Evaluation of Visualization Heuristics

Ryan Williams(✉), Jean Scholtz, Leslie M. Blaha, Lyndsey Franklin, and Zhuanyi Huang

Pacific Northwest National Laboratory, Richland, WA 99354, USA
{ryan.williams,jean.scholtz,leslie.blaha,lyndsey.franklin,
zhuanyi.huang}@pnnl.gov
https://www.pnnl.gov/

Abstract. Multiple sets of heuristic have been developed and studied in the Human Computer Interaction (HCI) domain as a method for fast, lightweight evaluations for usability problems. However, none of the heuristics have been adopted by the information visualization or the visual analytics communities. Our literature review looked at heuristic sets developed by Nielsen and Molich [7] and Forsell and Johansson [1] to understand how these heuristics were developed and their intended applications. We also reviewed heuristic studies conducted by Hearst and colleagues [2] and Väätäjä and colleagues [10] to determine how individuals apply heuristics to evaluating visualization systems. While each study noted potential issues with the heuristic descriptions and the evaluator's familiarity with the heuristics, no direct connections were made. Our research looks to understand how individuals with domain expertise in information visualization and visual analytics could use heuristics to discover usability problems and evaluate visualizations. By empirically evaluating visualization heuristics, we can identify the key ways that these heuristics can be used to inform the visual analytics design process. Further, they may help to identify usability problems that are and are not task specific. We hope to use this process to also identify missing heuristics that may apply to designs for different analytic purposes.

Keywords: Heuristics · Visualization · Heuristic evaluation
User study · Visual analytics · Information visualization

1 Introduction

Heuristics have been used since the early 1990's in the Human Computer Interaction (HCI) domain, for evaluating user interfaces. Heuristics, as described by Nielsen, are more rules of thumb rather then specific usability guides [6]. Heuristic evaluations are an approach used to discover usability problems, most commonly in software user interfaces [4]. Heuristics establish a common language around a prescribed definition to focus the evaluation of a user interface. These ten heuristics developed by Nielsen [6] have been used by the HCI community.

M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 208–224, 2018.
https://doi.org/10.1007/978-3-319-91238-7_18

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

Heuristic evaluations are an efficient method of discovering usability problems. They can be applied early on in the design process of a software user interface to discover potential usability issues before code is written. Forsell and Johansson [1] describe heuristic evaluations as a discount method that is easy to learn and apply during all phases of development. Alternatives to heuristic evaluations are eliciting qualitative opinions about the usefulness of a particular software user interface. This method is less useful because finding a consensuses on usability problems with unstructured opinions can be difficult.

At this point, neither the information visualization (InfoVis) community nor the visual analytics community has adopted a common set of heuristics to use early on in the design process. However, Forsell and Johansson [1] used the same methodology as [8] and developed a set of 10 heuristics for information visualization. Forsell and Johansson designed and executed a study to determine which heuristics of a set of 63 published heuristics could best explain a collection of 74 usability problems. Their study asked 6 participants with expertise in InfoVis and/or HCI to rate how well each heuristic explained a usability problem. The resulting set of 10 heuristics (Table 1) had the best explanatory coverage out of all possible heuristic combinations [1]. Explanatory coverage, as described by Nielsen [5], is a rating of a heuristic that explains a major part of a usability problem. The resulting set of ten heuristics from Forsell and Johansson cover the broadest set of their 74 usability problems.

While these heuristics have not been commonly used for evaluations in the InfoVis community, several studies have been conducted using these heuristics in an attempt to understand how they could be applied. Hearst and colleagues [2] found that using the heuristics for visualizations along with asking questions about the data resulted in complementary results as the questions about the data could compensate for heuristics that were difficult to understand and apply.

Foresell and Johannson noted that future work needed to be done to investigate any issues in applying their heuristics to find and explain usability problems [1]. Väätäjä and colleagues [10] found that heuristics related to interaction, veracity, and aesthetics needed to be added to the Forsell and Johansson set based on participant feedback captured during their study. Participants noted limitations in the heuristics to account for visualization libraries that are not flexible that could potentially leave out important information. The study also noted that if issues beyond basic usability issues are of interest, training domain

experts, with a good understanding of the data and information system being analyzed, to carry out the heuristics evaluation would likely provide insightful feedback [10].

