluations through qualitative discussions and assessments of visualization results. In contrast to UE, they do not involve actual end users or participants but instead ask the viewer of a resulting image to make an assessment for themselves. The following section gives details on why we chose to add this code.

3.3 Coding Method

Five coders (all co-authors of this paper) participated in the assessment of the literature. To calibrate, we started coding with Lam et al.'s [38] seven unique scenarios. We randomly picked the year 2009 to calibrate our codes. Each of the 54 papers in this year was assigned to a varying set of three coders. After the first coding pass the inter-coder reliability had reached 0.743 (Krippendorff's alpha [37, Ch. 12]). For each paper with a conflicted code, the coders discussed reasons for discrepancies and resolved them. After the first coding pass, all coders also met to discuss whether the initial code set needed to be extended. It was at this point that we decided to introduce the QRI code. We noted that a large number of papers in 2009 used QRI as an evaluation or proof for the quality of their results. A discussion among the authors of this paper ensued as to the validity of the code as an actual type of 'evaluation' and we will further reflect on the issue in Sect. 4. Yet, given that the coding of 2009 revealed an apparent prevalence of papers with QRI and because of past discussions of the approach in the literature [34, 47, 69], we included this code to quantitatively assess its actual prevalence and to be able to qualitatively discuss QRI practices.

After the first conflicts had all been resolved, the remaining papers were assigned to one coder each. In the process of coding the remaining papers all coders remained in close contact and communicated when choices were made as to which evaluations to include. For example, one point of discussion pertained to the amount of rigor required for an evaluation to be included. While some articles, for example, included multi-dataset comparisons of rendering speeds, some papers just reported them for a single example or in a rough manner ("less than 1 second"). We decided not to exclude evaluations based on rigor but chose to exclude cases that just reported anecdotal evidence. Papers that were unclear were marked and re-coded by a second coder.

4 Results and Discussion

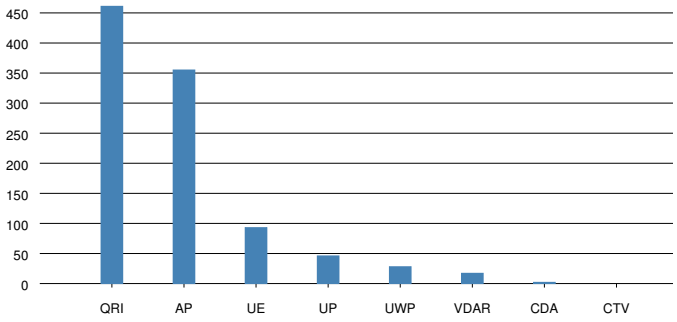
Collaboratively coding this broad set of papers led to many intensive discussions among the authors regarding current evaluation practices in our field, the meaning of rigorous and convincing evaluation with respect to what we observed in the papers, as well as the range of different evaluation approaches that are covered within the categories. In this section, we summarize both quantitative and qualitative observations from our literature analysis and our discussions about them.

In total we coded 581 papers from the IEEE Visualization conference as discussed in Sect. 3.[4] Out of these, 569 (97%) included at least one type of evaluation from our code set and 441 (76%) at least one out of the original seven scenarios [38]. In total, we coded 1002 scenarios spread across the 569 papers, meaning that many papers included several evaluations with differing goals. Fig. 1(a) shows a histogram of the total number of papers coded per scenario while Fig. 1(b) gives a historic overview of the spread of evaluation scenarios coded, in percent of all papers in a given year. Next, we discuss more detailed findings.

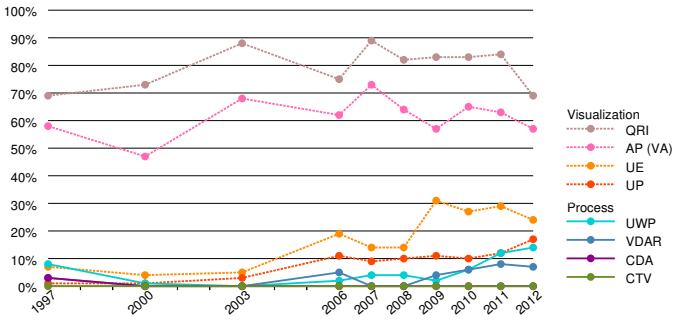
4.1 Evaluation Scenarios

The original seven scenarios were grouped by Lam et al. into two main categories: understanding *visualizations* (UP, UE, AP) and understanding data analysis *processes* (UWP, VDAR, CDA, CTV). Our new code

[4] The data can be found in a Google Spreadsheet at http://goo.gl/CGswy .



(a) Total number of papers per scenario (years: 1997, 2000, 2003, 2006–2012).



(b) Scenarios in percent of papers per year coded.

Fig. 1. Evaluation scenarios for IEEE Visualization conference papers.

QRI falls into the *visualization* category, leading to 955/1002 (95%) of all evaluation scenarios we found present in the coded papers to be of this category. We found only 47 scenarios (4.7%) that studied *processes* of data analysis. We found no instance of a study that assessed communicative value of a visualization (CTV).

The most common *visualization* scenario was QRI (46% of all scenarios) followed by AP (35% of all scenarios). Both scenarios together covered 81% of all the scenarios we coded. Evaluations in both QRI and AP were always conducted (by definition) without actual study participants. Interestingly, these two codes differ sharply from the other two in this category that involve studying participants: UE (9.3% of all scenarios) and UP (4.6% of all scenarios).

The following subsections discuss our most interesting findings and observations in regards to specific scenarios in more detail.

4.1.1 AP: The Importance of FPS and Memory Footprint

The reporting of performance of a (novel) algorithm, technique, or tool was, with 35% of all coded scenarios, the second most frequent type of evaluation we observed. Typically, authors reported computation times for processing or rendering speeds in frames per second, for a number of example datasets and a given platform. In the earlier papers, we also frequently observed the reporting of memory footprints, albeit less often in more recent papers—probably since memory shortage seems to be less of a concern these days. Such performance metrics are instructive because they inform the reader about the dataset sizes an implementation or technique is applicable for, on a given platform.

Other objective metrics to evaluate a technique or implementation are those that quantitatively assess the quality of a visualization algorithm. Their goal is, therefore, to measure what a user can see or observe without the need to study participants. In graph drawing, for instance, quality measures such as the number of edge crossings are used as criteria to assess the readability. For the visualization literature we analyzed, this subset of AP evaluations included, for example, the reporting of compression rates for shape compression techniques or error metrics for the generated visuals. Such objective metrics to assess a technique or its produced results was typically provided when some kind of ground truth or other established quality standards (visual and otherwise) existed against which results could be compared.

