Understand Users' Comprehension and Preferences for Composing Information Visualizations

HUAHAI YANG, YUNYAO LI, and MICHELLE X. ZHOU, IBM Research–Almaden

We are developing an automated visualization system that helps users combine two or more existing information graphics to form an integrated view. To establish empirical foundations for building such a system, we designed and conducted two studies on Amazon Mechanical Turk to understand users' comprehension and preferences of composite visualization under different conditions (e.g., data and tasks). In Study 1, we collected more than 1,500 textual descriptions capturing about 500 participants' insights of given information graphics, which resulted in a task-oriented taxonomy of visual insights. In Study 2, we asked 240 participants to rank composite visualizations by their suitability for acquiring a given visual insight identified in Study 1, which resulted in ranked user preferences of visual compositions for acquiring each type of insight. In this article, we report the details of our two studies and discuss the broader implications of our crowdsourced research methodology and results to HCI-driven visualization research.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms: Experimentation, Human Factors, Theory

Additional Key Words and Phrases: Visualization composition, user preference, chart comprehension, visual taxonomy, cognitive semantics, crowdsourcing

ACM Reference Format:
Huahai Yang, Yunyao Li, and Michelle X. Zhou. 2014. Understand users' comprehension and preferences for composing information visualizations. ACM Trans. Comput.-Hum. Interact. 21, 1, Article 6 (February 2014), 30 pages.
DOI: http://dx.doi.org/10.1145/2541288

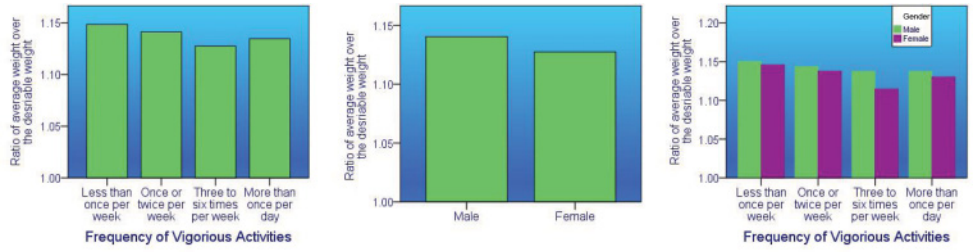
1. INTRODUCTION

The use of information visualization for data illustration and analysis used to be a privilege of dedicated scientists and professional data analysts. With the ever-increasing easy access to data and democratization of visualization tools (e.g., ManyEyes [Viegas et al. 2007] and GapMinder [Rosling 2009]), people are now able to use information graphics [Bertin 1983; Tufte and Howard 1983] in many aspects of their daily lives, such as examining airfare trends and comparing car insurance quotes. Not only do people use information visualization to explore various data aspects [Heer et al. 2007], but they also often want to *integrate* their stepwise analyses to obtain an overall picture or derive deeper insights [Collins and Carpendale 2007; Dörk et al. 2008; Thomas and Cook 2005].

Consider a person who is studying how various factors impact one's overweight ratio. To do so, this person first examines the impact of one's weekly exercise frequency (Figure 1(a)) and then the impact of gender (Figure 1(b)). To understand how *multiple factors*, in this case, one's exercise frequency *and* gender together, affect the overweight

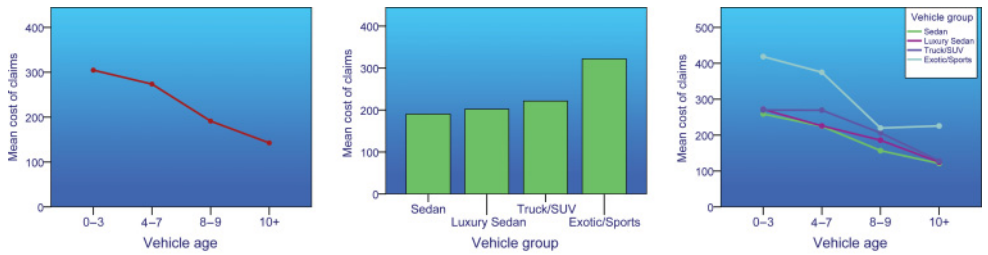
Authors' addresses: H. Yang, Y. Li, and M. X. Zhou, IBM Research–Almaden, San Jose, CA 95120; emails: hyang@us.ibm.com; yunyaoli@us.ibm.com; mzhou@us.ibm.com.

DOI: http://dx.doi.org/10.1145/2541288



(a) Overweight and exercise frequency (b) Overweight and gender (c) Overweight, gender, and exercise frequency

Fig. 1. Two simple information graphics and one of their possible compositions.



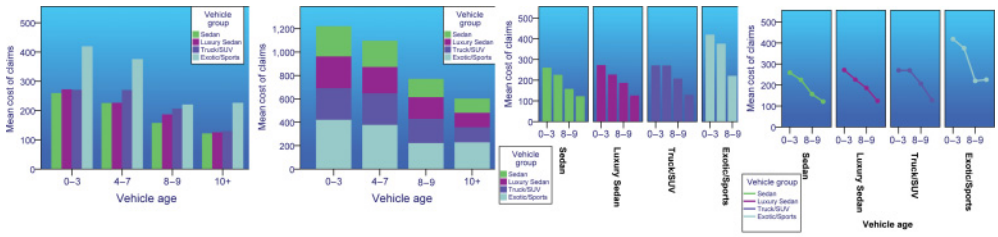
(a) Claim cost and vehicle age (b) Claim cost and vehicle type (c) Claim cost vehicle type, and age

Fig. 2. Another two simple information graphics and one of their possible compositions.

ratio, the person wants to *fuse* together the two individual displays into a composite one (e.g., Figure 1(c)).

