A Heuristic Approach to Value-Driven Evaluation of Visualizations

Emily Wall, Meeshu Agnihotri, Laura Matzen, Kristin Divis, Michael Haass, Alex Endert, and John Stasko

Abstract—Recently, an approach for determining the value of a visualization was proposed, one moving beyond simple measurements of task accuracy and speed. The value equation contains components for the time savings a visualization provides, the insights and insightful questions it spurs, the overall essence of the data it conveys, and the confidence about the data and its domain it inspires. This articulation of value is purely descriptive, however, providing no actionable method of assessing a visualization's value. In this work, we create a heuristic-based evaluation methodology to accompany the value equation for assessing interactive visualizations. We refer to the methodology colloquially as ICE-T, based on an anagram of the four value components. Our approach breaks the four components down into guidelines, each of which is made up of a small set of low-level heuristics. Evaluators who have knowledge of visualization design principles then assess the visualization with respect to the heuristics. We conducted an initial trial of the methodology on three interactive visualizations of the same data set, each evaluated by 15 visualization experts. We found that the methodology showed promise, obtaining consistent ratings across the three visualizations and mirroring judgments of the utility of the visualizations by instructors of the course in which they were developed.

Index Terms—Visualization evaluation, heuristics, value of visualization

1 Introduction

Evaluating the utility of visualizations is notoriously difficult [3, 20]. While the field of human-computer interaction has provided many techniques to assess the usability of an interactive system [28], determining the ability of a visualization to assist in understanding and analyzing data presents unique challenges [15, 30].

One approach to evaluating a visualization's utility is to measure accuracy and time in a study where participants perform benchmark tasks [5, 14]. These studies can be helpful to determine if people can manipulate the user interface and interpret the visualization to read data properly. However, they usually only assess a visualization's ability to communicate data "facts", that is, attributes of individual data elements and core statistical values such as correlations, distributions, and outliers. Many researchers seek to go beyond this evaluation approach in order to determine the potential utility or value of a visualization.

One approach to achieving a more in-depth assessment of a visualization's utility is the insight-based visualization evaluation methodology [24]. Using this approach, experts in the domain of a data set put a system to trial use to determine if the tool provides insights that are valuable to its end-users. Evaluators must determine how many insights about a data set the visualization inspired. An insight is defined to be complex, deep, qualitative, unexpected, and relevant [18]. While determining a visualization's ability to generate insights is clearly a big step toward determining its utility, this evaluation methodology can still be quite challenging. First, the study must be conducted with domain experts who have an appropriate level of knowledge about the data. Further, determining whether a unit of knowledge acquisition is an "insight" or not is still relatively subjective.

An alternative approach to determining utility is to deploy a visualization in the field and conduct a more in-depth, longitudinal evaluation. This type of study seeks to move beyond the limitations of short-term, lab-based evaluations. Perhaps the best known example of this evaluation methodology is the MILC (Multi-dimensional In-depth Long-term Case study) technique [27] that has been used to evaluate political analysis, biomedical research, and intelligence analysis [19]. System use is observed "in the field" as people apply it to real data and problems. The power and potential benefit of this approach for helping to determine the utility of a visualization is obvious. However, such an evaluation may be logistically challenging, very time-consuming, and pragmatically difficult to implement. Developers of new visualization techniques may seek evaluation methodologies that are lower cost but still achieve many of the same benefits.

In 2014, Stasko proposed a new framework for identifying the *value* of visualization [30]. In particular, this approach sought to move beyond the types of questions and tasks usually found in usability studies. As stated in the article, "[A measure of value] goes beyond the ability to support answering questions about data–it centers upon a visualization's ability to convey a true *understanding* of the data, a more holistic broad and deep innate sense of the context and importance of the data in 'the big picture'." The value framework contained four components corresponding to the time savings a visualization provides, the insights and insightful questions it spurs, the overall essence of the data it conveys, and the confidence about the data and its domain it inspires. The evaluation approach advocated in the work was largely descriptive. Each of the four components was explained, but no concrete techniques for assessing a visualization along those components was provided. To be more pragmatically beneficial, an accompanying evaluation methodology or corresponding prescriptive approach is also needed.

The goal of our research is to develop just such a methodology. We seek to provide an evaluation approach to estimate and even quantify the potential value of visualization for understanding a data set, centered on the four value components introduced in [30]. We also want this approach to be relatively "low cost" in terms of time and resources required to employ it. We fully acknowledge that longitudinal studies of deployed system usage are the hallmark for truly understanding a system's value. We similarly seek an approach that provides feedback about a system's utility, especially that beyond simple low-level task completion. But we seek an approach that is practical and relatively easy to utilize, one providing rapid feedback that also allows comparisons to be drawn between different visualization applications.

We intend the methodology to be useful for evaluations of the potential utility and value of both research and commercial visualization applications. Researchers and developers frequently desire feedback about new systems they develop and want help identifying the strengths and limitations of their systems. Other potential uses include evaluation and grading of academic class projects or visualization contests and providing information to decide between commercial tools. Our goal is not to replace traditional time and error usability evaluations, but to complement existing evaluation techniques with a higher-level,

- *Emily Wall, Meeshu Agnihotri, Alex Endert, and John Stasko are with Georgia Institute of Technology, Atlanta, GA, USA. E-mail: {emilywall, magnihotri6, endert, stasko}@gatech.edu.*
- *Laura Matzen, Kristin Divis, and Michael Haass are with Sandia National Laboratories, Albuquerque, NM, USA. E-mail: {lematze, kmdivis, mjhaass}@sandia.gov.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

value-driven evaluation focus.

In this paper, we describe the development of a methodology that enables a quantitative assessment of a visualization's value according to the value equation. Our approach to this challenge involved the identification of more specific guidelines under each component, and then a set of low-level heuristics to be judged under each guideline. All three levels of the value framework combine to create a form for use in evaluating a visualization.

This article describes the process we undertook to create the evaluation methodology and an initial assessment in which 15 visualization researchers evaluated three different visualizations of the same data set developed by student groups in an information visualization course project. Although our expert visualization participants expressed doubts about the evaluation instrument's ability to assess the value of the visualizations, their evaluation responses were consistent, achieving high inter-rater reliability. Their average ratings mirrored instructor feedback on the visualizations from course project evaluations. Thus, we believe that this evaluation methodology shows promise as a low-cost estimate of a visualization's value.

2 Related Work

Evaluating visualizations is an open and difficult research challenge [3, 20]. This complexity stems from the broad set of design goals that visualizations can be built to support. For many of these goals, specific evaluation methodologies have been presented [15].

For example, visualizations are commonly designed to help the user gain *insights* about a data set. In response, North et al. presented *insight-based evaluation* [18,24] as a methodology to assess how well a visualization supports people gaining insight into the data being shown. However, as discussed above, operationalizing the methodology is challenging due to the difficulty of defining, observing, and counting insights [4]. Further, insights may be dependent on the domain expertise or familiarity with the data set, making it difficult to use as a benchmark by which multiple visualizations can be compared.

Alternatively, task performance methodologies can be used. These approaches set up a series of tasks that users should be able to complete with the given visualization [5, 11, 14, 31]. Then, metrics such as task completion time and accuracy are used to evaluate how well a visualization performed. While these methodologies provide quantitative data which can ease comparisons, designing the set of tasks can be subjective, and the data sets require ground truth in order to evaluate accuracy. In addition to task performance and usability approaches, methodologies exist that evaluate visualizations based on user experience goals such as engagement, enjoyability, memorability, and others [23].