Our study wants to determine more specifically what was difficult about using these heuristics and how domain experience in creating and evaluating visualizations influences the applied use of heuristics for discovering usability problems in visualizations. We conducted a controlled experiment to determine what factors might impact an evaluators' capability to conduct a heuristic evaluation of a static information visualization. In particular, we wanted to understand how their familiarity with the visualizations and their experience level in conducting heuristic evaluations impacted their evaluations of visualizations using the Forsell and Johannnson heuristic set.

2 Study Overview

The goal of our research in heuristic evaluations for visualizations is to determine how a heuristic set can be used in the InfoVis community much in the same way Nielsen's heuristics have been used in the HCI community. Heuristics provide a fundamental benefit of discovering and communicating usability problems early in the development process. With a better understanding of which heuristics are best suited for visualizations and how heuristics can be applied to visualization evaluations, the InfoVis community can move closer to adopting a common set of heuristics to use to evaluate visualizations.

We conducted a controlled experiment to determine what factors might impact an evaluator's capability to conduct a heuristic evaluation of a static information visualization. We asked that each participant have experience in user experience design or research, user interface development and/or HCI research. In particular, we wanted to understand how their familiarity with the visualizations and the participant's experience level in conducting heuristic evaluations impacted their evaluations of visualizations.

We investigated the following three hypotheses about the usefulness of heuristics for finding usability problems:

Hypothesis 1. Participants would find the heuristics that involved interactive features less useful when given only static visualizations.

Hypothesis 2. Participants would find visualizations that they were less familiar with more difficult to evaluate using heuristics.

Hypothesis 3. Participants who were less familiar with heuristic evaluation would have a more difficult time using the heuristics overall.

2.1 Methods

Participants. Ten domain experts in the visualization field participated in the study. Each self-reported experience with one of more of the following: user experience design or research, user interface development, and HCI research.

The age range of the participants are: 4 ages 24–29, 2 ages 30–25, 3 ages 36–41, and 1 age 48–53. Because of the reliance on color visualizations in this study, all participants were screened for normal vision. All participants demonstrated normal (2 participants) or corrected-to-normal (6 glasses, 2 contacts) vision by achieving acuity of at least 20/20 using the Snellen eye chart [9]. Participants successfully completed the Ishihara color plates [3] with 100% accuracy indicating normal color vision. All volunteers consented to participate in accordance with the policies of the Institutional Review Board at the Pacific Northwest National Laboratory. Nine participants completed the evaluation for all visualizations; one participant evaluated only four of the five visualizations due to time constraints.

Visualization Heuristics. Participants were provided with a fixed set of heuristics with which they evaluated five common visualizations. We used the set of ten heuristics proposed by Forsell and Johansson [1], listed in Table 1.

Visualizations. Five common visualizations were selected for the study: Scatter Plot, Sunburst, Tree Map, Parallel Set, and Area Graph. To keep the visualized data a constant factor, all visualizations depicted data from the the VAST 2008 Mini Challenge Two: *Migrant Boats*[1]. The 2008 VAST Mini Challenge 2 comprises records dealing with the mass movement of persons departing a fictitious island to the United States during a two-year period. The resulting visualizations are illustrated in Fig. 1.

Usability Problems Ground Truth. Visualizations were used herein with default settings and layout options. Although we did not "plant" usability issues in the visualization, we did not adjust them to ensure that any usability problems were omitted. Two of the authors, JS and LF, who were quite familiar with conducting heuristic evaluations, performed a "ground truth" evaluation of the different visualizations. In this process, at least one usability problem was identified for each heuristic that the experts deemed relevant to the visualization. We do not claim that all of the usability problems were found in this set, but we felt that it was a good start towards predicting the usability problems that would be identified by our participants. The "ground truth" set of usability problems are given here for each visualization.

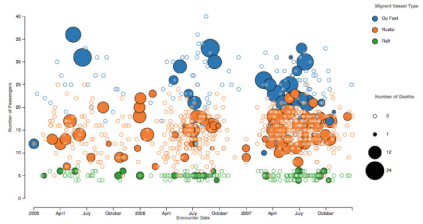
Scatter Plot

H1 Occlusion: Difficult to see individual points
H2 Date axis compressed and difficult to read
H5 Difficult to compare dot size from different months

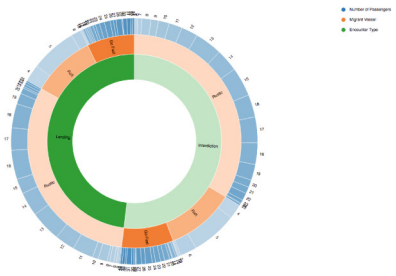
[1] See https://www.cs.umd.edu/hcil/VASTchallenge08/ for full Mini-Challenge details and data.

