Presented at EVIVA-ML workshop at IEEE VIS 2019

Inferential Tasks as a Data-Rich Evaluation Method for Visualization

Dylan Cashman* Tufts University
Yifan Wu† University of California, Berkeley
Remco Chang‡ Tufts University
Alvitta Ottley§ Washington University of St. Louis

ABSTRACT

Designing appropriate tasks in visualization evaluation remains challenging. Current evaluation tasks are often based on fact acquisition - e.g., asking the participants to find the minimum values or the correlation in a scatterplot. As a result, success in completing these tasks may not be indicative of the effectiveness of visualizations, especially those designed to support users in analyzing complex data. In this paper, we propose the use of "inferential tasks" for the evaluation of visualizations and visual analytics systems. Based on the concept of inferential learning, inferential tasks refer to tasks that require a user to draw conclusions not explicitly prompted by relying on their problem-solving and reasoning abilities. We demonstrate the effectiveness of inferential tasks in visualization evaluation through a pair of experiments. The results suggest that, due to the increased complexity of inferential tasks over fact-acquisition tasks, participants tend to perform more interactions with the visualization tool.

1 INTRODUCTION

Empirical evaluation is the underlying basis of the scientific method. When evaluating visualization systems, numerous approaches have been proposed over the years to ascertain the benefits of proposed visualization system (see the surveys by Lam et al. [19] and Carpendale [4]). Central to these approaches is the design of the *"tasks"* that the users would perform using the visualization. Measuring and observing the users' performance on these tasks provides the basis for evaluating the utility and effectiveness of the visualization.

Given that the design of tasks is critical to the success of an evaluation [25], researchers in the visualization community have conducted extensive surveys of the common tasks in visualization. These surveys range from taxonomies of analytic activities [1, 8], user interaction [36], management of insights [6, 26], multi-level tasks and abstractions [3, 29], and tasks in domain-specific applications (e.g. exploration and analysis of graphs [16], time-series [24], volume data [18], high-dimensional data [7] to name a few).

However, although there is a plethora of task taxonomies in the visualization community, the most common tasks used in the evaluation of visualization are still "fact acquisition" tasks. For example, Amar et al. proposed ten low-level analytic activities in visualization that have been frequently adopted as evaluation tasks [1], namely: *Retrieve Value*, *Filter*, *Compute Derived Value*, *Find Extremum*, *Sort*, *Determine Range*, *Characterize Distribution*, *Find Anomalies*, *Cluster*, and *Correlate*. With the exception of *Find Anomalies* and *Cluster*, the remaining eight tasks can be characterized as fact acquisition in that the users are tasked in acquiring a specific piece of information from reading the visualization. In each of these cases, a participant's success in performing the task does not signify the effectiveness of a visualization in helping the user reason about the data. Indeed, in these cases, the completion of the task can be carried out with a database query and does not require human expertise or knowledge about the data.

In this paper, we present an alternative type of evaluation task that is referred to as an *"inferential task"*. Based on the concept of inferential learning from psychology and education literature, where students construct knowledge by *inferencing* relations between learned concepts and new observations, inferential tasks similarly evaluate whether a participant can observe relations in a visualization and apply the observation to other parts of the data. For example, a typical inferential task used in evaluating visualization has the form: "In the visualization, observe that there is a relationship between data items A and B. Where else in the data does the same relationship occur?" In this example, the participant is not provided with what the relationship between A and B is. If such a relationship is provided in the task prompt, the task becomes fact acquisition. Instead, the participant needs to *infer* the possible relationships between A and B by posing a hypothesis and testing it by probing the data via the visualization. In the case that such a relationship does not exist elsewhere in the data, the participant could conclude that their initial hypothesis is incorrect. They would then need to produce another hypothesis in order to complete the task.