While frequently used as an evaluation approach, we saw a wide range of reporting rigor in AP scenarios. In fact, it was when coding this evaluation category that a discussion ensued among the authors at which point the report of AP results should be called an 'evaluation.' In particular, for time and memory performance we saw papers that simply reported a single frames-per-second number for a single, specifically selected dataset. Some papers just reported a range of rendering speeds without further specification of the dataset used or on which platform they were produced. A popular but notoriously imprecise assessment for rendering speeds was the term "interactive framerates" which can mean anything from 1 fps to 120 fps or more. We decided not to code such 'performance evaluations' as AP if they were limited to a single measurement without a clear platform or dataset. In contrast, typical evaluations gave rendering speeds for a number of different example datasets and reported the used platform. Good evaluations analyzed the behavior for a range of dataset sizes or used a number of additional metrics to assess different concepts of visual quality of the results. A good performance analysis was presented, for instance, by Lindstrom and Isenburg [39]; a nice example of objectively analyzing the quality of a proposed visualization result is Schultz and Seidel's work [57].

4.1.2 QRI: Qualitative Result Inspection

As discussed previously in Sect. 3, we added one category to the seven categories defined by Lam et al.: the prevalent 'qualitative results inspection' category (QRI). We included this code despite the fact that it is not an evaluation in the traditional sense. Put simply, a QRI addresses the reader of a paper and encourages him/her to agree on a quality statement by inspecting a result image. An example statement could be that "the figure shows that our tool can clearly depict structure x in the data which was impossible with previous approaches." While this is just one example, we encountered a variety of different approaches and factors of what we considered as QRI. While we decided to not further split this category during the coding, we next discuss these details for a better understanding of the breadth and rigor we saw in this scenario. In essence, we found three important types:

1. *Image Quality:* The classical form of QRI we found was the qualitative discussion of images produced by a (rendering) algorithm. A new algorithm was often targeted at producing images of a certain quality and it was common to show and assess visually that quality goals had been met (e. g., [48]).

2. *Visual Encoding:* Introducing a new visual encoding (e. g., a novel transfer function or novel glyphs for vector and tensor fields) was also quite common. An example was the introduction of superquadric glyphs for second-order tensors by Schultz and Kindlmann [56]. A QRI would in this case highlight what these new encodings could show and how.

3. *Walkthrough:* We intentionally, however, did not limit the scope of this category to visual encoding or image quality discussions, as we also found instances of qualitative discussions of system behavior (e. g., [72]) and interaction concepts (e. g., [8]). These discussions convincingly validated the proposed contributions.

We found two major approaches in how QRI were conducted: *comparative* and *isolated*. Comparative result inspections had clear state-of-the-art competitors that provided different solutions for the same problem. The goal was to improve upon these current state-of-the-art solutions. A typical approach was to compute output images with different algorithms, including the newly proposed one, for a range of different datasets. These images were then compared side-by-side and the authors walked the reader through them to explain the differences and benefits of the newly proposed algorithm (e. g., [48]).

Another approach was to qualitatively inspect the results in isolation; i. e., there was no clear competitor that addressed the same problem which could be used for comparison. In those cases, a solid description of the problem at hand as well as the justification of how the proposed new algorithm/technique/system addressed it was mandatory. Not doing so resulted in a *pure description* much like a manual that failed the purpose of evaluation—we did not code these descriptions.

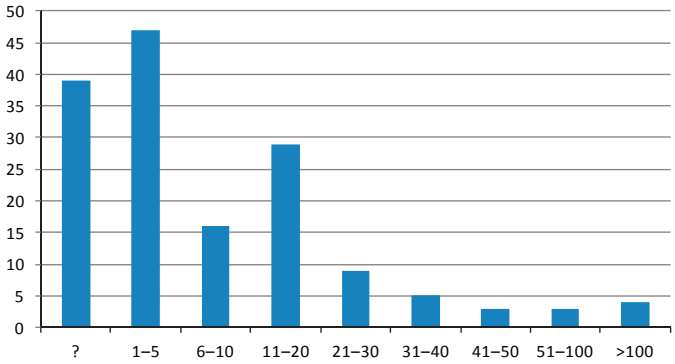


Fig. 2. Histogram of participants in evaluations, over all papers and years (only counting papers with at least one participant). For the papers with a known and positive participant number (i. e., excluding those marked with '?'), the mean is 23.8 and the median is 9 participants.

4.1.3 UE/UP: User Experience and Performance

User Experience and User Performance studies are probably the most common studies in the field of human-computer interaction. As such we were surprised to see that only 14% of all scenarios fell into one of the two categories and only 113 of the total 581 papers contained either one or both of the studies. Typical examples of UP studies assessed the time and/or errors of participants using a new technique (e. g., [40]) or compared the performance of human participants with that of an automatic technique (e. g., [73]). As the only two types of scenarios in the *visualization* category that involve participants, it is interesting to see that their prevalence was relatively low compared to the other two scenarios. Perhaps one of the reasons is that controlled experiments, the most common UP method, are typically time- and resource-intensive to design, conduct, and analyze. This could also explain why UE studies were much more common than UP studies. For UE we mostly observed reports of feedback gathered at informal demos to expert users but we also saw a few more detailed UE assessments (e. g., [42]).

Fig. 2 shows the number of participants we recorded for all coded scenarios. As can be seen in the figure, many studies involved 1–5 participants. Among these, studies with one or two participants were the most frequent (13 and 14, respectively). These were usually UE studies, reporting feedback from experts. The larger spike at 11–20 participants is explained by the fact that a large number of UP studies chose 12 participants (11 cases). Studies with very large numbers of participants were most commonly web-based UP or UE studies.

4.1.4 UWP/VDAR, or the Case of the Case Study

The scenarios UWP (Understanding Work Practices) and VDAR (Visual Data Analysis and Reasoning) are methodologically very similar as they both attempt to understand domain experts' analysis processes and practices. However, while UWP focuses on understanding the current practices and is therefore similar to a requirements analysis, VDAR scenarios (usually) focus on assessing the value of a newly introduced visualization tool for a group of domain experts.

Overall, we saw only very few UWP (28 papers, 4.8%) and VDAR (17 papers, 2.9%) scenarios. These numbers appear very low given that large parts of our community focus on application-driven research. We indeed found quite a few papers in our survey whose authors had obviously interacted with domain experts. In terms of UWP, however, they almost exclusively derived their motivations and requirements in an analytical fashion rather than from talking to real people or figuring out real needs in practice. Such papers are typically motivated by 'established facts' such as "in medicine, doctors need to ..." or "physicists need tools for ..." or simply refer to a previous state of the art in the literature to improve the performance of some algorithm. While these are also valid motivations, of course, they are not UWP studies. In terms of VDAR, we found a very similar situation. While many applications were built towards helping to solve a certain domain problem, we rarely found evaluation reports that tried to assess the resulting data analysis, decision making, or knowledge management/discovery when domain experts used the proposed tools. Instead, QRI studies were provided. Whether the presented findings were based on domain experts or the authors often remained ambiguous.