Table 1. Forsell and Johansson (2010) heuristics leveraged in our present study

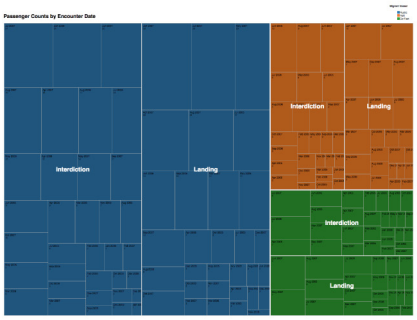
Number	Name	Description
H1	Spatial organization	Is related to the overall layout of a visual representation, which comprises analyzing how easy is to locate an information element on the display and to be aware of the overall distribution of information elements in the representation
H2	Information coding	Concerns the mapping of data elements to visual elements, as well as use of additional symbols or realistic characteristics that can be used either for building alternative representations (like groups of elements in clustered representations) or to aid in the perception of information elements
H3	Orientation and Help	This refers to functions that provide support for the user to control level of details, redo/undo of user actions and representation of additional information (for example, the path a user follows while navigating in a complex structure)
H4	Data set reduction	Features such as filtering allows reduction of information shown at a certain moment, leading more rapidly to adjustment of the focus of interest, and clustering allows the representation of a subset of data elements by means of special symbols, while pruning simply cuts off irrelevant information for the understanding a visual representation
H5	Recognition rather than recall	Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate
H6	Remove the extraneous	Is a principle that pushes the graphic designer to present the largest amount of data with the least amount of ink. Extra ink can be a distraction and take the eye away from seeing the data or making comparisons
H7	Prompting	Refers to the means available in order to guide the users towards making specific actions whether these are data entry or other tasks. This criterion also refers to all the means that help users to know the alternatives when several actions are possible depending on the contexts. Prompting also concerns status information, that is information about the actual state or context of the system, as well as information concerning help facilities and their accessibility
H8	Minimal actions	These refer to workload with respect to the number of actions necessary to accomplish a goal or a task. It is here a matter of limiting as much as possible the steps users must go through
H9	Consistency	Refers to the way interface design choices (codes, naming, formats, procedures, etc.) are maintained in similar contexts, and are different when applied to different contexts
H10	Flexibility	Is reflected in the number of possible ways of achieving a given goal. It refers to the means available to customization in order to take into account working strategies, habits, and task requirements



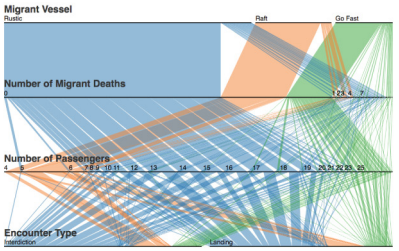
(V1) Scatter Plot. Analytical question: Which vessel type by the highest death per passenger ratio?



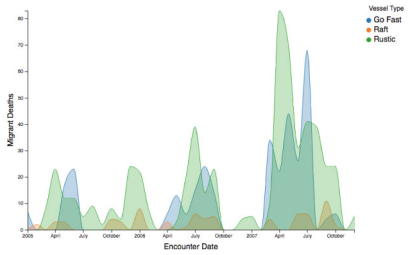
(V2) Sunburst. Analytical question: Which vessel type had the most cumulative passengers?



(V3) Tree Map. Analytical question: Which vessel type had the most cumulative passengers?



(V4) Parallel Sets. Analytical question: Which vessel type by the highest death per passenger ratio?



(V5) Area Plot. Analytical question: Which vessel type had the highest cumulative deaths?

Fig. 1. The five visualizations used in the present study. All depict the VAST 2008 Mini Challenge Two: *Migrant Boats* data.

H5 Difficult to verify that 24 is the largest number of deaths? (There seem to be sizes that do not correspond to the 0, 1, 12, 24 in the legend)
H2, H8 Scales do not help answer the question directly (which vessel type has the highest death per passenger ratio), requires mental math

Sunburst

H1 Vessel types are not contiguous which makes it difficult to count passengers (add from two parts in the display, both landing + interdiction)
H1 Numbers occlude, are difficult to read in Go Fast vessel type
H9 Blue is inconsistent within a vessel type and the variation does not mean anything (is it that the number of passengers $\sim$ opacity?)