The need for composing multiple information visualizations exists not only during an individual's analytic process but also during a *collaborative* visual analytic process [Danis et al. 2008; Heer et al. 2007]. Such a process often requires the fusion of multiple graphics created by *different* individuals to reach an integrated point of view. Consider a couple evaluating a set of used cars using information graphics. One of their main concerns is the potential insurance cost. Thus, they are examining the claim cost for various used vehicles. Whereas the husband is comparing the claim cost for vehicles of different ages (Figure 2(a)), the wife is examining the claim cost for different types of vehicles (Figure 2(b)). To make a decision together, they would like to combine their separately created information graphics into one that can provide them with an integrated view of their respective analyses (e.g., Figure 2(c)).

Composing multiple information graphics together, even simple ones, is nontrivial for two main reasons. First, a new information visualization must be created to encode *all* variables in each existing display. For the examples shown, the combined graphic in Figure 1(c) must now encode three variables—overweight ratio, gender, and weekly exercise frequency. Likewise, the composition in Figure 2(c) also must encode three variables—mean claim cost, vehicle type, and vehicle age. The increased number of variables then requires more effort in designing an effective information graphic [Cleveland and McGill 1985; Mackinlay 1986]. Second, individual visualizations often could be combined in multiple ways. For example, in addition to Figure 2(c), Figure 3 shows four more ways to combine the two original displays. Furthermore, the suitability of a particular composition may be decided by a number of factors, including characteristics of data and existing visual displays, user tasks, and preferences. Given



(a) Crossed Bar (b) Stacked Bar (c) Nested Bar (d) Nested Line

Fig. 3. Four sample composite visualizations.

these challenges, combining multiple existing information graphics is often a daunting task for visualization experts, let alone for novice users.

To address these challenges, we are developing an intelligent visual composition system that helps users easily compose multiple information visualizations on the fly, customized to their tasks and preferences. To guide our development effort, our first step is to acquire a thorough understanding of users' comprehension of composed information graphics and their preferences for such graphics under different conditions (e.g., data properties and tasks). Such studies directly help validate our two hypotheses, on which our automated visual composition system is based:

(1) People can obtain certain visual insights from composite graphics that are unobtainable from simple ones, and
(2) People prefer certain type(s) of compositions over others for acquiring a particular type of insight.[1]

Since we are unaware of any existing work satisfying our needs, we have designed and conducted our own study. Our current study focuses on investigating the composition of *relational graphics* [Tufte and Howard 1983], in particular, bar and line graphs, the two most commonly used business information graphics [Kosslyn 1989]. Corresponding to the two hypotheses presented, the goal of our study is to answer two sets of research questions:

(1) How do people comprehend simple and composite information graphics to derive insights?
    (a) What kind of visual insights does a user derive from a simple/composite information graphic? How do the derived insights statistically distribute?
    (b) How good are the user-derived visual insights?
    (c) How easy is it for a user to derive insights?
    (d) How useful is the presented graphic for obtaining the intended insights?
(2) What is the preferred order among the different compositions for people to acquire a given type of visual insight?

We designed and conducted two studies. The first study (Study 1) was to address the first set of questions. In this study, we designed and conducted a set of surveys on Amazon Mechanical Turk to assess people's comprehension of a given set of information graphics. In these surveys, participants expressed their understanding of the given visualization in free text. We then coded and analyzed the participants' descriptive input to uncover the underlying conceptual structures. From the content analysis, we derived a taxonomy of visual insights, which classifies the conceptual structures

[1]In the rest of the article, we equate acquiring a visual insight to achieving a visual task. We will use the two terms interchangeably.

induced by the given visualization. We then studied the distributional properties of these insights across the information graphics. The second study (Study 2) was to answer the second set of questions. We asked a set of participants to rank different composite visualizations in terms of their suitability for acquiring a given visual insight identified in Study 1.

The rest of the article is organized as follows: We first briefly discuss related work in Section 2 before introducing the types of information graphics used in our study in Section 3. We then present our first study in Section 4, followed by the second study in Section 5, including their methods, results, and analysis. We discuss the implications and limitations of our work in Section 6 and Section 7, respectively.

2. RELATED WORK

Our work is closely related to several research areas in HCI, information visualization, and cognitive semantics.

2.1. Automatic Generation of Visualization

Our work is directly related to research efforts on automated generation of visualization. Researchers have developed a number of approaches and systems in this area. For example, Mackinlay [1986] uses both expressiveness and effectiveness to guide the generation of an information graphic. More recently, Mackinlay et al. [2007] present an automated graphics system for commercial use, focusing on the user experience of such a system [Mackinlay et al. 2007]. In contrast, Casner [1991] uses a user's perceptual tasks to guide the design of information graphics [Casner 1991]. Roth et al. [1997] have extended the previous efforts and used both data characteristics and user tasks to guide the automatic creation of interactive information graphics [Roth et al. 1997]. To improve the extensibility of automated visualization systems, Zhou and Chen [2003] have explored automated generation of information graphics by examples [Zhou and Chen 2003]. To improve the generation quality of information graphics, researchers have also developed various algorithms and methods, including optimization-based approaches to dynamic generation of follow-up displays [Wen et al. 2005] and automated data transformation [Wen and Zhou 2008b], and different approaches to handle visual transitions between two different displays [Gotz and Wen 2009; Heer and Robertson 2008].

Similar to these efforts, we also aim at the automated creation of an information graphic tailored to a user's preferences and tasks. However, unlike previous efforts that either focus on generating an information graphic from scratch [Mackinlay 1986; Casner 1991; Mackinlay et al. 2007; Roth et al. 1997; Zhou and Chen 2003] or from *one* existing display [Wen et al. 2005; Gotz and Wen 2009; Heer and Robertson 2008], ours is on creating a new information graphic by automatically combining *two or more* existing ones. While learning from those previous works, we also systematically investigate users' comprehension and preferences of information graphics for the purpose of creating suitable visual compositions.

