

Narratives in Crowdsourced Evaluation of Visualizations: A Double-Edged Sword?

Evanthia Dimara, Anastasia Bezerianos, Pierre Dragicevic

▶ **To cite this version:**

Evanthia Dimara, Anastasia Bezerianos, Pierre Dragicevic. Narratives in Crowdsourced Evaluation of Visualizations: A Double-Edged Sword?. ACM Conference on Human Factors in Computing Systems (CHI) , May 2017, Denver, United States. 10.1145/3025453.3025870 . hal-01448095

HAL Id: hal-01448095
https://inria.hal.science/hal-01448095

Submitted on 31 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Narratives in Crowdsourced Evaluation of Visualizations: A Double-Edged Sword?

Evanthia Dimara
Inria, France
evanthia.dimara@gmail.com

Anastasia Bezerianos
Univ. Paris-Sud, CNRS, Inria, France
anastasia.bezerianos@lri.fr

Pierre Dragicevic
Inria, France
pierre.dragicevic@inria.fr

ABSTRACT

We explore the effects of providing task context when evaluating visualization tools using crowdsourcing. We gave crowdworkers *i)* abstract information visualization tasks without any context, *ii)* tasks where we added semantics to the dataset, and *iii)* tasks with two types of backstory narratives: an analytic narrative and a decision-making narrative. Contrary to our expectations, we did not find evidence that adding data semantics increases accuracy, and further found that our backstory narratives can even decrease accuracy. Adding dataset semantics can however increase attention and provide subjective benefits in terms of confidence, perceived easiness, task enjoyability and perceived usefulness of the visualization. Nevertheless, our backstory narratives did not appear to provide additional subjective benefits. These preliminary findings suggest that narratives may have complex and unanticipated effects, calling for more studies in this area.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI)

Author Keywords

crowdsourcing; evaluation; instructions; narrative; information visualization; decision making;

INTRODUCTION

Crowdsourcing platforms are a promising way of accessing a large and diverse pool of participants, allowing rapid evaluations of visualizations [21, 31, 37, 8, 11]. However, engaging crowdworkers and obtaining high quality responses can be challenging [28, 16]. In particular, task instructions in a remote study where the instructor has no way of helping or motivating the participants, should be designed with extra care. We thus need to better understand how task instructions affect the quality of responses in the evaluation of visualization tools.

There is evidence that humans can more easily make sense of the world through narratives, i.e., coherent sequences of events [19, 29, 20]. Researchers and practitioners have already started to use narratives in the context of data analysis and communication, in order to improve data understanding and engagement of the users [25, 45]. But narratives can also be used during visualization evaluation, in the form of a backstory in the task instructions, to help simulate the real use of a system and elicit a more representative user behavior. For example, Aseniero et al. [3] instructed participants to imagine taking the role of a project manager identifying an optimal plan in order to evaluate a system designed for software release plans. Narratives could also motivate participants by giving meaning to an otherwise abstract experimental task.

Nevertheless, the effects of adding narrative elements to task instructions are unclear, in particular in crowdsourcing settings where incentives vary across people, and attention and motivation are hard to control for. For example, what would be the difference between *i)* instructing participants to identify the data point with the minimum X value, and *ii)* instructing them to imagine they are trying to find the cheapest available house? Both versions are equivalent at the task level and consist of finding an extremum on a particular data dimension. The second version is possibly more salient and engaging, and with a context that is easy to understand, characteristics linked to good crowdsourcing performance [28, 36]. At the same time, the first version is more succinct and less demanding in terms of time and patience, aspects that have also been emphasized in crowdsourcing guidelines [16].

In this work, we investigate whether it is possible to get crowdsourcing participants to perform better in basic visualization tasks by enhancing task instructions with narrative components that provide a backstory. More specifically, we:

- investigate, in crowdsourced evaluation settings, the effect of moving from *i)* abstract task instructions that provide no contextual information for the dataset, to *ii)* adding minimum semantics to the dataset, and to further *iii)* adding a backstory narrative that justifies the purpose of the task.
- compare two popular types of narratives from the visualization literature: analytic narratives involving answering investigative questions about data, and decision making narratives involving making personal choices based on data.

We confirmed that adding data semantics can provide subjective benefits. However, we found no evidence that it increases accuracy, and even found that our backstory narratives could hurt accuracy. These findings may not extend to all cases, but they suggest that the effects of narratives in crowdsourced visualization evaluation need to be better understood.

Author's version, to appear in CHI 2017
CHI 2017, May 6-11, 2017, Denver, CO, USA.
http://dx.doi.org/10.1145/3025453.3025870

BACKGROUND

We next discuss studies on question wording, the use of narratives in information visualization, and provide motivations for using narratives in crowdsourced evaluations of visualizations.

Question Wording

Psychologists have long been interested in the effects of question wording. Some of the work in this area has focused on how question framing can affect reasoning and judgment, for example in terms of causal attribution [49]. Other work has focused on how to best design surveys to get reliable responses. Past work suggests that conciseness and context are both desirable [38], but it remains unclear how to strike the right balance between the two. In addition, guidelines for survey design may not directly translate to information visualization evaluation. For example, issues like desirability bias (i.e., respondents trying to give socially acceptable answers) are key in survey design [30] but likely less relevant to visualization evaluation.

Visualization Narratives

A currently popular line of research in information visualization suggests that complementing interactive visualizations with stories about data (visualization narratives) can turn data exploration into a more engaging and educational experience [27, 45, 46, 25, 47]. For example, narratives can be used for explaining changes in complex temporal networks [4], or for promoting user engagement during data exploration [10]. Two particularly compelling application areas are journalism [9] and science communication [35]. A number of tools have been proposed to help authors design visualization narratives and interleave textual stories with visual elements [45, 34, 26, 22].

Our work significantly differs from visualization narratives in that we explore the use of narratives during the *evaluation* of visualizations, not during their actual use. Thus our "end users" are study participants, not data consumers. Our narratives invite users to put themselves in a hypothetical situation (e.g., being a real-estate analyst, or a house buyer), which is not the case in typical visualization narratives. Finally, although the narratives we explore provide context about the datasets, they make no reference to trends and patterns in the data itself.