Tree Map

H1 Labels occlude each other a little bit
H3 Passenger count not labeled
H2 Shape of vessel of same passenger count is not consistent (comparing squares and rectangles)
H8 Have to add to get total passengers

Parallel Sets

H3] Number of deaths impossible to read (only labels 0, 1, 2, 3, 4, and 7, have to guess at the others)
H3 Number of passengers impossible to read (there are not enough labels at far right end)
H1 Connecting Death to Passengers requires following a thin line
H8 Still need mental math to answer ratio question
H1 Line crossings difficult to follow

Area Plot

H8, H6 Requires mental math to get cumulative deaths (sum area under curve)
H6 Labels are not precise enough for exact math
H2 Continuous lines and area makes it difficult to know how to do the addition (i.e., flat portion of raft for April 2005, how many 4 s do we add?)
H2 Viz better suited for relative comparison than precise reading

Procedure. Following informed consent and vision screening, participants were asked to complete a demographic survey in which they self-rated the following: (1) their experience using each visualization type; (2) their experience conducting heuristic evaluations for visualization; (3) the frequency at which they use each visualization type to extract information, and (4) their experience developing software to create each visualization type. Ratings were reported on a 6-point

Likert scale with a rating of 1 being very infrequently/inexperienced and 6 being very frequently/experienced.

The experience level of participants conducting heuristic evaluations were: 2 participants reported experienced, 4 participants reported somewhat experienced, and 4 participants reported that they were inexperienced as recorded in Table 2.

Table 2. Participant visualization experience levels

Visualization	Unfamiliar	Novice	Moderate	Expert	Advanced
Tree map	-	2	4	-	4
Parallel sets	3	3	3	-	1
Gantt chart	1	3	3	1	2
Sankey diagram	4	3	1	-	2
Donut chart	2	3	2	-	3
Stack bar	-	1	4	2	3
Heat map	-	1	3	1	5
Scatter plot	-	1	3	1	5
Sunburst	3	3	2	-	2
Time series	2	-	2	1	5

This assessment was not used to down select the from the 10 visualizations to the 5 used in our study but merely to understand how participants rated their experience using the visualizations listed. We inadvertently did not collect this information about the area graph though we felt that participants were most likely reasonably familiar with this one as well.

Participants were given a suggested analytical question related to each visualization to keep in mind as they applied the 10 heuristics. The analytical question for each visualization is given in the relevant figure captions in Fig. 1.

Participants were shown the visualizations in a random order. For each visualization, participants were asked to apply each heuristic, in the order listed in Table 1. For each heuristic, within a given visualization, participants were asked the following:

- Using this heuristic, do you see the potential for a usability problem with the visualization? [yes/no]
- Please describe the usability problem that you found. [free response]
- How well does this heuristic capture the usability problem that you identified? [6-point Likert scale]
- How relevant is this heuristic for evaluating this visualization? [6-point Likert scale]
- How confident do you feel in your ability to apply this heuristic to evaluate this visualization? [6-point Likert scale]

Following the visualization evaluations, participants provided qualitative feedback on the usability issues of each visualization. They rated the clarity of each heuristic description on a 6-point Likert scale, and were asked to provide (free response) feedback on which parts of the each description were easy to understand and use and which were difficult to understand and use.

3 Results

3.1 Usefulness of Heuristics

Hypothesis 1 posited that participants would find the heuristics that involved interactive features less useful when given only static visualizations. All the visualizations in this study were static with no interactions possible in the interfaces provided. The heuristics that refer to interactions are H3, H7, H8, H10; the heuristics that do not refer to interactions are H1, H2, H4, H5, H6, and H9. We operationalize useful in this context as relevance and confidence. Thus, we predict that, across all visualizations, for the set of heuristics referring to interactions, relevance and confidence ratings will be lower than for the heuristics that do not refer to interactions.

