Explaining the Gap: Visualizing One's Predictions Improves Recall and Comprehension of Data

Yea-Seul Kim
University of Washington
Seattle, WA, USA
yeaseul1@uw.edu

Katharina Reinecke
University of Washington
Seattle, WA, USA
reinecke@cs.washington.edu

Jessica Hullman
University of Washington
Seattle, WA, USA
jhullman@uw.edu

ABSTRACT

Information visualizations use interactivity to enable user-driven querying of visualized data. However, users' interactions with their internal representations, including their expectations about data, are also critical for a visualization to support learning. We present multiple graphically-based techniques for eliciting and incorporating a user's prior knowledge about data into visualization interaction. We use controlled experiments to evaluate how graphically eliciting forms of prior knowledge and presenting feedback on the gap between prior knowledge and the observed data impacts a user's ability to recall and understand the data. We find that participants who are prompted to reflect on their prior knowledge by predicting and self-explaining data outperform a control group in recall and comprehension. These effects persist when participants have moderate or little prior knowledge on the datasets. We discuss how the effects differ based on text versus visual presentations of data. We characterize the design space of graphical prediction and feedback techniques and describe design recommendations.

Author Keywords

Information visualization; self-explanation; prediction; internal representations of data; mental models

ACM Classification Keywords

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

INTRODUCTION

Information visualizations are common in news articles, scientific papers and other forms of digital resources, and serve to inform viewers by engaging with them in ways that text alone cannot. Visualizations often use interactivity, such as filtering, transformation and view navigation, to stimulate viewers' interest and enable user-driven querying of the data. Close studies of visualization comprehension indicate that users' interactions with their *internal representations*—mental models of what they know and are learning about a dataset—of the data are critical to the interpretation process [6, 20, 21, 33, 47]. Asking a person to express their internal representations can enhance critical thinking and prompt changes to existing beliefs to account for new information [16, 24, 46].

Most visualizations do not provide ways for users to explicitly incorporate their internal representations. The recent New York Times interactive visualization "You Draw It" [2] is a rare exception that attempts to prompt reflection by enabling a user to explicitly incorporate her prior knowledge. The interactive visualization asks viewers to draw their expectation of the data before presenting the observed data alongside their predictions. To prompt reflection on the difference, the interface provides feedback based on the accuracy of a user's expectation. However, it remains unknown whether the reflection on prior knowledge induced by such visualizations can positively impact recall and comprehension of data, or how to design such visualizations for maximal benefit.

We expand on prior work with five contributions: We contribute (1) a set of novel elicitation techniques for eliciting users' prior knowledge in visualization interaction. These include a graphical prediction technique for eliciting users' predictions of data, a feedback technique for presenting personalized feedback on the gap between predictions and observed data, and a self-explanation prompt to explicitly ask participants for self-explanations, which have been found to improve learning from texts and diagrams [1, 11, 12, 10], among others.

We contribute (2) a controlled experiment to test the effect of these techniques on recall and comprehension of data. We find that prompting participants to first predict data, or to self-explain presented data, or to do both, improves data recall and comprehension.

By further contributing (3) replications of our controlled study with datasets that differ in familiarity, we find that these techniques improve recall for datasets for which participants have moderate or little prior knowledge.

We also (4) evaluate how the impact of the techniques differs based on whether the information is presented using text or information visualization. We find that the visualization conditions benefit from predicting the data and viewing the gap between their prediction and the observed data whereas the text conditions do not.

Finally, we contribute (5) a characterization of the design space of data prediction and feedback techniques for information visualization and provide practical design recommendations.

CHI 2017, May 06 - 11, 2017, Denver, CO, USA.

ACM 978-1-4503-4655-9/17/05$15.00.
DOI: http://dx.doi.org/10.1145/3025453.3025592

HYPOTHESIS DEVELOPMENT

To generate hypotheses, we survey research in two main areas: (1) studies in cognitive psychology around the role of internal representations in information visualization comprehension, and (2) theories and studies on self-explanation in learning.

Internal Representations of Data in Visualizations

Researchers have developed multiple models describing how people interpret information visualizations (e.g., [30, 38, 44]). Though often overshadowed by research on lower level perceptual processes, "top-down" effects can guide interactions (e.g., eye movements) based on prior knowledge and beliefs, including knowledge about graphs and data analysis [30, 38, 48] and about the content of the data [7].

Studies in graph comprehension indicate that internal representations of relevant knowledge that one already possesses often play a critical role in reasoning with an external static or interactive visualization [20, 31, 33, 47]. Mayer et al. [35] compared well-designed animations to well-designed texts and diagrams for teaching, finding that performance on retention and transfer tests was better among the static media group that relied on internal representations to understand how concepts related. Similarly, Hegarty et al. [23] found that viewers who initially engaged in mental animation of a set of static views, then used an external animated visualization, understood the content better than those who used only the animated visualization. Other studies suggest that externalizing one's internal representations leads to better understanding of visualized information [15, 24, 37, 46]. For example, Stern et al. [46] found that individuals who had to construct a line graph of presented stock data were more accurate on transfer questions involving a new problem with a similar structure than participants who passively viewed a chart of the data. Constructing external representations is believed to help individuals to translate information between different representations, resulting in a more nuanced understanding of the concepts [16, 46]. Building on these results, we contribute various novel interactive elicitation techniques for information visualization that prompt a user to incorporate her prior knowledge in her interaction.

Self-Explanation in Learning

Self-explaining information to oneself is a constructive "metacognitive" learning activity in which a person actively reflects on the mechanism behind a given phenomena [11]. Self-explaining has been elicited by having learners explain the meaning of sentences [10, 11, 12], or diagrams [1], as they study a target domain. When quality self-explanations are generated, such as statements that link concepts from the text using tacit knowledge or attempt to fill in gaps through inferences [10], comprehension tends to be better than without self-explanation [10, 11, 12]. A mental model repair hypothesis is proposed to describe the benefits of self-explaining: by generating inferences to fill in missing information, integrating new information with prior knowledge, and monitoring and repairing faulty knowledge, learners who self-explain develop more accurate internal representations of a concept [10]. In this hypothesis, it is assumed that learners engage more with the process if they identify the discrepancy between their mental model and the presented information [10]. While students who spontaneously generate self-explanations perform better [11, 18, 39, 40, 41], most learners tend not to naturally self-explain [5, 43]. However, even simple prompts [12, 5, 43] have been found to be effective at triggering the self-explanation process [12]. Other studies have shown that immediate feedback on the accuracy of self-explanations can prevent wrong inferences in the explanation process [14, 3, 9]. We develop and test explicit and implicit prompts for self-explanation applied to information visualizations.

Self-explanation techniques have been widely applied in online learning environments, typically to text-based learning materials [13, 14, 3, 4, 40, 25]. A handful of studies have included a combination of diagrams and text [11, 42, 4]. However, studies that attempt to understand how self-explaining of visual representations is different from text are rare. An exception is Ainsworth et al. [1], who compared self-explaining across text and diagrams in a small group (n=20). They find that learners who used diagrams performed better on a comprehension test, generated more self-explanations, and appeared to benefit more from self-explaining than users of texts. We conduct a comparison of the benefits of multiple interactive elicitation techniques, including explicit self-explanation prompting, for data presented through information visualization versus text.