Illustrative Use Cases

Visualization designers and researchers have long used narratives in the form of illustrative use cases in order to convey a tool's functionalities in a way that is more accessible and persuasive than a factual description. In the research literature, typical narratives involve an expert who seeks to understand a dataset within a domain like cyber security [17] or business priority analysis [12]. For example, an ocean forecaster may want to analyze the Red Sea dataset for glider path-planning [23]. Alternatively, laymen can be involved, such as a person who seeks to grasp their nutrition habits [18]. Another common type of narrative involves decision making scenarios, such as a person seeking a house to buy [50], a prospective student choosing a university [12, 18], or a company executive choosing the location of a new factory branch [2].

Compared to the narratatives we study, illustrative use cases have in common the hypothetical situation component, but target different end users (article readers) and significantly differ in content (fictional data exploration activities). However, our study draws from the two types of narratives used in this context: *analytic narratives* and *decision-making narratives*.

Narratives in Visualization Evaluation

Narratives are sometimes used in information visualization evaluation. Minimalist forms of narratives that solely consist in attributing a meaning to the datasets, are the most common. For example, to evaluate HomeFinder, Williamson et al. [50] used questions such as "what neighborhood has the most expensive houses?". Narratives occasionally get more elaborate and can take the form of decision making tasks. For example, Yi et al. [51] asked their participants to choose a cereal brand, using the same narrative as in their illustrative case study.

Full narratives are commonly used in evaluations of domain-specific or decision-support visualization systems. For example, in order to evaluate a tool for software release plans, Aseniero et al. [3] asked participants to take the role of a project manager and choose the optimal plan. In order to evaluate a tool for preferential choices, Bautista and Carenini [5], immersed participants in shopping scenarios involving television sets, houses or cell phones, and put them in a situation of finding a hotel to stay in Vancouver. Similarly, Daradkeh et al. [13] asked participants to make hypothetical investments.

Although information visualization researchers sometimes use narratives in their evaluations, we are not aware of any study that has established their effectiveness, both in lab settings and in crowdsourcing settings.

Why Use a Narrative when Evaluating a Visualization?

One reason is that narratives presumably help simulate a "natural context" [49] and thus, a more representative use of the system, which is especially important when evaluating domain-specific and decision-support visualization systems as seen before. For example, if we want to carry out a crowdsourced evaluation of a system meant to help customers choose a car, plain instructions such as "select the best car" may not put crowdworkers in the right frame of mind. Therefore, it may seem more suitable to use a decision-making question framing and provide a narrative context that could help participants simulate a hypothetical purchase situation, or recall a similar situation from the past.

Another reason is that a narrative can possibly provide benefits in terms of enhanced motivation, attention and engagement, even if the evaluation's aim is to investigate how a generic visualization tool supports basic analytic tasks. These benefits could possibly translate into improved task comprehension and higher-quality responses. Improving the quality of responses is especially important in crowdsourced studies. Although crowdourcing is now widely accepted as an evaluation platform [31, 21, 42], the overall quality of responses can be low, which either leaves investigators with poor data to analyze or forces them to reject a large proportion of responses [28].

A number of strategies have been suggested to improve the quality of responses in crowdsourced studies. A common approach consists of only recruiting contributors with high reputation, possibly subjecting them to qualification tests [21],

and using verification questions to detect lack of diligence [40, 31, 42, 21]. Optimal payment strategies have also been explored [24, 36, 6], but studies suggest that higher monetary rewards increase the quantity but not the quality of responses [36, 6]. Other recommendations include using short task durations [31, 16] while avoiding breaking down tasks into meaningless chunks [36, 7]; paying attention to experiment design [32]; and providing sufficiently challenging, personalized and easy to understand tasks [28]. Even though much of previous work has emphasized the importance of providing clear, meaningful and engaging tasks, to our knowledge there is no study investigating whether the use of task narratives in visualization evaluation can yield measurable benefits.

EXPERIMENT

Our goal was to explore the effect of adding task context, in particular in the form of narratives, in a crowdsourced visualization evaluation. To identify what effects stem from adding minimal context as opposed to more complex backstories, we compared: *i)* providing no context whatsoever about the data, *ii)* providing minimal semantic context on the data (e.g., referring to houses rather than abstract data points), and *iii)* adding backstory narratives that also justify the purpose of the task. We used two types of narratives, as well as an additional control condition that will be explained later on.

Participants were assigned to one of the five context conditions and performed three basic visualization tasks using scatterplot visualizations. To assess the merits of the different context conditions, we used objective performance metrics that measured participants' ability to perform and understand the tasks, as well as subjective metrics based on self-reported impressions.

Dataset and Visualization

Our study involved simple datasets with two quantitative dimensions. The datasets were small-sized (21 data points each) artificial datasets created manually using spreadsheet software.

For the experimental stimuli we used a 2D scatterplot visualization, as it is a standard information visualization technique for presenting multiple data points along two dimensions [39]. The scatterplots supported basic interactions that depended on the task and will be described in the next subsection.

Tasks

We used three basic visualization tasks adapted from taxonomies of low-level information retrieval tasks [1, 48, 43]:

- An **Extremum** task (**Ext**), where participants had to find the data point with highest value according to the X dimension (see the leftmost scatterplot in Figure 1).
- A **Correlation** task (**Cor**), where participants had to find the scatterplot with the highest correlation among four different ones (see the second panel in Figure 1).
- A **Comparison** task (**Com**), where participants had to compare data points across their two dimensions simultaneously (see the third panel in Figure 1). The task consisted of finding a data point without any "competitor", a competitor being defined as a data point that has both larger X and smaller Y. The task had four possible correct answers.