Inferential tasks are sometimes used currently in visualization evaluations, and we do not presume to invent either the term or the task type. However, we offer both a formalization and an operationalization that have not existed previously in the literature. We define inferential tasks for visualization as a set of tasks whose solution has at least one free variable, thus requiring the participant to produce hypotheses in the space of free variables. Using this definition, we describe a method for designing inferential tasks that trades off task difficulty, richness of interaction data, and participant conversion rate. We demonstrate the effectiveness of the inferential task in two experiments using two types of visualizations (tabular data analysis using cross-filtering and a visualization of hierarchical data). In both experiments, participants were asked to perform both fact-acquisition tasks and inferential tasks. We show that when performing inferential tasks across three different visualizations, the participants consistently performed between 25%-250% more interactions and spent 35%-87% more time on their task. Together, the quantitative and qualitative data collected from the inferential tasks provide the researchers a richer sense of the performance and effectiveness of their visualization.

2 RELATED WORK AND BACKGROUND

How to evaluate hypotheses related to the usage of visualizations is a vibrant debate in the literature. In a 2004 call-to-arms, Plaisant suggested four categories of visualizations: controlled experiments comparing design elements, usability evaluations, controlled experiments between two tools, and case studies [28]. In 2008, Carpendale published a similar article that outlined the challenges in evaluating visualizations and proposed a number of evaluation methodologies from other domains that could be suitable for the visualization community [4]. In between the publication of these two seminal papers, researchers in the visualization community established the BELIV workshop [21] that focuses on the research of evaluation techniques, which remains an active and vibrant event today.

The call-to-arms by Plaisant and Carpendale and the establishment of BELIV have had a strong influence in the way that visualization papers are published today. In a survey by Isenberg et al. [15],

*e-mail: dcashm01@cs.tufts.edu
†e-mail: yifanwu@berkeley.edu¿
‡e-mail: remco@cs.tufts.edu
§e-mail: alvitta@wustl.edu

the authors found that in 2013, nearly all (97%) of the published papers in the IEEE VIS conference included an evaluation. This is a significant increase over the prior review by Perer and Shneiderman who found that as recently as 2007, only 42% of the papers published in IEEE InfoVis and VAST included an evaluation.

Beyond raising awareness, part of the reason for the increased focus on evaluation is the development of new evaluation methods and their acceptance and adoption. Methods such as insight-based evaluation [26], realization of expert-based feedback as viable metrics for evaluation [34], and the validation of crowdsourcing platforms (such as Amazon's Mechanical Turk) as a viable evaluation methodology [13] all contributed to the rise in evaluation in the visualization community.

However, while there has been tremendous growth in the development of new evaluation methods, there has been limited improvement of the *tasks* used in these evaluations. Notable exceptions are domain-specific tasks, either for specific data types or visualization designs [7, 16, 18, 24, 30], and tasks specific to particular application domains or usage contexts [5, 14, 23, 31]. To the best of our knowledge, our proposed use of inferential task in evaluating visualization is one of the few "complex" tasks that can be adopted generally for a wide range of data types, domains, users, and visualization designs. The closest previous work to our is the work by Lam et. al. which describes the set of tasks a participant must do to accomplish a goal [20] in a design study. The authors categorize tasks found in 20 design studies in the visualization community into single population analyses and multiple population analyses. Our work offers identifies these tasks with the concept of inferential learning, offers a formalization of the construction of these tasks, and presents experiments studying the richness of interaction data gathered from these tasks.

3 Inferential Learning

The principle behind the use of inferential task is rooted in inferential learning, a concept from psychology and education research. Seel suggests three broad categories of learning: procedural/acquisitive learning, experiential learning, and inferential learning [32]. Procedural learning (or acquisitive learning) refers to learning that occurs through repeating actions, turning the task into a procedure [33]. As the name suggests, experiential learning is the learning through experience and the reflection on that experience [17]. In contrast to procedural and experiential learning that involve *doing*, inferential learning is learning that occurs through *thinking*, in particular inductive, deductive, and abductive reasoning [32]. The key point is that a learner must reason on the task at hand, forcing them to build relationships between known and unknown information.