Curiosity theory describes a similar process to self-explanation. A person's curiosity is piqued when she perceives an information gap between her current knowledge and the information she is interested in [34]. Prompting people to guess is thought to be one way to arouse curiosity by making this gap explicit [34]. Awareness of the gap results in an increasing desire to seek knowledge to fill this gap [34], and inspires people to explore more information [29], including in crowdsourcing contexts [32]. We examine whether techniques aimed at implicitly prompt curiosity, such as by having people predict data before they see it, can improve comprehension and recall by enhancing their desire to see the true data.

Formulating Study Conditions & Hypotheses

Based on the previous research in psychology and education, we devise and evaluate several elicitation techniques for information visualization with associated hypotheses.

Study Conditions: Elicitation Techniques

Our techniques are based on three non-mutually exclusive mechanisms for eliciting reflection on prior knowledge and its relationship to presented data in a visualization.

1. Prompting a user to generate self-explanations of the observed data: In a digital setting, prompting a user to type in sentences explaining the data is an *explicit* way to elicit self-explanations [4].

2. Prompting a user to predict the data before seeing it: Asking a user to predict the data has two advantages for prompting reflection on prior knowledge: (1) Prior work showed that asking a user to actively construct an external representation of her prior knowledge about data results in a deeper understanding of the meaning of a dataset and its visual representation [16, 46]. Predicting may also trigger self-explanation [16]. (2) By asking the user to provide predictions

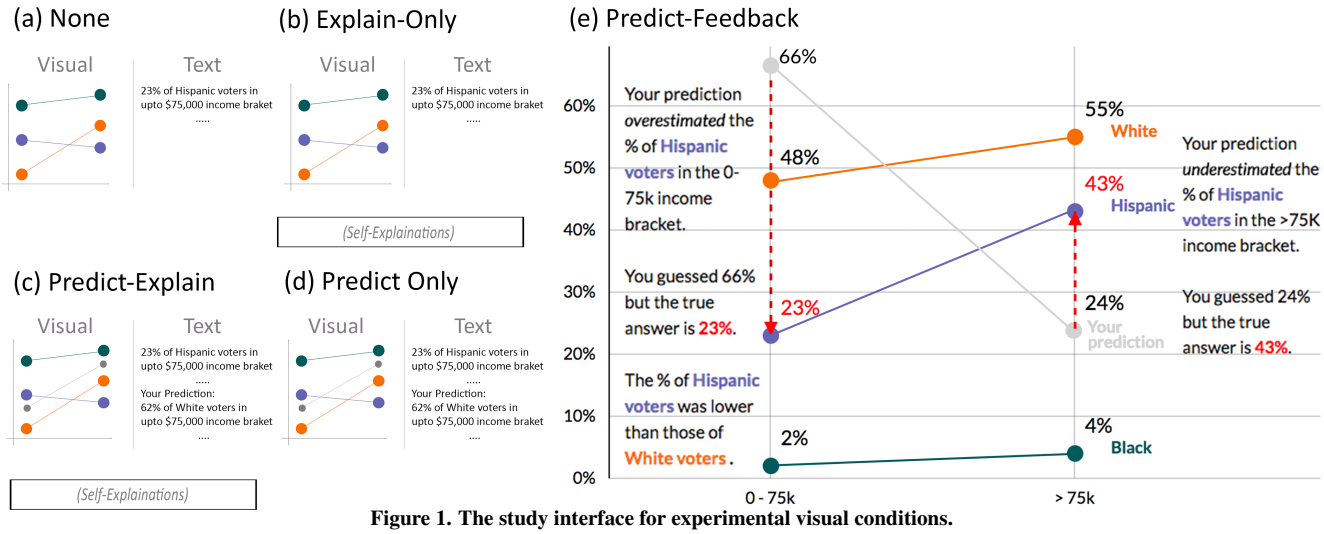

Figure 1. The study interface for experimental visual conditions.

of the data, an interface can then *visualize the gap* between the user's expectations and the observed data. Reviewing the gap may *implicitly* prompt self-regulated learning, in which the user becomes motivated to generate inferences to repair her knowledge [36].

3. Providing the user with feedback on her prediction: Providing direct feedback on the gap between the user's prior knowledge and the observed data may increase the likelihood that a user will recognize the gap and generate inferences to repair her knowledge [14, 3, 9].

Using these three mechanisms, we designed four elicitation techniques and one baseline condition. The observed data is presented using visualization in the visual conditions, and using text in the text condition.:

- **None (baseline)**: The user is prompted simply to examine the observed data (Fig. 1(a)).
- **Explain-Only**: The user is prompted to type self-explanations in a text box as she views the observed data (Fig. 1(b)).
- **Predict-Explain**: The user is shown the observed data, but with some of the data omitted. The user is prompted to predict the omitted data. After her prediction, the user is shown the observed data against her prediction. She is prompted to type in self-explanations about the gap between her prediction and the observed data in a text box (Fig. 1(c)).
- **Predict-Only**:The user is shown the observed data, but with some of the data omitted. The user is prompted to predict the omitted data. After her prediction, the user is shown the observed data against her prediction (Fig. 1(d)).
- **Predict-Feedback**: The user is shown the observed data, but with some of the data omitted. The user is prompted to predict the omitted data. After her prediction, the user is shown the observed data against her prediction. Textual and visual feedback are presented to annotate the difference between her prediction and the observed data to draw her attention to the gap (Fig. 1(e)).

Hypotheses

Since explicitly prompting self-explanations improves comprehension of information in texts and diagrams [1, 12, 11, 10], we expect a similar beneficial effect for data in visualizations:

H1: Participants in the Explain-Only conditions will recall data more accurately than participants in the None condition for visual and text modalities.

Based on the implicit prompting toward reflection by predicting, we expect that:

H2: Participants in the predict conditions (Predict-Explain, Predict-Feedback, Predict-Only) will recall data more accurately than participants in the None condition for visual and text modalities.

While reviewing their predictions and the observed data, participants in the text conditions must actively *seek and infer* the gap as opposed to the visual conditions where the gap is visually available [1]. We therefore expect that effects of predicting will be less pronounced in the text conditions compared to the visualization conditions:

H3: The effects of predicting using text (Predict-Explain-Text, Predict-Only-Text) on recall will be smaller than the effects of predicting using visualizations.

PRELIMINARY SURVEY: CHOICE OF DATASETS

To select a dataset for our main study, we conducted a preliminary survey on Amazon's Mechanical Turk (AMT). We sought a dataset with properties amenable to our elicitation techniques. If a user is extremely familiar with a dataset, prediction or explanation may not offer benefits over her prior knowledge. If a dataset is too unfamiliar, it may be too difficult for a user to make predictions about it. Additionally, our interest in comparing techniques across visualization and text modalities is best supported by a dataset that includes both a higher level relational structure (i.e., trends, which visualizations are often best at depicting) as well as individual data points, which can be remembered with greater numerical accuracy from text [27]. By quantitatively measuring the familiarity of multiple datasets, the preliminary survey also

enables us to later test the robustness of our results across datasets of varying familiarity.

