DOI: 10.1111/cgf.13009

Pacific Graphics 2016
E. Grinspun, B. Bickel, and Y. Dobashi
(Guest Editors)

Volume 35 (2016), Number 7

A Study On Designing Effective Introductory Materials for Information Visualization

Yuzuru Tanahashi, Nick Leaf, and Kwan-Liu Ma

University of California, Davis

Abstract

Designing introductory materials is extremely important when developing new information visualization techniques. All users, regardless of their domain knowledge, first must learn how to interpret the visually encoded information in order to infer knowledge from visualizations. Yet, despite its significance, there has been little research on how to design effective introductory materials for information visualization. This paper presents a study on the design of online guides that educate new users on how to utilize information visualizations, particularly focusing on the employment of exercise questions in the guides. We use two concepts from educational psychology, learning type (or learning style) and teaching method, to design four unique types of online guides. The effects of the guides are measured by comprehension tests of a large group of crowdsourced participants. The tests covered four visualization types (graph, scatter plot, storyline, and tree map) and a complete range of visual analytics tasks. Our statistical analyses indicate that online guides which employ active learning and the top-down teaching method are the most effective. Our study provides quantitative insight into the use of exercise questions in online guides for information visualizations and will inspire further research on design considerations for other elements in introductory materials.

Categories and Subject Descriptors (according to ACM CCS): H.5.m [Information Interfaces and Presentation (e.g. HCI)]: — Miscellaneous

1. Introduction

Information visualization is a rapidly evolving field, and must be to keep pace with the increasing complexity of available data. Recent research [LKH*16] has shown that parsing an unfamiliar visualization without some form of guide requires an extensive thought process and may cause some users to flounder and give up entirely. Demanding that visualizations be self-explanatory to any potential user would unacceptably limit either the target audience or the visualization itself—i.e. the author of a visualization often wants to reach new (potentially novice) users with new visualization techniques. Thus, a visualization author may have to include introductory materials that teach their users how to use the visualization to derive insights about the underlying data.

The introductory materials for a visualization usually begin with some sort of text description. This text is often supplemented by one or more questions that help the user engage with the visualization, and which are designed to guide the user's learning process. The text-plus-questions format is equally applicable to static, dynamic, passive, and interactive visualizations. We refer to introductory materials using the text-plus-questions format as InfoVis Guides (IVGs) in the context of this paper. In order to ensure that their visualization successfully conveys the meaning inherent in the data, the author requires an effective IVG.

This paper describes a study on the qualities of different IVGs and their quantitative impact on user comprehension of visualizations. Our IVGs were developed using two concepts from educational psychology: teaching method and learning type. Each IVG utilizes either a top-down or bottom-up teaching method, and caters to either the active or passive learning type. We use these concepts to vary the exercise questions, yielding four different IVG types based on the combination of teaching method and learning type, plus one extra IVG type without exercise questions to serve as a control.

Our study covers four types of visualization—scatter plot, graph, storyline [TM12], and tree-map [Shn92]. Some of the visualizations are fairly new (e.g. storyline), and we would not expect a general audience to be more than vaguely familiar with them. While the study participants likely have some experience with the more traditional visualizations (e.g. scatter plots), our results show that their understanding of the visualizations was not perfect, leaving room for IVGs to further increase user comprehension. Moreover, the more traditional visualizations often serve as the building blocks for new visualizations, so that guidelines for how to teach users about the traditional visualization types can likely be extended to guidelines for teaching new visualizations built upon them.

Our data was collected from a large number of crowd-sourced participants. Each participant was given a task that can be logi-

cally divided into a learning segment and a testing segment. The participants were evenly distributed between combinations of the five IVG types and four visualization types. The collected data was then analyzed using a robust set of statistical tests. Our analyses show that, in general, participants exposed to IVGs designed to use the top-down teaching method and catering to the active learning type showed the greatest improvement during the test segment. Ultimately, this suggests that visualization authors should design their IVGs to utilize the top-down and active concepts to maximize the benefit for their users.

2. Related Work

Research on the more general form of IVGs, computer-based tutorials, can be split into two overarching categories: skill-training, and knowledge development. There is a large variety of topics in skill-training tutorials, including new forms of tutorial presentation techniques [HT07, KPS03], evaluating the effects of different tutorial formats [Har95, PEB91], and automating the generation of tutorials [CAR*12, RHAH11, WCC*14]. More recent studies on skill-training tutorials have also explored the benefits of using crowd-sourced comments for incorporating contextual information into the tutorial [BDL*14, LBLT13, LB14]. These kinds of tutorials are more common in the literature, and focus more on the acquisition of a specific skill.

Knowledge-development tutorials focus on the development of the theoretical or practical understanding of a subject rather than the cultivation of a skill. While the research on knowledge-development tutorials is less common, some studies suggest that these types of tutorial are beneficial to users. For example, [SOM96] presented a case study which showed students given supplementary computer-based tutorials for physics course lectures performed better than students who used the lectures alone. The tutorial discussed in this study provided the students the ability to explore and learn the different aspects of the subject at their own paces. Steinberg's study suggests that computer-based knowledge-development tutorials can significantly improve students' understanding of the subject matter.

Discussion of computer-based knowledge-development tutorials can also be found in the literature on E-learning [Gar11]. E-learning is a field of study that explores different methods of education and knowledge development with the support of computers and other digitized media [MM98]. Within this vast field of study, there have been a few studies that discuss knowledge-development techniques and designs that may enhance computer-based tutorials. Berthold and Renkl [BR09] investigated the effects of incorporating different assisting features to multiple (external) representations (MERs) when teaching students the concept of probability. In their study, they found that, if applied correctly, a variety of features were effective in enhancing the student's conceptual understandings. In a different study, Boyer *et al.* [BPW*08] studied what characteristics of tutorial dialogs between two humans improve learning and found that positive feedback, such as reassuring and encouraging comments, can improve learning. Based on their findings, they derived design implications for improving computer-based tutorial dialogs. Both of these studies suggest that incorporating different learning techniques can further improve knowledge-development when using computer-based tutorials.

Computer-based knowledge-development tutorials, or IVGs in our context, are also common as a means of teaching new information visualizations. IVGs for new information visualizations can clarify the purpose of the visualization, orient users within its visual landscape, and avoid potential misinterpretations of the displayed information. Despite the potential benefits of IVGs, there has been limited research on how to design effective IVGs that implement useful learning techniques.