2.2. Empirical Studies on Visualization and Graphic Comprehension

To understand various aspects of visualization design and their impact on users, researchers have conducted many empirical studies. These studies include understanding low-level analysis tasks for given datasets [Amar et al. 2005], examining specific visualization techniques (e.g., Chen and Czerwinski [2000] and Stasko et al. [2000]), evaluating the quality of visualization [North 2006], studying the impact of visualization quality on users [Wen and Zhou 2008a], assessing visualization design from various aspects [Heer and Bostock 2010; Heer et al. 2009], and exploring the language

used in describing visualizations [Metoyer et al. 2012]. Similar to these studies, ours is also on the understanding and analysis of users' visual perceptual behavior. However, our work specifically focuses on examining users' comprehension and preferences of compositions of multiple existing graphics.

Directly related to our work, there is a rich body of research on understanding people's comprehension of information graphics [Culbertson and Powers 1959; Anscombe 1973; DeSanctis 1984; Kosslyn 1989; Curcio 1989; Friel et al. 2001]. This line of work considers graphic comprehension to be a logical progression of mental information processing [Bertin 1983; Pinker 1990; Carpenter and Shah 1998]. Empirical studies thus often were aimed at testing hypotheses related to this process [Shah et al. 2005]. For example, some studies investigated the effect of display format on graph comprehension [Simkin and Hastie 1986; Shah et al. 1999; Canham and Hegarty 2010; Shah and Freedman 2011]. However, they do not directly address the design issues of composing graphics, such as the appropriate graphic compositions for a given type of visual task. To achieve our goal of creating automated graphics composition systems, we designed our studies to target a specific set of questions that have not yet been addressed. For example, what types of insights do people gain from graphics? How do these insights occur in different graphics? How are the insights related to different types of graphic compositions?

It is worth noting that researchers have started to employ Amazon Mechanical Turk to crowdsource visualization studies [Heer and Bostock 2010], which inspired us to leverage Mechanical Turk for our own. However, unlike the previous study, which aims at validating the viability of Mechanical Turk as a platform for visual perception experiments (e.g., studying visual aspects such as chart sizing and gridline spacing), ours is on studying users' higher-level semantic comprehension of graphics *as a whole*. As described later, we thus must deal with new challenges raised in designing our crowdsourced studies and analyzing the complex crowdsourced results.

2.3. Taxonomies for Visualization

Researchers have developed various taxonomies for describing various aspects of visualization (e.g., Chuah and Roth [1996], Kosslyn [1989], Shneiderman [1996], Zhou and Feiner [1998], and Amar et al. [2005]). Compared to these efforts, our work derives a task-oriented visual insight taxonomy by analyzing more than 1,500 users' written descriptions of information graphics. Moreover, our analysis also reveals the distributional properties of the taxonomy, including the frequency of insight occurrences across different types of information graphics. Essentially, this empirical crowdsourced approach to visualization taxonomy is a departure from the usual theory-based taxonomy development.

Our method of deriving visual taxonomy is inspired and guided by the approach of cognitive semantics [Talmy 2000; Croft and Cruse 2004; Evans and Green 2006], where the conceptual structures in people's minds are manifested as linguistic meanings. By collecting and examining a large amount of linguistic data describing a set of visualizations, we uncover the common conceptual structures induced by visualization and thus produce a taxonomy of visual insights. In particular, we argue that these visual insights are specialized image schema [Johnson 1990], which are abstract conceptual representations arising from our sensory experience of measuring things.

3. INFORMATION GRAPHICS COMPOSITIONS

Before reporting on our studies in details, here we define the general terminology used in our work and explain the scope of our studies.

3.1. Terminology

In this article, we work with *relational graphic* [Tufte and Howard 1983], which encodes two or more data variables, excluding time series[2] and map-based visualizations. We further controlled the scope of our study to focus on the two most commonly used relational graphics: bar and line graphs, including their variants (e.g., stacked bars and side-by-side bars in Mackinlay et al. [2007] and Wilkinson [2005]).

We adopt the terminology from the Grammar of Graphics (GoG) [Wilkinson 2005] to describe a relational graphic in our study. GoG generalizes or serves as a basis for a range of frameworks that specify the slicing, dicing, and rendering of a multivariate dataset [Wilkinson 2012], such as Trellis layout [Becker et al. 1996], product plots [Wickham and Hofmann 2011], and ggplot2 [Wickham 2009].

According to the grammar, a data variable encoded in a relational graphic belongs to one of two types. The first type is known as *measure* or *analysis* variable, of which statistics is calculated. The second type is called *category* variable, which divides the values of measure into groups.

We consider a bar or line graph a *simple graphic* if it encodes only two data variables—one measure variable and one category variable (e.g., Figure 2(a) and (b)). In contrast, it is a *composite graphic* if it encodes three or more data variables. A composite graphic (Figure 3) can be composed by combining two or more simple graphics through data merging and repartitioning.

Three algebraic operators are defined in GoG [Wilkinson 2005] to describe the possible compositions:

(1) *Cross*, which "crosses all of the values of one data variable with all of the values of another variable." Figures 3(a), 3(b), 1(c), and (2c) are examples of crossed composite graphics;
(2) *Nest*, which "nests all of the values of one data variable in all of the values of another variable" and results in paneled graphics, also known as small multiples of graphics. Figure 3(c) and (d) provides examples of nested composite graphics; and
(3) *Blend*, which "combines all of the values of one variable with all of the values of another variable." Such database union like operation does not produce distinct types of graphics; therefore, the current study does not use this operator.[3]

In the broader context of combining multiple visualizations, a variety of methods have been reported in the literature, such as juxtaposition, superimposition, overloading, nesting, and integration [Javed and Elmqvist 2012]. Some of these methods may leverage the composition operators in GoG, and others may not. In this article, we focus only on the compositions operators presented earlier in order to control the number of experimental conditions.