We selected datasets with a range of topics from the results of a scientific experiment to the average smart phone price from different manufacturers (all datasets are available in the supplementary materials.) The datasets had the same format, consisting of two categorical dimensions and one continuous measure, resulting in six total data points (Fig. 2). This format is commonly used in the social and physical sciences [8] and allows us to evaluate how well people observe and recall the higher level patterns (e.g., trend lines) as well as how well people observe and recall individual data points. We formulated three measures that approximate prior familiarity with each dataset:

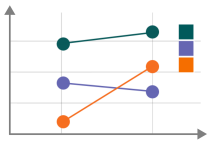

Figure 2. Multivariate data format to allow value and trend estimation.

Perceived familiarity: How familiar participants perceive themselves to be with the data after seeing a short description and a visualization with labeled axes but without data points.

Value familiarity: The absolute error of participants' predictions for each data point. We calculated the absolute difference between the participant's prediction of each data point and the observed data. We normalized the values by dividing by the maximum value on the *y*-axis to allow for comparison of the value familiarity across the datasets.

Trend familiarity: The difference between participants' predicted slopes and the true slopes for each line in the visualization. We calculated the absolute difference between a participant's slope and the true slope for each of the three groups in each dataset.

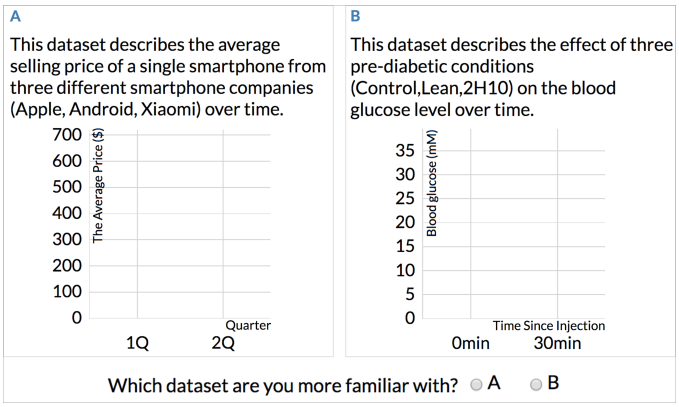

Figure 3. A pair of visualizations with the data omitted. Participants were asked to choose which dataset is more familiar based on labels.

Procedure

The survey consisted of two parts. In the first part, we examined *perceived familiarity* by presenting participants with 15 visualization-pairs (resulting from the pairwise combinations of the six datasets), one at a time in a randomized order (Fig. 3).

The visualizations did not reveal any data; instead, participants only saw labeled axes with a short description of each dataset. Participants were asked to select the dataset that they were more familiar with using a radio button. After watching a short tutorial video on how to enter a data point in a chart, participants were asked to predict the values of all six data points for each of the six datasets (Fig. 4). Each dataset appeared on a separate screen in randomized order. Each screen presented an empty chart area with labels. Three buttons appeared to the right of the chart labeled with each of the three groups for that dataset (e.g., french fries, coke, shake).

The range of the *y*-axis of the visualizations was set to [0, 1.2*max_data_value]. To input a prediction, participants selected a group by clicking the button for that group, then clicked the chart area to set the position of the two data points for that group. Participants could drag the circles to adjust their prediction for each group.

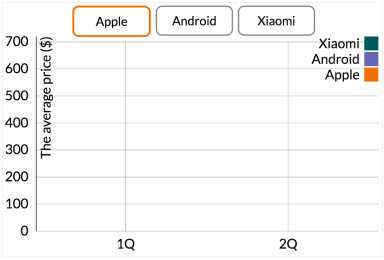

Figure 4. Example of a prediction interface.

Results

We recruited 100 workers from AMT. To calculate the perceived familiarity of each dataset, we summed the number of votes per dataset from the first part of the study.

Table 1 shows the ranking of the datasets by these three measures. While perceived familiarity and value familiarity are highly correlated (Spearman's $\rho = .84$, $p < .001$), perceived familiarity and trend familiarity are more weakly correlated (Spearman's $\rho = .42$, $p < .001$). We see evidence of this in the dataset on median house prices, where participants' perceived familiarity is relatively low (rank 5), but they do well on the prediction tasks (rank 1 for predicting trend). We suspect that various factors might contribute to a difference between perceived familiarity and prediction accuracy. In addition to having prior knowledge specific to the dataset, heuristics may allow participants to guess reasonably accurately because they have some knowledge of the domain. In particular, domain specific knowledge (e.g., the average price of houses is cheaper in Colorado than those in New York), and domain general knowledge (e.g., prices tend to go up over time) may allow participants to make reasonable guesses even when they feel they have little expertise on a topic.

We aggregated the three familiarity rankings and sorted the datasets by the aggregated familiarity (the order of the Table 1). For our main study, we chose the Colorado voting results data, since this dataset was neither clearly familiar nor unfamiliar to participants. To ensure that recalling the data is sufficiently challenging in our main study, we included two visualizations of voting results, for Connecticut and Colorado. We used the more familiar fast food calorie content dataset and the more unfamiliar pre-diabetic experiment dataset to later check the robustness of results from our main study in partial replications.

Table 1. Three familiarity measures for the six datasets. The order of rows in the table is by the mean rank across the three measures. The number indicates the ranking of the dataset with the actual measure in parentheses. Perceived familiarity votes are out of 500 (100 workers * 5 maximum votes per dataset).

	Perceived Familiarity	Value Familiarity	Trend Familiarity
Fast Food Calorie Content	1 (348)	1 (0.14)	2 (0.11)
Smartphone Price	2 (312)	2 (0.15)	3 (0.13)
Median House Price	5 (188)	3 (0.18)	1 (0.10)
Voting Result	3 (297)	4 (0.21)	4 (0.20)
National Budget	4 (283)	5 (0.26)	5 (0.25)
Pre-diabetic Experiment	6 (72)	6 (0.28)	6 (0.36)

STUDY DESIGN

Study Objectives & Experimental Conditions

To understand how different techniques for eliciting prior knowledge with visualization impact data recall and comprehension, we designed a between-subjects factorial study. Participants were assigned to a baseline condition (**None**) or one of four elicitation techniques: **Explain-Only, Predict-Explain, Predict-Only, Predict-Feedback**.

Additionally, to better understand the effects of the elicitation techniques with a visualization, we varied whether a participant interacts with (and makes predictions about) the data using text or visualization (**Text** or **Vis**).

We crossed the elicitation techniques with modality, with the exception of Predict-Feedback, resulting in 9 possible conditions: None-Vis, Explanation-Only-Vis, Predict-Explain-Vis, Predict-Only-Vis, Predict-Feedback-Vis, None-Text, Explanation-Only-Text, Predict-Only-Text, Predict-Explain-Text. We excluded a Predict-Feedback text modality treatment due to the difficulty of generating personalized feedback based on freeform text predictions.