Our research focuses on the effects of incorporating practice testing, one of the most effective learning techniques [DRM*12], in IVGs. This particular learning technique is commonly practiced in IVGs in the form of example demonstrations of data analytics scenarios to supplement the base-line description of the visual language. Different IVGs present these example cases in different styles. Some styles can be more effective than others. However, there has not yet been a study on what approach is most effective in teaching novice users about the visualization. We believe that the findings in this study will guide future visualization developers on how to design IVGs that can effectively communicate their advancements to a wide range of potential users.

3. Research Design

The goal of the user study was to measure and compare the impact of different IVG types on the participants' understanding of the visualization techniques. Towards this end, we conducted a large-scale user study using participants sourced from Amazon's Mechanical Turk program. Each participant's task was divided into two segments. The learning segment presented the participant with a visualization and accompanying IVG. The test segment presented visualizations and required the participant to answer a series of questions about them. The first five questions were devoted to the same visualization that the participant saw during the learning segment. The remaining five questions covered a new visualization of the same type, but using different data. The participants were evenly divided to cover every combination of IVG and visualization type. A diagram of the study process that each participant completed is shown in Figure 1. Full study materials and an example IVG plus comprehension test is avaialable online at https://vidi.cs.ucdavis.edu/Projects/infovis_guide_design.

3.1. IVG Designs

Each of the IVGs that we tested contained a visualization, which varied depending on the visualization type being tested against, and two or four slides. The example visualization is shown throughout the IVG, the first two slides contain text describing the example visualization, and the last two slides, if present, contain one exercise question, each. The first slide's text explains the visualization technique in general and how information is visually encoded. The second slide explains the example visualization more specifically, including what pieces of information are currently visible and what type of interaction is supported. These two slides are the same for all IVGs which test the same visualization type. Each of the last two

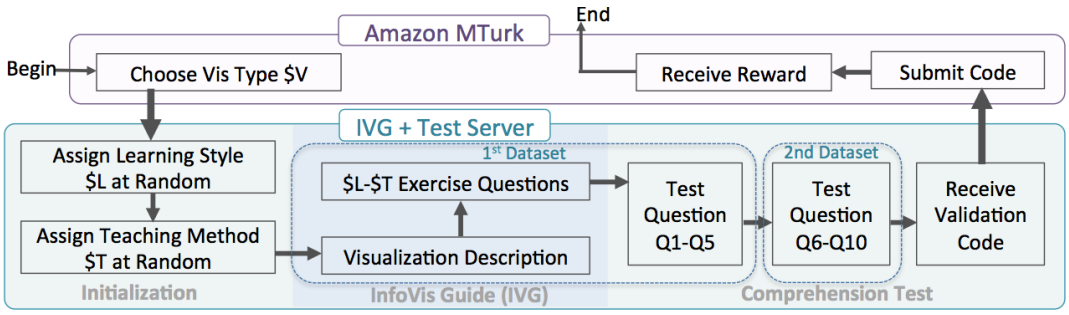

Figure 1: Flow chart of the participant's task. There was one Human Intelligence Task (HIT) for each visualization type **$V**, and it was available until the desired number of participants (200) completed it. After selecting their HIT, the participants were assigned an IVG with some combination of learning type **$L** and teaching method **$T**. The participants were uniformly distributed amongst the possible IVGs so that there were 40 participants for each. The participants are shown a visualization of type **$V** throughout the IVG and for the first five test questions. For the last five questions, they were shown a type **$V** visualization of a different dataset.

slides contains an exercise question relating to the example visualization, which is meant to reinforce the participants' understanding of the visualization. The presentation style and content of these exercise questions varied depending on the combination of learning type and teaching method employed for the IVG. The control IVG had no exercise questions, and contained only the two text slides.

We strove to maintain a balanced approach when designing the IVGs. We could have easily increased the amount of introductory text or added more exercise questions to better educate the participants. Increasing the length of the IVGs has the potential to hurt participant engagement, however. Longer IVGs might even cause potential participants to avoid our task, which carries a risk of biasing our demographic. This risk carries over into real-world scenarios, where a visualization author hoping to capture a general audience must be stingy with their participants' time and attention. Thus, our study focuses on how to achieve the greatest teaching impact within a constrained time budget.

3.1.1. Learning Type

Learning type denotes the degree of active participation from the student when learning new concepts. There are two primary learning types, passive and active [FEM*14]. Passive learning indicates that the students are only receiving the information and that there is no participatory dialog between the student and the IVG. Active learning, on the other hand, indicates that the participant will need to actively participate in a corresponding dialog in order to proceed with the lesson.

In our IVGs, these two learning types determine whether the student needs to actively respond to exercise questions. For active IVGs, the students are given exercise questions and prompted to select a multiple (four) choice response. When the student selects an incorrect answer, the IVG displays a hint for solving the question and highlights the correct answer from the multiple choice. Once the student clicks on the correct response, the IVG proceeds to the next page. In contrast, passive IVGs display the hint and correct answer immediately. The participant is not required to participate in an active dialog, and can simply read through the exercise.

3.1.2. Teaching Method

The teaching method determines whether the student is taught in a bottom-up or top-down method [Opr94, ZKDC11]. Bottom-up is an inductive method that focuses on small, detailed pieces of information which the student then incorporates together for comprehensive understanding. Top-down teaching is a deductive method which presents a broad overview to help students understand the abstract, high-level parts of an idea which then provide context for understanding its components in detail.

These two methods determine the content of the exercise questions in our IVGs. The bottom-up exercises focus on interpreting information about individual visual entities. The top-down IVGs, in contrast, ask participants to draw more advanced, less direct inferences from the data, and are designed with the presumption that the participant can already interpret the visual representations. The visualization task categorization presented in Section 3.2 is used to classify bottom-up and top-down visualization tasks.

3.1.3. IVG Types

The combinations of learning type and teaching method determine the five unique IVG types used in our study, shown in Figure 2. The first four IVG types all contain two exercise questions, with the learning type determining whether they require participant interaction or not, and the teaching method determining whether the questions are based on basic tasks or advanced tasks. Figure 3 shows examples of an exercise question for scatter plot visualization. All IVGs for the same visualization type share a common dataset, while how the participants proceed to the next page and the contents of the questions depend on the IVG's learning type and teaching method. The NoEx IVG serves as a control, and contains only the introductory description without any accompanying exercise questions.