Particularly within the visual analytics community, contests have been used to help evaluate data visualization systems [7, 21, 26]. Developing data sets, problems, and scenarios for such contests is extremely time-consuming and difficult [6], however, and each focuses on a very specific type of data.

Deployment studies, where a system is used for everyday tasks in context outside the lab, provide a deeper look into a visualization system's utility. The MILC technique [27] is one example of this approach. Such evaluations generally are viewed as powerful instruments of assessment, but they can be logistically challenging and time-consuming.

Grounded in research from the human-computer interaction community, heuristic-based evaluation methodologies [9, 16, 17] for visualization have been proposed [8, 10, 22, 25, 32, 35]. For example, Amar and Stasko [1] identify heuristics designed to cover the known "gap" in visual analytics processes. However, these heuristics are fairly high level, suggestive, and provide limited guidance on improving specific visual or interactive aspects of a visualization tool. Conversely, Zuk and Carpendale [35] suggest a set of ten "Cognitive and Perceptual Heuristics" for designing visualizations. But their high specificity in wording leads to less flexibility in interpretation from one visualization to another. Forsell and Johansson [10] instead compiled 63 published heuristics and tested them on a collection of 74 usability problems from previous information visualization evaluations to identify the top 10 heuristics that covered 87% of the 74 problems. However, as Tarrell et al. [32] point out, by broadly wording such heuristics, they may be misinterpreted by different evaluators. Furthermore, as these heuristics have solely been tested on usability issues, they might not be effective for visual data analysis and reasoning evaluations. Some researchers have compared heuristic evaluation of visualizations to alternatives like usability evaluation with benchmark tasks [29] or having evaluators answer questions about the data [12]. These studies revealed that heuristic evaluation can complement other evaluation alternatives.

Tory and Möller [33] adapted an expert review process with heuristics to get feedback on design alternatives for specific visualizations. They found that experts can provide quick and useful feedback on specific design goals and heuristics. Ardito et al. [2] provided additional context and guidance around heuristics to assist less skilled inspectors in the evaluation of domain-specific visualization tools. Perhaps most closely related to our work is the heuristic-based methodology by de Oliveira and da Silva [22]. They presented a set of 15 heuristics based on common visualization design goals distilled from a literature review. While these heuristics are meaningful, a method for translating them into an operational methodology is missing. Our proposed value-based methodology includes not only heuristics, but realizes them in a full methodology.

Finally, using value or utility as a metric to characterize visualizations has been previously explored, though in a markedly different approach than we follow. van Wijk proposed an economic value model [34] that mathematically represents and calculates the value of a visualization in purely numerical terms. We posit that the value of a visualization is often difficult to define through strictly mathematical terms, and thus adopt a heuristic-based approach for determining value.

3 Value of a Visualization

In 2014, Stasko explored some of the different objectives of evaluating a visualization [30]. Potential goals include improving a system, comparing two systems, or simply determining the quality or "goodness" of a system. While all are helpful applications of evaluation, he argued for a broader, more encompassing notion of the value of a visualization.

Some of the motivation for this focus on value was to move beyond evaluations involving participants performing low-level benchmark tasks and answering specific questions about a data set. While this type of evaluation can help determine whether a visualization is learnable and comprehensible, it fails to examine some of the larger benefits of visualization. Stasko felt that these larger benefits are what makes visualization unique among data analysis approaches. He states, "Visualization should ideally provide broader, more holistic benefits to a person about a data set, giving a "bigger picture" understanding of the data and spurring insights beyond specific data case values." [30]

He described a simple value equation

$$\mathbf{V = T + I + E + C}$$

with the following four components:

- T - A visualization's ability to minimize the total **time** needed to answer a wide variety of questions about the data
- I - A visualization's ability to spur and discover **insights** and/or **insightful questions** about the data
- E - A visualization's ability to convey an overall **essence** or takeaway sense of the data
- C - A visualization's ability to generate **confidence**, knowledge, and trust about the data, its domain and context.

The article introducing this value equation was limited to a qualitative discussion of its details and the four components. While examples of applying the equation to specific visualizations were provided, they were simply narrative descriptions. No accompanying methodology or quantitative breakdown was provided, thus it lacked prescriptive power to evaluate visualizations and compare their potential values. Hence, our goal in this work is to provide an actionable methodology to accompany the value equation.

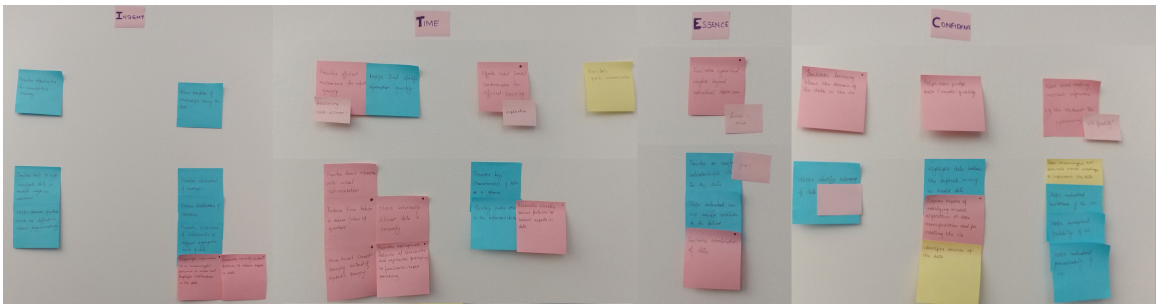


Fig. 1: View of materials from the affinity diagramming exercise to create the initial version of the three-level value framework.

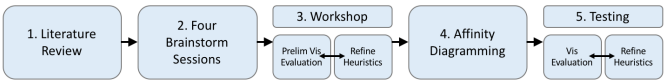


Fig. 2: The five stage process used for developing the value heuristics.

4 Developing the Heuristics

We developed the value-driven visualization evaluation heuristics through an iterative process, shown in Figure 2. As a starting point, we surveyed many visualization evaluation and design heuristics papers to help generate an initial list of 17 heuristics of value, each falling under one of the high-level components from the original value equation. We held four additional brainstorming sessions, one for each high-level component, resulting in the expansion of the list of heuristics to 70. At this point, our objective was to be inclusive of all reasonable heuristics.

With the list of heuristics expanded, we conducted a half-day workshop with three high-level goals: (1) refine the list of heuristics, (2) assess the rateability of each heuristic, and (3) test the heuristics on visualizations. We first removed any heuristics that were very similar to others. Each member of the research team also selected the five heuristics that they viewed as most important for each high-level component in order to establish a type of prioritization.

Next, we discussed the rateability of each heuristic. We each assessed all of the heuristics in two ways: as rateable with a yes/no judgment and through a more nuanced low/medium/high judgment. Heuristics that could not be rated with either approach were eliminated or reworded.

Then the research team was joined by two additional visualization experts who had not been involved in the process of developing the heuristics. The group studied a sample visualization and each individual rated each heuristic. We examined the resulting ratings to assess their consistency. We discussed these ratings at length to understand the causes of any particularly noteworthy disparities.

We found that different individuals interpreted some heuristics in different ways, so we rephrased them. For example, we changed "The visualization facilitates learning more broadly about the domain of the data" to "The visualization promotes understanding data domain characteristics beyond the individual data cases and attributes." The initial phrasing was more abstract and led raters to focus on specific data points or attributes they may not have previously known about. However, our goal with this heuristic was to promote a higher-level understanding of the domain (the "forest") rather than small details of knowledge about specific data points or attributes (the "trees").