Participants

A prospective power analysis was performed for sample size determination based on the effect size and standard error of each technique and modality in pilots using a mixed effects model. We achieved 0.8 power under $\alpha = 0.05$ with 42 participants per condition. We then recruited 378 participants (42 per condition) from AMT, rewarding their participation with $1.50.

Procedure

Fig. 5 shows an overview of the study procedure. The study started with an *introduction* (Fig. 5(1)), in which we explained the data domain as the percentage of voters of different ethnicities (Hispanic, White, Black) who voted Republican in the 2008 presidential election, for several income brackets and states. To eliminate possible difficulties with the interactive nature of data entry, participants in the visual conditions watched a tutorial video to learn how to set and adjust a value in an related line chart. On the next page, participants in the prediction conditions (Predict-Explain, Predict-Feedback, and Predict-Only) were asked to *predict* the voting percentages for one randomly selected ethnic group across two income levels for each state (Colorado, Connecticut) in randomized order.

To ensure that participants interacted to a similar degree across treatments, participants who were not prompted to predict were asked to *retype* a paragraph about elections in the U.S. (Explain-Only and None condition) (Fig. 5(2)).

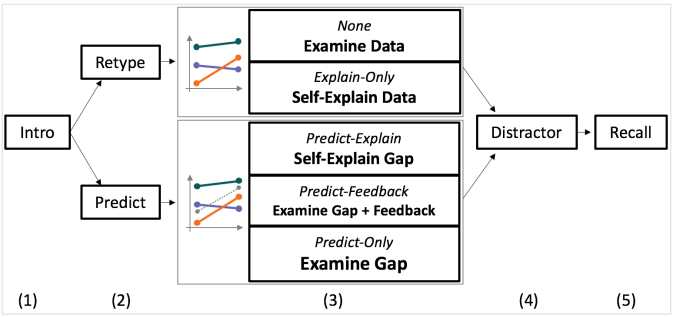

Figure 5. Overview of the study procedure. If participants were not asked to predict, they were asked to retype a general text on elections. If participants were not ask to generate self-explanations, they examined either the data or feedback depending on the condition.

On the next page, all participants *examined* the observed data (Fig. 5(3)), with prompts and feedback varying by condition. Participants in the None conditions were asked to examine the observed data several times (Fig. 1(a)).

Participants in the Explain-Only conditions were asked to generate and type in a few sentences of explanations to help themselves understand the data (Fig. 1(b)).

Participants in the Predict conditions saw their predictions in a lighter color against the observed data in the visual conditions. In the text conditions, the textual predictions that the participant made were shown with the observed data presented in text.

Participants in the Predict-Explain conditions were asked to self-explain the difference between their prediction and the observed data (Fig. 1(c)).

Those in the Predict-Feedback condition saw accuracy feedback based on their predicted values (Fig. 1(e)). Feedback contained 1) the directionality of the participant's error (e.g., "Your prediction *overestimated* the percentage of Hispanic voters."), 2) verbal statements that reiterated the participant's prediction and the observed data (e.g., "You guessed 66%, but the true answer is 23%"), and 3) comparative information that indicated high-level patterns (e.g., Higher portions of white voters vote for John McCain then hispanic voters.), if the participant violated the pattern (e.g., "The percentage of Hispanic voters was lower than those of White voters").

Participants in the Predict-Only condition were asked to examine their predictions and the observed data several times (Fig. 1(d)).

As a distractor task, all participants then completed as many questions on a 10 question digital paper folding test as they could in three minutes [17]. The task also served to gather information on participants spatial visualization abilities, which have been shown to correlate with effective use of internal representations [23, 20, 22], (Fig. 5(4)).

After completing the paper folding test, participants in all conditions were asked to recall the percentage of voters of different ethnicities (Fig. 5(5)). Recall interfaces for each state were provided on separate pages and presented in reverse order from that in which the data was examined. Participants used an interface that matched the modality by which they viewed the data (text or visualization).

Participants were asked to respond to demographic questions, including age, education level, gender, and ethnicity, and were asked about their experience with visualizations.

RESULTS

Data Preliminaries

The average time to complete the experiment was 19.4 minutes (SD=8.4), with no differences in response time across the conditions ($F(4) = 1.073, p = .37$) There were no significant differences between participants' demographic responses or relevant experience across the conditions (see supplementary materials for a detailed analysis). We excluded 3 participants that did not specify predictions in the text conditions, and 2 participants who participated multiple times.

Analysis Approach

We used two mixed effects models implemented in R's lme4 package to evaluate H1, H2, and H3. We used the normal approximation to calculate p-values of fixed effects using t-scores produced by lme4. The R code and detailed analyses are available in the supplemental materials [1].

Dependent Variables

We considered two types of error indicating how well participants could recall the observed data. The accuracy in recalling individual data points was measured using the *absolute error*, i.e., the absolute difference between the recalled value and the observed value. To measure accuracy in recalling the higher level structure of datasets (e.g., trends within each group), we calculated the *trend error*, i.e., the absolute difference between the recalled and the actual slope of each line (set of values for an ethnicity) in the visualization.

Model Specification

In each mixed effects model, we included the four elicitation techniques (Explain-Only, Predict-Explain, Predict-Only, and Predict-Feedback), modality, and the interaction terms between modality and the techniques as fixed effects, with the control/baseline condition as the omitted reference condition. We included the participant id and the ethnicity group (e.g., Hispanic, etc.) as random effects. The spatial ability score, calculated as the number of correct answers out of 10 on the paper folding task, was included as a fixed effect. We centered the spatial ability score by its mean so that fixed effects describe a participant of average spatial ability. For easier interpretation, we report the intercepts for each elicitation technique separately for visual versus text with 95% confidence intervals. Coefficients are expressed in terms of the actual units used in the datasets (i.e., percentage).

[1] https://github.com/yeaseulkim/ExplainingTheGap

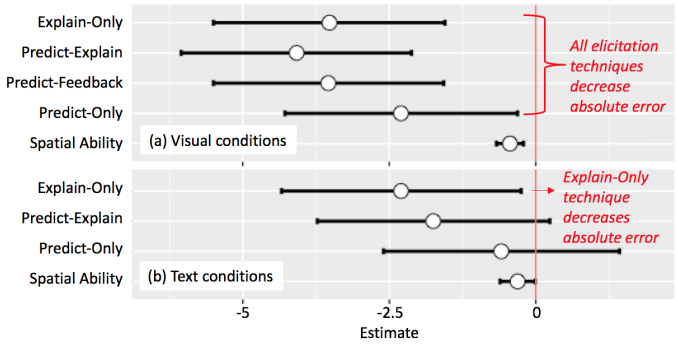

Figure 6. Estimated fixed effect coefficients from analyzing absolute errors for (a) visual and (b) text conditions for the voting result dataset. The error bars indicate 95% confidence intervals. Intervals that do not include include zero imply that we can be reasonably sure that some effect exists.