3.2. Scope of Study

To further contain the scope of our study, we considered only the composition of two simple graphics for the following reasons. First, people often have difficulty in comprehending high-dimensional graphics involving three or more data variables [Wen and Zhou 2008a]. Second, we wanted to avoid unnecessary complications involved in

[2]Although a few datasets used in our study have time-valued variables, they are not time series data per se, since these variables just happen to be related to the temporal scale, such as "the highest year of school completed."

[3]It also has relatively limited scope of applicability because two variables can only be sensibly blended if they are of the same semantic nature (e.g., both represent time). If two variables represent semantically different dimensions, they cannot be blended (e.g. blending money and time is not meaningful).

Table I. Ten Experimental Conditions and Composite Graphics Used in Study 1

Single Graphics	Crossed Line	Nested Line	Crossed Bar	Stacked Bar	Nested Bar
Line : Line	Fig. 6(a)	Fig. 6(e)	—	—	—
Bar : Bar	—	—	Fig. 6(i)	Fig. 6(q)	Fig. 6(m)
Line : Bar	Fig. 6(c)	Fig. 6(g)	Fig. 6(k)	Fig. 6(s)	Fig. 6(o)

designing a composite information graphic. For example, we would have to consider the ordering of their compositions if we had three or more simple information graphics. Furthermore, it is desirable to hold the number of variables constant so that we could isolate the effect of composition operators.

More formally, we have the following: given two simple graphics $A = \langle m_1, c_1 \rangle$ and $B = \langle m_2, c_2 \rangle$, where m_1, m_2 are measure variables and c_1, c_2 are category variables, the composition of A and B is denoted as $A : B \mapsto C$, where $C = \langle m_1, m_2, c_1, c_2 \rangle$.

If m_1 and m_2 are two completely different types of variables (e.g., *salary* and *years of education*), it is often difficult to encode them in a conventional bar/line graph or its common variants (e.g., stacked or aligned graphs).[4] Hence, in our experiments, we considered only composing two simple graphics with the same measure, $m_1 = m_2$, so the composite graphics used in our experiments contained three variables—one measure variable and two category variables, and the composition operators apply to category variables.

Since we consider two composition operators (*Cross* and *Nest*) and two types of graphics (bar and line graphs), we now have four types of composite graphics: *Crossed Bar*, a common graphic with a set of clustered bars (Figure 3(a)); *Nested Bar*, a paneled set of bar graphics (Figure 3(c)); *Crossed Line*, a multiline graphic (Figure 2(c)); and *Nested Line*, a paneled set of line graphics (Figure 3(d)). In addition to these four compositions, we added the fifth one, *Stacked Bar* graph (Figure 3(b)), a commonly used variant of crossed bar composition.

Given that we have three possible pairs of input (`Line:Line`, `Line:Bar`, `Bar:Bar`) and five types of output composition (`Crossed Bar`, `Nested Bar`, `Crossed Line`, `Nested Line`, and `Stacked Bar`), we could have had a total of 15 cases to test. However, some of the cases are not desirable in practice—for example, composing two bar graphs into a line graph, which completely differs from its sources and is likely to disrupt a user's visual momentum [Woods 1984]. We thus removed 5 such cases. Table I summarizes the remaining 10 cases.

4. STUDY 1: UNDERSTANDING INFORMATION GRAPHICS

Study 1 aims at answering our first set of research questions—that is, understanding how users comprehend information graphics to derive visual insights. To achieve this goal, we designed and conducted a set of online surveys on Amazon Mechanical Turk.[5] It is not the intent of this study to formally compare and contrast the experiment conditions. Instead, the 10 conditions were designed to systematically cover the practical variations of the studied graphics and their compositions so that we have a better chance of observing a full range of visual insights within the scope of the study.

4.1. Study Design

The study was a between-subjects design. Each participant was given one of the 10 online surveys, each corresponding to one of the 10 experimental conditions (Table I).

[4]Researchers have shown how to create unconventional information graphics encoding multiple measure variables (e.g., Roth and Mattis [1991]). Moreover, we currently focus on composing common bar and line graphs that can be generated using widely available tools such as Excel or SPSS.

[5]http://mturk.com.

Table II. Datasets Used to Generate Graphics

Graphic	Dataset	Relevant Literature
Fig. 6(i)	Obesity [National Center for Health Statistics 2000]	[Frank et al. 2004]
Fig. 6(q)	Education [National Opinion Research Center 1991]	[Gillborn and Mirza 2000]
Fig. 6(a)	Education [National Opinion Research Center 1991]	[Gillborn and Mirza 2000]
Fig. 6(c)	Labor [National Opinion Research Center 1991]	[Altonji and Blank 1999]
Fig. 6(e)	Family [National Opinion Research Center 1991]	[Downey 1995]
Fig. 6(k)	Environment [National Opinion Research Center 1993]	[Blocker and Eckberg 1997]
Fig. 6(o)	Environment [National Opinion Research Center 1993]	[Blocker and Eckberg 1997]
Fig. 6(m)	Infant Mortality [United Nation 1995]	[Fain et al. 1997]
Fig. 6(g)	Car Insurance [McCullagh and Nelder 1989]	[Lemaire 1985]
Fig. 6(s)	Ship Damage [McCullagh and Nelder 1989]	[Kitamura et al. 1998]

Measures were taken to ensure that a participant took a survey only once. However, it was technically difficult to prevent one from participating in multiple surveys. Therefore, each survey used a different data-topic-variable combination to avoid the potential learning effects that may negatively impact the diversity of observed visual insights. Because we did not plan to formally compare the graphic types in this study, a fully factored design involving different combinations of datasets, data variables, and graphic types was not attempted in order to make the study manageable.