Condition	Task	Page 1	Page 2
ABS	Ext	You will be asked to answer a few questions about data. In the next page you will see many data points displayed in a diagram.	Which is the data point with the largest value of X?
	Cor	Now you will see four diagrams with data points. You will be asked to compare them.	In which of these four diagrams is Y most related to X?
	Com	Now you will see one of the previous diagrams again. You will be asked a question that requires identifying "competitors". In our case, a data point is a competitor of another data point if it has both larger X and smaller Y.	Select a data point that has no competitor.
SEM	Ext	You will be asked to answer a few questions about houses. In the next page you will see many houses displayed in a diagram.	Which is the biggest house?
	Cor	Now you will see four diagrams with houses. You will be asked to compare them.	In which of these four diagrams is price most related to size?
	Com	Now you will see one of the previous diagrams again. You will be asked a question that requires identifying "competitors". A house is a competitor of another house if it is both bigger and cheaper.	Select a house that has no competitor.
AN-NAR	Ext	You will be asked to answer a few questions about houses. Imagine that you are a real estate analyst and you need to understand the house market. You focus on extremely rich customers who seek to buy a house that is as big as possible. In the next page you will see the houses currently on the market, displayed in a diagram.	Given what you read, which house would be the most attractive to your customers?
	Cor	Now you want to focus on regular customers who are not necessarily very rich. You want to investigate how reliable some real estate agencies are. You will see four diagrams with houses. Each diagram shows the houses proposed by a different agency. An agency that sets arbitrary prices is NOT reliable. While in a reliable agency, price is very related to size.	Given what you read, which of these four real estate agencies is the most reliable?
	Com	Now you will see one of the previous diagrams again. It shows the houses offered by the best agency. You need to report on their best deals. A good deal is a house that has no "competitor". A house is a competitor of another house if it is both bigger and cheaper.	Given what you read, select a house that is a good deal.
DM-NAR	Ext	You will be asked to make a few decisions about houses. Imagine you are moving to a new city and you need to buy a house. You are extremely rich and you want your house to be as big as possible. In the next page you will see the houses currently on the market, displayed in a diagram.	Given what you read, which house would you buy?
	Cor	You don't have as much money as you initially thought. So before buying a house, you need to find a reliable real estate agency. You will see four diagrams with houses. Each diagram shows the houses proposed by a different agency. An agency that sets arbitrary prices is NOT reliable. While in a reliable agency, price is very related to size.	Given what you read, which of these four real estate agencies would you choose?
	Com	Now you will see one of the previous diagrams again. It shows the houses offered by the best agency. You will finally get to choose your house. A good choice is a house that has no "competitor". A house is a competitor of another house if it is both bigger and cheaper.	Given what you read, which house would you buy?

Table 1. The instructions text in each experiment condition. The condition O-NAR **has identical page 1 with** AN-NAR **and identical page 2 with** SEM **condition.**

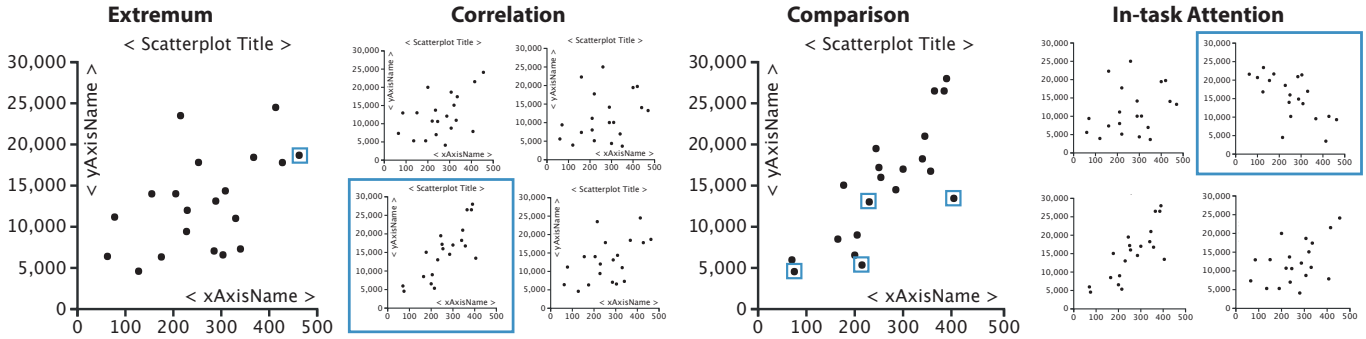


Figure 1. Stimuli used in each task (Ext, Cor and Com), and in the in-task attention test. Correct answers are annotated in blue. Axes were labeled (*X*, *Y*) for ABS, and (*size* m^2, *price* (\$)) in all other context conditions. The title was *Diagram Z* : *Datapoints* in ABS, and was *Diagram Z* : *Houses* in SEM (all tasks) and DM-NAR (Ext, Cor tasks). In all other conditions the title was *Agency Z* : *Houses*. *Z* was an integer (1, 2, 3, or 4) identifying the scatterplot.

As said before, the scatterplots supported basic interactions. In the Ext and Com tasks, hovering over a data point highlights it in light gray, displays horizontal and vertical projection lines, and overlays the data point's *X* and *Y* values on the axes. Participants gave their answer by clicking on a data point, after which its color changed to green. For the Cor task, scatterplots were highlighted in light gray when hovered. Participants selected their answer by clicking on one of the four plots, after which the selected plot changed to green. In all tasks, participants could either confirm their choice by clicking on a "next" button, or change their selection.

Context conditions

As context for our datasets, we decided to use scenarios involving the real estate market. Our choice is based on both the nature of the house price / house size tradeoff that is easy to understand, and its use in previous evaluations of both analytical and decision making visualization systems [50, 5, 14].

For each task, instructions were split in two pages on the Web form: **page 1** displayed introductory and background information relevant to the task; and **page 2** showed the task question and the visualization. Participants were allowed to navigate back-and-forth between the two pages.

Each task came in five different variants, one per context condition. Scatterplots were identical or had minor label differences (see caption of Figure 1), while the major differences were in the text instructions on page 1 and page 2. Table 1 shows the complete text instructions for all of the context conditions except O-NAR, covered later on. The study employed:

- An **Abstract** (ABS) condition, where the dataset has no specific meaning. Both page 1 and page 2 use abstract wordings. An example for a question is *"Which is the data point with the largest value of X?"*.
- A **Simple Semantics** (SEM) condition, where the data points were houses and dimensions were price and size. Questions were of the type *"Which is the biggest house?"*.
- An **Analytic narrative** (AN-NAR) condition, where page 1 contains a narrative that asks participants to put themselves in the situation of a real estate analyst and to find answers to analytical questions. An example of question on page 2 would be *"Given what you read, which house would be the most attractive to your customers?"*.
- A **Decision making narrative** (DM-NAR), where page 1 contains a narrative that asks participants to put themselves in the situation of a house buyer and, given some criteria and constraints, to make choices. Questions were of the type *"Given what you read, which house would you buy?"*.