While they have not been a primary subject of study, inferential tasks have previously been used to evaluate visualizations. Green et. al. used inferential tasks in a between-subjects experiment to evaluate a visual analytics system and a web-based list (tabular) view of the same data [9, 10]. The authors found improved performance in participants using the visualization for inferential tasks, but no significant difference for simple tasks, suggesting that the value of a visualization may be more noticeable for inferential tasks. Further, the authors found that the participants' interactions with the visualization when performing the inferential task can be analyzed to predict the participants' personality traits. This finding gives credence to the belief in the richness of the interaction data when using inferential tasks in an evaluation.

However, while there is evidence that inferential tasks can be useful, there does not exist a clear definition of inferential tasks for evaluating visualizations. It is also unclear how these tasks can be created to adjust for task difficulty. In this section, we provide the formalism for inferential tasks and give examples of how these tasks can be constructed. In the following sections, we present two controlled experiments to evaluate the value of inferential tasks in visualization evaluation.

Origin	Date	Airline	Weather	Operation	Deg. of Freedom
Calif.	31st	UA	Rainy	Read Value	0
Calif.	31st	-	Rainy	Find Max	0
Calif.	31st	-	-	Find Max	0
Calif.	31st	-	Rainy	-	1
Calif.	-	UA	Rainy	-	1
Calif.	31st	-	-	-	2
Calif.	30-31	-	-	-	$2-3^*$

Table 1: Examples of tasks on the visualization in Figure 1 with different degrees of freedom. Empty cells denote unspecified parts of a query. When the Operation is unspecified, it can be considered the same as "find something interesting". In the final row, brackets correspond to asking the user to compare between two values across the other two dimensions, which has between 2 and 3 degrees of freedom (see text).

3.1 Defining Inferential Tasks

Using the concept of *degrees of freedom* of a task, we present a unifying framework for considering both fact acquisition and inferential tasks and suggest ways in which these tasks can be constructed that reflect their level of difficulty.

First, we observe that the each user interaction with a visualization can be considered as the generation and the evaluation of a question/hypothesis. For the purpose of our definition, we presuppose that the outcome of testing the question or hypothesis does not need to be a "true" or "false", but can instead be a number, a string, or some fact about the data. With this assumption, a user's interaction with a visualization (such as clicking on a button or filtering by some values) can equate to generating a question or hypothesis (e.g. "is there an abnormally high number of delayed flights originating from California?"). In return, when the user views the resulting visualization, the user is evaluating the question or hypothesis by looking to see if the bar in the barchart that corresponds to delayed flights from California is higher than others.

Using this definition, a fact-acquisition task is one in which a hypothesis is *given* to the participant (in plain words as part of the instruction of an evaluation task). The participant's role is therefore to determine the sequence of interactions to generate that hypothesis within the visualization, and to evaluate the hypothesis by reading from the visualization. In this regard, we say that the "degree of freedom" of this task is zero because the participant does not need to perform any inferential reasoning. In contrast, inferential tasks are tasks whose degrees of freedom is one or above. When performing inferential tasks, the participant would need to consider multiple data items or data dimensions to formulate hypotheses or questions and to evaluate them. An example is illustrated in Figure 1

In the examples given above, any dimension not specified in the task prompt but present in the visualization is an additional degree of freedom. If the visualization featured many different features of the data, each feature could be part of the solution. The participant is required to test out a combinatorially large number of hypotheses, and may become bored or disengaged with the task. It is also likely that there are many valid solutions, which can frustrate the participant and also make it difficult to validate the participant's responses. Care must be taken to limit the number of degrees of freedom of each task implied by the visualization being used.