3.2. Visualization Task Categories

Brehmer and Munzner recently presented a typology of abstract visualization tasks [BM13] which defined information query as the

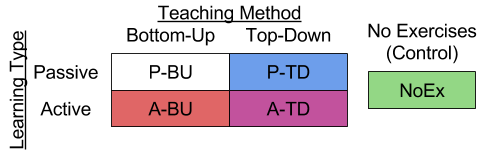

Figure 2: Info Vis Guide (IVG) Types. Excercise questions are formulated using a combination of learning type and teaching method to create the four categories tested in our study. A fifth categorie containing no exercises, NoEx, is used as a control.

one fundamental participant activity in the application of information visualization. In their study, they further divided this information query into three distinct categories: *identify*, *compare*, and *summarize*. *Identify* involves decoding a specific piece of information, typically from a single entity in the visualization. Queries in this category will tend to use information in a local fashion, and will require a fairly precise result. *Compare* involves determining the relative value or finding differences between a small set of points. *Summarize* corresponds to inferring broad trends and distributions from all or a large subset of the visualization entities, and thus has a more global scope.

This task categorization concept was applied to both the IVG and the test designs. Exercises using different teaching methods have natural analogs to different task categories. *Identify* tasks are inherently bottom-up, requiring the participant to draw information from the most fundamental parts of the visualization. *Compare* and *summarize* tasks require the participant to infer more abstract knowledge, and are thus naturally top-down in nature. When designing our tests, we made sure to cover all task categories to test the participant's comprehension of all aspects of the visualization.

3.3. Comprehension Test Design

The test segment consisted of ten multiple choice questions, each with four possible responses. Figure 4 shows examples of these test questions. The ten questions were split into two sets, each dedicated to a particular dataset. We chose ten as a reasonable compromise between the time required to take the test and the degree of thoroughness we desired. Our previous experiences suggest that participants become discouraged when the time required to complete the test exceeded their expectations. Too long of a test can cause participants to hurry, which reduces the quality of the resultant data. The questions varied according to visualization type, but not IVG. Thus, there were four sets of questions in total, and the all of the tests for a given visualization type used the same set of questions, regardless of the IVG.

The number of multiple choice options was similarly chosen as a compromise. We provided four choices for each question in order to lower the expected score of complete guessers without overwhelming our participants with options. Limiting the number of choices also helps to guide the participants to promising lines of inquiry; more open-ended questions with too many choices often lead participants to dead ends, which can discourage them from performing well. The first dataset that participants encounter in the test is the same one used in the IVG. We switch to a visualization produced by the same technique on a similar but distinct dataset for the second half of the questions. This is done to make sure that the IVG successfully explained how to interpret the visualization technique, and did not simply teach the participant how to interpret a single visualization. Participants were shown just one question at a time. As soon as they selected their answer, they were shown the next question. They were unable to revisit previous questions, nor were they able to review their answers once they were submitted. We chose not to show the participants their final answers to limit in-test learning and reduce cheating, opting instead to show them their final score at the end of the test.

We could have chosen to gather data about the participants' prior knowledge of the visualization that they were being tested on. A questionnaire which allowed participants to self-report their level of comfort with the visualization technique runs afoul of problematic biases. How would novice participants know enough to correctly gauge their own understanding? Issuing a separate test before the IVG may cause extra learning, or even give participants a false understanding of the visualization. Both methods would increase the time requirements of the study, as well, which was a constraint that we tried to minimize. Ultimately, we chose not to collect prior knowledge data.

4. User Study

We drew our study participants from Amazon's Mechanical Turk (MTurk), which gave us access to a large population with a broad range of backgrounds. MTurk is a popular web site that allows *requesters* to post simple jobs that can be carried out online. These jobs are referred to as *Human Intelligence Tasks* (*HITs*) and a set number of registered *workers* can perform these HITs in order to receive a monetary reward.

For our user study, we collected results 200 participants for each visualization type, for a total of 800 (after excising invalid results as described in Section 4.1). The groups of 200 participants were further divided into five groups of 40 participants. Each group of 40 participants was assigned a particular IVG type, either one of the four with exercises or the control, NoEx. The participants were randomly assigned to a group upon accessing the online test page.

We collected the following data for each participant:

- the participant's answer to each question,
- the time the participant took to answer each question, and
- the time the participant took to complete the IVG.

Note that even though the IVG and per-question completion times were captured, they were not included in our analyses. We did not control the testing environment of the participants beyond the web page that they accessed, and thus had no knowledge about unrelated factors (e.g. distractions or slow connection speed) which might increase a participant's response time. The completion time metrics that we captured were only used to filter the participants, as described in section 4.1.

4.1. Mechanical Turk Configuration

The standard task designs available on Mechanical Turk do not support tests that use custom JavaScript. Therefore, we used the "ex-