In the two narrative conditions AN-NAR and DM-NAR, participants had to read the narrative on page 1 to be able to interpret the question on page 2. Thus, participants who do not read the text on page 1 carefully enough will see their performance negatively impacted. In the SEM condition, in contrast, the question is self-contained for tasks Ext and Cor (but not Com, see Table 1). Because the narrative conditions differ from SEM in two respects (the presence of a narrative, and the necessity to read page 1 to be able to answer any question), we introduced a fifth, intermediary condition:

- **Optional Narrative** (O-NAR) condition, a hybrid control condition where page 1 is identical to AN-NAR and page 2 is identical to SEM. An example of a question would be *"Which is the biggest house?"*. Thus, despite the presence of a narrative on page 1, participants did not need to read it to answer the task question.

In order to rule out poorly framed narratives, we tested and refined them through crowdsourced and in-person pilot studies.

Objective Performance Metrics

In this section we describe how we measured performance. All metrics were devised before data was collected.

Accuracy

For all tasks we used a normalized measure of *accuracy* ranging from 0 to 1. We preferred quantitative to binary metrics because of their higher statistical power. We assigned 1 to correct answers (Figure 1, in blue). For other answers, we gave a score depending to how close they are to the right answer.

In the Ext task, where participants needed to find the data point with the largest X, each of the 21 data points got a score of $S = (\frac{X - X_{min}}{X_{max} - X_{min}})^2$, where X is x-coordinate of the chosen data point, X_{min} is the minimum x-coordinate of the plot, and X_{max} is the x-coordinate of the correct answer.

In the Cor task, where participants needed to identify the highest correlation, we assigned a score of $S = \frac{C - C_{min}}{C_{max} - C_{min}}$, where

C stands for the correlation of the selected plot, C_{min} is the lowest correlation and C_{max} if the correlation of the correct plot. This time we did not square the score because incorrect correlations were much lower than the correct one.

In the Com task, where users needed to identify a data point without competitors (i.e., a non-dominated point), we assigned a score of $S = (\frac{D_{max}-D}{D_{max}-D_{min}})^2$, where D stands for the number of points that dominate the selected point, D_{max} is the maximum number of points that dominate any point in the dataset, and D_{min} is the minimum number (zero in our case).

In-task Attention

Since a lack of diligence or a poor understanding from participants may not always translate into incorrect responses, we used attention as a secondary measure of performance. As a proxy for in-task attention, we measured participants' ability to recall the options presented to them in the correlation task (see the rightmost panel in Figure 1). The test was administered after all tasks were completed. We asked participants to identify which plot was not presented to them before. As we can see in Figure 1, the correct answer has a negative correlation, whereas all options presented previously had a positive correlation. The in-task attention measure is likely linked to other factors such as task comprehension.

Since all incorrect answers were about equally wrong, for the in-task attention metric we assigned a binary score of 1 for the correct answer and 0 for all other answers.

Post-Task Attention

Because researchers may want to conduct longer experiments than ours and because narratives may yield participant fatigue (or alternatively, abstract tasks could cause a loss of interest), we also measured participant's attention after the tasks. We administered at the end of the experiment an independent instructional manipulation check where participants needed to read instructions very carefully to get a correct answer [41]. As before, we assigned a binary score of 1 for the correct answer and 0 for all other answers.

Metrics not Considered

We did not consider task completion time as a metric, as it would be difficult to interpret in the context of our study. This is because depending on the context condition, longer task completion times could be either an indication of lower performance (e.g., in the ABS condition) or an indication of higher motivation and engagement (e.g, in the DM-NAR condition).

Subjective Metrics

We used subjective metrics as a complement to the previous metrics. All responses were reported on a 7-point Likert item.

- *Confidence:* After each task, we asked participants to report their confidence in their answer.
- *Easiness:* We also asked them to rate the perceived difficulty of each task. Since we wanted all scores to reflect a positive direction, we referred to easiness rather than difficulty.
- *Enjoyability:* After all tasks were completed, we asked participants to report how much they enjoyed the job overall.
- *Usefulness:* We asked participants to report to what extent they thought the diagrams would be useful if they wanted to buy a product. The goal was to examine if a richer context makes tasks more meaningful and change participants' perspective on the utility of the visualization tested.

Experiment Design

The experiment followed a mixed design. The independent between-subjects variable was context (ABS, SEM, O-NAR, AN-NAR, DM-NAR). The independent within-subjects variable was the task (Ext, Cor, Com).

Each participant performed all three tasks in the same order: Ext, Cor and finally Com, accounting for what we thought was an increasing level of difficulty. Since each participant was assigned to a unique context condition, they each saw the three tasks with the same type of context provided.

Procedure

We ran the experiment as a Crowdflower job[1]. Participants opened an external 12-page Web form. They first performed the Ext task, consisting of two pages as previously mentioned. They selected their answer as described previously. On the following page, they rated their confidence and task easiness. They followed the same process for the Cor and Com tasks. Once they finished the 3 tasks, participants rated the enjoyability of the job, the usefulness of the diagram, and were given the instructional manipulation check on the same page. On the next page, they were given the in-task attention test. On the last page, they provided basic demographic information, and were finally given a completion code to paste in crowdflower to complete their job. Participants spent on average 7 minutes on the job and were given a reward of 60 US cents.

Crowdsourcing Quality Control

Although a common crowdsourcing practice is to reject jobs from participants whose performance is abnormally poor or who failed attention tests, we accepted and analyzed all jobs[2]. The reasons are twofold: *i)* since different conditions are expected to yield different levels of attention and performance, excluding low-quality jobs would bias our results; *ii)* we seek to improve the overall quality of *all* submitted jobs, with the hope that less jobs will need to be rejected in the future.

Participants

Our total sample consisted of 405 highly rated (level 3) Crowdflower contributors. Sample size per condition ranged from n=80 to n=83 (for a planned sample size of n=80). Figure 2 summarizes participants' self-reported demographics.

Research questions

Prior to data collection we framed four research questions and hypotheses. Since our hypotheses were not derived from a theory, we refer to them as "expectations" [15].

Q1 *Does adding minimal semantics help?* Assuming we find an effect of narrative, the purpose was to determine how much of the effect is simply due to the fact that the narrative assigns a meaning to the dataset and its dimensions. We expected benefits when adding minimal semantics alone.