3.2 Task Operations

Every task on a visualization requires the user to execute an *operation*. The operation could be as simple as *read value*, or it could

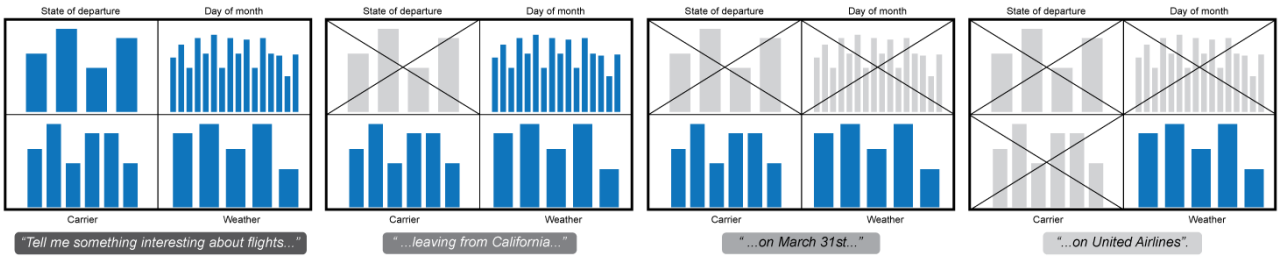

Figure 1: An illustration of the degrees of freedom in a task on a cross-linked set of four bar charts. The data represents the number of flights in four dimensions: *State of departure*, *Day of month*, *Carrier*, *Weather*. The initial task prompt, *Tell me something interesting about flights*, has four degrees of freedom, requiring the user to make complex hypotheses across all four visualized dimensions. By adding qualifiers onto the statement, the number of degrees of freedom is decreased. By the end of the statement, the only degree of freedom is in the *Weather* dimension. An *inferential task* has at least one degree of freedom in its solution. A task with no degrees solution is called a *fact acquisition task*. In this case, such a task would specify all four dimensions and ask the user to read the value corresponding to a single bar.

require the user to *find max*. The choice of operation may have an effect on the degrees of freedom of the task. Consider the cross-filtered visualization depicted in Figure 1. A task, "*Tell me something interesting about flights leaving from California on March 31st on a Rainy day*", leaves one degrees of freedom open (see Row 4 in Table 1). The *operation* of the task, "tell me something interesting", is open-ended.

If, instead, the operation in the previous task is changed to "find the carrier with the greatest number of flights on a Rainy day leaving California on March 31st" (a *Find Extremum* task [1]), the degrees of freedom of the task would be reduced to zero, since the hypothesis is completely described by the task's operation. The user can filter to the correct state, date, and weather condition, and visually ascertain the max in a single atomic interaction (see row 2 in Table 1).

4 Experiments

While the use of inferential tasks in visualization evaluation holds promise and has been shown to be effective [10, 11], previous experiments used a different definition of inferential tasks. In this section, we describe the two experiments we conducted using tasks constructed from our method to better understand the effect of an inferential task on user interaction data.

Our two experiments test two types of visualizations: (1) a traditional cross-coordinated visualization for tabular data and (2) visualizations for exploring hierarchical data. We measure interactions with the visualization, such as mouse clicks or mouse movements, throughout the user's completion of the task. We hypothesize that inferential tasks provide richer interaction data; specifically, we hypothesize that across each visualization, and each dataset, inferential tasks will: **(H1)** have more interactions, and **(H2)** participants will spend more time on inferential tasks.

4.1 Cross-Linked Histograms

In our first experiment, we asked participants a sequence of questions about flights to Hawaii in the month of December. The data was a modified version of a dataset collected by the Bureau of Transportation Statistics consisting of flight delay information in the united states [27], modified so that the questions asked had easy answers (e.g. by injecting additional flights on a certain day). The participant is provided a set of histograms corresponding to four attributes of the flights: *State of origin*, *Day of month*, *Carrier*, and whether the flight was delayed or not. By clicking on a bar or brushing over several bars, the participant is able to filter the data. The data that doesn't match the filter is grayed out after a filter is applied. The interface is based on the visualization used by Liu and Heer [22] used to study latency effects, and can be seen in Figure 2.