Some heuristics were not difficult to understand but turned out to be very difficult to rate. For example, we altered "The visualization highlights potential data issues like unexpected, duplicate, missing, or invalid data" to become "If there were data issues like unexpected, duplicate, missing, or invalid data, the visualization would highlight those issues." The first phrasing proved difficult because it presumed problems in the data that might not be present. If a rater did not spot such data issues, was it because the visualization failed to highlight them or because none were present?

Some heuristics that were difficult to rate were discarded or rephrased. Despite rephrasing, a few remained difficult to rate, but we were reluctant to remove them because we felt that they ultimately captured an important aspect of a visualization's value. For example, the guideline "The visualization provides opportunities for serendipitous discoveries" proved difficult for assigning a rating, but we felt that it captured a core element of insight. This ultimately led us to restructure the value framework into a three-level hierarchy, adding a set of mid-level guidelines to each of the four components.

To form the hierarchy, we conducted an affinity diagramming exercise to organize the heuristics into their new structure (Figure 1). We then repeated the process of rating a visualization, analyzing inconsistencies, and rephrasing, removing, or adding to the hierarchy of components, guidelines, and heuristics. The resulting hierarchy is presented in the next section.

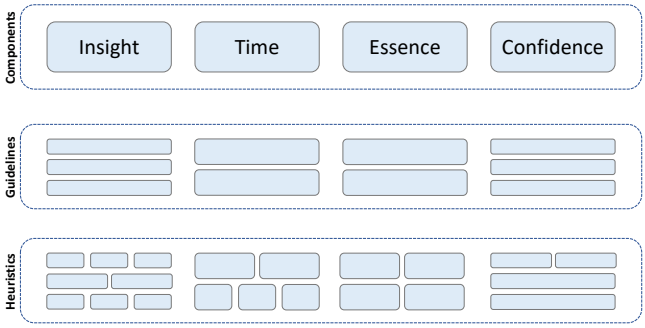


Fig. 3: The structure and terminology used to describe the hierarchical value framework. Each *component* is made up of *guidelines* which describe important aspects of the high-level component. Each *guideline* is then comprised of a small set of low-level *heuristics* that are designed to be actionable, rateable statements reflecting how a visualization achieves that guideline.

5 Value-Driven Evaluation Methodology

The value framework consists of three hierarchical levels (Figure 3). The top level contains the original four *components*: insight, time, essence, and confidence. Within each component, a small set of mid-level *guidelines* capture the core concepts of the high-level *components*. Finally, each guideline contains one to three low-level *heuristics*. We developed these heuristics to be actionable, rateable statements that embody the core concepts of the guidelines and components in the hierarchy above them. Hence, the upper-level guidelines and components themselves are not intended to be directly rated in this methodology. Instead, the ratings of the individual heuristics are aggregated up the hierarchy to form the overall score for a visualization (described in more detail later). We informally refer to the methodology as ICE-T, an anagram of the four value components (**I**nsight, **C**onfidence, **E**ssence, and **T**ime).

5.1 Framework Realization

We present the entire value hierarchy in Figure 7. Below, we briefly describe the contents of each component.

Insight – This component is comprised of three mid-level guidelines, which are roughly intended to capture how a visualization supports *intentional* and *incidental* insights. Intentional insight refers to tasks or questions a person sets out to address, while incidental insight refers to serendipitous discoveries where the user may have stumbled upon an unexpected piece of knowledge.

Time – This component is comprised of two mid-level guidelines, intended to capture how a visualization facilitates faster, more efficient understanding of data with respect to both *searching* and *browsing* of data. Searching refers to a user's deliberate task to locate particular information within a data set, while browsing refers to a user's more casual scanning of a data set to find potentially interesting information.

Essence – This component is comprised of two mid-level guidelines, intended to capture how a visualization communicates the essence of the data set with respect to *overview* and *context*. Overview refers to a high-level view or summarization of the data set, while context refers to relevant information surrounding the data set.

Confidence – This component is comprised of three mid-level guidelines, intended to capture how a visualization helps a user feel confident in his/her understanding of the data set with respect to the *quality of the data* and *quality of the visualization*. Confidence in the quality of the data refers to an understanding of potentially missing or erroneous data, while confidence in the quality of the visualization refers to an understanding of the accuracy of the representation of the data (e.g., does the visualization mislead?).

5.2 Implementation

We intend the methodology to be administered using a survey (available in the supplemental materials). Each heuristic should be individually rated for a visualization along a 7-point scale ranging from 1-*strongly disagree* to 7-*strongly agree*, or N/A-*not applicable*. All the heuristics are stated in a positive manner, that is, a higher score (strongly agree) aligns to a visualization being more valuable.

We performed a trial of the methodology in the early stages of developing and refining the heuristics with static and minimally interactive visualizations. We found that many heuristics were not applicable to these types of visualizations because they assumed that the rater could interact with the data. For example, the heuristic "The interface supports using different attributes of the data to reorganize the visualization's appearance" is not applicable to a static visualization. Thus, the methodology is intended to be applied to interactive visualizations.

To evaluate a visualization, a small number of visualization-knowledgeable raters should interact with the visualization and complete the survey. We recommend five raters based on our analysis in Section 7.1.2. These raters should have knowledge about and experience working with visualizations. Domain knowledge is also relevant, so the raters should have at least some familiarity with concepts from the domain of the data set being visualized.

5.3 Score Aggregation

In order to achieve an overall value rating for a visualization, we propose an initial approach of aggregating scores at each level of the hierarchy using a simple average.

Let s_h be the score for a heuristic h ranging from 1-7 identified by a rater. Each mid-level guideline is scored by averaging its corresponding j low-level heuristics from the hierarchy: $s_g = \frac{1}{j}\sum_{i=1}^{j} s_{h,i}$. Each high-level component is then scored by averaging its corresponding k guideline scores: $s_c = \frac{1}{k}\sum_{i=1}^{k} s_{g,i}$, where $c \in \{insight, time, essence, confidence\}$. Finally, a visualization's overall score is defined as $s = \frac{1}{4}(s_{insight} + s_{time} + s_{essence} + s_{confidence})$. This method serves as an initial aggregation approach, not favoring any one component, guideline, or heuristic over another. In a subsequent section, we discuss alternative ways that scores might be aggregated.

6 Assessing the Methodology

To assess this value-driven evaluation approach, we conducted a user study with visualization experts who we asked to use the methodology to rate three visualizations. In this assessment, our primary goals include (1) assessing the inter-rater reliability of the evaluators' ratings and the corresponding statistical power, (2) understanding how heuristic ratings map to properties of individual visualizations, (3) gauging evaluators' confidence in assigning scores to heuristics, and (4) gathering overall impressions of the methodology from the visualization experts. This section describes the design of the assessment and the results are presented in the following section.

6.1 Participants

We sent a recruitment email to 23 people, all of whom hold a Ph.D. and perform visualization-related research, so thus can be considered visualization experts. A total of 15 participants (12 male, 3 female) ultimately completed the study. The participants included six research staff, eight professors, and one software engineer. The participants had a range of 7-30 years of professional experience in the field of visualization (mean = 14 years).

6.2 Materials

For the experiment, we selected three visualizations developed by student groups in an undergraduate information visualization course at Georgia Tech (Visualization A[1], B[2], and C[3]). Rather than choosing existing published or publicly available visualizations, this ensured that participants in the study would not have prior exposure to the visualizations. Further, the three visualizations all utilize the same data set (information about U.S. colleges) to ensure that there were no confounding differences between visualizations in terms of the participants' familiarity with the data sets.

Figure 4 shows the three visualizations. Visualization A used a map to show college locations and parallel coordinates for comparing attribute values. Visualization B used a scatterplot, two focus views, and extensive filtering and interaction. Visualization C employed a bubble clustering view along with a scatterplot.