4.1.1. Datasets. We used several datasets to generate the graphics in the surveys. Since our ultimate goal is to create a graphic composition system that helps people in their real-world tasks, we used four criteria to choose the datasets. First, the selected data should be real instead of synthetic. Second, the selected data should be easily understandable by people. Third, the selected data should be meaningful—for example, they provide answers to questions that people care about in the real world. Finally, the data should be easily obtainable. We selected six sample datasets coming with SPSS, a popular statistical software (Table II).[6] All of the chosen datasets were accompanied by publications that illustrated the use of the data in real-world analytic tasks. The publications provided the relevant real-world analytic questions and sometimes even included corresponding visualizations similar to what we used in the study.

4.1.2. Information Graphics. Per the scope of our study, our goal is to understand how a user perceives graphics to derive visual insights. To test all 10 composition conditions, we designed a total of 10 surveys to test participants' comprehension (Table I). For each survey, we used the implementation of GoG in SPSS to generate three information graphics: two simple graphics and one of their compositions. For the sake of consistency, we customized the graphics so that they had the same physical (418 $\times$ 334 pixels) and font sizes (12 point for dimension labels and 9 point for tick labels). We also applied a styling theme[7] to ensure a uniform look and feel for all graphics.

4.1.3. Participants. We recruited turkers from Amazon Mechanical Turk as our participants. *Turkers*, who perform tasks posted on Amazon Mechanical Turk for a monetary payment, are "relatively representative of the population of US Internet users" [Ipeirotis 2010; Berinsky et al. 2012] and are shown to be reliable experimental subjects for perceptual visualization research [Heer and Bostock 2010]. To recruit turkers for our survey, we posted the task description on Amazon Mechanical Turk, which directed the turkers to an online survey site built for this study.

[6]It bundles with 207 sample datasets. Although most datasets are hypothetical, some are real-world data, from which we chose. The six datasets were used for our 10 cases to be tested. If a dataset was used more than once, we made sure that a different set of data variables was used in different cases.
[7]The Marina theme in SPSS Statistics 18.

4.1.4. Survey Instrument. Each survey included four pages. Page 1 started with a brief description about the dataset (e.g., the dataset on examining car insurance) and the data variables (e.g., the car type and cost) used in the tested graphics. Page 2 displayed the first simple graphic, followed by three questions. The first two were open-ended questions, asking a participant to describe in free text the displayed graphic and the obtained insights, respectively. The third question was a 7-point Likert scale question, asking the participant how useful the displayed graphic is in general to answer the first two questions. Page 3 displayed the second simple graphic encoding the same dataset but with a different categorical variable, followed by the same three questions as those on Page 2. After displaying two simple graphics, Page 4, the final page, presented a composite graphic, one of the compositions of the two simple graphics. This page had five associated questions. The first two were open-ended questions asking for the participant's perception about the composite graphic, same as the first two questions on Pages 2 and 3. The third question was a multiple choice question, asking the participant to assess the truthfulness of a list of five statements that describe the composite graphic. This question was intended to assess the level of agreement between participants' understanding and our own understanding of the composite graphic. The fourth question was a 7-point Likert scale question, asking the participant how difficult it was to use the composite graphic to answer the previous question (Question 3). The fifth question asked the participant to explain in free text the type of difficulties encountered. Appendix A contains the details of the survey questions.

4.1.5. Procedure. We initially launched a pilot study with 24 participants on Mechanical Turk. Upon the completion of the pilot and verification of the quality of the results, we deployed all 10 surveys and recruited 50 turkers for each survey. In addition to paying attention to quality control issues such as work acceptance rate, geographic location, reward, and punishment [Kim et al. 2012], we required a turker to enter a minimum of 15 words for all open-ended questions. The answers were automatically validated to meet this condition as the first round of filtering. Later on, each individual answer was read by the experimenters, and surveys with random answers were discarded. Results of incomplete surveys were also discarded. Each participant was allotted 20 minutes per survey. On average, it took approximately 10 minutes for a participant to complete each of our surveys. Each approved completion was rewarded 1.50 US dollars. For the 10 surveys deployed, we collected a total of 514 completed surveys,[8] with between 48 and 55 completed results per survey.

4.2. Content Analysis Methodology

Participants' generated textual descriptions of visualization is the primary data collected for the study. An extensive and thorough methodology was developed to analyze the content of these data.

4.2.1. Procedure. As described earlier, each survey contained three pairs of open-ended questions aiming to collect the participants' *description* and perceived *insights* for the three displayed graphics, respectively. However, from the collected results, we found that the participants did not distinguish between the two questions. They often wrote insights for the description question or vice versa. Therefore, in our analysis, we chose to treat each pair of the questions as one single question and read their answers together as the participants' description of the graphic. We collected a total of $514 \times 3 = 1{,}542$ pieces of free-text descriptions in 72,312 words, with each description containing 46.9

[8]Since turkers were directed to the surveys via a Web link, the number of completed surveys is different from the number of assigned work due to possible confusions and mistakes by turkers.

words on average. All of the descriptions were individually read and coded by three investigators.

After an initial period of independent reading, the investigators convened multiple times to discuss the findings. We also performed formal intercoder reliability test on the coding results by computing Krippendorff's α, a statistical measure of the agreement achieved in the coding results. We first tested on the coding results of two investigators on one survey. The Krippendorff's α ranged from 0.25 to 1 [Hayes and Krippendorff 2007] for different codes, where 0 indicates the absence of reliability, 1 indicates the perfect reliability, and a score higher than 0.7 is considered achieving acceptable reliability in practice.[9] By the testing score, the coding results were further examined, more discussions on the disagreements ensued, and the coding scheme was further refined. Detailed description of the coding scheme can be found in Appendix B.