The methodological approach prevalent for VDAR and common for UWP studies are case studies. While we did not find many VDAR or UWP scenarios, we found many papers that validated their results using 'case studies.' In visualization, case studies are commonly defined as "detailed reporting about a small number of individuals working on their own problems, in their normal environment" [62]. In doing so, they can be a particularly strong form of evaluation as they show how a new visualization would fit into existing work practices (UWP) or how it can help to conduct VDAR. Analyzing our set of visualization papers, however, we found a variety of interpretations of what a case study is and a large variance in the rigor of reporting them. We categorize these different interpretations of case studies into four categories:

1. *Case study from domain expert:* Reports on how a domain expert used a new visualization approach to analyze his/her data. The new visualization improves upon previous practices of this expert. In-depth reports of such case studies have, for instance, been conducted on knowledge discovery in high-dimensional data [60]. In the papers we coded this could be found in several of the case study papers of the earlier years when domain experts themselves reported on the use of a tool or technique.

2. *Case study from close collaboration:* Specific domain problems have been tackled by an ongoing, intensive collaboration between visualization researchers and domain experts. The analysis of a problem reported in a case study has been the result of this collaborative, often iterative endeavor. In particular, such forms of case studies are common in participatory design [6], or when following a design study methodology [59]. Domain expert collaborators often co-author a resulting paper. An example for such a design study is Kok et al.'s work [33].

3. *Case studies from visualization researchers:* Here, visualization researchers report on how they used a new visualization approach to solve/improve upon a certain problem without a (strong) involvement of domain experts. These are particularly relevant forms of case studies when the 'epistemic distance' between the problem at hand and the visualization researcher is small; i. e., the visualization researcher either is an expert in the problem domain addressed, or the problem is easily accessible and understandable without the necessity to have in-depth domain knowledge. Albeit published at the InfoVis conference, an example of the latter can be found in the BallotMaps design study [74].

4. *Usage scenario from visualization researchers:* We found many papers in our survey, that reported on 'case studies' yet which did not fall into one of the previous three categories. Here, authors reported simply on how a new visualization approach could be used by a hypothetical domain expert, as opposed to reporting on new domain-specific findings on the problem at hand or on how it improved upon domain practices. Instead of case studies we call these 'usage scenarios,' echoing a previous call to distinguish usage scenarios from the more formal case study method [59].

We often found it hard to understand in which of these bins a case study validation fell. Many of the papers we analyzed simply did not provide enough detail on what was done and how. A typical example is "this work is based on a collaboration" without any further details. Questions on how many collaborators, who they are, what their previous practices were, and how they participated in a project/case study often remained unclear. The number of participants was inconclusive not only in the "case study section" but also in many descriptions of UE scenarios. Fig. 2 shows that for 39 papers we could not deduce from the text either how many participants were involved or if the participants were even real or imagined (left-most bar in Fig. 2).

Understanding these details, however, is particularly important as the four categories described above vary significantly in terms of how

strong and convincing the evaluation is. While we consider case studies with domain experts (categories 1 and 2) the strongest, followed by case studies on problems that are known to visualization researchers (cat. 3), usage scenarios (cat. 4) are the weakest and least convincing. Usage scenarios are still a legitimate form of evaluation, however, the different weightings are important when the strength of an evaluation has to be judged. Doing a real case study but not reporting on the details leaves the reader questioning the approach and eventually decreases the perceived strength of an evaluation. Note that for case studies, as opposed to lab studies or surveys, it is not necessary to have a large number of domain expert participants. Three to five is very convincing, however, often even 1–2 is sufficient if interesting conclusions can be derived and if the study is comprehensive and detailed.

4.1.5 CTV/CDA: Communication and Collaboration

In our analysis, we found no example for the scenarios 'Communication through Visualization' (CTV), and only two for 'Collaborative Data Analysis', both from the year 1997. While visualization for collaborative analysis has just recently gained more attention [25], it has not been a very strong topic in the conference—perhaps explaining the lack of evaluations. The lack of CTV, however, is somewhat surprising as the types of questions covered by a CTV study (how does a tool or visualization aid in learning, teaching, presentation, casual use?) are of potential importance to a number of research results and at the very least those that are more application oriented. What could be an explanation for this finding? We are not sure, but perhaps these types of studies are reported elsewhere or by researchers other than the ones who developed and designed the underlying visualization capabilities.

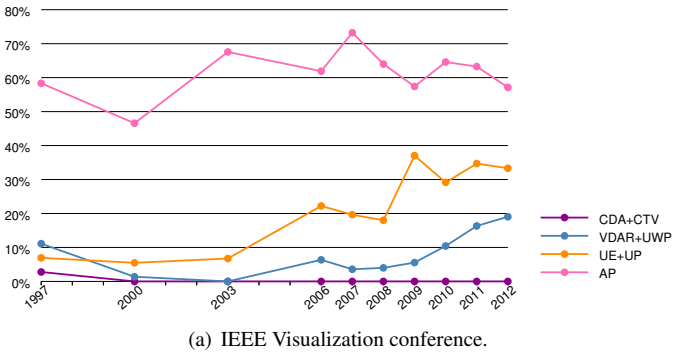
4.2 Historical Development

Examining Fig. 1(b), we find that the use of QRI has been rather consistent over 15 years, hovering around 80% of all papers. This is really not that surprising for a visualization conference. We tend to communicate our results visually, even if only as a qualitative backup to some quantitative evaluation measures. To get a better idea of the trends in the data, in Fig. 3 we leave out QRI. We also combine UE+UP (understanding user feedback and performance), UWP+VDAR (understanding user needs and reasoning), as well as CDA+CTV (collaboration).

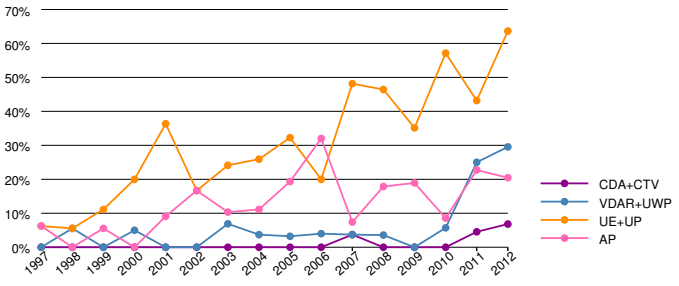
Focusing on Fig. 3(a) (results for the IEEE Visualization conference), we also notice that the use of AP has hovered around 60% of all papers and, unquestionably, has played the dominant role over the years and no visible change in trend can be detected. However, since 2003, there is a clear trend for the increased importance of engaging the user. We believe this is a clear success of a number of soul-searching events in the early 2000s in our community. Lorensen's "death of visualization" [41] clearly left its mark together with a number of panels and keynotes about the future of visualization. Most notably among them was Johnson's "top scientific visualization research problems" [28]. While it is often remembered as the starting point of uncertainty visualization, it also considers human-centered design as well as better visual abstractions as some of the challenges. In fact, what had topped the list was a category named "think about the science" that calls for a closer collaboration with domain experts.

A further trend is the steady increase of the reporting of VDAR+UWP over the last four years. It was 2009 that the five paper types (technique, system, design study, evaluation, model papers) were introduced at the IEEE Visualization conference. While these paper types were pioneered at the IEEE Information Visualization conference in 2003 and in 2007 Kwan-Liu Ma organized the IEEE Visualization panel "Meet the Scientists," it was only in 2009 that they were formalized at the IEEE Visualization conference. We believe that this formalism helped some of the authors to recognize that it is important to more rigorously report their work with users. Some of this effect is also noticeable for the UE+UP evaluation scenarios.