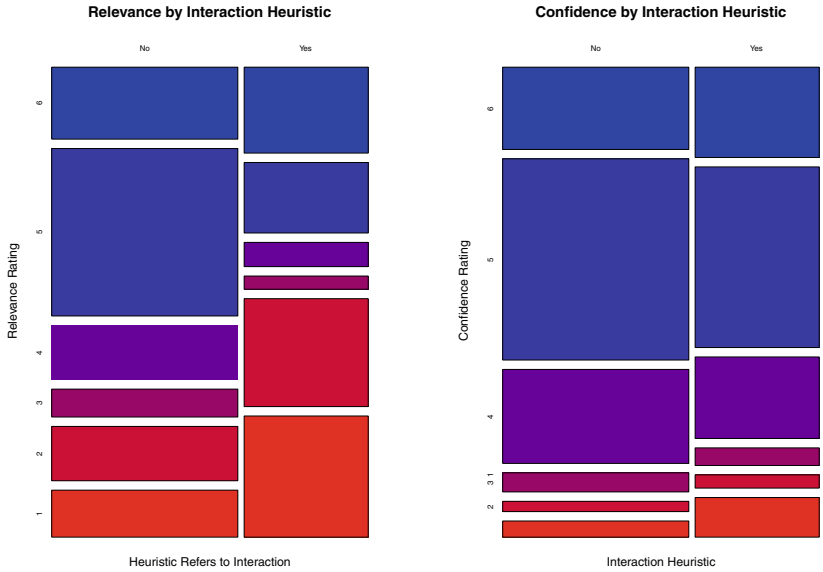


Fig. 2. Relevance (left) and confidence (right) ratings for the heuristics referring to interaction (Yes column) versus the those for heuristics no referring to interaction. These plots depict the data associated with evaluation hypothesis 1.

We compared relevance ratings between the two groups with a linear mixed ordinal regression model with a logit link function, using the fixed factor interaction heuristic (2 levels: yes, no) and the random factor observer. The Type I error rate was $\alpha = .05$. The relevance ratings for heuristics referring to interactions were significantly lower than the relevance ratings for heuristics not referring to interactions (Parameter $\beta_{\text{Yes}} = -1.10$, $z = -5.95$, $p < .001$). The relevance ratings are shown in Fig. 2.

Confidence ratings were compared with a linear mixed ordinal regression model with a logit link function, using the fixed factor interaction heuristic (2 levels: yes, no) and the random factor observer. The Type I error rate was $\alpha = .05$. The confidence ratings for heuristics referring to interactions were significantly lower than the confidence ratings for heuristics not referring to interactions (Parameter $\beta_{\text{Yes}} = -0.2$, $z = -123.4$, $p < .001$). The confidence ratings are shown in Fig. 2.

Hypothesis 2 posited that participants would find visualizations with which they were less familiar more difficult to evaluate using heuristics. According to their self-reported ratings, participants were less familiar with the Sunburst (V2) and Parallel Set (V4) visualizations. They were more familiar with the Scatter Plot (V1), Tree Map (V3), and Area Plot (V5) visualizations as recorded in Table 2. We operationalized difficult to be confidence ratings in using the heuristics. The prediction is that participants will report lower confidence across all heuristics between more and less familiar visualizations.

Confidence ratings were compared with a linear mixed ordinal regression model with a logit link function, using the fixed factor familiarity with visualizations (2 levels: low, high) and the random factor observer. The Type I error rate was $\alpha = .05$. There was no difference in confidence ratings between familiarity groups (Parameter $\beta_{\text{Low}} = -0.08$, $z = -0.51$, $p = .61$).

Hypothesis 3 posited that participants who were less familiar with heuristic evaluation would have a more difficult time using the heuristics overall. We split participants into two groups, with low familiarity with heuristic evaluation defined as those participants who self-reported their experience with heuristic evaluations as level 1–3, and the high familiarity participants defined as those who self-reported their experience as level 4–6. We operationalized difficult to mean confidence in using the heuristics. We predicted that more difficulty would be lower confidence in the participants reporting less experience with heuristic evaluations (Fig. 3).

Confidence ratings were compared with a linear mixed ordinal regression model with a logit link function, using the fixed factor familiarity with heuristic evaluations (2 levels: low, high) and the random factor observer. The Type I error rate was $\alpha = .05$. There confidence ratings were significantly lower for participants with less experience with heuristic evaluations (Parameter $\beta_{\text{Low}} = -1.3973$, $z = -7.09$, $p < .001$).

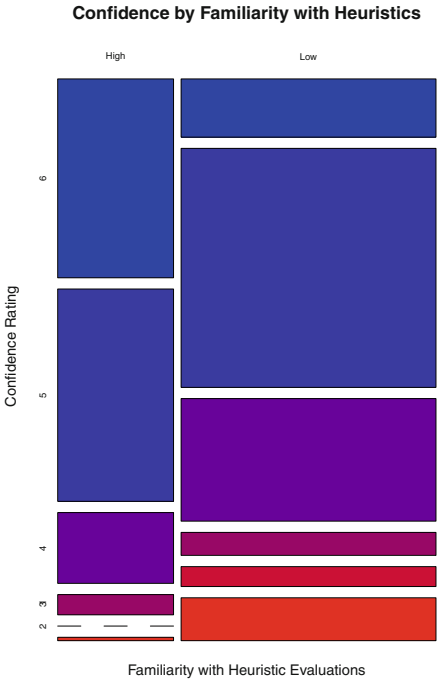


Fig. 3. Confidence ratings for the participants with low familiarity with heuristic evaluation (Low column) versus the those with high familiarity with heuristic evaluation (High column). Plot depicts data associated with evaluation hypothesis 3.