Using the refined coding scheme, all three investigators independently coded all of the collected data. After all investigators completed their coding processes, review meetings were held again to go through all three sets of codings and discuss discrepancies among the three sets. We found two main reasons causing the discrepancies. First, the majority of discrepancies were caused by simple human oversights mostly due to fatigue, as coding more than 1,500 pieces of free text with about 50 words in each piece was an exhausting process, and it took more than a full month for all three coders to finish their job independently. These oversights were reconciled quickly in the review meetings. Second, the remaining discrepancies were the results of different interpretations of the text, often due to inherent ambiguities in the text. For example, one graphic showed the average years of education received by three races, respectively, "black," "white," and "other." One participant stated *"white received more years of education than other."* It was unclear whether to interpret "other" in this sentence as in "other" race, or as in "other" than "white" (i.e., "black" + "other" combined). In such cases, the discrepancies were not reconciled and were left as they were.

After reconciling the codings of three coders, intercoder reliability scores were calculated for all of the coding categories. The achieved α scores ranged from 0.76 to 1, with an average of 0.95 across all coding categories, indicating strong consistency among the coders.

4.2.2. Cognitive Semantics-Based Coding of Visual Insights. The development of our coding scheme for visual insights followed a grounded theory influenced approach [Glaser and Strauss 1967], where codings were developed and refined iteratively. The theoretical basis of the analysis came from an emerging school of linguistics called *cognitive linguistics* [Croft and Cruse 2004; Evans and Green 2006], in which a cognitive approach to semantics [Allwood and Gärdenfors 1999; Talmy 2000] is one of the main thrusts. This cognitive semantics approach proceeds by utilizing language as a key methodological tool for revealing conceptual structure in people's minds. Image schema is proposed as such an abstract conceptual structure [Johnson 1990]. Space, Containment, Force, and Locomotion are examples of the common image schemas arising from our interaction with the physical world.

According to cognitive semantics, we should be able to uncover people's understanding (i.e., conceptual structure) of visualization by analyzing the linguistics descriptions of visualization. The development of our coding scheme of visual insights is therefore an investigation of the specialized image schemas arising from people's interaction with visualization. As such, the common properties of image schema were adhered to in

[9]To give readers an intuitive feeling of α: for binary data, a single disagreement out of 55 pairs of judgments may bring it from 1 down to as low as 0.66 when data distribution is highly skewed, which is the common case in our data.

developing the visual insight types (see Table III). For example, the insights types are all abstract (not specific to any particular visualization), analogue (resemble sensory experience rather than symbolic), internally complex (consist of many subcomponents), schematic (not detailed like mental images), multimodal (may involve not only visual but also motor experience), subject to transformation (from one insight type into another due to the shift of attention), and produced without conscious effort (consequently require careful reading and introspection to uncover).

Since natural language text is prone to diverse interpretations due to its inherent ambiguities or impreciseness, we developed and then followed a few general principles to ensure coding consistency and to implement the cognitive semantics methodology.

(1) **Binary and permissive scoring**. If part of an answer matched a coding category, the entire answer was counted as a match. For example, the description *"incidents spiked upward starting in 1960 and peaking during the 1965–69. After that, there was a slight downward trend until 1975"* contains information matching multiple coding categories (Table III), including *Identify Extrema* (e.g., *"peaking during the 1965–69"*) and *Characterize Distribution* (e.g., *"downward trend until 1975"*). This rule enhanced coding consistency among the coders.
(2) **Text based scoring**. The coding should rely only on semantic interpretation of the sentences *alone*. Consider the statement, *"I learned that the number of damage incidents in the period 1975–79 was 7 times less than that of 1965–69."* Although the two data points mentioned happened to represent two extreme values (highest and lowest) in the graph, we treated such a case only as a value comparison but not as an identification of extrema to avoid relying on data outside of linguistics.
(3) **Semantic interpretation**. Anchoring interpretation on the semantics rather than on surface wording—for example, if a participant did not use superlative expressions such as *"Ship B has the highest number of incidents"* but wrote *"Ship B has more incidents than any other ships."* We treated the description as identifying an extreme value.
(4) **Concreteness requirement**. The text description must be concrete enough for coders to mentally "picture" what is being depicted—that is, to invoke a specific type of image schema of visualization. For instance, the description *"the infant mortality rate is directly related to the predominate climate of the region where the baby was born"* does not invoke any of the concrete insight types. In contrast, the description *"more infants are more prone to death in the tropical regions when compared to other regions. desert regions stand at second place, maritime third and temperate fourth. death rate is much lesser in arctic region"* depicts a data distribution of which a coder could draw a rough shape.

4.3. Results

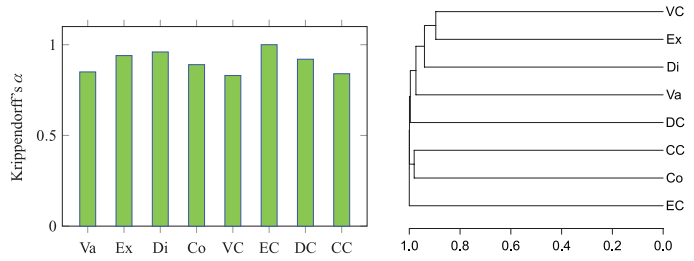
In this section, we present our key findings of Study 1 and discuss them in the context of the first set of research questions posed in the Introduction.

4.3.1. What Insights Are Derived? In an attempt to answer research question 1(a), we developed a taxonomy of visual insights. The taxonomy is the result of reading and discussing the participants' descriptions of graphics based on the principles of cognitive semantics. During the reading and coding process, new types of visual insights were discovered, brought to discussion, and added to the coding category. In the end, eight types of insights were cataloged. They can be organized into two groups: basic insights and comparative insights.