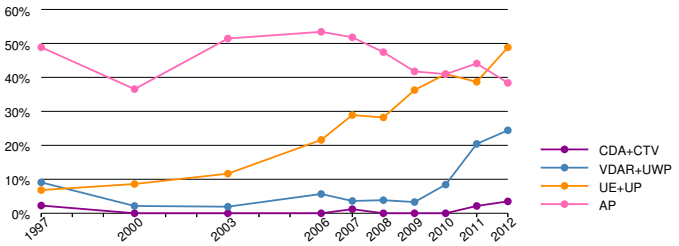
One main caveat exists in reading these numbers—as we chose not to subjectively judge the rigor of an evaluation, many of the cases included in the above numbers do not count as 'formal' evaluations.



(a) IEEE Visualization conference.



(b) IEEE Information Visualization conference.



(c) Both conferences combined.

Fig. 3. Evaluation scenarios in percent of all papers in a year.

4.3 Practices in the Community as a Whole

Comparing our analysis of the IEEE Visualization conference with the results of Lam et al.'s [38] for the IEEE Information Visualization conference,[5] a quite different picture emerges. Here we constrain our discussion to the coding results of the IEEE Information Visualization conference, depicted in Fig. 3(b). The dominating evaluation scenario is UE+UP—as one would expect from a strong HCI influence. A steady increase of these scenarios is visible for the past 10–15 years. The historically second most important evaluation scenario is AP. Its influence has been somewhat steady over the last ten years. In this regard it is worthwhile to mention the outlier in 2006; a year with an atypically high number of graph drawing papers. Here, the focus was more on algorithms than human-centered design which could explain the reversal of AP/UP+UE scores. The influence of VDAR+UWP has been similar to the IEEE Visualization conference—there has been little such work until only recently. Yet, the increasing trend that we saw for this scenario at the IEEE Visualization conference can only be found at IEEE Information Visualization since 2011.

To understand the difference between the IEEE Visualization conference and the IEEE Information Visualization conference even better, we also look at the percentage of evaluations per year that included human participants (see Fig. 4). It is clear from this figure that a steady 80% of all evaluations in the IEEE Information Visualization conference incorporate participants (with the just mentioned exception of 2006). For the IEEE Visualization conference, this percentage has been steadily rising for the past ten years and arrived at about 50%

[5] We ourselves coded the years 2011 and 2012 of the IEEE Information Visualization conference (an additional 88 papers) to understand the full history.

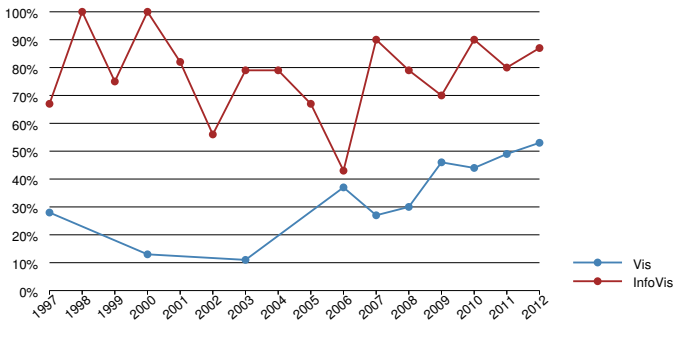


Fig. 4. Evaluations with human participants in % of all evaluations, considering only Lam et al.'s 7 scenarios as the baseline, excluding QRI.

in 2012. This observation suggests a much stronger HCI influence in IEEE Information Visualization than for IEEE Visualization.

Finally, it is worth to examine the visualization community as a whole. Fig. 3(c) shows the results as if there was just a single conference, simply combining all papers in each year. The clear trend, in our community, is to increasingly understand the performance and experience of users (UE+UP) and the diminishing (yet until 2011 still dominating) focus on algorithmic performance (AP). Overall, this is not too surprising and had been our gut feeling all along. This is also influenced by an increasing number of published papers at the IEEE Information Visualization conference and a diminishing number of papers being published at the IEEE Visualization conference.

5 Considerations for Evaluation in Visualization

Next, we discuss some of our observations in the study and own past experiences. Some of these consideration are similar to previous prescriptive advice on evaluation in visualization [9, 47, 50, 58, 62, 65]; here, we provide further evidence for these arguments by grounding them in our study, extend upon them, and add further considerations.

5.1 Analyzing and Reporting Real Problems

Our systematic review demonstrated that the process evaluation scenarios UWP, VDAR, CTV, and CDA are rarely conducted or, at least, are rarely reported in IEEE Visualization papers. This observation is similar to what Lam et al. [38] found for the 'information visualization' papers they had surveyed. These process-based evaluation scenarios, however, are relevant for virtually any visualization research. In our field we need these types of evaluations to understand what the problems are that our target audience faces; to understand how visualization tools support their analysis and reasoning; to understand how they can communicate about their insights using visualizations; and to understand how visualizations can support collaborative data analysis. Without investigating these questions we risk working in an Ivory Tower and, ultimately, facing the "death of visualization" [41]. As Lorensen [41] emphasized in 2004, as a community we must talk more to our "customers" to understand their work practices and visualization uses. Our papers should reflect this work by an increased visibility of insights from UWP, VDAR, CTV, and CDA evaluation scenarios.

While we do not advocate that every visualization paper needs to include one of these evaluation scenarios to be a solid paper, it would be beneficial if each paper draws a clear connection to what open scientific problem it contributes or which real-existing tool it improves. Beyond that, Fig. 3(a) shows that the percentage of UWP+VDAR evaluation papers have steadily been increasing since Lorensen's call to action. Moreover, our subjective impression of papers published in more recent years is that many authors, in fact, do talk to their clients and target audiences. The insights from such investigations, however, are often stated matter-of-factly as opposed to explaining how certainty about these insights was achieved and to invite further investigation. To improve this situation we need to encourage the reporting of UWP, VDAR, CTV, and CDA evaluation scenarios and need to make them first class citizens in the visualization literature.

5.2 Statistical Significance & Qualitative Expert Feedback

Evaluation methods used to study UWP, VDAR, CTV, and CDA are often qualitative in nature, such as interviews with domain experts, observations of work practices, or longitudinal case studies of newly proposed tools. The goal behind these evaluations is to maximize the *realism* of the findings, as opposed to other methods that seek to maximize *generalizability* or *precision* [9, 43]. If we, as a community, want to engage in understanding real problems of real people and how to solve them, we also have to understand methodological characteristics that come with qualitative approaches designed for studying the real world. Examples of such characteristics include the small number of participants to derive valid conclusions, and the goal of a rich and grounded qualitative understanding of complex systems rather than statistical significance of quantitative measures.