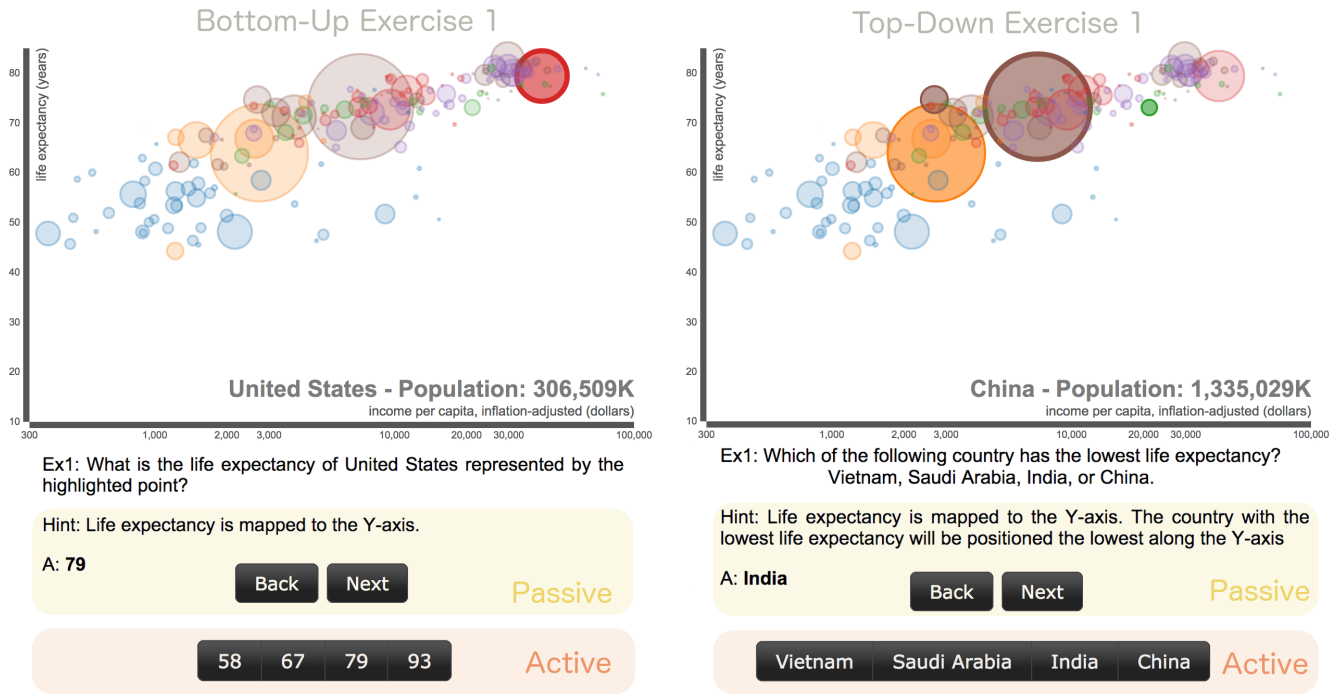

Figure 3: Examples of exercise questions used for scatter plot IVGs. The left image shows an *identify* question, which is used in bottom-up IVGs. The right image shows an example *compare* question that would be used in a top-down IVG. The texts and buttons below the visualization show examples of the passive and active exercise types. The same exercise questions are provided for both learning types, but the required interaction differs. The active exercises required the participant to select one of the four possible answers, and notified them whether they were correct. The passive exercise, however, immediately gave the answer to the participant and allowed them to move to the next slide without any further action.

ternal HIT" hosting method to host the test on our own web server. Each unique visualization type was set up with its own web page for testing and was assigned its own HIT project.

We applied a scaling incentive system where participants were paid more for correct answers to encourage participant engagement. The reward was set at \$0.05 for 0–3 correct answers, and scaled quadratically to \$0.80 for 9-10 correct answers. The HIT expiration time was set to 20 minutes. Participants at the top end of the scale were therefore were rewarded at a rate close to \$2.40 per hour, which is significantly higher than the reward rate of \$1.66 per hour suggested by [PCI10]. On the other hand, poor performance could result in a rate as low as \$0.15 per hour.

Note that the expected score produced by simple guessing, 2.5 points, falls within the lowest reward range. This significantly reduced, but did not eliminate, the number of participants who simply clicked through the test questions to finish the HIT as quickly as possible. The few participants who did click through the test were easy to excise from the data—they spent only a fraction of a second on many questions which most participants required at least 10 seconds to answer. The data collected from these participants were manually removed from the results, and valid results were obtained from new participants. We also relied on the MTurk rating system and winnowed the pool of participants to those with at least 50 HITs completed and that hold an approval rate greater than or equal to 95%.

5. Results

Figures 5 and 6 show the summary of the collected results. Figure 5 shows the average test scores for each combination of IVG type and visualization. Figure 6 shows four line charts each depicting the rates of correct participant response for individual questions. Here, the question numbers are mapped to the x-axis, and correct response rates are mapped to the y-axis. Vertical strips of white, yellow, and pink respectively indicate question categories *identify*, *compare*, and *summarize*. Based on these test scores, we carried out three different analyses: 1) to examine whether incorporating exercise questions actually improve participant performance, 2) to derive design implications for effective IVGs regarding learning types and teaching methods, and 3) to investigate what visualization task categories are most affected by IVG types.

5.1. Analysis 1: Effects of Exercise Questions

In order to examine whether there is any statistical significance to how different IVG types (i.e., exercise questions) affect participant performance, we applied one-way analysis of variance (ANOVA)