[1] https://www.crowdflower.com

[2] Three jobs however had to be rejected because their duration went over the 30-min limit imposed by the crowdsourcing platform.

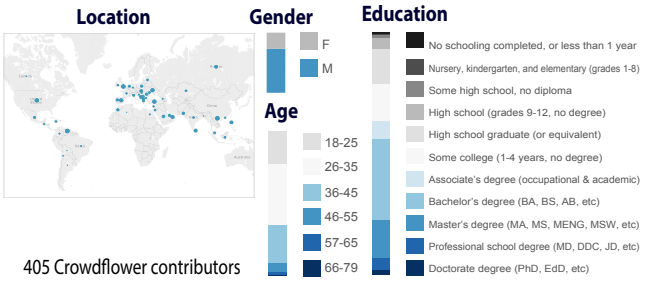


Figure 2. Self-reported demographics of our participants.

Q2 *Does adding a narrative (on top of minimal semantics) help?* This was our main research question. We expected that the benefits of narratives (e.g., higher engagement) would outweight their costs (e.g, higher attention demands).

Q3 *Should the question refer to the narrative?* The purpose was to better understand the reason behind any effect of narrative we may find. For example, if narratives happen to yield poorer performance but the control condition O-NAR does not, it could mean that the problem comes from participants not reading the narratives. In addition, if O-NAR alone yields improvements, it could mean that task-irrelevant narratives are sufficient to motivate participants.

Q4 *Is a decision making narrative better than an analytical narrative?* The purpose of this question was to better understand which type of narrative yields the most benefits. Although we did not expect large differences for the Ext and Cor tasks, we expected that DM-NAR would outperform AN-NAR for the Com task, since this task involves mental operations (dominance recognition) typical of everyday decision making tasks.

Overview of Results

We analyze, report and interpret all our inferential statistics using interval estimation [15]. Experimental stimuli, data and analyses are available at `http://www.aviz.fr/narratives`.

Before we turn to our main research questions, we first give an overview of all our results. We report the sample mean for each condition according to our objective performance metrics (accuracy, in-task attention and post-task attention) and our subjective metrics (confidence, perceived easiness, overall job enjoyability and perceived usefulness of the visualization).

In addition to sample means, we report 95% confidence intervals (CIs) indicating the range of plausible values for the population mean [15]. See Figure 3 for help on how to interpret overlaps in CIs. For in-task and post-task attention, we use Wilson's confidence intervals for a single proportion. For all other metrics, we use BCa bootstrap confidence intervals.

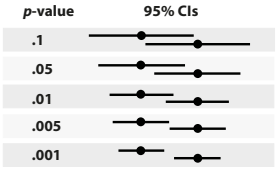


Figure 3. Indicative chart showing the correspondence between degree of CI overlap and *p*-values for independent samples (after [33]).

Accuracy

Mean accuracy scores are shown in Figure 4, with tasks on columns and conditions on rows. The first column shows scores averaged across all three tasks. As we can see on this column, crowdsourced participants were fairly accurate overall (scores of 0.7–0.8 out of 1). However, it appears that participants who were given the DM-NAR narrative performed less accurately on average than those who were only given minimal context (SEM) or no context at all (ABS). The other narrative AN-NAR may have also performed worse than SEM, but the evidence is much weaker.

Regarding the extremum (Ext) task, both narrative conditions AN-NAR and DM-NAR appear less accurate on average than all other conditions. For the correlation (Cor) task, AN-NAR appears worse than SEM, while DM-NAR and O-NAR may also be worse than SEM, but the evidence is weaker. For the comparison (Com) task, DM-NAR is clearly worse than AN-NAR. For this task, AN-NAR appears to outperform ABS.

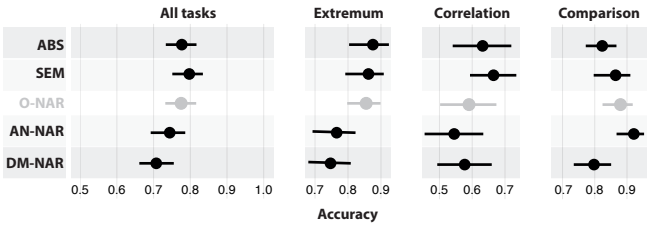


Figure 4. Accuracy per task and condition. Error bars are 95% CIs.

In-task Attention

As we can see in Figure 5, participants exhibited a better recall of the correlation task when given minimal semantics (SEM) than no context (ABS), suggesting they were paying more attention. However, adding a narrative (AN-NAR or DM-NAR) to the semantics decreased their recall. The decrease is less evident but possible for the control condition O-NAR where the narrative was not required to perform the task.

Post-task Attention

As we can see in Figure 5, the results are mostly inconclusive regarding post-task attention. There is however some weak evidence that adding a narrative when it is not needed (O-NAR) may make people less attentive after they perform the task compared to providing only minimal semantic context (SEM).

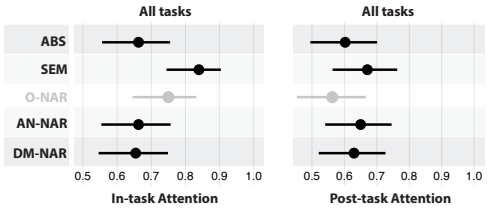


Figure 5. In-task and post-task attention. Error bars are 95% CIs.

Confidence

Figure 6 reports confidence scores normalized between 0 and 1. Confidence was overall high (0.7–0.8), but participants were on average less confident when provided no context (ABS). We

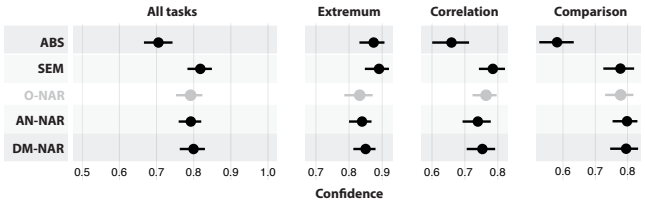


Figure 6. Reported confidence scores. Error bars are 95% CIs.

did not observe this differences for the Ext task, but it is clear for Cor and remarkably large for Com, with the remaining conditions yielding comparable confidence scores.