We explicitly chose visualizations with varying quality and design decisions to try to capture a larger range of ratings in the value heuristics. Project grades from the course, assessed by the professor and teaching assistants, suggested that Visualizations A and B would receive higher scores than C, with Visualization B slightly ahead of A. This ordering corresponded to the research team's assessment of the relative value of the three visualizations, based on a qualitative assessment of their features and design choices. Therefore, we predicted that if the value-driven evaluation methodology is effective, Vis B would receive the highest overall score and Vis C would receive the lowest.

Each participant rated all three visualizations via a web-based survey form. They scored the 21 low-level heuristics using the 1-7 rating scale described earlier. We further augmented the survey for the purposes of this assessment so that each low-level heuristic rating was accompanied by a rating of the participant's confidence in assigning the score, judged from 1 (very low) - 4 (very high). We gave evaluators no specific directions beyond assessing the potential value of each visualization. We were confident that all knew the data domain well because of their background with universities and higher education.

6.3 Procedure

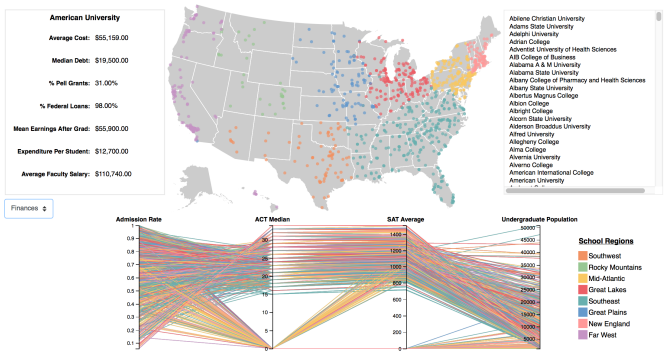
We emailed participants an electronic consent form to sign and return. Upon receiving signed consent, we emailed participants instructions for completing the study, including a background questionnaire, a link to the online survey containing the value hierarchy and heuristics, and links to the three interactive visualizations, each with an annotated screenshot to inform participants about each visualization's affordances.

We asked the participants to examine the annotated screenshot for a visualization, then use the visualization to familiarize themselves with

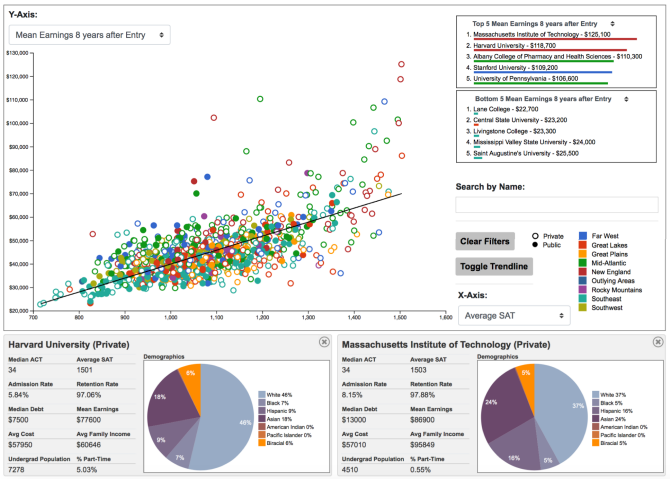
[1] http://vis.gatech.edu/demo/value/vis133/

[2] http://vis.gatech.edu/demo/value/vis460/

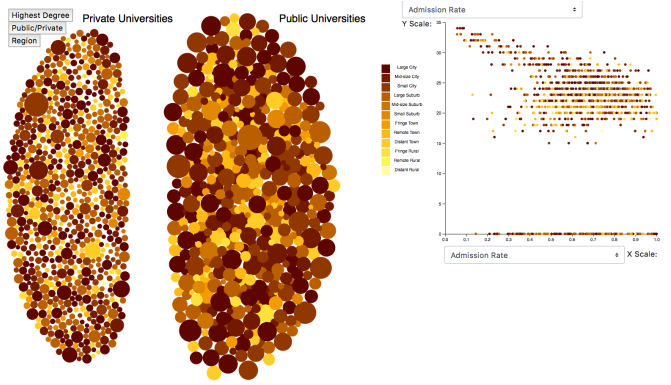
[3] http://vis.gatech.edu/demo/value/vis745/



(a) Visualization A



(b) Visualization B



(c) Visualization C

Fig. 4: The visualizations of U.S. university data used in the assessment of the value-driven methodology for evaluating visualizations.

its representation, interactions, and data. Finally, we asked them to complete the heuristic survey for the visualization. In addition to rating each heuristic and denoting their confidence in that score, participants also had the option of typing comments about each heuristic.

We used a pseudorandom order of visualizations to minimize potential ordering effects. There were six possible visualization orderings and all six were used for at least two participants. We gave participants two weeks to complete the study with no explicit time limit for how long to spend familiarizing themselves with a visualization or completing the heuristic survey. Therefore we do not know how long each

	Vis A	Vis B	Vis C	Average
P15	6.09	6.01	5.00	5.70
P14	5.08	5.51	4.94	5.18
P10	4.45	5.99	4.74	5.06
P5	5.05	6.24	3.69	4.99
P1	5.11	5.30	3.95	4.79
P4	4.39	5.24	4.50	4.71
P3	4.52	5.71	3.76	4.66
P13	5.60	5.90	2.49	4.66
P8	4.08	5.89	3.55	4.51
P9	3.96	5.37	4.05	4.46
P2	4.20	4.58	4.44	4.41
P7	4.24	4.78	3.62	4.21
P11	4.42	4.11	4.10	4.21
P6	4.78	4.68	2.81	4.09
P12	4.23	4.06	3.98	4.09
Avg.	4.67	5.30	3.96	

Fig. 5: Summary (total) ratings of the three visualizations by the 15 study participants. Cells highlighted in green identify a participant's highest rated visualization, and those highlighted in yellow indicate a person's lowest rated visualization. Rows are sorted vertically by overall rater difficulty.

	Vis A				Vis B				Vis C			
Components	I	T	E	C	I	T	E	C	I	T	E	C
P15	6.11	5.75	6.50	6.00	6.89	6.42	5.75	5.00	4.83	4.92	5.25	5.00
P14	5.72	5.00	5.25	4.33	6.22	5.42	4.75	5.67	5.17	4.92	5.50	4.17
P10	4.56	4.58	4.50	4.17	6.39	6.50	5.75	5.33	5.61	4.17	5.00	4.17
P5	5.61	5.08	5.00	4.50	6.28	6.67	6.50	5.50	4.50	3.50	3.25	3.50
P1	4.78	5.33	5.00	5.33	5.61	5.42	5.00	5.17	4.22	3.83	4.25	3.50
P4	4.72	3.50	4.00	5.33	4.61	5.92	5.25	5.17	4.67	3.83	3.50	6.00
P3	4.67	3.17	5.25	5.00	5.67	6.42	6.75	4.00	4.89	3.17	3.00	4.00
P13	6.22	4.58	5.25	6.33	5.94	7.00	6.00	4.67	2.78	2.17	2.00	3.00
P8	4.17	4.50	3.50	4.17	5.89	6.17	6.00	5.50	4.11	3.42	3.50	3.17
P9	4.00	4.17	3.50	4.17	5.56	5.50	5.75	4.67	4.61	2.25	5.00	4.33
P2	5.06	3.50	4.25	4.00	5.00	4.17	4.50	4.67	4.78	4.75	4.25	4.00
P7	4.61	4.00	4.00	4.33	5.72	4.67	4.75	4.00	4.06	3.17	3.25	4.00
P11	4.94	3.50	5.25	4.00	4.44	3.25	4.75	4.00	5.22	4.25	4.25	2.67
P6	4.61	5.58	4.75	4.17	5.39	4.08	4.75	4.50	2.83	2.67	3.25	2.50
P12	4.67	3.50	3.75	5.00	4.33	4.17	3.75	4.00	4.67	4.25	4.00	3.00

Fig. 6: Participants' ratings of the different visualizations, broken down by the four value components. The color mapping is red (1) to green (7) with white (4) being neutral. This table shows the overall strength of each visualization with respect to each of the four components. Scanning the values in a column shows how all the different raters scored a visualization with respect to a specific component.

person spent on each evaluation. Once they completed the study, we sent each participant a thank-you email that solicited their summative thoughts about the methodology and study.