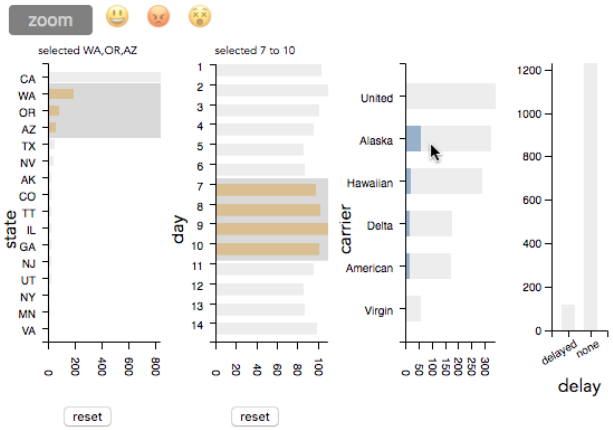

Figure 2: A cross-linked set of histograms used in an experiment comparing user interactions between fact acquisition tasks and inferential tasks. We found that participants spent more time and had more interaction with the visualization on the inferential tasks.

Tasks Participants completed a set of three fact acquisition tasks and three inferential tasks. The fact acquisition tasks asked participants to apply various filters and make judgements about the resulting histograms. Inferential tasks had at least one degree of freedom, as described in Section 3.1. An example of an inferential task given that had three degrees of freedom (State, Delayed, and Carrier) is given:

On the 25th, there is something unusual about the flights from the carrier Alaska. Which carrier had a similar pattern in their flights on the 31st?

Participants 16 participants were gathered via HITs on Amazon's Mechanical Turk. 3 participants exited the survey before they were able to complete the demographics survey. Of the remaining 13 participants, there were 3 women, 9 men, and one declined to answer. The ages ranged from 22 to 42 ($\mu = 29.17$ and $\sigma = 5.89$). Participants were compensated per task completed, a bonus for completing all tasks, and a bonus related to their accuracy on the tasks.

Procedure Participants were first presented with a consent form outlining their compensation, and then were given a tutorial with a training task of each type before being tested on 3 inferential and fact acquisition tasks.

	p	μ_{acq}	μ_{inf}
Interactions	< 0.01	15.93	40.13
Time (s)	< 0.001	93.46	174.56

Table 2: The effect of inferential tasks vs. data acquisition tasks on task time and number of interactions with a cross-filtered histogram visual tool across 10 participants.

4.1.1 Results

10 participants successfully completed all 6 experimental tasks; the remainder were excluded from analysis. For each participant, an average number of interactions (i.e. mouse clicks or brushes on the histogram) was calculated across their three fact acquisition tasks, and a separate average was calculated across their three inferential tasks. The effect was significant for both number of interactions and length of interactions using a two-tailed within-subjects t test. The results are summarized in Table 2.

4.2 Hierarchical Data Exploration

We also investigated the behavioral impact of task type for hierarchical data exploration. We analyzed interaction data from a controlled user study in which participants completed search and inferential tasks on two hierarchical visualizations: an indented tree (V1) and a nested boxes representation (V2). Figure 3 shows the stimuli used in the study. The datasets were taken from the National Center of Biotechnology Information's Genome database and showed the evolutionary relationships between species. The icons, text sizes, and interaction styles were constant for the two views, and participants explored the tree by clicking on nodes to expand the various subtree. Clicking on an already expanded node collapses the subtree.

Tasks Participants completed a control search task and an inferential task for each visualization. The search task was a single fact acquisition task that instructed participants to find a specific species, and read a single value, therefore having no degrees of freedom in the task.

Comparison tasks were used as inferential tasks:

> *Under "Anura," find the classification "Bufo" and note the subclasses it contains. There is another classification under "Mesobatrachia" that has something notable in common with "Bufo." Find that classification.*

Participants 299 participants completed the task via an external link on Amazon's Mechanical Turk. There were 143 women and 155 men (with 1 not answering) in the subject pool with ages ranging from 18 to 65 ($\mu = 31.8$ and $\sigma = 9.60$).

Figure 3: Two hierarchical visualizations used for measuring interactions on fact acquisition tasks and inferential tasks. On the left, a textual visualization where indenting indicates nesting. On the right, a window-style hierarchical visualization. Each participant was given both types of tasks on each visualization, using the same dataset.