Easiness

Figure 7 provides some evidence that without context (ABS) the tasks appear harder overall. Although there is no visible difference for the Ext task, for the Com task the difference is clear. There is also some evidence that participants found the control condition O-NAR a bit harder overall, especially for the Ext and Cor tasks. Finally, for the Ext task, the use of a DM-NAR narrative may have made the task appear easier compared to the use of a AN-NAR narrative.

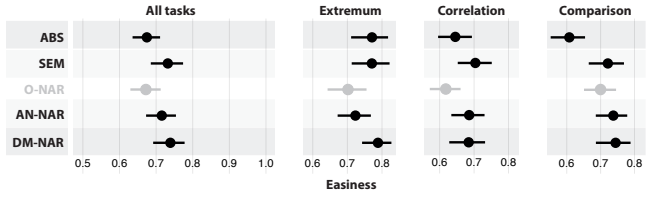


Figure 7. Reported easiness scores. Error bars are 95% CIs.

Enjoyability and Usefulness

Figure 8 provides good evidence that when no context is provided on the visualization tasks (ABS), participants find the overall job less enjoyable and rate the visualization as less useful than when any type of context is provided.

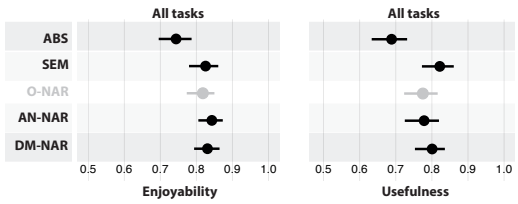


Figure 8. Enjoyability and Usefulness

Planned Analyses

In the previous section we gave an overview of all our results and identified several patterns, but due to the many comparisons involved some of these patterns may not be reliable. In this section we report on more focused comparisons based on our previously stated research questions. All the analyses in this section were planned before data was collected.

As before, we report sample statistics with 95% CIs. For dichotomous variables (in-task and post-task attention), we report proportion differences and compute CIs using score intervals for difference of proportions and independent samples. For continuous variables (all other metrics), we report differences in means and BCa bootstrap confidence intervals.

Q1: Does adding minimal semantics help?

To answer this question, we compared the ABS and SEM conditions for each objective performance metric: accuracy (averaged across all tasks), in-task attention and post-task attention.

The results are shown in Figure 9. The shaded areas indicate our initial expectations, i.e., a positive effect of SEM on all metrics. Contrary to our expectations, we found no evidence that adding semantics has a noticeable effect on participants' accuracy. We also found no evidence of a strictly positive effect on post-task attention, although the uncertainty on this metric is rather large. Nevertheless, participants better recalled the correlation task, suggesting that adding minimal semantic context can have positive effects on in-task attention.

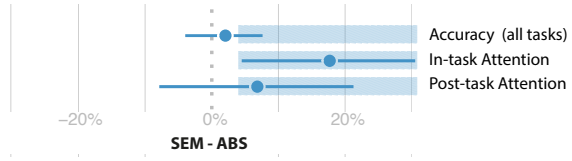


Figure 9. Mean differences in accuracy, in-task attention and post-task attention between SEM and ABS. Positive values indicate a benefit for SEM. Error bars are 95% CIs.

Q2: Does adding a narrative help overall?

To answer this question, we performed a contrast between the (SEM) condition and all narrative conditions (O-NAR, AN-NAR, DM-NAR) combined. The metrics were the same as above.

Here too, we expected a positive effect of narrative across all metrics. However, the results in Figure 10 go contrary to our expectations. Adding a narrative on top of dataset semantics makes participants less accurate. It also makes them less able to recall the correlation task, suggesting lower in-task attention. We do not have enough data to conclude that narratives also reduce post-task attention, but this remains a possibility.

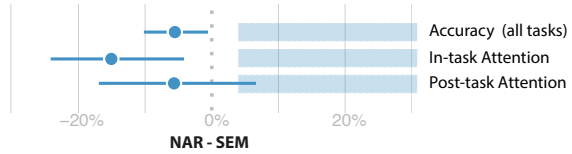


Figure 10. Mean differences in accuracy, in-task attention and post-task attention between all NAR conditions combined and SEM. Positive values indicate a benefit for narratives. Error bars are 95% CIs.

Q3: Should the question refer to the narrative?

To answer this question, we compared O-NAR and AN-NAR.

We expected that participants would be more accurate when reading the narrative is not required to carry out the task (O-NAR). As we can see in Figure 11, the results do not indicate a clear direction for an effect, and only suggest that the difference is rather small. We also thought participants would pay less attention when the narrative is not required. The data is mostly inconclusive. There is only very weak evidence that this could have been the case for post-task attention, but that the opposite pattern may have occured for in-task attention.

Q4: Is a decision-making framing better than an analytic one?

To answer this question, we compared AN-NAR and DM-NAR in terms of their performance on the comparison (Com) task.

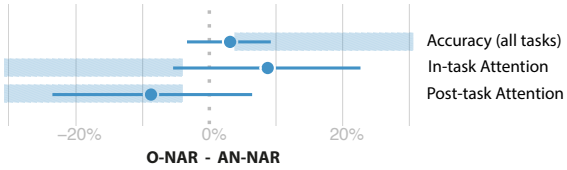


Figure 11. Mean differences in accuracy, in-task attention and post-task attention between O-NAR and AN-NAR. Positive values indicate a benefit for O-NAR. Error bars are 95% CIs.

We previously justified our focus on the Com task by explaining that we did not expect large differences between the two narratives for Ext and Cor. For context, Figure 12 shows, in gray, the differences across all tasks and for the tasks Ext and Cor. The results are consistent with our conjecture, although there is a larger uncertainty concerning the correlation task.

The rest of the figure (in blue) shows the differences for the Com task, a task that mimics a real decision task and for which we expected the decision making narrative (DM-NAR) to outperform the analytic narrative (AN-NAR). But contrary our expectations, participants performed remarkably worse when given a DM-NAR narrative. Concerning in-task or post-task attention, our results are inconclusive.