7 Results

7.1 Participant Ratings

We aggregated the scores from each participant as described earlier in Section 5.3 to identify an overall value score for each visualization. Total scores computed from all participants are shown in Figure 5. The rows are sorted top-to-bottom by the average value each participant gave across all three visualizations, so the most favorable raters appear at the top and the most difficult raters appear at the bottom. Vis B received the highest average score of 5.30, Vis C received the lowest score at 3.96, and Vis A received an intermediate score of 4.67. These ratings aligned with the relative ranking of the visualizations that we received from the instructors of the course in which they were created, as well as our own assessment of their relative value.

Scores from the individual raters were generally consistent with the group average, with 11 of the 15 participants scoring Vis B highest and no participant scoring it lowest. Similarly, Vis C received the lowest score from 11 participants and never received the highest score.

Figure 6 drills down a level on the data to show participants' component ratings for each visualization. Here, we use a green-red color map to highlight regions of positive (green) and negative (red) views of the

visualizations. The patterns of scores were somewhat more variable for the mid-level guidelines and the low-level heuristics. We discuss some observations at these lower levels and their relationship to specific characteristics of the visualizations in more detail below.

7.1.1 Inter-Rater Reliability

We assessed inter-rater reliability using the rater vs. group approach. We calculated the mean rating for each visualization on each heuristic, and then calculated the correlation between each participant's scores and the mean scores. The mean rater-to-group correlation was significant for all three visualizations (for Vis A: $r = 0.68$, $t(13) = 3.33$, $p < 0.01$; for Vis B: $r = 0.75$, $t(13) = 4.06$, $p < 0.01$; for Vis C: $r = 0.54$, $t(13) = 2.29$, $p < 0.05$), indicating that there was substantial agreement among the raters. This suggests that although raters each had their own backgrounds and individual differences, the overall ratings were consistent for the three visualizations.

We also calculated inter-rater reliability at the component level to assess whether the participants' scores were more consistent for some components than for others. For this analysis, the participants' scores were collapsed across all three visualizations to ensure that the number of ratings for each participant was sufficient to produce a meaningful correlation. The analysis revealed that the mean rater-to-group correlation was significant for three of the four components and marginally significant for the fourth. There was significant inter-rater reliability for the insight component ($r = 0.56$, $t(13) = 2.46$, $p < 0.05$), the time component ($r = 0.58$, $t(13) = 2.55$, $p < 0.05$), and the confidence component ($r = 0.55$, $t(13) = 2.40$, $p < 0.05$). However, the rater-to-group correlation did not quite reach significance for the essence component ($r = 0.49$, $t(13) = 2.03$, $p = 0.06$).

7.1.2 Power Analysis

Given the size of the correlations observed in this evaluation, we conducted a power analysis to calculate the number of raters that would be required to achieve consistent results using this methodology. Using the average rater-to-group correlation for the overall scores across all three of the visualizations ($r = 0.66$), and the conventional values for Type I and Type II errors ($\alpha = 0.05$ and $\beta = 0.20$, respectively), we estimate that five raters would be sufficient.

7.2 Relationships Between Scores and the Characteristics of the Visualizations

Figure 7 shows the scores for each visualization on every heuristic. Vis B received the highest average score on all but two of the low-level heuristics. For the insight heuristic "The visualization shows multiple perspectives about the data," Vis A had the highest average score at 5.4 while Vis B and Vis C were tied at a slightly lower score of 5.2. In this particular evaluation, all three visualizations showed multiple perspectives, so this heuristic does not do much to distinguish between them. On the confidence heuristic "If there were data issues like unexpected, duplicate, missing, or invalid data, the visualization would highlight those issues," Vis A received the highest average score at 4.07, with Vis B and Vis C receiving lower scores of 3.33 and 3.29, respectively. Some of the participants (P7 and P13) who provided comments for this heuristic noted that missing data was evident due to the zero values in the parallel coordinates plot in Vis A. In Vis B, zero values do not appear in the scatterplot, making it less obvious that there is missing data. In Vis C, zero values are shown in the scatterplot, but as one participant noted, a user would have to go through all of the dimensions, one by one, to understand which data is missing.

Another illustrative case is the time heuristic "The interface supports reorganizing the visualization by the data's attribute values." This heuristic has the biggest differences in average scores across the three visualizations, with Vis B receiving the best average score at 6.07, Vis C receiving a score of 4.93, and Vis A receiving a very poor score of 2.73. In this case, Vis A suffers due to the lack of flexibility in the parallel coordinates plot. The features of Vis B, including filtering, search, and the highly-flexible scatterplot, lead to a very high score.

At the guideline level, Vis B had the highest average score for all of the guidelines except for the confidence guideline: "The visualization helps understand data quality." This guideline has only one heuristic underneath it, and Vis B scores relatively poorly on this heuristic because it does not make missing data readily apparent, as discussed above. This reveals a potential weakness in our method of using a simple average to aggregate the scores at each level of the hierarchy. Since some of the mid-level guidelines have more low-level heuristics than others, some of the heuristics get weighted more heavily in the aggregation process.

Vis C had the lowest average scores on all of the guidelines except for two. It outperformed Vis A on the time guideline "The visualization provides mechanisms for quickly seeking specific information" and on the essence guideline "The visualization provides an understanding of the data beyond individual data cases." For the heuristics under both of these guidelines, participants remarked that the parallel coordinates plot in Vis A was too limited. The scatterplot in Vis C provided more support for these goals.

7.3 Confidence in Ratings

In addition to collecting a rating for each heuristic, we also gathered a confidence level for each, ranging from 1-very low confidence to 4-very high confidence. In general, the participants reported that they were confident in their responses, with a mean confidence rating of 3.22 and a standard deviation of 0.70. None of the heuristics had an average confidence rating lower than 3.

One or more participants gave a confidence rating of 1 to a total of five heuristics, three related to insight and two related to time. For the heuristics related to insight, one participant (P13) had low confidence in the heuristic "The visualization promotes exploration of relationships among different aggregation levels of the data" and commented that it was unclear what "aggregation" meant for this data set. Another participant (P3) had low confidence in their ratings for both of the heuristics that fell under the guideline "The visualization provides a new or better understanding of the data." P3 commented "If I were a school administrator I suspect that this would generate more questions."

For the heuristics related to time, two different participants (P2 and P13) had low confidence in their ratings for the heuristic "The visualization supports smooth transitions between different levels of detail in viewing the data." P2 commented that there was not enough information to rate this heuristic, and P13 commented that they were unsure of what levels of detail the question referred to. Another participant (P11) had low confidence in their ability to rate the heuristic "The visualization avoids complex syntactic querying by providing direct interaction" and commented that they did not understand what this heuristic meant.