Four type of basic visual insights were identified: Read Value (Va), Identify Extrema (Ex), Characterize Distribution (Di), and Describe Correlation (Co). From a logic point of view, one may argue that some of these are more basic than others. For example, Va

Table III. Coding Schema: Taxonomy of Task-Oriented Visual Insights

Types	Description
Basic Insights	
Read Value (Va)	Explicitly specify the measure variable value or its range for one or more clearly identified data points. Example: "The cost of claims for vehicles of age 0–3 years old is 310." "The mean occupational prestige scores for all racial groups are between 35 and 45."
Identify Extrema (Ex)	Explicitly state the identities of the data points possessing extreme values of the measure variable. Example: "Women who exercise three to six times a week have the lowest ratio of weight over desirable weight." "On average, white male respondents have the highest level of education attained."
Characterize Distribution (Di)	Explicitly describe the variation of measure variable values across all or most of the values of a category variable. Example: "The racial groups ordered by their mean highest average years of school received are white, other, and black." "The vehicle groups A, B, and C have similar mean average cost of claims, whereas vehicle group D has a much higher mean average cost of claims than others."
Describe Correlation (Co)	Explicitly describe the relationships between the values of the measure variable and those of a category variable. Example: "I learned that the older the vehicles are the lower the mean average cost of claims seem to be." "As the number of siblings increases, the mean highest years of school completed by the respondents decreases."
Comparative Insights	
Compare Values (VC)	Explicitly contrast the values of the measure variable at some identified data points in the display. Example: "Males have slightly higher mean occupational prestige score than females." "The mean average cost of claims for group D is higher than that of groups A, B, and C."
Compare Extrema (EC)	Explicitly compare values of measure variables at two data points both identified as extrema. Example: "Even the most environmentally conscious conservative is less concerned with the environment than the least environmentally conscious liberal." "The highest mean occupational prestige score for male and female respondents are similar."
Compare Distribution (DC)	Explicitly compare the characterizations of the measure variable across all or most of the values of two category variables. Example: "Male respondents, regardless of race, have a higher mean occupational prestige score than their female counterparts." "Across all level of education received, female respondents are more concerned about the environment than male respondents."
Compare Correlation (CC)	Explicitly compare the degree of associations of the measure variable with one category variable in relation to that of another category variable. Example: "Regardless of race of the respondents, as the number of siblings increases, the highest level of education attained reduces." "The average cost of claims for vehicles tends to decrease as vehicles get older for all groups, except for the oldest vehicles in group D."



(a) Coding reliability of eight insight types

(b) Dendrogram showing correlations among insight types, x axis measures dissimilarity, with 1.0 being the most dissimilar.

Fig. 4. Coding statistics of eight insight types. Refer to Table III for the definition of the insight codes.

may be argued to be the most basic type. However, from an image schema point of view, these four types of insights have equal status, as each represents a unique conceptual structure that pays attention to a different aspect of the experience of measuring some quantities.

Comparative insights are higher-order insights that compare the results of the four types of basic insights. Accordingly, there are four types of them: Compare Values (VC), Compare Extrema (EC), Compare Distribution (DC), and Compare Correlation (CC). The definitions and examples of these visual insights are listed in Table III. It is very interesting to note that comparative insights are the only combinational insights discovered. Other logically possible combinations, such as "Read value of extrema," "Identify extrema of correlation," and so on, have not been found in our data. It is possible that these novel combinations require conscious effort to generate and thus do not fall into the category of image schema of visualization.

Figure 4(a) shows the intercoder reliability values for all insight types. As can be seen, the α values range from 0.83 to 1, signaling a high level of agreement among the coders.

To verify whether these eight types of insights are orthogonal dimensions in a taxonomy, we used a variable clustering technique [Sarle 1990] to group the insight types by their degrees of correlations. The resulting dendrogram suggests that the eight types of insights are highly independent from one another (Figure 4(b)). Since they do not form obvious clusters, we now have a taxonomy consisting of eight orthogonal types of visual insights.

4.3.2. How Are the Insights Distributed? To better understand how each type of graphic correlates with different types of insights, we further examined how the types of insights distribute among the graphics. To do so, we consider each participant's description of a graphic a binary-valued eight-element vector, where each element denotes the presence (with value 1) or absence (with value 0) of an insight type. As a result, there are a total of $256 = 2^8$ possible insight patterns for a graphic. How ever, we found only 56 patterns from our collected descriptions. The distribution of the frequency of these patterns by frequency rank (Figure 5) looks similar to that of Zipf's law [Zipf 1949]. Fitting a Zipf's law curve to the data yielded a good fit, $R^2 = 0.91$, whereas other estimates, such as log-log scale and exponential decay curve, did not show a better fit. These results suggest that Zipf's law is a good model of insight pattern distribution, especially when a larger number of patterns are present. The three most frequently occurred insight patterns all contain a single insight type, and these are *Compare Values*, *Identify Extrema*, and *Describe Correlation*, in descending order of frequency.