3.2 Usability Problems Reported by Participants

Table 3 shows the ground truth usability problems that were reported by our participants, regardless of the heuristic participants attributed them to. In a number of instances, the description of the problem provided by a participant was vague, and at times they described several things that would fit under different heuristics. At times the usability problem they found did not fit with the heuristic to which they attributed it. Several participants reported "failed" if the heuristic required an interaction to determine if there was a usability issue. Therefore, we created confusion matrices for each of the five visualizations to tease out the usability problems that were actually attributable to the specific heuristics. Tables 4, 5, 6, 7 and 8 contain the confusion matrices for each of the five visualizations. The first column notes the expert ground truth, with a star denoting if the experts found a usability problem related to that particular heuristic. The subsequent columns show whether each participant found a usability problem using the heuristics listed there.

Table 3. Ground truth usability issues described by participants

Visualization	Issue	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Scatter plot	Occlusion (H1)	X	X	-	-	X	X	X	X	X	-
	Compressed axis (H2)	-	X	-	-	-	X	X	-	-	-
	Dot compare (H5)	X	-	-	-	X	-	X	-	-	-
	Dot size (H5)	-	X	-	-	X	X	X	X	-	-
	Indirect scales (H8, H2)	-	-	-	-	X	-	-	-	-	-
Sunburst	Non-contiguous vessel types (H1)	X	-	X	X	X	X	-	-	-	X
	Occlusion (H1)	-	X	-	-	-	X	X	X	-	X
	Inconsistent color (H9)	-	-	-	X	X	X	X	X	-	X
Tree map	Occlusion (H1)	X	-	X	X	-	-	-	-	-	X
	Missing labels (H3)	-	-	-	-	-	-	-	X	-	-
	Mental addition (H8)	-	X	-	X	X	-	X	X	X	X
	Inconsistent shapes (H9)	-	-	-	X	-	X	X	X	-	-
Parallel	Labels unreadable (H3)	-	-	X	X	X	X	X	X	X	-
	Line crossings (H1)	X	X	X	X	X	X	-	X	X	-
	Line following (H1)	X	X	X	X	X	-	-	X	-	-
	Missing labels (H3)	-	-	X	-	X	X	X	-	X	-
	Mental math (H8)	-	X	-	X	X	-	-	-	-	-
Area	Continuous lines vs addition (H2)	-	-	X	X	X	-	X	-	-	X
	Vis-task mismatch (H2)	-	-	X	X	X	-	-	-	-	X
	Labels not precise (H6)	-	X	-	-	-	-	X	-	-	X
	Imprecise labels (H6)	-	X	-	-	-	-	X	-	-	X
	Mental math (H8)	-	-	-	-	X	-	-	-	-	-

In Table 4, we can see that 4 participants found a usability problem that was comparable to the one found in our expert evaluation for heuristic 1. Only one participant found usability problems attributed to heuristic 2. No participants found usability problems attributable to heuristics 3 and 4. Six participants found 1 or 2 usability problems attributable to heuristic 5. One participant found a usability problem using heuristic 8. There were a total of 7 usability problems described that did not fit in any of the heuristics but were reported by participants (total of the row labeled "other"). If you compare the results in Tables 3 and 4 you see that in Table 3, in the Scatter plot section, participants 5, 7, 8 and 9 reported a usability problem attributable to heuristic 1. However, the description they provided did not match with that heuristic. In the confusion matrix, such a description would be attributed to another heuristic or "other". Looking at the various confusion matrices, it seems that more participants were able to find usability problems that were explained by heuristics 1 and 2.