Core Results

Visual Conditions

Absolute Error: Participants who used one of the four elicitation techniques recalled individual data points more accurately than those in the None-Vis condition (Fig. 6(a)). Proceeding by magnitude of effect, the Predict-Explain-Vis condition had the lowest absolute error relative to the None-Vis condition by -4.08 (i.e., participants in the Predict-Explain-Vis condition were, on any given recalled point out of the 12 total, more accurate at recalling the voting percentage by 4.08% out of 100% compared to the those in the None-Vis condition: $t = -4.02, p < .0001$) The Predict-Feedback-Vis condition had the next lowest effect (-3.54; $t = -3.49, p < .001$), followed by the Explain-Only-Vis condition (-3.52; $t = -3.47, p < .0001$), and the Predict-Only-Vis condition (-2.30; $t = -2.25, p < .05$). Hence, interactive elicitation appears to be reliable way to improve absolute recall of data presented through visualizations, even with variations in how prior knowledge is elicited.

Participants' scores on the spatial ability test also predicted a lower absolute error, as we would predict from prior work indicating the relationship between spatial visualization ability and visualization comprehension [23, 20, 22]. With each additional correct answer in the paper folding task, participants' expected absolute recall error decreased by 0.44 ($t = -3.63, p < .001$).

We observed no difference in absolute error between the four elicitation techniques.

Trend Error: Only participants in the Predict-Explain-Vis and the Predict-Feedback-Vis conditions had a lower trend error compared to the None-Vis condition. Specifically, being in the Predict-Explain-Vis condition lowered errors by -2.06 relative to the None-Vis condition ($t = -2.07, p < .05$), while being in the Predict-Feedback-Vis condition lowered errors by -2.79 relative to the None-Vis condition ($t = -2.80, p < .05$). We observed no effect of the spatial ability score on participants' trend errors ($t = -0.53, p = .596$).

Comparing Effects of Techniques: Visual vs Text

Absolute Error: Overall, we observed 3.76 percentage points (36%) more errors on average per recalled value among visual conditions compared to text. This aligns with prior research

indicating that text is better for exact value retention than visualization [27].

In comparing the absolute error between the text conditions, the Explain-Only-Text condition was the only condition that led to lower absolute errors than the None-Text condition, by an average of -2.00 ($t = -2.07, p < .05$). We observed no effect from the Predict-Explain and the Predict-Only techniques in the text modality conditions when we compared them to the None-Text condition (Fig. 6(b)). Text presentations may not provide the same type of natural support for prompting implicit reflection and correction of one's prior knowledge compared to visualizations.

Participants with higher spatial ability scores again had a lower absolute error on recall by 0.31 ($t = -2.04, p < .05$).

Trend Error: Overall, we did not observe differences in average trend errors between visual conditions and text conditions. In comparing the trend error between the text conditions, we found no effect of any of the elicitation techniques (i.e., Explain-Only-Text, Predict-Explain-Text, or Predict-Only-Text) compared to the None-Text condition. We saw no apparent decrease in trend error from spatial ability ($t = -1.30, p = .193$).

Anchoring in the Prediction Conditions

Participants in the Predict-Only-Vis, Predict-Explain-Vis, and Predict-Feedback-Vis conditions may have a tendency to recall aspects of their own prediction due to the deliberate attention required to generate the prediction. We observed a weak positive correlation between the values that participants predicted and the values that they recalled ($R^2 = 0.176$, intercept$= -1.28$, $slope = 0.18$) Additionally, we found that for 75.9% of the data points from participants who were asked to predict, their recalled value showed a bias in the same direction as their predicted value: if they underestimated the value in the prediction phase, they tended to underestimate the value in the recall phrase. The same pattern could be observed when they overestimated. Hence, a slight anchoring effect appears to be present.

Quantity and Quality of Self-Explanations

The quantity and quality of self-explanations have been shown to affect comprehension in prior work [12, 42]. We analyzed the correlation between self-explanation quantity and quality and recall performance for participants in the Explain-Only condition. As a proxy of quantity we counted each sentence as a self-explanation (mean:3.2, range: 1-7). We tallied the total number of self-explanations generated by a participant and regressed the average recall error made by the participant on the sum. We observed no effect of the number of explanations on recall error ($R^2 = 0.001, F = 0.39, p < .533$).

To measure the quality of each self-explanation, we devised criteria informed by Chi's approach to distinguish high and low quality explanations [10]. We differentiated between two factors that characterize the quality of self-explanations: the level of prior knowledge involved in inference (inference with prior knowledge, inference with no prior knowledge, no inference), and the level of detail of the inference (high, low).

Inference without prior knowledge, High detail: "Generally, people with incomes over 75k were less likely to vote for John McCain in 2008. Blacks who made over 75k were slightly more likely to vote for him, but it was a very small increase to 3% meaning not many blacks voted for McCain in any income category."

Inference without prior knowledge, Low detail: "Majority of people no matter ethnicity voted for the Democrats and not Republicans."

Inference with prior knowledge, High detail: "There was a slight increase in the White voting population with the higher income bracket, I could assume that this is due to McCain's policies which benefit the wealthier. "

Inference with prior knowledge, Low detail:"People are more conservative in Colorado."

No inference: "Each colored line is a different race. Each point is a different income bracket."

Two researchers coded the set of 42 explanations (*Cohen*ś $kappa = 1$). We conducted a two way factor analysis on the average absolute error and the trend error, but observed no effect of either quantity or quality on factors. It may be that the difference between participants' self-explanation quality was smaller overall than in educational studies of spontaneous self-explanation, perhaps due to the incentives to work quickly on AMT. Detailed results are available in the supplemental materials.

Replication on Low & High Familiarity Datasets

We conducted two additional partial replications of our study to evaluate the effect of elicitation techniques on datasets that our preliminary survey identified as more and less familiar on average. We replicated all visual conditions.

Low Familiarity Dataset: Scientific Experiment Results

We created two visualizations depicting results from a scientific experiment on the blood glucose level of various groups of mice after antibody injection [19]. Each visualization differentiated two amounts of time since injection (0 and 30mins, and 60 and 120mins). Each visualization included lines for three groups (Lean, 2H10, and control) similar to the three groups of Hispanics, White, and Black voters in the main study. Hence, the two visualizations replicated the structure of the voting results data across two states.

Absolute Error: We observed a similar pattern of effects of the techniques on decreasing absolute error as for the voting result data, with the exception of the Predict-Only-Vis conditions (Fig. 7(a)). The Explain-Only-Vis condition had lower errors by 1.78 ($t = -3.75, p < .001$), and the Predict-Explain-Vis conditions had lower errors than the None condition by 1.27 ($t = -2.62, p < .01$). The Predict-Feedback-Vis condition had lower errors than the None condition by 1.31 ($t = -2.69, p < .01$). Predict-Only-Vis condition had no effect compared to the None condition ($t = -1.43, p = .152$).

We observed no difference in recall performance between the Explain-Only-Vis, the Predict-Explain-Vis, and the Predict-Feedback-Vis condition.

Trend Error: We also observed no effects of the elicitation techniques on decreasing trend error compared to the None condition.