7.4 Qualitative Feedback

We subsequently invited study participants to share their feedback and comments about the evaluation methodology. Although the results of the evaluation indicated that the participants were fairly consistent in their responses, many of those who offered feedback were skeptical about this approach. The feedback from the participants fell into two general categories: comments about specific heuristics, and comments about the evaluation process itself.

The participants' comments about specific heuristics indicate that the wording of the heuristics was confusing in some cases. For example, two participants (P10 and P14) were unsure of what was meant by the phrase "data cases." We simply used this term to refer to a single item or instance in the data set; in our study, this would be a university. Others felt that specific heuristics were too broad, too subjective, or too multi-faceted, making them difficult to evaluate.

The comments about the evaluation process itself revealed three general themes. First, two participants (P8 and P13) felt that the evaluation process would have been more effective if they were given a persona or a task to complete using each visualization. A frequent comment was that the ratings for each visualization might differ for different kinds of tasks. Second, three participants (P1, P10, and P11) noted that it was difficult to rate some of the heuristics when the visualizations provided multiple views of the data. One view might score well on the heuristic while another might score poorly, and the participants were unsure of how to coalesce those differences into a single score. Finally, the most

terminology	*data case*- refers to an instance of the data set; synonymous with data item or data point *attribute*- refers to properties of the data cases in the data set; synonymous with feature, dimension, or variable *relationship in the data*- refers to attributes among the data, such as correlations, clusters, or distributions	Vis A		Vis B		Vis C	
		μ	σ	μ	σ	μ	σ
Insight	The visualization facilitates answering questions about the data The visualization exposes individual data cases and their attributes	6.07	1.03	6.33	1.11	5.27	1.22
	The visualization facilitates perceiving relationships in the data like patterns & distributions of the variables	5.33	1.18	6.27	0.96	5.27	1.28
	The visualization promotes exploring relationships (between individual data cases as well as different groupings of data cases) ~~(among different aggregation levels of the data)~~	3.60	1.76	4.60	1.76	3.33	1.63
	The visualization provides a new or better understanding of the data The visualization helps generate data-driven questions	4.73	1.03	5.73	1.03	4.33	1.45
	The visualization helps identify unusual or unexpected, yet valid, data characteristics or values	5.27	0.96	5.27	1.16	4.13	1.41
	The visualization provides opportunities for serendipitous discoveries The visualization provides useful interactive capabilities to help investigate the data in multiple ways	4.40	1.55	6.07	0.80	4.67	1.29
	The visualization shows multiple perspectives about the data	5.40	1.12	5.20	1.61	5.20	1.32
	The visualization uses an effective representation of the data that shows related and partially related data cases	4.87	1.25	5.40	1.35	3.73	1.39
Time	The visualization affords rapid parallel comprehension for efficient browsing The visualization provides a meaningful spatial organization of the data	5.40	1.40	5.47	1.19	3.33	1.35
	The visualization (shows) ~~(provides)~~ key characteristics of the data at a glance	4.60	1.40	5.00	1.81	3.27	1.62
	The visualization provides mechanisms for quickly seeking specific information The interface supports (using different attributes of the data to reorganize the visualization's appearance) ~~(reorganizing the visualization by the data's attribute values)~~	2.73	1.71	6.07	1.28	4.93	1.83
	The visualization supports smooth transitions between different levels of detail in viewing the data	2.92	1.93	5.00	1.73	3.27	1.53
	The visualization avoids complex (commands and textual queries) ~~(syntactic querying)~~ by providing direct interaction (with the data representation)	5.53	1.25	5.93	1.10	4.00	1.65
Essence	The visualization provides a big picture perspective of the data The visualization provides ~~(an effective)~~ a comprehensive and accessible overview of the data	4.80	1.01	5.07	1.28	3.40	1.18
	The visualization presents the data by providing a meaningful visual schema	5.21	0.80	5.29	1.14	3.57	1.34
	The visualization provides an understanding of the data beyond individual data cases The visualization facilitates generalizations and extrapolations of patterns and conclusions	4.27	1.53	5.60	0.91	4.40	1.40
	The visualization helps understand how variables relate in order to accomplish different analytic tasks	4.47	1.55	5.53	1.06	4.40	1.45
Confidence	The visualization helps avoid making incorrect inferences The visualization uses meaningful and accurate visual encodings to represent the data	5.07	1.16	5.53	1.06	3.67	0.90
	The visualization avoids using misleading representations	5.07	1.00	5.33	1.18	4.00	1.84
	The visualization facilitates learning more broadly about the domain of the data The visualization promotes understanding data domain characteristics beyond the individual data cases and attributes	4.87	1.06	5.60	0.83	4.13	1.41
	The visualization helps understand data quality If there were data issues like unexpected, duplicate, missing, or invalid data, the visualization would highlight those issues	4.07	1.73	3.33	1.54	3.29	1.54
	Cumulative Vis Score	4.67		5.30		3.96	

Fig. 7: This figure depicts the entire value evaluation hierarchy and framework, including the four components, the guidelines under each component, and the constituent heuristics for each guideline, both as used in our study (crossed-out text) and as updated afterwards. The figure also shows summary (average) ratings for the three visualizations on each of the heuristics, as well as the standard deviation of each rating. We employ a red-green color map to help communicate at a glance lower/poorer ratings (red) to higher/better ratings (green). Both versions of the hierarchy, in addition to other study materials, can be found at `visvalue.org`.

common comment, offered by four different participants (P1, P2, P8, and P9), was that this type of evaluation might be more effective if a small set of visualizations were rated relative to one another, rather than applying the heuristics to one visualization in isolation.

8 Refining the Methodology and Heuristics

We used the results of the study and feedback from the evaluators to revise the methodology and heuristics. As two of the participants noted, the efficacy of a visualization is highly dependent on its context of use. Thus, we recommend that the intended users and task be communicated to evaluators prior to an evaluation. Further, participants sometimes found it difficult to decide how to rate a heuristic for a visualization with multiple views. We suggest that they be rated according to the best view for the task, as discussed further in Section 10.

Some participants indicated that they were unsure of the meaning of specific terms, such as "data cases," that were used in the heuristics. To address this concern, we added a terminology table to the beginning of the heuristic questionnaire to clarify common language.

In addition to clarifying common terminology, we rephrased five of the individual heuristics based on participant feedback. Specifically, participants were confused by the usage of "aggregation levels" in the insight heuristic "The visualization promotes exploration of relationships among different *aggregation levels* of the data." We rephrased the heuristic to read "The visualization promotes exploring relationships between *individual data cases as well as different groupings of data cases*." Participants were also confused by the use of the term "syntactic querying" in the time heuristic "The visualization avoids complex *syntactic querying* by providing direct interaction." We rephrased this heuristic to read "The visualization avoids complex *commands and textual queries* by providing direct interaction *with the data representation*." Evaluators also commented that there were too many concepts to evaluate in the essence heuristic "The visualization provides an effective, comprehensive and accessible overview of the data." To simplify this heuristic, we removed the word "effective."

We rephrased two of the heuristics because the evaluators' ratings had a high standard deviation, indicating disagreement among the evaluators that seemed to be caused by differing interpretations of the heuristics. The first was the time heuristic "The visualization *provides* key characteristics of the data at a glance." We believe this is because the use of the word "provides" was unclear, so it was replaced with "shows." The second was the time heuristic "The interface supports reorganizing the visualization by the data's attribute values" We rephrased this heuristic to read "The interface supports using different attributes of the data to reorganize the visualization's appearance."