In our own previous work as researchers and reviewers, however, we have found that often reviewers with a strong positivist background are very quick in putting qualitative approaches down as non-rigorous research without being aware of the methodological differences and characteristics. Based on these preconceptions, papers get rejected with the justification that the authors should have employed a quantitative approach with a statistical analysis. As noted by Greenberg and Buxton [23], such practices can be very harmful, muting creative ideas and discouraging qualitative endeavors of studying real users with real problems. A notable exception from our community is, for instance, Lundström et al.'s [42] qualitative work on understanding multi-touch visualization based on in-depth interviews with five domain experts.

Rather than blindly assuming and expecting certain evaluation methods, we argue for a more thorough consideration of research questions/goals and the right evaluation approach for this question/goal, echoing previous calls [23, 47, 50]. In particular, we argue against considering evaluation ***only*** as controlled quantitative studies with null hypothesis significance tests (NHST). While these studies can be helpful for research questions focusing on low-level effects, statistical significance cannot be the only goal for a controlled study. Essentially, significance can almost always be achieved eventually if one uses a high-enough number of participants, making it imperative to *interpret* performance differences in the study context. Similarly, non-rigorous yet common practices of allowing flexibility in data collection, analysis, and reporting make it easy to falsely find evidence for effects that actually do not exist [64]. Moreover, the usefulness of NHST is heavily being questioned in the statistics literature [12, 32] and alternatives exist for reporting the results of controlled studies [13].

Also, controlled studies are simply not the right tool for a number of evaluation goals. In visualization we often deal with ill-defined, fuzzy, and broad domain problems which cannot easily be broken down into low-level tasks [59]; consider, e. g., the goal of comparing different genomes [45]. We then have to engage in better understanding of such complex problems to design complex visualization systems/tools/algorithms helping to solve them. In such situations, however, the usage of controlled quantitative studies is often questionable. First, there are rarely low-level tasks that can be tested. Second, even if there are low-level tasks and a controlled experiment with domain experts can be run, the findings might not be *actionable* [20]. For instance, the finding that a newly proposed visualization tool makes low-level tasks of domain experts faster might be interesting from an application domain point of view. Such findings, however, can hardly be broken down into a cause-effect relation and questions such as what has caused the performance increase remain unclear and speculative [58].

5.3 "... and they really liked it."

One prevalent pitfall we observed that relates to the issue of rigor was to consider positive subjective judgments from domain experts as a sufficient form of evaluation. This pitfall most often occurred in UE evaluations reporting on anecdotal feedback from interviews with potential target users. While subjective positive feedback from experts on a new technique or tool is encouraging and can be a valid tool of evaluation, simply stating "... and they really liked it" is not sufficient. Such statements are usually prone to *demand characteristic effects*. Well-known in psychology [24, 51] and in human-computer interaction

[7, 14], this phenomenon explains experimental artifacts that arise from the unconscious wish of participants to align with the researcher's hypotheses. When interviewing potential target users, especially if they are involved in a collaboration [59], it is thus very likely that they will offer positive and favorable feedback. While positive feedback is a helpful condition for tool adoption in a target domain, it is by no means a rigorous and convincing stand-alone evaluation. Instead, qualitative evaluations could be carried out, e. g., in the form of in-depth observations of a new tool in use by domain experts; or interviews with a (semi-)structured protocol could be used with a set of pre-defined questions that ask, for instance, also for a critique of the new tool.

5.4 On the Question of How Many Study Participants

Perhaps one of the central questions is the one on how many participants should be involved before a study is representative or a design is well tested. The answer depends on the goals of the work. If we are testing the performance of participants in low-level tasks and want to reduce the probability of committing a Type II error a power analysis can help to determine the right number of participants [5]. If we are testing the usability of a tool, there is an ongoing debate of how many participants to test. Nielsen and Molich [49] found that with six people one can find, on average, about 80% of all usability problems. This result has since been revised and criticized and the debate is still looming [55].

On the other hand, if we are working with specific domain experts, we often do not want to make the claim of generality, but rather that of transferability to people with the same needs. A qualitative evaluation or observational study can be very useful in this case even with a 'low' number of expert participants (e. g., [42]). It is common in qualitative work to sample participants using nonprobability sampling [44]. Purposeful sampling (or criterion-based selection) methods are those that are often used when one wants to collect the most insightful data from carefully chosen experts. Sample sizes (number of participants) are in these cases often determined based on emerging results where minimizing information redundancy is the primary criterion: One basically engages in a cycle of data collection and analysis and stops studying participants once the collected information becomes redundant [44].

Extending our discussion in Sect. 5.2, we have noticed that colleagues with a mathematical background often have difficulties accepting studies that are based on only few participants. Hence it is worthwhile to point out that algorithmic studies are often based on very few data sets (often less than four) with the goal to generalize to all data sets. This argument is no different from qualitatively analyzing a new technique or tool with, e. g., four domain experts—if the evaluation is done rigorously. Both cases are not unreasonable, however, since we should keep in mind that there is no number of data sets or participants that could validate any specific scientific theory. Instead we aim for testing a tool or an algorithm as thoroughly as we can [54].

5.5 Evaluation Reporting Rigor

In our literature analysis, we encountered many publications that lacked important methodological details in reporting their evaluations. Without these details, however, it is difficult, if not impossible, to assess the quality and impact of an evaluation. Based on these insights about current practices, we call for more rigorous descriptions of evaluation activities and summarize some specific recommendations below:

Reporting who participated: Especially in qualitative endeavors such as QRI, UWP, and VDAR as well as, at times, UE it is important to provide details on who did what. Most importantly, it should be clear whether evaluation data (or insight on common practices or typical problems in a domain) is based on one or more real experts or on a hypothetical expert imitated by the authors. If no details are provided, readers will tend to assume the latter which might reduce the impact of strong field work. Consider, for instance, a task analysis for a certain domain problem. This task analysis might be based on work with real domain experts, on a literature review, or simply on the intuitions of the researchers. If work with domain experts has been conducted it is not only important to state that fact but also to provide further specific details about these experts and how the engagement between researcher and experts took place. For instance, what is their training, what is their work context, and what constitutes their expertise?

Reporting on collaboration details: Many projects in visualization are based on a collaboration between domain experts with driving problems/data and researchers with visualization/data analysis expertise. Such collaborative projects bear a great potential of having a real impact in an application domain by including different angles of expertise in the design process. Yet, for a reader of a paper that reports on the results of such design studies it is still crucial to learn about methodological specifics to be able to judge its potential impact. Statements such as "this work is based on a collaboration" or "collaborators are co-authors of this paper" are very important but are by far not enough information to methodologically judge a project. Guidelines (and pitfalls) on how to report collaborative visualization projects can be found in Sedlmair et al.'s work on design study methodology [59].

Reporting on study protocols: Especially for AP, UP, and UE studies it is important to follow established reporting protocols to facilitate reproducibility and comparability. For controlled laboratory studies with participants there are various sources that give advice on how to report them (e. g., [18]). For AP studies it is considered good practice to test them on both synthetic and real data. In this context we also noticed that, while the Vis/InfoVis/VAST contests have been helpful, well-curated benchmark datasets for many problems are still lacking.