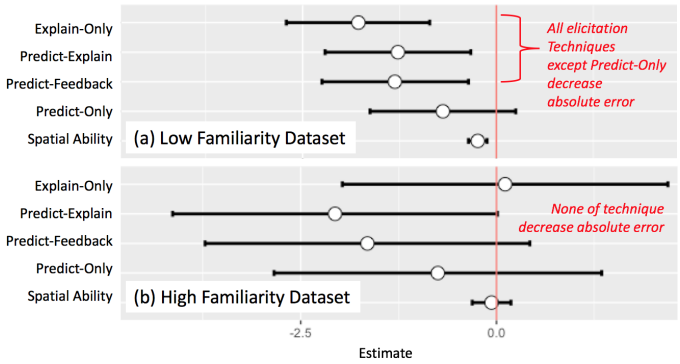

Figure 7. Estimated fixed effect coefficients from analyzing absolute errors for visual conditions (a) for the scientific experiment dataset, and (b) the fast food calories dataset.

High Familiarity Dataset: Calorie Content of Fast Food

We created two visualizations using data on the calorie content of three fast foods (Milkshake, Coke, French fries) for two serving sizes (small, large) at McDonald's and Burger King.

Absolute and Trend Error: Results of the mixed effect model for absolute error (Fig. 7(b)) and trend error indicate no differences between any conditions for the fast food dataset. Detailed results are available in supplemental material.

We conclude from these additional partial replications that prediction, when combined with additional mechanisms like explicit self-explanation or feedback, can lead to lower absolute error even for very unfamiliar data where prediction might be difficult. Perhaps because the data was already too familiar, we saw no replication of effects for any elicitation technique on the high familiarity dataset.

DISCUSSION

Our results suggest the promise of incorporating mechanisms for eliciting prior knowledge in visualization.

First, our work extends prior work on self-explanation by showing that prompting users to self-explain information visualizations can improve their ability to recall specific data points later.

Additionally, our work is the first to show that incorporating prediction tasks, as in our Predict-Explain-Vis, Predict-Only-Vis, Predict-Feedback-Vis also improves users' ability to recall specific data. We hypothesize that predicting focuses a user's attention on their prior knowledge, making them more likely to attend to the gap between their prior knowledge and the observed data when it appears. We expected that predicting would be less effective in the text modality than in the visual modality. In fact, we did not observe an effect of any of the prediction techniques on improving recall in the text conditions. This may be because participants in the visual conditions are shown the gap in a visual form, reducing the deliberative effort required to process the gap in a text format. Using a visualization of the gap is likely to free up participants' cognitive resources in contrast to text, so that they can focus on more meaningful activities [45] such as updating and repairing their mental model.

Except for the Predict-Only-Vis conditions, we were able to replicate the effect of the elicitation techniques for decreasing absolute error with a less familiar dataset, the scientific experiment results. These results suggest that prior knowledge about a dataset is not necessarily required for a user to benefit from prediction, explanation, and feedback techniques. The fact that we were not able to replicate the effects in the Predict-Only condition may indicate that when familiarity is lower, a user needs additional reinforcement to recognize the gap, such as being prompted to explain or being given visual feedback on the gap.

We could not replicate the effects of the elicitation techniques on the high prior familiarity dataset, the calorie content of fast food. One possible reason that we did not see an effect here is that participants were already relatively accurate in estimating the values, such that the initial 'gap' is too small to see much positive impact from the elicitation techniques.

We observed that the Predict-Feedback-Vis and Predict-Explain-Vis techniques decreased the trend error in recalling the visualized trends for the voting results data. However, we did not see similar improvements of trend errors with either of the other datasets. Varying graphical complexity may be one reason for this discrepancy: in the voting results data, two of the lines intersect (thus adding complexity), whereas in the other two datasets all three lines have similar, non-intersecting trends.

We believe that prompting predictions and providing feedback, which we found to reduce both the absolute and relative recall error, would work well when presenting visualizations in practice. Compared to explicitly prompting a user to provide self-explanations, first prompting a user to make predictions is likely to engage a user's curiosity. Once the user has "invested" attention by predicting, they may be more open to feedback that can further direct their attention to needed adjustments in their prior knowledge.

Though we asked participants in the predict conditions to predict only select data points, we observed a decrease in recall error across all data points. In the Predict-Explain-Vis conditions, we see evidence that participants are generating explanations associated with all three ethnic groups despite being prompted to explain only the difference between the predicted group and observed data. For example, participants wrote in the comments:

> "I really overestimated the numbers of black people. I expected more of them to vote than the Hispanics."

> "I may have underestimated the population of Hispanics (and Blacks) in Colorado. I assumed both were minorities who held liberal viewpoints, since traditionally minorities tend to prefer Democratic candidates."

This suggests that a designer need not require a user to predict every data point in the visualization to engage users with the entire dataset.

Task	Detail Task	Manipulation Component	Encoding	Possible Interaction
Predict Continuous (Quantitative) Variable	Predict Data Value	Mark (bar)	Bar chart	Drag up a bar to set height
		Mark (line)	Line chart	Draw a line
		Mark Attribute (color of areas)	Map (choropleth)	Brush on color over an area
Predict Categorical (Nominal, Ordinal) Variable	Predict Categorical Membership	Mark Attribute (color of areas)	Area chart	Brush on color over an area
			Pie chart	Brush on color over a sector
Predict Data Structure and Model	Predict Correlation / Fit	Mark (line)	Scatter plot	Draw a regression line
			Line chart	Draw a line
	Predict Cluster		Scatter plot	Draw a contour
			Dendrogram	Drag and drop element to the cluster
	Predict Connectivity		Node-link diagram	Draw an edge
	Predict Confidence Interval		Box plot	Draw a line to mark confidence interval

Figure 8. Possible tasks.

Design Space for Graphical Prediction & Feedback

Our study shows the benefits of eliciting users' prior knowledge, such as their expectations of data, and prompting them to reflect on how their knowledge relates to the data. However, the design space for applying graphical prediction and feedback techniques to information visualization remains relatively unexplored. In the following, we characterize key considerations in applying these techniques to visualizations.

We informed our elaboration of the design space through several forms of evidence: observations from our studies; examples in the media, primarily from the New York Times [2]; and our own development of prototypes applying the techniques to visualizations like bar charts and line charts.

Based on these experiences, we differentiate three considerations that influence the effectiveness of graphical prediction and feedback applications: the *prediction task and graphical elicitation technique* (for what tasks and in what ways can the user express her prior knowledge?), the *contextualization mechanism* (how does the interface provide clues to constrain the user's prediction?), and the *feedback technique* (how does the interface draw the user's attention to the gap between her predictions and the observed data?).

Prediction Task & Elicitation Technique

A first question in designing a visualization that elicits predictions is "What should the user predict?". A visualization can elicit value predictions for quantitative or nominal (categorical) variables, or the outcome of a model or analysis.

Direct manipulation is a natural way to implement the first one, *value prediction*. For example, a user might click to add a mark, or drag a mark from an axis to set or update its position in a 2D plot like a scatter plot. Other marks require different interactions: a user might drag a bar to set its height or click to position the top of the bar in a bar chart, and use a smooth dragging operations to position a line in a line chart.