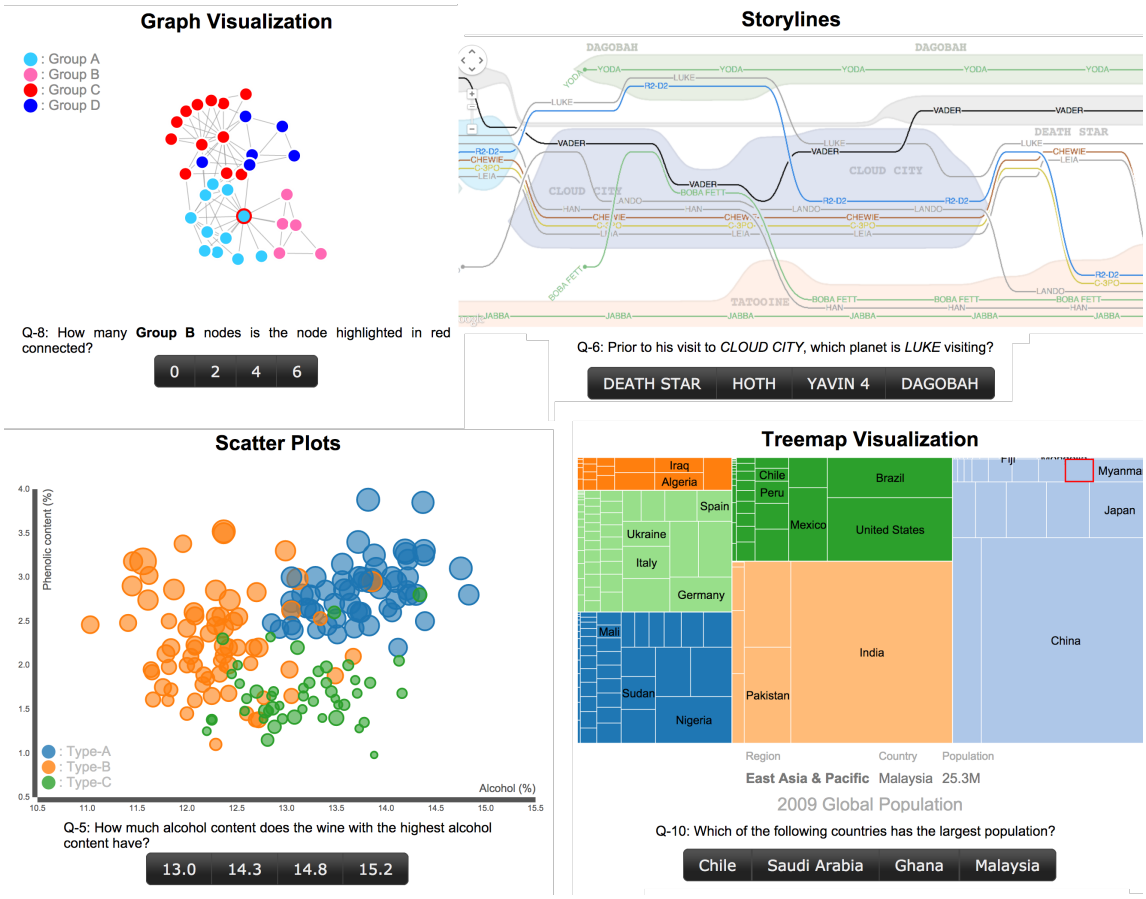

Figure 4: Test segment examples. Each visualization type–graph (top-left), storyline (top-right), scatter plot (bottom-left, and tree map (bottom-right)–is shown, along with example test question. Below, we give an example of the expected solution process for each question.

- Graph: This is an example of a *compare* question, since it pertains to the relationship between the highlighted node and all of its Group B neighbors. The participant is expected to find the highlighted node, recognize the members of Group B by the color given in the legend, and count how many of them are connected to the highlighted node. There are three obvious connections, but they must also recognize the connection that is hidden due to over-plotting in order to arrive at the correct answer—4.
- Storyline: This test question is an *identify* question because the participant must identify the membership value at a single point. To correctly answer this question, the participant must find the particular storyline for Luke in Cloud City, and follow it back to the previous location, and then find the label for that location, which is Dagobah.
- Scatter plot: This example is an *identify* question which requires the participant to identify the value of a particular point. In it, the participant is expected to find the axis for alcohol %, and determine which node lies furthest to the right on that axis. After they have located the node, they must estimate its value from its location on the axis. The highest alcohol content in this dataset is 14.8%.
- Tree map: This test example is a *compare* question. The participant must find which of the four indicated countries has the largest area. Saudi Arabia was the correct answer at the time of the study.

on the test scores for each visualization type. Based on previous research [DRM*12], our hypotheses were:

H1a. IVG types will have a significant effect on participant performance,
H1b. IVGs without exercise questions are least effective.

The results of these analyses suggested that IVG type has a significant effect on participants' test performances for scatter plots ($F_{4,195} = 3.09$, $p = 0.017$) and storyline visualizations ($F_{4,195} = 2.50$, $p = 0.044$), confirming our first hypothesis, H1a, for these two visualizations. However, the extended analyses based on post-hoc Tukey tests revealed that A–TD was the only IVG type that was more effective ($\alpha = 0.05$) than NoEx for scatter plots and P–BU for scatter plots, storylines, and tree maps. Hence, though the results implicated that A–TD may be the best IVG, they did not confirm our second hypothesis, H1b, about NoEx being the least effective. Moreover, these results implicated a strong possibility that P–BU may be the least effective IVG type for teaching visualizations to novice users.

We also noticed that the difference in IVG types had little effect on the test scores for graph visualization. Although this would re-

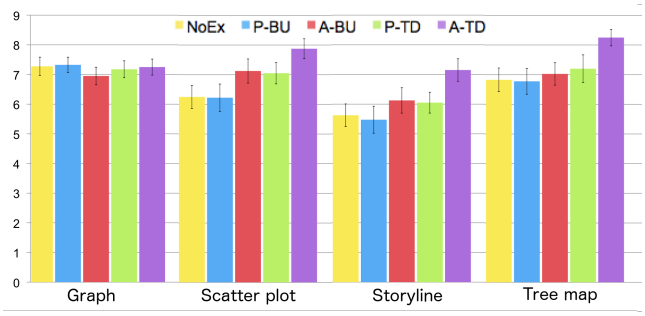

Figure 5: Average test scores for each visualization and IVG. The error bar shows standard error. The scores for the graph visualization are fairly even for all visualization types, which we attribute to the participants likely being more familiar with graphs. The A–T-D IVG shows a clear improvement in average score for the other visualization types.

quire further investigation, we believe that this consistency in user performance is because the participants were already familiar with the graph visualization, leaving little room for the exercise questions to improve the participants' comprehension. For example, star constellations [Goo] and train/subway maps [GB94] are both very similar to graph visualizations and are integrated with people's everyday lives across various cultures. This prior knowledge of similar representations may have lent the participants an intuitive understanding of the visualization, necessary for a variety of analytic tasks [Shu90], before they completed the IVG.

5.2. Analysis 2: Learning Type and Teaching Method

We next applied two-way ANOVA with respect to the two defining attributes of the IVGs, learning type and teaching method, in order to analyze the relative effect of the IVG types. The analysis was applied for each type of visualization. Based on previous research [FEM*14, ZKDC11] and the results of Analysis 1, we defined the following hypotheses:

H2a. Active IVGs will be more effective than Passive IVGs,
H2b. Top–Down IVGs will be more effective than Bottom–Up IVGs,
H2c. The interaction between learning type and teaching method will be significant, and that the A–TD IVGs will be the most effective.

Tables 1, 2, 3, and 4 show the results of these analyses for the graph, scatter plot, storyline, and tree map visualizations. In these tables, a p value in bold font indicates a statistical significance at the $\alpha = .05$ level.

Here, the results from the two-way ANOVA for the graph visualization (Table 1) showed no evidence that IVG designs affect test performance. This is expected, since our previous analysis had already indicated that IVG type had no significant effect on test scores.