Reporting how many people participated: As part of the reporting of study protocols, we also found several examples that specifically lacked the number of participants/experts involved in an evaluation (Fig. 2, left-most bar). Phrases such as "some" or "several" are too vague to be able to make concrete judgments about the work that has been done. Reporting numbers of participants is particularly important for UP, UE, UWP, VDAR, CTV, and CDA.

Reporting controlled experiments with rigor: Many evaluations that we coded as UP would clearly fail the set standards for properly reporting the results of quantitative controlled evaluations in, for example, the HCI community. Many papers only report averages of performance measurements and then reason on observed differences, without the statistical tests that would signify that the found differences are not due to chance. While we caution people to not overly rely on null hypothesis significance tests as mentioned in Sect. 5.2, without proper inferential statistical analysis the implications become even more meaningless. The APA (American Psychological Association) provides general guidelines for reporting studies and statistical test results [71].

Reporting qualitative result inspections with more rigor: An important question is what problems should be addressed in a QRI study. An aspect of this question is which datasets should be used and whether they are cherry-picked to underline the proposed solution's strength or if they are a representative sample of the targeted problem space. While there are no set standards, it is important that an algorithm is tested using several *benchmark* datasets and with different measures and therefore 'proves its mettle' in Popper's [54] words. While there are certainly a number of commonly used datasets available online, we emphasize that there is a need for well-designed and thought-through datasets that exhibit a number of relevant characteristics.

The popular approach of using QRI as an evaluation method also raises the more general question of how much rigor and formalism is needed for such an evaluation approach to be valid in general. Certainly, it is not enough to assume that some example pictures will speak for themselves without any further discussion as we observed in some cases (which we did not code as QRI). It is typically also better to go beyond pure manual-like descriptions and to clearly discuss how a new approach improves the state of the art. The commonly employed technique of a walk-through may be good as a QRI evaluation but there are better techniques in form of real case studies (see Sect. 4.1.4).

Paper size restrictions: We are aware that adding more methodological details to a paper often might conflict with required page limits. However, even one paragraph about these details is better than nothing. Providing and referencing supplemental material is a valuable approach if there are many details about an evaluation.

6 Conclusion

Our systematic review of evaluation practices in the IEEE Visualization conference has shown an overall emphasis on evaluations of algorithmic performance (AP) and qualitative result inspections (QRI) through images. However, there is an increasing trend in the evaluation of user experience (UE) and user performance (UP). Further, the steady increase of reporting environment and work practices (UWP) as well as how new visualizations help in data analysis and reasoning (VDAR) is particularly encouraging, especially after the soul-searching that happened about ten years ago in our community. Also, it became clear that a major difference between the use of evaluation in the IEEE Information Visualization conference and the IEEE Visualization conference is the emphasis of overall user performance vs. algorithmic performance, the latter being much more common in IEEE Visualization.

A general conclusion of our work is that, while it has improved over the last years, the general level of rigor of reporting evaluations is still too low. Many authors did not report details on who collaborating domain experts are and how they worked with them. Being clear about an employed methodology, however, can greatly improve the impact of ones research results. We, thus, believe that there is great room for improvement—even by just including 1–2 paragraphs about the protocol that was followed when interacting with the domain experts. The considerations provided in Sect. 5 can serve as a starting point.

There are several avenues to continue this work. Of course, the mentioned question of how to rigorously and formally evaluate visualization work using qualitative results inspection needs to be discussed further. It will also certainly be interesting to extend this analysis to papers published at other visualization venues such as the Eurographics Conference on Visualization (EuroVis) as a place that does not make a dedicated difference between 'scientific' and 'information' visualization, to see whether similar trends exist. Also, a separate analysis of the IEEE Conference on Visual Analytics Science and Technology would be of value to see whether there is a stronger emphasis on UWP and VDAR processes as well as on collaborative visualization (CTV/CDA), as suggested by its name and agenda. We did not include this data in our analysis as VAST is still a relatively young field (it started in 2006), paper counts are still relatively small, and its practices are not yet as established as for the two fields we compared. Still, in a few years comparing VAST to our data will be able to given an even more complete picture of evaluation practices in the visualization community.

In summary, by coding 581 papers from over ten years of IEEE Visualization we provided a systematic overview of the evaluation practices in this visualization community, discussed observations and quantitative results, and gave a historical perspective and comparison to IEEE Information Visualization. We hope that our work encourages the community to keep up the increasing trend of reporting evaluations and, in particular, also qualitative evaluations with experts.

Acknowledgments

We thank, in particular, Lam et al. [38] for being able to use their data and compare it to our findings. We also thank Hans-Christian Hege for initial discussions on the topic of evaluation in visualization and Pierre Dragicevic for his valuable comments. This work was supported in part by NSF IIS-1302755, ABI-1260795 and EPS-0903234.

References

[1] D. Acevedo, C. D. Jackson, F. Drury, and D. H. Laidlaw. Using visual design experts in critique-based evaluation of 2D vector visualization methods. *IEEE Transactions on Visualization and Computer Graphics*, 14(4):877–884, July/Aug. 2008. doi> 10.1109/TVCG.2008.29

[2] N. Ahmed, Z. Zheng, and K. Mueller. Human computation in visualization: Using purpose driven games for robust evaluation of visualization algorithms. *IEEE Transactions on Visualization and Computer Graphics (Proc. SciVis)*, 18(12):2104–2113, Dec. 2012. doi> 10.1109/TVCG.2012.234

[3] K. Andrews. Evaluating information visualisations. In *Proc. BELIV*, pp. 1:1–1:5. ACM, New York, 2006. doi> 10.1145/1168149.1168151

[4] I. Babuska and J. T. Oden. Verification and validation in computational engineering and science: Basic concepts. *Computer Methods in Applied Mechanics and Engineering*, 193(36–38):4057–4066, 2004. doi> 10.1016/j.cma.2004.03.002

[5] R. Bausell Barker and Y.-F. Li. *Power Analysis for Experimental Research*. Cambridge University Press, 2002.

[6] K. Bødker, F. Kensing, and J. Simonsen. *Participatory IT Design: Designing for Business and Workplace Realities*. The MIT Press, Cambridge, MA, USA, 2004.