	Viz	p	μ_{acq}	μ_{inf}
Clicks	V1	< 0.001	35.85	58.93
Clicks	V2	< 0.005	32.76	40.52
Moves	V1	< 0.001	1852	3051
Moves	V2	< 0.005	2320	2927
Time(s)	V1	< 0.001	116.5	193.0
Time(s)	V2	< 0.001	146.0	202.1

Table 3: Results of an experiment measuring the effect of inferential tasks vs. data acquisition tasks on task time and number of interactions with two hierarchical visualizations seen in Figure 3 over 299 participants.

Procedure The study consisted of two sessions, one for each visualization. For each session, participants completed a search task and an inferential task. The order of the tasks were counterbalanced to prevent ordering effects. Once the four tasks were done, they then completed a brief demographic survey.

Data Collection and Cleaning During the study, we captured every mouse click and mouse move event. We recorded the time of the event, the data element, and coordinates of the cursor. All participants who answered each question were included in the analysis; no effort was made to remove participants that clicked fewer than 5 items, for example. This was intentional, as we don't want to remove the effect of a participant being frustrated with one type of task (for example, inferential) over another (fact acquisition).

4.2.1 Results

For each visualization, we tested a difference of means with a within-subjects two-tailed t-test for the amount of time spent on the task, the number of clicks, and the number of mouse moves. All six cases were significant differences, with exact values given in Table 3.

4.3 Analysis

The effect of inferential tasks vs. fact acquisition tasks was tested across three different visualizations (cross-linked histograms, textual indented tree, nested boxes) and two different scenarios (cross-filtering aggregations, hierarchical data exploration). In all cases, the difference in number of interactions and amount of time spent on the task was large and significant. In the context of our two hypotheses, we confirm the hypothesis **(H1)** that participants perform more actions in inferential tasks than in data acquisition tasks. This is evident from the fact that participants performed between 25% and 250% more interactions in both experiments across all three visualizations. Likewise, we also confirm the hypothesis **(H2)** that participants spend more time on inferential tasks. In the two experiments, this difference ranged from 35% to 87% more time.

5 Discussion and Conclusion

In this work, we offer a definition for inferential tasks that can be used to construct such tasks for visualization evaluation. But there are many other considerations for implementing such tasks. The difficulty of tasks must be considered; leaving open degrees of freedom in the task means that the task is more open-ended, more difficult, and can result in frustration for the experimental participant. Computational models of task difficulty could be built based on data size and schema, and they could even be used to automatically generate training tasks for a new dataset. The relationship between cardinality, dimensionality, and degrees of freedom closely mirrors other combinatorial notions of data that have been very impactful in both information visualization and machine learning [2, 12].

In a recent work in the Journal of Experimental Psychology, Westfall et. al. identified that when a mismatch between the specificity of experimental stimuli and the lack of specificity in the conclusion of an experiment, our statistical tests no longer have validity [35].

In a visualization experiment, this would correspond to testing the participant on an easy task like fact acquisition, and then generalizing the efficacy of the visualization onto more difficult tasks. In our work, we propose using inferential tasks in experiments using visualizations, as the interactions that a participant makes with the visualization better match those of the complex tasks which visualizations are used for in the real world. We performed two experiments to validate the effectiveness of the use of inferential task. The results confirm our claim that the increased complexity of the inferential tasks (when compared to fact acquisition tasks) perform more interactions over longer task completion times. When used within the appropriate context of an experimental design, inferential tasks can be improve the community's ability to evaluate visualizations and visual analytics systems.

ACKNOWLEDGMENTS

Support for the research is partially provided by DARPA FA8750-17-2-0107 and NSF CAREER IIS-1452977. 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] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pp. 111–117. IEEE, 2005.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[3] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[4] S. Carpendale. Evaluating information visualizations. In *Information visualization*, pp. 19–45. Springer, 2008.

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

[6] Y. Chen, J. Yang, and W. Ribarsky. Toward effective insight management in visual analytics systems. In *Visualization Symposium, 2009. PacificVis' 09. IEEE Pacific*, pp. 49–56. IEEE, 2009.