Data encoded in the visual attributes of marks, such as color or shape may also be predicted. For example, nominal data (categories) might be encoded by the color hue of marks in a scatter plot. *Predicting categorical membership* can be instrumented with interactions like brushing. For example, some points may remain uncolored in a scatter plot where color encodes the categorical membership of data. The designer can have a user select a color from an interactive legend, and drag across points to assign that category. In The New York Times' elicitation of users' predictions for the 2014 senate election, a user was able to cycle through different binned probability levels for the voting percentage for each party by repeatedly clicking on a state in the map [28].

Similar to categorical membership, the designer of a visualization may ask a user to *predict clusters* in a scatterplot or network diagram. For example, given an interactive network diagram, the user can draw a contour around the nodes or use brushing interactions similar to those described for categorical membership to designate clusters of related nodes. *Predicting connectivity* could also be applied to support edge prediction in a network diagram.

Alternatively, the designer can ask the user to visualize her expectations for the outcome of more complex analyses applied to raw data. For example, graphical prediction techniques could be used to elicit predictions on multivariate correlations, uncertainty, or other results attained through statistical modeling. For example, the New York Times 'You Draw It' interactive prompts the user to *predict a regression line* representing the relationship between parents' income percentile and percent of children who attended college [2]. To facilitate understanding of uncertainty in data, a user might be prompted to *predict a confidence interval or region* given a 2D presentation of bars, points, or lines denoting sample statistics.

Regardless of what is predicted, *the directness of the prediction interaction and degrees of freedom* provided to the user by the interface are important design considerations. Freeform interactions can be used to allow the user greater flexibility in drawing, such as providing a high degree of resolution (i.e., space of possible fits) to users drawing regression lines in 2D visualizations. More constrained forms of interaction can be realized through snapping functions (e.g., snapping a predicted regression line to the nearest grid point). Similarly, a designer may choose to only allow the user to manipulate certain parameters of components (e.g., drag a curve or slider to change line curvature, drag the edge of a circle to increase size while maintaining shape). While more constrained interactions may serve to reduce error and focus user's attention on key parameters (e.g., slope or magnitude alone), they are likely to add abstraction. Our own experimentation with interactive prototypes suggest that many users enjoy the novelty of using an interactive visualization interface to draw with few constraints.

Contextualization Mechanism

Contextualization mechanisms can be used to guide the user's prediction as they form a guess. For example, the amount of effort required for the user to make a guess can differ based on the number of *reference marks* (e.g., dots, bars, lines) that the visualization initially presents. In our study, we presented two of the three ethnic groups by default, which provided cues to guide users' predictions of the remaining group. How much data to reveal through reference marks can be decided based on how familiar users are expected to be with the dataset (less familiar=more reference marks). Or, the reference marks can be selected dynamically through personalization. For example, for datasets that depict regional data, identifying and initially presenting marks depicting the user's region based on their IP address may increase engagement while providing useful context for the user.

Prediction hints provide more direct guidance, either through text or visual annotation, on where a user's prediction should be made, helping educate users about the meaning of the encodings. As the New York Times' visualization "You Draw it" [2] demonstrates, one or more data points can be presented as a hint that the user's prediction line should pass through.

Designers should consider the scale of the x and y-axes in 2D charts. In piloting our preliminary study for choosing a dataset, we observed that users' predictions were quite sensitive to the axis range. When we presented the full 0-100% percentage range for percentage variables (e.g., the percentage of the U.S national budget of health care), users' estimates showed a bias toward the center in the plotting range. This effect was lessened when we trimmed the axis range based on the maximum value of the dataset, suggesting that users implicitly view the axis range as a clue to the data scale.

Feedback Technique

After a user draws her prediction, feedback on how the user's prediction compares to the observed data can help prompt reflection on prior knowledge.

For example, personalized feedback can provide information on the *accuracy of a prediction*, as we provided in our study. Feedback may take the form of aggregated, quantified accuracy information (e.g., "Overall, you were 80% right in guessing the amount of CO_2 emission in U.S") or information on the directionality of biases (e.g., "You over-estimated the overall trend."). Feedback may also occur at a more granular level, encouraging the user to adjust her expectations of individual data points: "Your guess on year 2001 was 6 points off; a little higher and you would be correct."

As users' predictions are collected, *social feedback* may serve to further engage users to think about the data and their own expectations. However, social feedback may also overshadow user's own interpretations; hence social feedback might be withheld until after the user has provided their own prediction [26] The interface can prompt social comparisons by visualizing other users' predictions alongside the user's.

General design considerations affecting feedback include how specific and in what modality feedback is presented (e.g., visual, text, etc.). Based on our study finding that textual presentations are less effective for drawing attention to the gap, we except visual feedback or a combination of visual and text feedback to be more powerful. Animating feedback, such as by dynamically moving marks added by the user to their true positions, or adding textual feedback to prediction errors point by point, may be particularly effective for drawing a user's attention to the gap.

Limitations and Future Work

Our study focused on data with a particular structure, chosen to allow us to examine effects for both absolute data values and trends. We also deliberately chose to test on two sets of visualization with 12 data points to ensure that the recall task was challenging. Future work should evaluate the same techniques applied to other datasets with varying formats and sizes. In our study, all predict conditions predicted one group (e.g., ethnicity) out of three. We did not systematically vary the contextualization mechanism (e.g., number of reference marks provided). However, we suspect that the effects of the techniques can be increased or decreased by adding or removing forms of context.

Elicitation techniques may provide an engaging way to helping novices to learn about data analysis and statistics. Our future work will test whether people are more likely to interact with the interface longer or in more engaged manner, if they prompted to predict and see feedback [32].

Finally, eliciting users' predictions of data opens the possibility for more sophisticated forms of modeling of users' evolving mental conceptions as they use an interactive visualization. As the interface maintains a model of the user's current knowledge, dynamic suggestions can be provided on what content to view next, or on how to correct errors in mental conceptions before continuing to explore data. Developing natural and effective elicitation interfaces and studying how elicitation and feedback can be incorporated into existing visual analysis pipelines are important tasks for future work.

CONCLUSION

Our work began by asking, "What if visualizations integrated users' prior knowledge about the dataset?" Informed by prior work in cognitive and educational psychology, we developed multiple novel elicitation techniques for incorporating users' prior knowledge in visualization interaction. We tested the effects of these techniques, including eliciting users' predictions of data, presenting personalized feedback on predictions, and explicitly prompting self-explanations with a visualization. We observed that providing opportunities for users to interact with their prior knowledge improves recall of data values, and is more powerful when used with visualization than with text. Our findings pave the way for a new paradigm of interactive visualization that enables users to interact with their internal representations to deepen their understanding of data.

REFERENCES

1. Shaaron Ainsworth and Andrea Th Loizou. 2003. The effects of self-explaining when learning with text or diagrams. *Cognitive Science* 27, 4 (2003), 669–681.