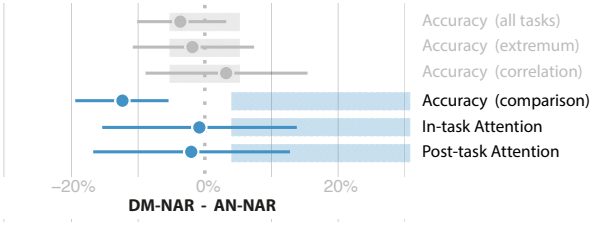


Figure 12. *In gray:* Mean differences in accuracy between DM-NAR and AN-NAR across tasks and for Ext and Cor; *In blue:* Mean differences in accuracy, in-task and post-task attention between DM-NAR and AN-NAR. Positive values indicate a benefit for DM-NAR. Error bars are 95% CIs.

DISCUSSION

Our study participants were fairly accurate overall, with average scores of 0.7–0.8 out of 1 across all tasks and narrative conditions (see Figure 4). Subjective task easiness scores were between 0.6–0.8 out of 1. Thus it seems that the difficulty of the tasks was properly calibrated overall.

We thought that providing task context in the form of data semantics or narratives would improve the overall quality of responses. Our findings suggest it is not necessarily the case: in terms of accuracy, data semantics do not seem to help much, and the narratives we used can even harm.

Our experiment only allows us to speculate about the reasons for these findings. First, as we discussed before, crowdworkers generally appreciate succinct instructions [16]. An otherwise simple task can appear more demanding in attention and time if it requires reading a long (in crowdsourcing standards) piece of text beforehand. Second, experienced contributors are generally used to performing abstract and mechanical tasks since these abound on crowdsourcing platforms. The fairly good performances we observed for abstract conditions do suggest contributors were overall able to understand the context-less tasks and willing to carry them out.

It is unclear whether contributors simply skipped the narratives. On the one hand, results of our post-task attention test confirm that not all our contributors read all instructions carefully (only 50–70% passed the test, see Figure 5). On the other hand, we did not find evidence that this was the reason for the lower task accuracy (research question **Q3**, Figure 11). However, Figure 4 does suggest that for the Ext task, asking the question in a way that does not require reading the narrative can help.

Despite these results, we have strong evidence that adding data semantics improves subjective experience on a range of metrics (confidence, perceived easiness, enjoyability, and perceived usefulness of the visualization). However, our backstory narratives did not yield measurable subjective benefits compared to data semantics alone. Thus, even though crowdsourcing contributors appreciate working with meaningful data, they may not be particularly interested in more elaborate narratives and may prefer to focus on carrying out their task.

Overall, our study provides compelling reasons for incorporating data semantics in crowdsourced evaluations of visualizations, i.e., stating what the datasets and their dimensions mean. But until further studies are carried out to nuance or contradict our findings, it seems safer to use elaborate narratives parsimoniously, unless there are clear reasons to do so. Such reasons include the evaluation of domain-specific and decision-support visualization systems as discussed previously.

Finally, we contribute a finding that has implications for the evaluation of visualizations for decision making. We found that the decision making narrative was less accurate than the analytic narrative for a task (Com) that has elements of real-life decision making. Most likely, the decision making framing caused participants to focus more on subjective preferences and less on giving a correct answer. This seems to imply that decision making tasks are more error-prone than equivalent analytic tasks, and that evaluating a decision-support system with standard analytic questions may not reflect a realistic use of the system and may overestimate its performance.

LIMITATIONS AND FUTURE WORK

Our study does not attempt to explain how to design good narratives. Its goal was rather to answer the question: if a researcher adds a narrative when evaluating a visualization (as is done sometimes), should she expect performance to improve? This goal leaves room for imperfections in the wording of the narratives. More studies are however needed to understand the effect of narrative design, and whether better narratives exist that could be successful at improving job quality.

Although our study found several clear effects (e.g., the accuracy drop caused by narratives (Figure 10), the lower performance of the decision-making framing for comparison tasks (Figure 12), and the negative subjective experience with abstract task framing (Figures 6 and 8)), other effects are less conclusive, calling for follow-up studies. Furthermore, while our study used a large sample (n=405) to test a range of conditions and questions, our findings can be made more robust with additional studies testing alternative narratives, scenarios, visual encodings, datasets, tasks, and performance metrics (such as open exploration and insight evaluation [44]).

Our study uncovered what could be termed a "double-edged sword effect" of narratives, but does not provide detailed definitive explanations for all the effects observed. Future research will need to investigate why and how different types of narratives affect task performance and subjective experience. This research could involve, for example, interviewing crowdworkers. Finally, investigating the effect of narratives in lab settings would be another compelling route to explore.