[7] R. Etemadpour, L. Linsen, C. Crick, and A. Forbes. A user-centric taxonomy for multidimensional data projection tasks. In *IVAPP*, pp. 51–62, 2015.

[8] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009.

[9] T. M. Green and B. Fisher. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pp. 203–210. IEEE, 2010.

[10] T. M. Green, D. H. Jeong, and B. Fisher. Using personality factors to predict interface learning performance. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pp. 1–10. IEEE, 2010.

[11] T. M. Green, W. Ribarsky, and B. Fisher. Visual analytics for complex concepts using a human cognition model. In *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, pp. 91–98. IEEE, 2008.

[12] P. Hanrahan. Vizql: a language for query, analysis and visualization. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pp. 721–721. ACM, 2006.

[13] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 203–212. ACM, 2010.

[14] P. Isenberg, T. Zuk, C. Collins, and S. Carpendale. Grounded evaluation of information visualizations. In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, p. 6. ACM, 2008.

[15] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013.

[16] N. Kerracher, J. Kennedy, and K. Chalmers. A task taxonomy for temporal graph visualisation. *IEEE transactions on visualization and computer graphics*, 21(10):1160–1172, 2015.

[17] D. Kolb. *Experiential learning: experience as the source of learning and development*. Prentice Hall, Englewood Cliffs, NJ, 1984.

[18] B. Laha, D. A. Bowman, D. H. Laidlaw, and J. J. Socha. A classification of user tasks in visual analysis of volume data. In *Scientific Visualization Conference (SciVis), 2015 IEEE*, pp. 1–8. IEEE, 2015.

[19] 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, 2012.

[20] H. Lam, M. Tory, and T. Munzner. Bridging from goals to tasks with design study analysis reports. *IEEE transactions on visualization and computer graphics*, 24(1):435–445, 2018.

[21] B. Lee, C. Plaisant, C. S. Parr, J.-D. Fekete, and N. Henry. Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pp. 1–5. ACM, 2006.

[22] Z. Liu and J. Heer. The effects of interactive latency on exploratory visual analysis. *IEEE Transactions on Visualization & Computer Graphics*, (1):1–1, 2014.

[23] D. Lloyd and J. Dykes. Human-centered approaches in geovisualization design: Investigating multiple methods through a long-term case study. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2498–2507, 2011.

[24] S. Miksch and W. Aigner. A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics*, 38:286–290, 2014.

[25] T. Munzner. A nested process model for visualization design and validation. *IEEE Transactions on Visualization & Computer Graphics*, (6):921–928, 2009.

[26] C. North. Toward measuring visualization insight. *IEEE computer graphics and applications*, 26(3):6–9, 2006.

[27] B. of Transportation Statistics. On-time performance. http://www.transtats.bts.gov/Fields.asp?Table ID=236. Accessed: 2018-03-29.

[28] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, pp. 109–116. ACM, 2004.

[29] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE transactions on visualization and computer graphics*, 20(12):1604–1613, 2014.

[30] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization & Computer Graphics*, (1):402–412, 2018.

[31] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization & Computer Graphics*, (12):2431–2440, 2012.

[32] N. M. Seel. Inferential learning and reasoning. In *Encyclopedia of the Sciences of Learning*, pp. 1550–1555. Springer, 2012.

[33] L. R. Squire and S. Zola-Morgan. Memory: brain systems and behavior. *Trends in Neurosciences*, 11(4):170 – 175, 1988. doi: 10.1016/0166-2236(88)90144-0

[34] M. Tory and T. Moller. Evaluating visualizations: do expert reviews work? *IEEE computer graphics and applications*, 25(5):8–11, 2005.

[35] J. Westfall, D. A. Kenny, and C. M. Judd. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5):2020, 2014.

[36] J. S. Yi, Y. ah Kang, J. T. Stasko, J. A. Jacko, et al. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization & Computer Graphics*, (6), 2007.