We also have a number of problems found by participants classified as "other". These include participant observations about the lack of interaction or controls for the visualization ("There are no user controls to filter, reduce, or

Table 4. Confusion matrix for the Scatter plot visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1*	5	1	-	1	-	2	-	-	-	-
EH2*	-	2	-	-	-	-	-	-	-	-
EH3	-	-	-	-	-	-	-	-	-	-
EH4	-	-	-	-	-	-	-	-	-	-
EH5*	1	2	2	-	1	1	-	-	1	-
EH6	-	-	-	-	-	-	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8*	-	1	-	-	-	-	-	-	-	-
EH9	-	-	-	-	-	-	-	-	-	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	-	-	1	1	-	1	1	-	2	1
Failed	-	-	1	-	-	-	1	1	-	1
No response	4	5	6	8	9	6	8	9	7	8

Table 5. Confusion matrix for the Sunburst visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1*	6	2	-	-	2	3	-	-	-	-
EH2	1	-	-	-	-	-	-	-	-	-
EH3	2	-	1	-	1	-	-	-	-	-
EH4	-	-	-	-	-	-	-	-	-	-
EH5	-	-	-	-	-	-	-	-	-	-
EH6	-	-	-	-	-	-	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8	-	-	-	-	-	-	-	-	-	-
EH9*	-	3	-	-	-	1	-	-	5	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	-	-	3	2	-	1	2	2	-	2
Failed	-	-	-	1	-	-	1	1	-	1
No response	3	5	6	7	7	5	7	7	5	7

re-order the data."). Vague color critiques ("Lots of color and text.") that did not describe a specific problem also fell into this category. It may be that additional heuristics are needed or if current heuristic descriptions need to be improved. Color in particular may be deserving of its own heuristic, given the number of time participants described it across all visualization and heuristics.

Table 6. Confusion matrix for the Tree map visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1*	1	1	-	-	-	1	-	-	-	-
EH2	-	-	-	-	-	-	-	-	-	-
EH3*	-	-	2	-	2	4	-	-	1	-
EH4	-	-	-	1	-	-	-	-	-	-
EH5	1	2	2	-	1	1	-	-	1	-
EH6	-	-	-	-	-	-	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8*	6	1	-	1	1	-	-	1	-	-
EH9*	3	1	-	-	-	-	-	-	2	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	-	3	2	3	2	1	3	1	2	-
Failed	-	-	-	-	-	-	1	1	-	-
No response	1	4	6	5	5	4	4	7	5	7

Table 7. Confusion matrix for the parallel coordinates visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1*	5	1	1	1	1	3	-	-	-	-
EH2	-	-	-	-	-	-	-	-	-	-
EH3*	1	1	2	-	-	-	-	-	1	-
EH4	-	-	-	-	-	-	-	-	-	-
EH5	-	-	-	-	-	-	-	-	-	-
EH6	-	-	-	-	-	1	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8*	-	1	-	-	-	-	-	-	-	-
EH9	-	-	-	-	-	-	-	-	-	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	1	3	2	2	1	2	1	-	-	2
Failed	-	-	-	-	1	-	1	1	-	-
No response	3	4	5	7	7	4	8	9	9	8

The experts who built the ground truth set of usability problems also helped to code participant responses. In a number of cases, participants described a usability problem using what the experts felt was the appropriate heuristic. However, the problem was not a usability problem that was called out in the ground truth set of usability problems. This suggests that there may be some subjectivity with regards to what is considered a "problem" and what is not, and

Table 8. Confusion matrix for the area visualization

Expert heuristic	PH1	PH2	PH3	PH4	PH5	PH6	PH7	PH8	PH9	PH10
EH1	-	-	-	-	-	-	-	-	-	-
EH2*	3	1	-	1	1	2	1	1	2	-
EH3	1	-	-	-	-	-	-	-	1	-
EH4	-	-	-	-	-	-	-	-	-	-
EH5	-	-	-	-	-	-	-	-	-	-
EH6*	1	-	-	-	-	1	-	-	-	-
EH7	-	-	-	-	-	-	-	-	-	-
EH8*	-	-	-	-	-	-	-	-	-	-
EH9	-	-	-	-	-	-	-	-	1	-
EH10	-	-	-	-	-	-	-	-	-	-
Other	-	4	4	1	2	1	1	-	-	1
Failed	-	-	-	-	-	-	1	1	-	1
No response	5	4	6	8	7	6	7	8	6	8

just how much of a problem something must be before a visualization begins to suffer. Minor issues to one person may in fact be usability breakdowns to another.