[7] B. Brown, S. Reeves, and S. Sherwood. Into the wild: Challenges and opportunities for field trial methods. In *Proc. CHI*, pp. 1657–1666. ACM, New York, 2011. doi> 10.1145/1978942.1979185

[8] S. Bruckner, V. Šoltészová, M. E. Gröller, J. Hladůvka, K. Buhler, J. Y. Yu, and B. J. Dickson. BrainGazer – Visual queries for neurobiology research. *IEEE Transactions on Visualization and Computer Graphics (Proc. Vis)*, 15(6):1497–1504, Nov./Dec. 2009. doi> 10.1109/TVCG.2009.121

[9] S. Carpendale. Evaluating information visualizations. In *Information Visualization: Human-Centered Issues and Perspectives*, volume 4950 of *LNCS*, pp. 19–45. Springer, Berlin, 2008. doi> 10.1007/978-3-540-70956-5_2

[10] C. Chen and M. P. Czerwinski. Empirical evaluation of information visualizations: An introduction. *International Journal of Human-Computer Studies*, 53(5):631–635, Nov. 2000. doi> 10.1006/ijhc.2000.0421

[11] C. Chen and Y. Yu. Empirical studies of information visualization: A meta-analysis. *International Journal of Human-Computer Studies*, 53(5):851–866, Nov. 2000. doi> 10.1006/ijhc.2000.0422

[12] J. Cohen. The earth is round ($p < .05$). *American Psychologist*, 49(12):997–1003, Dec. 1994. doi> 10.1037/0003-066X.49.12.997

[13] G. Cumming. *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge, New York, 2012.

[14] N. Dell, V. Vaidyanathan, I. Medhi, E. Cutrell, and W. Thies. "Yours is better!": Participant response bias in HCI. In *Proc. CHI*, pp. 1321–1330. ACM, New York, 2012. doi> 10.1145/2208516.2208589

[15] D. L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385–388, July 2010. doi> 10.1093/biostatistics/kxq028

[16] T. Etiene, L. G. Nonato, C. Scheidegger, J. Tierny, T. J. Peters, V. Pascucci, R. M. Kirby, and C. T. Silva. Topology verification for isosurface extraction. *IEEE Transactions on Visualization and Computer Graphics*, 18(6):952–965, June 2012. doi> 10.1109/TVCG.2011.109

[17] T. Etiene, C. Scheidegger, L. G. Nonato, R. M. Kirby, and C. T. Silva. Verifiable visualization for isosurface extraction. *IEEE Transactions on Visualization and Computer Graphics (Proc. Vis)*, 15(6):1227–1234, Nov./Dec. 2009. doi> 10.1109/TVCG.2009.194

[18] A. Field and G. Hole. *How to Design and Report Experiments*. Sage Publications, Ltd., London, 2003.

[19] C. Forsell. A guide to scientific evaluation in information visualization. In *Proc. Information Visualization*, pp. 162–169. IEEE Computer Society, Los Alamitos, 2010. doi> 10.1109/IV.2010.33

[20] M. Gleicher. Why ask why? Considering motivation in visualization evaluation. In *Proc. BELIV*, pp. 10:1–10:3. ACM, New York, 2012. doi> 10.1145/2442576.2442586

[21] A. Globus and S. Uselton. Evaluation of visualization software. Technical Report NAS-95-005, NASA Advanced Supercomputing Div., Feb. 1995.

[22] B. N. Green, C. D. Johnson, and A. Adams. Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine*, 5(3):101–117, Fall 2006. doi> 10.1016/S0899-3467(07)60142-6

[23] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *Proc. CHI*, pp. 111–120. ACM, New York, 2008. doi> 10.1145/1357054.1357074

[24] M. J. Intons-Peterson. Imagery paradigms: How vulnerable are they to experimenters' expectations? *Journal of Experimental Psychology: Human Perception and Performance*, 9(3):394–412, June 1983. doi> 10.1037/0096-1523.9.3.394

[25] P. Isenberg, N. Elmqvist, J. Scholtz, D. Cernea, K.-L. Ma, and H. Hagen. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization*, 10(4):310–326, Oct. 2011. doi> 10.1177/1473871611412817

[26] B. Jackson, D. Coffey, L. Thorson, D. Schroeder, A. M. Ellingson, D. J. Nuckley, and D. F. Keefe. Toward mixed method evaluations of scientific visualizations and design process as an evaluation tool. In *Proc. BELIV*, pp. 4:1–4:6. ACM, New York, 2012. doi> 10.1145/2442576.2442580

[27] C. D. Jackson, D. Acevedo, D. H. Laidlaw, F. Drury, E. Vote, and D. Keefe. Designer-critiqued comparison of 2D vector visualization methods: A pilot study. In *ACM SIGGRAPH Sketches & Applications*. ACM, New York, 2003. doi> 10.1145/965400.965505

[28] C. Johnson. Top scientific visualization research problems. *IEEE Computer Graphics and Applications*, 24(4):13–17, July/Aug. 2004. doi> 10.1109/MCG.2004.20

[29] D. F. Keefe, D. Acevedo, J. Miles, F. Drury, S. M. Swartz, and D. H. Laidlaw. Scientific sketching for collaborative VR visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 14(4):835–847, July 2008. doi> 10.1109/TVCG.2008.31

[30] D. F. Keefe, D. B. Karelitz, E. L. Vote, and D. H. Laidlaw. Artistic collaboration in designing VR visualizations. *IEEE Computer Graphics and Applications*, 25(2):18–23, Mar./Apr. 2005. doi> 10.1109/MCG.2005.34

[31] R. M. Kirby and C. T. Silva. The need for verifiable visualization. *IEEE Computer Graphics and Applications*, 28(5):78–83, Sept. 2008. doi> 10.1109/MCG.2008.103

[32] R. B. Kline. What's wrong with statistical tests—and where we go from here. In *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*, chap. 3, pp. 61–91. APA, Washington, DC, 2004. doi> 10.1037/10693-003

[33] P. Kok, M. Baiker, E. A. Hendriks, F. H. Post, J. Dijkstra, C. W. G. M. Löwik, B. P. F. Lelieveldt, and C. P. Botha. Articulated planar reformation for change visualization in small animal imaging. *IEEE Transactions on Visualization and Computer Graphics (Proc. Vis)*, 16(6):1396–1404, Nov./Dec. 2010. doi> 10.1109/TVCG.2010.134

[34] R. Kosara. Visualization criticism – the missing link between information visualization and art. In *Proc. Information Visualization*, pp. 631–636. IEEE Computer Society, Los Alamitos, 2007. doi> 10.1109/IV.2007.130

[35] R. Kosara, F. Drury, L. E. Holmquist, and D. H. Laidlaw. Visualization criticism. *IEEE Computer Graphics and Applications*, 28(3):13–15, May/June 2008. doi> 10.1109/MCG.2008.63

[36] R. Kosara, C. G. Healey, V. Interrante, D. H. Laidlaw, and C. Ware. User studies: Why, how, and when? *IEEE Computer Graphics and Applications*, 23(4):20–25, July/Aug. 2003. doi> 10.1109/MCG.2003.1210860

[37] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc., Thousand Oaks, CA, USA, 3rd ed., 2013.

[38] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, Sept. 2012. doi> 10.1109/TVCG.2011.279

[39] P. Lindstrom and M. Isenburg. Fast and efficient compression of floating-point data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1245–1250, Sept./Oct. 2006. doi> 10.1109/TVCG.2006.143

[40] M. A. Livingston, J. W. Decker, and Z. Ai. Evaluation of multivariate visualization on a multivariate task. *IEEE Transactions on Visualization and Computer Graphics (Proc. SciVis)*, 18(12):2114–2121, Dec. 2012. doi> 10.1109/TVCG.2012.223

[41] B. Lorensen. On the death of visualization. In *Position Papers of the NIH/NSF Fall 2004 Workshop on Visualization Research Challenges*, 2004.