Other heuristics had ratings with relatively high standard deviations because they did not apply to particular visualizations. Evaluators had the option of choosing "not applicable" (N/A) for any given heuristic, but they were inconsistent in their use of that option. When a heuristic did not apply, some evaluators instead gave it a low or neutral score. Rather than changing the wording of the heuristics in such cases, we suggest explicitly instructing evaluators to use the N/A option whenever they question the applicability of a heuristic to a particular visualization.

Both the initial heuristics used in our assessment and their revisions are shown in Figure 7 and included in the supplemental materials.

9 Applying the Methodology

Moving forward, we believe that the ICE-T evaluation methodology shows promise for future visualization evaluations. In this section, we provide guidelines for applying and implementing the methodology.

Recruiting Evaluators. Rating the heuristics requires thought about the holistic design and implementation of a visualization, how it applies principles of perception, appropriate use of visualization techniques, and so on. As a result, the methodology is best applied by individuals who have experience in and knowledge about developing visualizations. However, it is difficult to specify a precise ideal level or duration of experience. It may be desirable in some cases to have evaluators who have designed and developed multiple systems over many years. Alternatively, for some scenarios, students who have completed a course on data visualization may suffice. Furthermore, depending on the evaluation goals, other criteria may be important for identifying evaluators. For example, visualizations of data in a specific domain may require that the evaluators also have knowledge about that domain.

Administering the Survey. We have deployed guidance and materials, including both electronic and printable versions of the heuristics survey, at `visvalue.org`. The evaluators should first familiarize themselves with the visualization tool being assessed and the data it depicts. We recommend accompanying the visualization with a short overview or tutorial as we did in our study. Furthermore, a description of the potential users and the context of use is also recommended. The evaluators should complete the heuristic form, being permitted to refer back to the visualization throughout.

Determining and Reporting Scores. Once evaluators have completed their ratings of the visualizations, the scores can be compiled into a succinct report summarizing the value of the visualization from the point of view of this methodology. It may be useful to numerically and visually report the averaged scores for each heuristic (and potentially the variance of those scores as well). A color-coded table, similar to those shown in Figures 5-7, could be used to visually indicate strengths of a visualization and areas for improvement. This can be used by the developers of the visualization to refine the design and functionality of the visualization and increase its overall value.

Interpreting the Scores. Within the 7-point Likert scale ratings of heuristics, a score of 4 indicates a neutral rating. The statements are phrased positively, so higher scores are considered "better." While obviously there is no set quality level or scale, from our initial assessment of the methodology, we find that a visualization with an average score of 5 or greater for a particular heuristic across all evaluators represents a strength of the visualization, while a score of 4 or lower represents a heuristic for which the visualization has a weakness. Based on our initial assessment of the methodology, we find that valuable, good visualizations should be earning an overall cumulative average score of 5 or higher. Visualizations earning an overall cumulative score of 4 or less are candidates for redesign and further thought. The establishment of more specific score guidelines is possible with additional usage and testing of the methodology.

Unlike some other evaluation methodologies, the ICE-T approach does not produce an actionable list of design problems and suggested modifications. However, the scores from a visualization's evaluation could be used to create actionable suggestions for areas of improvement in a visualization. For example, a visualization that has a low score on the insight heuristic "The visualization facilitates perceiving relationships in the data like patterns & distributions of the variables" could be improved by adding a representation of the data that can show potential correlations or clusters.

Potential Applications. In our assessment of the methodology, the heuristics were used by experts to rate three visualizations. However, we believe the methodology usage is not limited to comparative scenarios. Since the methodology results in quantitative measures of a visualization, it can be used to evaluate a single visualization in isolation. As discussed above, developers may seek to achieve a particular score level, and evaluators could establish score zones corresponding to outstanding, satisfactory, or poor performance. Potential uses of the ICE-T methodology include early evaluations of the efficacy of a research or commercial system in order to find relative strengths and weaknesses, much like that proposed for MILC evaluations [27], academic project evaluation and grading, decisions among alternatives for commercial or application-driven contexts, or similar scenarios.

10 Discussion

Reflections on the Assessment. Our analysis of the study results indicated that the evaluation methodology shows promise for identifying the value of a visualization. While participants voiced skepticism about some aspects of the methodology, their ratings were highly consistent nonetheless. Furthermore, the average scores for the visualizations corresponded to the relative rankings provided by the course instructors

and our own prior assessment of their relative quality. At the lower levels of the hierarchy there were occasional discrepancies between the average scores of the visualizations and the order of the overall scores, but these patterns could always be traced back to specific design features in the visualizations, such as the affordances of a scatterplot as opposed to a parallel coordinates plot. Another important outcome was the power analysis indication that a consistent result could be achieved with as few as 5 raters, suggesting the potential for our approach to be an effective, relatively "low-cost" evaluation methodology.

Aggregating Scores. In our assessment of the heuristic-based evaluation methodology, we aggregated scores using a top-down approach. That is, the visualization's score is comprised of a simple average of its score for each high-level component. The component scores are a simple average of the associated mid-level guideline scores, and the mid-level guideline scores are a simple average of the ratings for the low-level heuristics. The implication of this choice is that some low-level heuristics will ultimately carry more weight in the final rating of a visualization. For example, two of the guidelines under confidence each only have a single low-level heuristic, while two of the guidelines under insight each have three low-level heuristics. The guidelines with fewer low-level heuristics (and the components with fewer guidelines) will ultimately have a greater impact on a visualization's rating.

This approach could be modified according to individual evaluation goals, however. One alternative could include a bottom-up scoring approach, where each low-level heuristic is given equal weight. The tradeoff then would be that guidelines and components with more heuristics beneath them in the hierarchy would have a greater impact on a visualization's score. Fully custom heuristic weightings could also be employed, defined by the visualization developer or the evaluators. By applying higher or lower weights to specific heuristics, different capabilities could be emphasized toward the particular evaluation goals for a visualization. Furthermore, the evaluators themselves could be given control to increase or decrease importance of different components.

Multiple Views. One source of confusion that became clear in the assessment of the methodology was how to rate a single heuristic when a visualization contains multiple views, where one view might do something well while the other one does it poorly. For example, P1 commented "it was challenging to choose an answer because of the use of multiple views in the visualization (...) I found myself taking a mean of the answers for the multiple views to answer the questions." P10 said "I wished I could specify different answers for different parts of the visualization. Because in the same visualization there were several views that would perform quite differently on these scales." Evaluators may rate the heuristic according to the *best-case* (the best view for that heuristic determines the rating), the *worst-case* (the worst view for that heuristic determines the rating), or the *average-case* (some overall impression given multiple views determines the rating).

This issue could lead to inconsistent ratings among evaluators. For example, when rating Vis A under essence "the visualization helps understand how variables relate in order to accomplish difference analytic tasks," P13 noted that it was true for one view (parallel coordinates) and gave a rating of 6. On the other hand, P8 commented that the ability to understand relationships in the data using the vis as a whole was too limited and hence gave a rating of 3. This disparity can be mitigated by prescribing either best-case, worst-case, or average-case ratings to be used by evaluators. The purpose of having additional views is often to capture an aspect of the data or provide an analytic capability not well-supported by other views. Hence, we suggest that the intuitive choice is to prescribe that evaluators utilize best-case ratings. That is, if any one view of a visualization satisfies a heuristic well, then the entire visualization itself should be considered to do it as well.

Validating the Methodology. The visualization ratings that our study evaluators produced aligned with those that we received from the instructors of the class in which the visualizations were created. While this gives us confidence that the ratings from the study were appropriate, it is not a formal validation of the study results. Ideally, one should more rigorously confirm that the evaluation methodology produces accurate ratings of visualizations, a so-called "ground truth" [13].