The analysis results for other visualization types tell a different story. The results for the scatter plot visualization (Table 2) indicate that both attributes, learning type and teaching method, have significant effects on the participants' performance. Here, the test

Table 1: Two-way ANOVA of IVG Types (Graph)

	SS	df	MS	F	p
Learning	0.90	1	0.90	0.29	.59
Teaching	0.23	1	28.18	0.07	.79
Learning:Teaching	2.02	1	2.02	0.65	.42
Error	483.95	156	3.10		
Total	487.10	159			

Analysis 1 showed that the IVG type has little effect on the scores for the graph visualization. Analysis 2 for the graph visualization, using two-way ANOVA with respect to learning type and teaching method, confirms this result. We suspect this is due to participants' prior knowledge of graph visualization from interaction with graphs in common real-world settings (e.g. subway maps).

Table 2: Two-way ANOVA of IVG Types (Scatter Plot)

	SS	df	MS	F	p
Learning	29.76	1	29.76	4.82	**.030**
Teaching	24.81	1	24.81	4.02	**.047**
Learning:Teaching	0.04	1	0.04	0.01	.921
Error	963.63	156	6.18		
Total	1018.24	159			

The two-way ANOVA for scatter plots shows that both learning type ($p = .030$) and teaching method ($p = 0.047$) had a statistically significant effect on participant performance in the comprehension test. There was no statistical evidence ($p = 0.921$) to suggest that the interaction between these two attributes had any additional effect, however.

scores for active learning ($\mu_A = 7.5$) showed an improvement over the scores for passive learning ($\mu_P = 6.6$), and top-down teaching ($\mu_{TD} = 7.5$) also improved the test scores compared to bottom-up teaching ($\mu_{BU} = 6.7$). The analysis results for storyline visualization (Table 3) also indicated that the test scores for active learning ($\mu_A = 6.6$) were significantly higher than the scores for passive learning ($\mu_P = 5.8$). The results for the tree map visualization (Table 4) tell a similar story, with the top down IVG ($\mu_{TD} = 7.7$) scoring significantly higher than bottom-up ($\mu_{BU} = 6.9$). These statistical observations supported two of our hypotheses, H2a and H2b. This indicates a strong possibility that active learning and the top-down teaching method are beneficial for IVGs.

None of the results, however, showed any statistical significance for the interaction between learning type and teaching method, thus failing to confirm our third hypothesis, H2c Though active learning and top-down teaching seem to work well together—A–TD was the highest performing IVG type—they do not provide any special advantage due to their interaction.

5.3. Analysis 3: Effects of IVGs on Visualization Tasks

Finally, we conducted a series of post-hoc chi-square tests aimed at determining which visualization task categories were most affected by the parameters of the IVG type. For these analyses, we used the responses from the NoEx group as the expected values and the re-

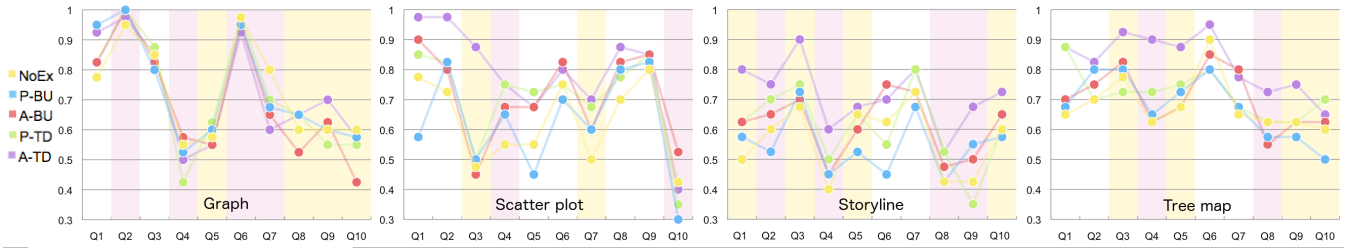

Figure 6: Rate of correct answers for individual questions. The background indicates corresponding task category of the question: pink for *summarize*, yellow for *compare*, and white for *identify*. This figures give some hints at the underlying trends—the A–T-D has some small visible advantage for scatter plot, storyline, and tree map—but the effect is too subtle to be able to read anything concrete directly from the line plot.

Table 3: Two-way ANOVA of IVG Types (Storylines)

	SS	df	MS	F	p
Learning	30.63	1	30.63	4.65	**.033**
Teaching	25.60	1	25.6	3.89	.050
Learning:Teaching	2.02	1	2.02	0.31	.579
Error	1027.35	156	6.59		
Total	1085.60	159			

The two-way ANOVA (using teaching method and learning type) for the storyline visualization shows that only learning type ($p = 0.033$) had a statistically significant effect on comprehension test performance. The teaching method ($p = 0.050$) did not have a significant impact either way.

Table 4: Two-way ANOVA of IVG Types (Tree map)

	SS	df	MS	F	p
Learning	16.90	1	16.90	2.70	.102
Teaching	27.23	1	27.23	4.34	**.039**
Learning:Teaching	6.39	1	6.39	1.02	.314
Error	977.85	156	6.27		
Total	1028.38	159			

The two-way ANOVA for the tree map visualization shows that the learning type ($p = 0.102$) had no statistically significant impact on comprehension test performance, but that the teaching method ($p = 0.039$) did have a positive impact.

sponses from the other IVGs as the observed values. Each response was translated into a binary value representing either correct or incorrect and was tested against the expected values to derive whether the use of a particular IVG had a significant effect on the participants' performance.

Table 5 shows the p values of the chi-square test results. Each cell shows whether a specific IVG type had a significant effect on the participants' ability to derive the correct answer for the corresponding task category. Here, the p in bold font indicates a statistical significance at the $\alpha = .01$ level. All effects that were confirmed as significant were in the positive direction (i.e., there was an improvement in the participants' performances).

As expected from the previous analyses, the chi-square test results for P–BU IVGs showed no significant effects on any visualization task categories. On the other hand, the rest of the analysis results showed that P–TD, A–BU, and A–TD IVGs all have significant effects on *identify* tasks and that A–TD IVGs also have significant effects on both *compare* and *summarize* tasks. The confirmation of the effect of A–BU IVGs on *identify* tasks and the effects of A–TD IVGs on *compare* and *summarize* tasks is unsurprising given the question categories present in those IVGs. We can interpret these indications as the result of prior exposure to similar problem solving scenarios preparing participants for the questions. However, the statistical evidence which suggests that both P–TD and A–TD IVGs improve participants' performance in answering *identify* questions cannot be interpreted in the same way. This evidence indicates that exposing participants to advanced *compare* and *summarize* exercises, instead of simply training them to answer similar questions, could actually improve the participants' overall ability to interpret visual data.