REFERENCES

1. Robert Amar, James Eagan, and John Stasko. 2005. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, 111–117.
2. Toshiyuki Asahi, David Turo, and Ben Shneiderman. 1995. Using treemaps to visualize the analytic hierarchy process. *Information Systems Research* 6, 4 (1995), 357–375.
3. Bon Adriel Aseniero, Tiffany Wun, David Ledo, Guenther Ruhe, Anthony Tang, and Sheelagh Carpendale. 2015. STRATOS: Using Visualization to Support Decisions in Strategic Software Release Planning. In *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1479–1488.
4. Benjamin Bach, Natalie Kerracher, Kyle Wm Hall, Sheelagh Carpendale, Jessie Kennedy, and Nathalie Henry Riche. 2016. Telling Stories about Dynamic Networks with Graph Comics. In *Proc. of the Conference on Human Factors in Information Systems (CHI)*. ACM, New York, United States.
5. Jeanette Bautista and Giuseppe Carenini. 2008. An empirical evaluation of interactive visualizations for preferential choice. In *Proc. of the working conference on Advanced visual interfaces*. ACM, 207–214.
6. Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM* 58, 8 (2015), 85–94.
7. Sarah E Bonner, Reid Hastie, Geoffrey B Sprinkle, and S Mark Young. 2000. A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research* 12, 1 (2000), 19–64.
8. Nadia Boukhelifa, Anastasia Bezerianos, Tobias Isenberg, and Jean-Daniel Fekete. 2012. Evaluating Sketchiness as a Visual Variable for the Depiction of Qualitative Uncertainty. *TVCG* 18, 12 (2012), 2769–2778.
9. Liliana Bounegru, Tommaso Venturini, Jonathan Gray, and Mathieu Jacomy. 2016. Narrating Networks: Exploring the affordances of networks as storytelling devices in journalism. *Digital Journalism* (2016), 1–32.
10. Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete. 2015. Storytelling in Information Visualizations: Does it Engage Users to Explore Data?. In *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1449–1458.
11. J. Boy, R.A. Rensink, E. Bertini, and J.-D. Fekete. 2014. A Principled Way of Assessing Visualization Literacy. *TVCG* 20, 12 (Dec 2014), 1963–1972.
12. Shenghui Cheng and Klaus Mueller. 2016. The Data Context Map: Fusing Data and Attributes into a Unified Display. *TVCG* 22, 1 (2016), 121–130.
13. Mohammad Daradkeh, Clare Churcher, and Alan McKinnon. 2013. Supporting informed decision-making under uncertainty and risk through interactive visualisation. In *Proc. of the Fourteenth Australasian User Interface Conference-Volume 139*. Australian Computer Society, Inc., 23–32.
14. Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. 2017. The Attraction Effect in Information Visualization. *TVCG* 23, 1 (2017).
15. Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*. Springer, 291–330.
16. Serge Egelman, Ed H Chi, and Steven Dow. 2014. Crowdsourcing in HCI Research. In *Ways of Knowing in HCI*. Springer, 267–289.
17. Fabian Fischer and Daniel Keim. 2013. Vacs: Visual analytics suite for cyber security-visual exploration of cyber security datasets. In *IEEE VIS*.
18. Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. 2013. Lineup: Visual analysis of multi-attribute rankings. *TVCG* 19, 12 (2013), 2277–2286.
19. Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *JPSP* 79, 5 (2000), 701.
20. Melanie C Green, Jeffrey J Strange, and Timothy C Brock. 2003. *Narrative impact: Social and cognitive foundations*. Taylor & Francis.
21. Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 203–212.
22. Jeffrey Heer, Fernanda B Viégas, and Martin Wattenberg. 2007. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proc. of the SIGCHI conference on Human factors in computing systems*. ACM, 1029–1038.
23. Thomas Hollt, Ahmed Magdy, Peng Zhan, Guoning Chen, Ganesh Gopalakrishnan, Ibrahim Hoteit, Charles D Hansen, and Markus Hadwiger. 2014. Ovis: A framework for visual analysis of ocean forecast ensembles. *TVCG* 20, 8 (2014), 1114–1126.

24. John Joseph Horton and Lydia B Chilton. 2010. The labor economics of paid crowdsourcing. In *Proc. of the 11th ACM conference on Electronic commerce*. ACM, 209–218.
25. Jessica Hullman and Nick Diakopoulos. 2011. Visualization rhetoric: Framing effects in narrative visualization. *TVCG* 17, 12 (2011), 2231–2240.
26. Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. 2013a. Contextifier: Automatic generation of annotated stock visualizations. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2707–2716.
27. Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. 2013b. A deeper understanding of sequence in narrative visualization. *TVCG* 19, 12 (2013), 2406–2415.
28. Jessica R Hullman. 2011. Not All HITs Are Created Equal: Controlling for Reasoning and Learning Processes in MTurk. In *Proc. of the ACM CHI Workshop on Crowdsourcing and Human Computation.* Citeseer.
29. Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
30. Graham Kalton and Howard Schuman. 1982. The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society. Series A (General)* (1982), 42–73.
31. Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proc. of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.
32. Ron Kohavi, Randal M Henne, and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 959–967.
33. Martin Krzywinski and Naomi Altman. 2013. Points of significance: error bars. *Nature methods* 10, 10 (2013), 921–922.
34. Bongshin Lee, Rubaiat Habib Kazi, and Greg Smith. 2013. SketchStory: Telling more engaging stories with data through freeform sketching. *TVCG* 19, 12 (2013), 2416–2425.
35. Kwan-Liu Ma, Isaac Liao, Jennifer Frazier, Helwig Hauser, and Helen-Nicole Kostis. 2012. Scientific storytelling using visualization. *IEEE Computer Graphics and Applications* 32, 1 (2012), 12–19.
36. Winter Mason and Duncan J Watts. 2010. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter* 11, 2 (2010), 100–108.
37. Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. 2012. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *TVCG* 18, 12 (2012), 2536–2545.
38. Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. Survey research in HCI. In *Ways of Knowing in HCI*. Springer, 229–266.
39. Tamara Munzner. 2014. *Visualization Analysis and Design*. CRC Press.
40. Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009a. Instructional manipulation checks: Detecting satisficing to increase statistical power. *JESP* 45, 4 (2009), 867–872.
41. Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009b. Instructional manipulation checks: Detecting satisficing to increase statistical power. *JESP* 45, 4 (2009), 867–872.
42. Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5, 5 (2010), 411–419.
43. Steven F Roth and Joe Mattis. 1990. Data characterization for intelligent graphics presentation. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 193–200.
44. Purvi Saraiya, Chris North, and Karen Duca. 2005. An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. *TVCG* 11, 4 (July 2005), 443–456.
45. Arvind Satyanarayan and Jeffrey Heer. 2014. Authoring Narrative Visualizations with Ellipsis. In *Proc. of the 16th Eurographics Conference on Visualization (EuroVis '14)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 361–370.
46. Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *TVCG* 16, 6 (2010), 1139–1148.
47. Fernanda B Viégas, Danah Boyd, David H Nguyen, Jeffrey Potter, and Judith Donath. 2004. Digital artifacts for remembering and storytelling: Posthistory and social network fragments. In *System Sciences, 2004. Proc. of the 37th Annual Hawaii International Conference on*. IEEE, 10–pp.
48. Stephen Wehrend and Clayton Lewis. 1990. A problem-oriented classification of visualization techniques. In *Proc. of Visualization'90*. IEEE Computer Society Press, 139–143.
49. Bernard E Whitley Jr and Irene Hanson Frieze. 1986. Measuring causal attributions for success and failure: A meta-analysis of the effects of question-wording style. *Basic and Applied Social Psychology* 7, 1 (1986), 35–51.
50. Christopher Williamson and Ben Shneiderman. 1992. The Dynamic HomeFinder: Evaluating Dynamic Queries in a Real-estate Information Exploration System. In *Proc. of the 15th Annual International ACM SIGIR Conference (SIGIR '92)*. ACM, New York, NY, USA, 338–346.
51. Ji Soo Yi, Rachel Melton, John Stasko, and Julie A Jacko. 2005. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization* 4, 4 (2005), 239–256.