Looking across Tables 4, 5, 6, 7 and 8, it is clear that within each visualization, all participants reported similar numbers of usability problems. This illustrates that participants found similar numbers of issues, regardless of their confidence or ratings of relevance of the heuristics. But given that a number of the issues were found in the "other" category, we examined the ratings participants gave for the clarity of the heuristics themselves. We note that participants 1–3 did not complete the clarity ratings, so the analysis was conducted on the remaining seven participants. We split participants into two groups, with low familiarity with heuristic evaluation defined as those participants who self-reported their experience with heuristic evaluations as level 1–3, and the high familiarity participants defined as those who self-reported their experience as level 4–6. This is because we found that familiarity with heuristic evaluations influenced ratings of relevance.

Clarity ratings were compared with a linear mixed ordinal regression model with a logit link function, using the fixed factor familiarity with heuristic evaluations (2 levels: low, high) and the random factor observer. The Type I error rate was $\alpha = .05$. There was not a significant difference in clarity ratings between the participant groups (Parameter $\beta_{\text{Low}} = -1.01$, $z = -1.74$, $p = .08$).

4 Conclusion

Our study shows that having familiarly with heuristic evaluations increases the confidence in the ability to apply heuristics for evaluating visualizations. This suggests that those who are asked to conduct a heuristic evaluation should first have some experience or training in using heuristics to evaluate visualizations. A training tool could be developed to help those novice to heuristic evaluations learn best practices in heuristic application for visualization evaluations. Or the evaluators who will be conducting the heuristic reviews could use the visualizations in this paper to see what usability issues they find and compare them with the ground truth. Another idea would be to hold some workshops at various information visualization and visual analytics conferences to introduce the technique of heuristic evaluation to the communities.

Our data suggests that conducting heuristic evaluations for visualizations does not require any experience using the visualization itself. This is actually a good outcome as heuristic evaluations in information visualization and visual analytics will certainly involve novel visualizations.

While the heuristic clarity ratings did not differ significantly based on the participant's experience with heuristic evaluations, we did find that participants had difficulty attributing a usability problem to a specific heuristic. In particular, participants had difficulty attributing usability issues relating to color to a heuristic. This could be addressed by creating a heuristic that specifically addresses color usability issues. Or through training that demonstrates how evaluators can operationalize usability problems so that developers can understand how to fix the usability issue. Future studies could help determine specifically what evaluators found difficult with attributing a usability problem to a heuristic by asking the participants to describe the found usability problems aloud.

We definitely recommend that visualizations on which a heuristic evaluation is going to be conducted include the necessary information so that interactions can be properly evaluated. To simulate the interaction of a visualization, interactive interface elements such as buttons, icons or menus could be displayed. Another static image could show the result of the different interactions allowing evaluators to assess heuristics concerned with interactions and processes.

Our study has shown that while some further investigations into heuristics for visualizations and visual analytics should definitely be done at this time, using these heuristics for information visualization and visual analytics should definitely help researchers and developers to produce early visualizations with fewer usability issues. We encourage the communities to use the heuristic review technique using this set of heuristics to see if their visualizations are improved.

Acknowledgments. The research described in this document was sponsored by the U.S. Department of Energy through the Pacific Northwest National Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

1. Forsell, C., Johansson, J.: An heuristic set for evaluation in information visualization. In: Proceedings of the International Conference on Advanced Visual Interfaces, pp. 199–206. ACM (2010)
2. Hearst, M.A., Laskowski, P., Silva, L.: Evaluating information visualization via the interplay of heuristic evaluation and question-based scoring. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 5028–5033. ACM (2016)
3. Ishihara, S.: Tests for Colour-Blindness. Hongo Harukicho, Handaya (1917)
4. Nielsen, J.: Finding usability problems through heuristic evaluation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 373–380. ACM (1992)
5. Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 152–158. ACM (1994)
6. Nielsen, J.: 10 usability heuristics for user interface design. https://www.nngroup.com/articles/ten-usability-heuristics/
7. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 249–256. ACM (1990)
8. Plaisant, C., Grinstein, G., Scholtz, J., Whiting, M., O'Connell, T., Laskowski, S., Chien, L., Tat, A., Wright, W., Görg, C., et al.: Evaluating visual analytics at the 2007 vast symposium contest. IEEE Comput. Graph. Appl. **28**(2) (2008)
9. Snellen, H.: Test-types for the Determination of the Acuteness of Vision. PW van de Weijer, Utrecht (1862)
10. Väätäjä, H., Varsaluoma, J., Heimonen, T., Tiitinen, K., Hakulinen, J., Turunen, M., Nieminen, H., Ihantola, P.: Information visualization heuristics in practical expert evaluation. In: Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization, pp. 36–43. ACM (2016)