Table 5: Chi-Square Test for Visualization Task Categories

	Identify	*Compare*	*Summarize*
P–BU	0.5068	0.7955	0.7787
P–TD	**0.0005**	0.8628	0.3027
A–BU	**0.0079**	0.0697	0.3027
A–TD	**0.0001**	**0.0001**	**0.0003**

The results of the chi-square test show statistically significant results for all IVG types except P–BU. The P–TD and A–BU only achieve statistically significant improvements on *identify* questions. The A–TD IVG type, however, yielded a statistically significant improvement for each of the three task categories: *identify*, *compare*, and *summarize*.

5.4. Summary of the Analyses

One of the findings consistent throughout the analyses is that all IVG types with exercise questions had, if any, a positive effect on the participants' performance. This suggests that incorporating exercise questions into the introductory materials for information visualizations may or may not enhance user comprehension, but would not impede their understanding. That being said, our analyses also revealed some interesting aspects in the way that visualiza-

tions benefit from IVGs, and naturally lead to design considerations for effective IVGs.

In terms of visualization types, our Analyses 1 and 2 both indicated that there is no statistically significant effect when employing exercise questions in IVGs for graph visualization. We believed that this might be because the participants were able to intuitively understand its visual language based on their prior knowledge of similar concepts that are well integrated into their daily lives such as star constellations and train maps. While this would require further research, it is very possible that for visualizations which are truly intuitive, the type of the employed IVG would not affect the participant's ability to carry out visual data analytics.

In terms of IVG types, our Analysis 1 indicated that A–TD IVGs are the most effective and P–BU IVGs are the least effective. Analysis 2 then showed that the differences in effectiveness of these IVGs are not necessarily based on the unique interaction between learning type and teaching method, and that these attributes independently contribute to the effectiveness of IVGs. Finally, Analysis 3 revealed that there is little effect to be expected from P–BU IVGs and that, furthermore, A–TD IVGs are the only type that provides a comprehensive advantage for enhancing a user's ability to conduct visual data analytics.

6. Discussion and Future Work

Our four visualization types were picked to represent a sampling of both standard and newer, more complex visualization techniques. We also tried to select some visualizations which might serve as building blocks for more advanced visualizations, or which might be used in combination with other visualization techniques. At the same time, we limited our selection in order to maintain a good sample size (40 participants per combination of IVG and visualization type) without drastically increasing the number of required participants (800 participants total for our 20 unique IVG-visualization combinations). Conducting further study on IVGs for other visualization types would verify whether our results generalize to a broader set of visualizations. A typology of fundamental visualization techniques, similar to the task typology [BM13] that we utilized for designing our tests, would be useful for defining a set of visualizations which could yield more objectively generalizable test results.

Mechanical Turk, as a source for study participants, presents a number of tradeoffs. It is an excellent way to access a large pool of participants from a broad array of backgrounds, which is particularly beneficial considering that our study compared 20 combinations of IVG and visualization. However, this meant we were unable to control the participants' environment while they were completing the study, and thus we could not include completion time in our analysis. We also chose to limit the total completion time of the task to just 20 minutes in order to attract the broadest range of participants.

The limitations of our crowd-sourced study prevented us from conducting pre-tests and background surveys for our participants. A pre-test in the vein of [BRBF14] to measure participants' visual literacy would allow us to more precisely quantify the learning effect of the IVG. User backgrounds have been shown to have a sizeable impact on visual understanding [ZK09, RG14]. Our study relied on randomized assignment of participants and a relatively large sample size to avoid bias due to participant background, but collecting rich demographic information could yield new insights and enable us to place our study within a broader context. A smaller, in-person study could include a pre-test and background survey, allowing us to paint a more thorough picture of the performance variation between our subjects.

We intentionally limited the scope of our study in order to reduce the number of potential confounding factors and thereby increase the reliability of our results. Consequently, many factors which are important for visualization as it is currently practiced are outside the scope of this work. Interaction is one of the most prominent aspects of visualization which was outside our scope. The visualizations in our study allowed limited interaction where it was warranted (e.g. highlighting a particular node in a graph or displaying a label for a scatter plot point). The participants were taught about this interaction indirectly through explanations in the IVG text, but there was no interactive tutorial section where the participant was required to interact directly with the visualization as part of the learning process. The only interaction supported by the IVG itself was to go to the next slide and, for active learning IVGs, to select an answer for the exercise question. Based on our results regarding active learning, we suspect that tutorials which require users to directly interact with the visualization will have an even greater positive impact on user comprehension. Further research on interactive tutorials is required to determine whether this effect exists and what its magnitude might be.

Both the decision to source participants from Mechanical Turk and our choice to focus on text-plus-questions tutorials limited the types of visualization tasks that we could consider. For instance, most of the *how* tasks in Brehmer and Munzner's typology require some form of interaction with the visualization, which we explictly avoided as we describe in the previous paragraph. The *why* tasks are ultimately built from a small set of base tasks–*identify*, *compare*, and *summarize*. Due to our imposed time limit, and because we wanted the tasks to be open to as broad an audience as possible, we chose to only include the base-level tasks. While these choices may over-simplify the Brehmer and Munzner's typology, they were necessary to fit the typology to our study design. A more expansive study could consider a broader set of tasks to improve the generalizability of the results.

7. Conclusion

We conducted an extensive investigation into the effectiveness of different InfoVis Guides (IVGs) for introducing visualizations to audiences with little or no prior experience. We used existing literature to create IVGs based on four classifications which utilized questions with either a Bottum-Up (BU) or Top-Down (TD) teaching method and a passive or active learning type. There were four IVG types–plus one control IVG type that contained no questions–which were tested on four common visualizations. The effectiveness of the IVGs was measured through an extensive user study of 800 participants, evenly divided amongst all combinations of IVG and visualization. We then conducted a thorough analysis of the study results to quantitatively compare IVG effectiveness.