[42] C. Lundström, T. Rydell, C. Forsell, A. Persson, and A. Ynnerman. Multi-touch table system for medical visualization: Application to orthopedic surgery planning. *IEEE Transactions on Visualization and Computer Graphics (Proc. VIS)*, 17(12):1775–1784, Dec. 2011. doi> 10.1109/TVCG.2011.224

[43] J. E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In R. M. Baecker, J. Grudin, W. Buxton, and S. Greenberg, eds., *Readings in Human-Computer Interaction: Towards the Year 2000*, pp. 152–169. Morgan Kaufmann, 2nd ed., 1995.

[44] S. B. Merriam. *Qualitative Research: A Guide to Design and Implementation*. Jossey-Bass, San Francisco, 2nd ed., 2009.

[45] M. Meyer, T. Munzner, and H. Pfister. MizBee: A multiscale synteny browser. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis)*, 15(6):897–904, Nov./Dec. 2009. doi> 10.1109/TVCG.2009.167

[46] E. Morse, M. Lewis, and K. Olsen. Evaluating visualizations: Using a taxonomic guide. *International Journal of Human-Computer Studies*, 53(5):637–662, Nov. 2000. doi> 10.1006/ijhc.2000.0412

[47] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis)*, 15(6):921–928, Nov./Dec. 2009. doi> 10.1109/TVCG.2009.111

[48] B. Nelson, R. M. Kirby, and R. Haimes. GPU-based interactive cut-surface extraction from high-order finite element fields. *IEEE Transactions on Visualization and Computer Graphics (Proc. Vis)*, 17(12):1803–1811, Dec. 2011. doi> 10.1109/TVCG.2011.206

[49] J. Nielsen and R. Molich. Heuristic evaluation of user interfaces. In *Proc. CHI*, pp. 249–256. ACM, New York, 1990. doi> 10.1145/97243.97281

[50] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, May/June 2006. doi> 10.1109/MCG.2006.70

[51] M. T. Orne. Demand characteristics and the concept of quasi-controls. In R. Rosenthal and R. L. Rosnow, eds., *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books*, chap. 5, pp. 110–137. Oxford University Press, New York, 2009.

[52] A. Perer and B. Shneiderman. Integrating statistics and visualization for exploratory power: From long-term case studies to design guidelines. *IEEE Computer Graphics and Applications*, 29(3):39–51, May/June 2009. doi> 10.1109/MCG.2009.44

[53] C. Plaisant. The challenge of information visualization evaluation. In *Proc. AVI*, pp. 109–116. ACM, New York, 2004. doi> 10.1145/989863.989880

[54] K. Popper. *The Logic of Scientific Discovery*. Routledge Classics, Taylor & Francis, London, 2010.

[55] M. Schmettow. Sample size in usability studies. *Communications of the ACM*, 55(4):64–70, Apr. 2012. doi> 10.1145/2133806.2133824

[56] T. Schultz and G. L. Kindlmann. Superquadric glyphs for symmetric second-order tensors. *IEEE Transactions on Visualization and Computer Graphics (Proc. Vis)*, 16(6):1595–1604, Nov./Dec. 2010. doi> 10.1109/TVCG.2010.199

[57] T. Schultz and H.-P. Seidel. Estimating crossing fibers: A tensor decomposition approach. *IEEE Transactions on Visualization and Computer Graphics (Proc. Vis)*, 14(6):1635–1642, Nov./Dec. 2008. doi> 10.1109/TVCG.2008.128

[58] M. Sedlmair, P. Isenberg, D. Baur, and A. Butz. Information visualization evaluation in large companies: Challenges, experiences and recommendations. *Information Visualization*, 10(3):248–266, July 2011. doi> 10.1177/1473871611413099

[59] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis)*, 18(12):2431–2440, Dec. 2012. doi> 10.1109/TVCG.2012.213

[60] J. Seo and B. Shneiderman. Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework. *IEEE Transactions on Visualization and Computer Graphics*, 12(3):311–322, May/June 2006. doi> 10.1109/TVCG.2006.50

[61] H. Sharp, Y. Rogers, and J. Preece. *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons, Chichester, UK, 2nd ed., 2007.

[62] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proc. BELIV*, pp. 8:1–8:7. ACM, New York, 2006. doi> 10.1145/1168149.1168158

[63] B. Shneiderman, C. Plaisant, M. Cohen, and S. Jacobs. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson/Prentice Hall, 5th ed., 2010.

[64] J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, Nov. 2011. doi> 10.1177/0956797611417632

[65] M. Tory and T. Möller. Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 10(1):72–84, Jan./Feb. 2004. doi> 10.1109/TVCG.2004.1260759

[66] M. Tory and T. Möller. Evaluating visualizations: Do expert reviews work? *IEEE Computer Graphics and Applications*, 25(5):8–11, Sept./Oct. 2005. doi> 10.1109/MCG.2005.102

[67] S. P. Uselton, G. Dorn, C. Farhat, M. W. Vannier, K. Esbensen, and A. Globus. Validation, verification and evaluation. In *Proc. Visualization*, pp. 414–418. IEEE Computer Society, Los Alamitos, 1994. doi> 10.1109/VISUAL.1994.346285

[68] J. van Wijk, T. Isenberg, J. B. Roerdink, A. C. Telea, and M. Westenberg. Evaluation. In D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, eds., *Mastering the Information Age: Solving Problems with Visual Analytics*, chap. 8, pp. 131–144. Eurographics Assoc., Goslar, Germany, 2010.

[69] J. J. van Wijk. The value of visualization. In *Proc. Visualization*, pp. 79–86. IEEE Computer Society, Los Alamitos, 2005. doi> 10.1109/VISUAL.2005.1532781

[70] J. J. van Wijk. Views on visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):421–432, July/Aug. 2006. doi> 10.1109/TVCG.2006.80

[71] G. R. VandenBos, ed. *Publication Manual of the American Psychological Association*. APA, Washington, DC, 6th ed., 2009.

[72] J. Waser, R. Fuchs, H. Ribičić, B. Schindler, G. Blöschl, and M. E. Gröller. World lines. *IEEE Transactions on Visualization and Computer Graphics (Proc. Vis)*, 16(6):1458–1467, Nov./Dec. 2010. doi> 10.1109/TVCG.2010.223

[73] A. Wiebel, F. Vos, D. Foerster, and H.-C. Hege. WYSIWYP: What you see is what you pick. *IEEE Transactions on Visualization and Computer Graphics (Proc. SciVis)*, 18(12):2236–2244, Dec. 2012. doi> 10.1109/TVCG.2012.292

[74] J. Wood, D. Badawood, J. Dykes, and A. Slingsby. BallotMaps: Detecting name bias in alphabetically ordered ballot papers. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis)*, 17(12):2384–2391, Dec. 2011. doi> 10.1109/TVCG.2011.174