It may be tempting to use other established visualization evaluation techniques (i.e., time & error-focused benchmark tasks, long-term deployment studies, etc.) to perform such a validation. However, we suggest that those techniques capture somewhat different aspects of a visualization's quality and utility than what our approach is intended to capture. We would expect results from the different methods to broadly align, but they might produce slightly different findings due to the different goals of each method.

In future work, we would like to better understand the effectiveness of the methodology compared to alternative evaluation approaches. For example, would the results of an insight-based evaluation [24] correlate to a rating produced by our I(nsight) component? By directly comparing evaluation results using our methodology to other approaches, one could gain a better understanding of the tradeoffs and appropriateness of the value-driven evaluation methodology.

Limitations. While our assessment shows promise for the ICE-T methodology, it is not without its limitations. Our assessment only addressed three specific visualizations. To better understand the generalizability of our methodology, we must examine its use on a wider range of visualization types with varying data domains, representations, and intended task support. Furthermore, we employed only visualization *experts* in our study. We do not know whether other evaluators with less visualization expertise would achieve similar results. Finally, the heuristics themselves require subjective interpretation by the evaluators, which may be unsettling to those people seeking more objective, precise assessments. However, we believe that the subjectivity is inherent to evaluating the overall *value* of a visualization and is hence a part of this methodology.

11 Conclusion

Numerous past papers have noted that evaluating visualizations is difficult. The process of developing a survey to quantify the value of visualizations confirmed this trope, but it also helped to pinpoint some of the reasons *why* evaluation is so difficult. It is hard to define the value of a visualization in terms that multiple raters can understand and apply with consistency. To be effective, the heuristics that raters will use must be easy to evaluate, but they must also be meaningful and able to differentiate between different design choices in visualizations.

Throughout the process of developing the ICE-T methodology, we created, eliminated, and refined numerous heuristics. The evaluation study showed that the resulting set of heuristics does a good job of distinguishing between three visualizations, ranking their potential value, and identifying particular points of strength or weakness. Although the expert raters were somewhat skeptical about the methodology, the results revealed that they were highly consistent with one another. The pattern of scores conformed to our own qualitative assessment of the value of the three visualizations. Furthermore, the effect size achieved by this evaluation indicates that a consistent score could be achieved with only 5 raters, which would make this kind of evaluation feasible for real-world use.

In summary, we have described the development of a new methodology for evaluating interactive visualizations. Our initial assessment shows promise for the methodology as a low-cost, but effective evaluation approach. The methodology is intended to identify a visualization's holistic value, and thus presents a complementary approach to existing evaluation techniques such time & error, insight-based, or deployment studies. The full value-driven methodology, including the heuristic survey and guidelines for use, is available online at `visvalue.org`.

Acknowledgments

This work was partially supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energys National Nuclear Security Administration under contract DE-NA0003525.

REFERENCES

[1] R. Amar and J. Stasko. Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):432–442, 2005.

[2] C. Ardito, P. Buono, M. F. Costabile, and R. Lanzilotti. Systematic inspection of information visualization systems. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, BELIV '06, pp. 1–4, 2006.

[3] S. Carpendale. Evaluating information visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, eds., *Information Visualization*, pp. 19–45. Springer-Verlag, Berlin, Heidelberg, 2008.

[4] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky. Defining insight for visual analytics. *IEEE Computer Graphics and Applications*, 29(2):14–17, 2009.

[5] C. Chen and Y. Yu. Empirical studies of information visualization. *International Journal of Human-Computer Studies*, 53(5):851–866, Nov. 2000.

[6] K. Cook, G. Grinstein, and M. Whiting. The vast challenge: History, scope, and outcomes. *Information Visualization*, 13(4):301–312, Oct. 2014.

[7] L. Costello, G. Grinstein, C. Plaisant, and J. Scholtz. Advancing user-centered evaluation of visual analytic environments through contests. *Information Visualization*, 8(3):230–238, June 2009.

[8] B. Craft and P. Cairns. Beyond guidelines: What can we learn from the visual information seeking mantra? In *Proceedings of the Ninth International Conference on Information Visualisation*, IV '05, pp. 110–118, 2005.

[9] C. Forsell. Evaluation in information visualization: Heuristic evaluation. In *Proceedings of the 16th International Conference on Information Visualisation (IV)*, pp. 136–142, 2012.

[10] C. Forsell and J. Johansson. An heuristic set for evaluation in information visualization. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, pp. 199–206, 2010.

[11] C. M. Freitas, M. S. Pimenta, and D. L. Scapin. User-centered evaluation of information visualization techniques: Making the hci-infovis connection explicit. In *Handbook of human centric visualization*, pp. 315–336. Springer, 2014.

[12] M. A. Hearst, P. Laskowski, and L. Silva. 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*, CHI '16, pp. 5028–5033, 2016.

[13] S. Hermawati and G. Lawson. Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? *Applied ergonomics*, 56:34–51, 2016.

[14] A. Kobsa. An empirical comparison of three commercial information visualization systems. In *Proceedings of the IEEE Symposium on Information Visualization 2001*, InfoVis '01, pp. 123–, 2001.

[15] 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.

[16] J. Nielsen. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, pp. 373–380, 1992.

[17] J. Nielsen and R. Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, pp. 249–256, 1990.

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

[19] A. Perer and B. Shneiderman. Integrating statistics and visualization: Case studies of gaining clarity during exploratory data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 265–274, 2008.

[20] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, pp. 109–116, 2004.

[21] C. Plaisant, J.-D. Fekete, and G. Grinstein. Promoting insight-based evaluation of visualizations: From contest to benchmark repository. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):120–134, Jan. 2008.

[22] M. Rossi de Oliveira and C. Guimaraes da Silva. Adapting heuristic evaluation to information visualization - a method for defining a heuristic set by heuristic grouping. In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, VISGRAPP '17, pp. 225–232, 2017.

[23] B. Saket, A. Endert, and J. Stasko. Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, pp. 133–142. ACM, 2016.

[24] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456, July 2005.

[25] J. Scholtz. Developing guidelines for assessing visual analytics environments. *Information Visualization*, 10(3):212–231, July 2011.

[26] J. Scholtz, C. Plaisant, M. Whiting, and G. Grinstein. Evaluation of visual analytics environments: The road to the visual analytics science and technology challenge evaluation methodology. *Information Visualization*, 13(4):326–335, 2014.

[27] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, pp. 1–7, 2006.

[28] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, 6th ed., 2016.

[29] B. Sousa Santos, S. Silva, B. Quintino Ferreira, and P. Dias. An exploratory study on the predictive capacity of heuristic evaluation in visualization applications. In *International Conference on Human-Computer Interaction*, pp. 369–383. Springer, 2017.

[30] J. Stasko. Value-driven evaluation of visualizations. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, BELIV '14, pp. 46–53, 2014.

[31] J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human-Computer Studies*, 53(5):663–694, Nov. 2000.

[32] A. Tarrell, A. Fruhling, R. Borgo, C. Forsell, G. Grinstein, and J. Scholtz. Toward visualization-specific heuristic evaluation. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, BELIV '14, pp. 110–117, 2014.

[33] M. Tory and T. Möller. Evaluating visualizations: Do expert reviews work? *IEEE Computer Graphics and Applications*, 25(5):8–11, Sept. 2005.

[34] J. van Wijk. Views on visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):421–432, July 2006.

[35] T. Zuk, L. Schlesier, P. Neumann, M. S. Hancock, and S. Carpendale. Heuristics for information visualization evaluation. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, pp. 1–6, 2006.