Our analysis confirms that the inclusion of exercise questions improves participant performance, and that questions requiring active user participation (i.e. active learning IVGs) had the most beneficial effect. The analysis also shows that Top-Down exercises were more effective than Bottom-Up, and that IVGs utilizing the active learning type with top-down tasks were the most effective. These results suggest useful, practical, and evidnce-based guidelines for writing text-plus-questions introductory tutorials for information visualizations. We also expect that our findings could be generalized to other tutorial types. Ultimately, we hope that such work serves to lower barriers to entry into visualization for a broad potential audience.

8. Acknowledgements

This research was sponsored in part by the UC Davis RISE program, US National Science Foundation via grants DRL-1323214, IIS-1528203, and IIS-1320229, and U.S. Department of Energy via grant DE-FC02-12ER26072.

References

[BDL*14] BUNT A., DUBOIS P., LAFRENIERE B., TERRY M., CORMACK D.: Tagged comments: Promoting and integrating user comments in online application tutorials. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), ACM, pp. 4037–4046. 2

[BM13] BREHMER M., MUNZNER T.: A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2376–2385. 3, 9

[BPW*08] BOYER K., PHILLIPS R., WALLIS M., VOUK M., LESTER J.: Learner characteristics and feedback in tutorial dialogue. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (June 2008), p. 53–61. 2

[BR09] BERTHOLD K., RENKL A.: Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology 101*, 1 (2009), 70–87. 2

[BRBF14] BOY J., RENSINK R. A., BERTINI E., FEKETE J.-D.: A principled way of assessing visualization literacy. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 1963–1972. 9

[CAR*12] CHI P.-Y., AHN S., REN A., DONTCHEVA M., LI W., HARTMANN B.: Mixt: Automatic generation of step-by-step mixed media tutorials. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (2012), pp. 93–102. 2

[DRM*12] DUNLOSKY J., RAWSON K. A., MARSH E. J., NATHAN M. J., WILLINGHAM D. T.: Improving studentsâĂŹ learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest 14*, 1 (2012), 4–58. 2, 6

[FEM*14] FREEMAN S., EDDY S. L., MCDONOUGH M., SMITH M. K., OKOROAFOR N., JORDT H., WENDEROTH M. P.: Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences 111*, 23 (2014), 8410âĂŞ8415. 3, 7

[Gar11] GARRISON D. R.: *E-learning in the 21st century: A framework for research and practice*. Taylor & Francis, 2011. 2

[GB94] GARLAND K., BECK H. C.: *Mr Beck's Underground Map*. Capital Transport, 1994. 7

[Goo] Google Sky. https://www.google.com/sky/. 7

[Har95] HARRISON S. M.: A comparison of still, animated, or nonillustrated on-line help with written or spoken instructions in a graphical user interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1995), ACM Press/Addison-Wesley Publishing Co., pp. 82–89. 2

[HT07] HUANG J., TWIDALE M. B.: Graphstract: Minimal graphical help for computers. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology* (2007), ACM, pp. 203–212. 2

[KPS03] KANG H., PLAISANT C., SHNEIDERMAN B.: New approaches to help users get started with visual interfaces: Multi-layered interfaces and integrated initial guidance. In *Proceedings of the 2003 Annual National Conference on Digital Government Research* (2003), Digital Government Society of North America, pp. 1–6. 2

[LB14] LOUNT M., BUNT A.: Characterizing web-based tutorials: Exploring quality, community, and showcasing strategies. In *To appear in Proceedings of SIGDOC 2014, ACM Conference on the Design of Communication* (2014), ACM. 2

[LBLT13] LAFRENIERE B., BUNT A., LOUNT M., TERRY M. A.: Understanding the roles and uses of web tutorials. In *ICWSM* (2013), The AAAI Press. 2

[LKH*16] LEE S., KIM S.-H., HUNG Y.-H., LAM H., KANG Y.-A., YI J. S.: How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 499–508. 1

[MM98] MAYER R. E., MORENO R.: A cognitive theory of multimedia learning: Implications for design principles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1998), pp. 1–11. 2

[Opr94] OPRANDY R.: Listening/speaking in second and foreign language teaching. *System 22*, 2 (1994), 153 – 175. 3

[PCI10] PAOLACCI G., CHANDLER J., IPEIROTIS P. G.: Running experiments on amazon mechanical turk. *Judgment and Decision Making* (2010), 411–419. 5

[PEB91] PALMITER S., ELKERTON J., BAGGETT P.: Animated demonstrations vs written instructions for learning procedural tasks: a preliminary investigation. *International Journal of Man-Machine Studies 34*, 5 (1991), 687 – 701. 2

[RG14] REINECKE K., GAJOS K. Z.: Quantifying visual preferences around the world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), ACM, pp. 11–20. 9

[RHAH11] RAMESH V., HSU C., AGRAWALA M., HARTMANN B.: Showmehow: Translating user interface instructions between applications. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (2011), ACM, pp. 127–134. 2

[Shn92] SHNEIDERMAN B.: Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph. 11*, 1 (1992), 92–99. 1

[Shu90] SHUELL T. J.: Phases of Meaningful Learning. *REVIEW OF EDUCATIONAL RESEARCH 60*, 4 (1990), 531–547. 7

[SOM96] STEINBERG R. N., OBEREM G. E., MCDERMOTT L. C.: Development of a computerâĂŘbased tutorial on the photoelectric effect. *American Journal of Physics 64*, 11 (1996), 1370–1379. 2

[TM12] TANAHASHI Y., MA K.-L.: Design considerations for optimizing storyline visualizations. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2679–2688. 1

[WCC*14] WANG C.-Y., CHU W.-C., CHEN H.-R., HSU C.-Y., CHEN M. Y.: Evertutor: Automatically creating interactive guided tutorials on smartphones by user demonstration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), ACM, pp. 4027–4036. 2

[ZK09] ZIEMKIEWICZ C., KOSARA R.: Preconceptions and individual differences in understanding visual metaphors. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 911–918. 9

[ZKDC11] ZEID A., KAMARTHI S., DUGGAN C., CHIN J.: Capsule: An innovative capstone-based pedagogical approach to engage high school students in stem learning. In *Proceedings of the ASME International Mechanical Engineering Congress & Exposition* (2011), pp. 305–314. 3, 7