2. Gregor Aisch, Amanda Cox, and Kevin Quealy. 2015. You Draw It: How Family Income Predicts Children's College Chances. *The New York Times*, May 28, 2015, *http://nyti.ms/1ezbuWY*,. (2015).
3. Vincent Aleven and Kenneth R Koedinger. 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive science* 26, 2 (2002), 147–179.
4. Vincent Aleven, Octav Popescu, and Kenneth R Koedinger. 2001. Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of Artificial Intelligence in Education*. Citeseer, 246–255.
5. Katerine Bielaczyc, Peter L Pirolli, and Ann L Brown. 1995. Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and instruction* 13, 2 (1995), 221–252.
6. Sally Bogacz and J Gregory Trafton. 2002. Understanding static and dynamic visualizations. In *International Conference on Theory and Application of Diagrams*. Springer, 347–349.
7. Matt Canham and Mary Hegarty. 2010. Effects of knowledge and display design on comprehension of complex graphics. *Learning and instruction* 20, 2 (2010), 155–166.
8. Patricia A Carpenter and Priti Shah. 1998. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied* 4, 2 (1998), 75.
9. Andrea Cheshire, Linden J Ball, and CN Lewis. 2005. Self-explanation, feedback and the development of analogical reasoning skills: Microgenetic evidence for a metacognitive processing account. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society, ed. BG Bara, L. Barsalou & M. Bucciarelli*. Citeseer, 435–41.
10. Michelene TH Chi. 2000. Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in instructional psychology* 5 (2000), 161–238.
11. Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science* 13, 2 (1989), 145–182.
12. Michelene TH Chi, Nicholas Leeuw, Mei-Hung Chiu, and Christian LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive science* 18, 3 (1994), 439–477.
13. Cristina Conati and Kurt VanLehn. 1999. Teaching meta-cognitive skills: Implementation and evaluation of a tutoring system to guide self-explanation while learning from examples. In *Artificial Intelligence in Education*. IOS Press, 297–304.
14. Cristina Conati and Kurt Vanlehn. 2000. Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education (IJAIED)* 11 (2000), 389–415.
15. Leda Cosmides and John Tooby. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *cognition* 58, 1 (1996), 1–73.
16. Richard Cox. 1999. Representation construction, externalised cognition and individual differences. *Learning and instruction* 9, 4 (1999), 343–363.
17. Ruth B Ekstrom, John W French, Harry H Harman, and Diran Dermen. 1976. Manual for kit of factor-referenced cognitive tests. *Princeton, NJ: Educational testing service* (1976).
18. Monica GM Ferguson-Hessler and Ton de Jong. 1990. Studying physics texts: Differences in study processes between good and poor performers. *Cognition and Instruction* 7, 1 (1990), 41–54.
19. Carolina E Hagberg, Annika Mehlem, Annelie Falkevall, Lars Muhl, Barbara C Fam, Henrik Ortsäter, Pierre Scotney, Daniel Nyqvist, Erik Samén, Li Lu, and others. 2012. Targeting VEGF-B as a novel treatment for insulin resistance and type 2 diabetes. *Nature* 490, 7420 (2012), 426–430.
20. Mary Hegarty. 2004. Diagrams in the mind and in the world: Relations between internal and external visualizations. In *Diagrammatic representation and inference*. Springer, 1–13.
21. Mary Hegarty and Marcel-Adam Just. 1993. Constructing mental models of machines from text and diagrams. *Journal of memory and language* 32, 6 (1993), 717–742.
22. Mary Hegarty and Sarah Kriz. 2008. Effects of knowledge and spatial ability on learning from animation. *Learning with animation: Research implications for design* (2008), 3–29.
23. Mary Hegarty, Sarah Kriz, and Christina Cate. 2003. The roles of mental animations and external animations in understanding mechanical systems. *Cognition and instruction* 21, 4 (2003), 209–249.
24. Mary Hegarty and Kathryn Steinhoff. 1997. Individual differences in use of diagrams as external memory in mechanical reasoning. *Learning and Individual differences* 9, 1 (1997), 19–42.
25. Dianne E Howie and Kim J Vicente. 1998. Making the most of ecological interface design: The role of self-explanation. *International Journal of Human-Computer Studies* 49, 5 (1998), 651–674.
26. Jessica Hullman, Eytan Adar, and Priti Shah. 2011. The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1461–1470.

27. Sirkka L Jarvenpaa. 1990. Graphic displays in decision making the visual salience effect. *Journal of Behavioral Decision Making* 3, 4 (1990), 247–262.
28. Wilson Andrews Josh KatZ and Jeremy Bowers. 2014. Elections 2014: Make Your Own Senate Forecast. *The New York Times*, Sep 2, 2014, *http://nyti.ms/1plfIyv*,. (2014).
29. Joshua Klayman and Young-Won Ha. 1987. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review* 94, 2 (1987), 211.
30. Stephen M Kosslyn. 1989. Understanding charts and graphs. *Applied cognitive psychology* 3, 3 (1989), 185–225.
31. Jill H Larkin and Herbert A Simon. 1987. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science* 11, 1 (1987), 65–100.
32. Edith Law, Ming Yin, Kevin Chen Joslin Goh, Michael Terry, and Krzysztof Z Gajos. 2016. Curiosity Killed the Cat, but Makes Crowdwork Better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4098–4110.
33. Zhicheng Liu and John T Stasko. 2010. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *Visualization and Computer Graphics, IEEE Transactions on* 16, 6 (2010), 999–1008.
34. George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75.
35. Richard E Mayer. 2014. Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning* 43 (2014).
36. Danielle S McNamara, Tenaha OâĂŹReilly, Michael Rowe, Chutima Boonthum, and IB Levinstein. 2007. iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. *Reading comprehension strategies: Theories, interventions, and technologies* (2007), 397–421.
37. Hedwig M Natter and Dianne C Berry. 2005. Effects of active information processing on the understanding of risk information. *Applied Cognitive Psychology* 19, 1 (2005), 123–135.
38. Steven Pinker. 1990. A theory of graph comprehension. *Artificial intelligence and the future of testing* (1990), 73–126.
39. Peter Pirolli and Margaret Recker. 1994. Learning strategies and transfer in the domain of programming. *Cognition and instruction* 12, 3 (1994), 235–275.
40. Alexander Renkl. 1997. Learning from worked-out examples: A study on individual differences. *Cognitive science* 21, 1 (1997), 1–29.
41. Alexander Renkl, Robin Stark, Hans Gruber, and Heinz Mandl. 1998. Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary educational psychology* 23, 1 (1998), 90–108.
42. Marguerite Roy and Michelene TH Chi. 2005. The self-explanation principle in multimedia learning. *The Cambridge handbook of multimedia learning* (2005), 271–286.
43. R Ryan. 1996. Self-explanation and adaptation. *Psychology* (1996).
44. Priti Shah, Richard E Mayer, and Mary Hegarty. 1999. Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology* 91, 4 (1999), 690.
45. Keith Stenning and Jon Oberlander. 1995. A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive science* 19, 1 (1995), 97–140.
46. Elsbeth Stern, Carmela Aprea, and Hermann G Ebner. 2003. Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction* 13, 2 (2003), 191–203.
47. J Gregory Trafton, Susan B Trickett, and Farilee E Mintz. 2005. Connecting internal and external representations: Spatial transformations of scientific visualizations. *Foundations of Science* 10, 1 (2005), 89–106.
48. Jeff Zacks and Barbara Tversky. 1999. Bars and lines: A study of graphic communication. *Memory & Cognition* 27, 6 (1999), 1073–1079.