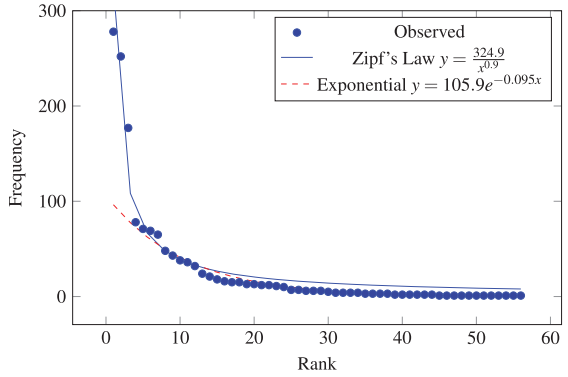


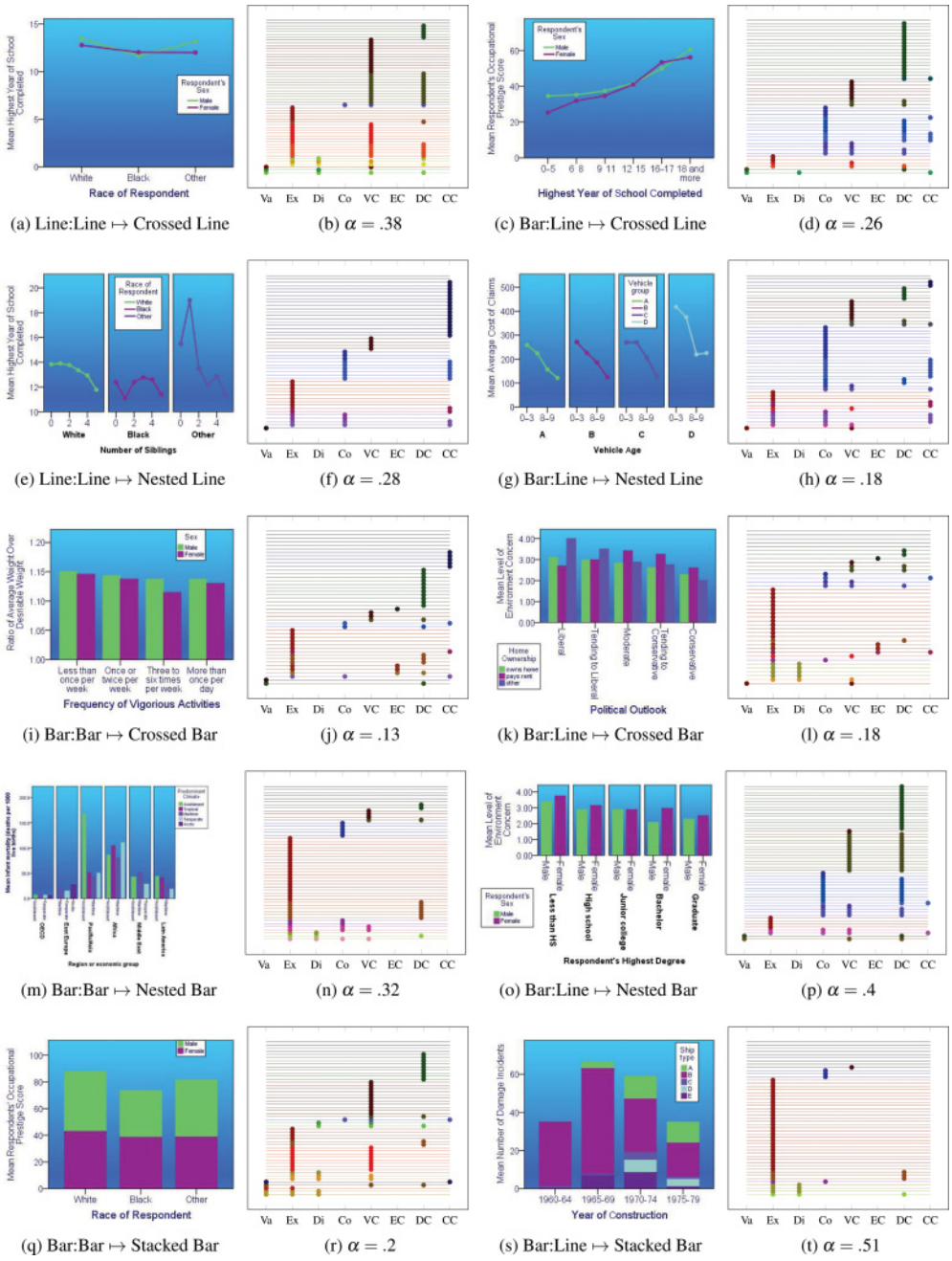
Fig. 5. Distribution of insight patterns: frequency versus rank.

In addition, since our goal is to create composite graphics, we further examined how insights distribute among different types of composite graphics. To see the patterns of insight distribution for all participants, we visualized the coded data using an expanded variation of parallel coordinate visualization [Inselberg and Dimsdale 1991], taking advantage of the fact that our insight codings are binary. Figure 6 contains such visualizations of insight distribution in each of the 10 compositions examined in our study. In the visualization, each horizontal line represents one participant, eight insight types are laid out along the x axis, a dot appearing on a position of a line indicates the presence of the corresponding insight in that particular participants' answers, lines exhibiting the same insight pattern are drawn in the same color, and lines with the same color are grouped together.

Using this visualization, we can detect certain patterns as to how insights are distributed and correlated with each type of composite graphics. For example, we can easily identify the most common insight pattern for a composite graphic, as it is just the largest group of lines with the same color in the chart. For instance, we can see that a single insight type of *Identify Extrema* is dominating in Figure 6(t). We can also see which insight type is the most common for a composite graphic by summing vertically the total number of dots for that insight. For example, in Figure 6(p), almost all participants have the insight of *Compare Distribution*. In addition, the overall diversity of insights patterns for a graphic can be easily seen by the total number of colors used. For example, Figure 6(r) has 15 insight patterns. Across different charts, we can also see where a type of insight is likely to appear. For example, *Compare Correlation* is most likely to occur in *Nested Line*, as in Figure 6(f) and (h), and *Compare Extrema* happens almost only in *Crossed Bar*, as in Figure 6(j) and (i).

A Kruskal-Wallis test showed that the effect of graphic types on the likelihood of insight occurrence was significant, $p < 0.01$, for all but *Read Value*. The likelihood of *Read Value* is low for any type of composite graphics. Pairwise comparisons showed that Nested Line is significantly more likely than others to generate both *Correlation* and *Compare Correlation* insights; *Crossed Line* is likely to be associated with *Compare Values*; *Stacked Bar* is more likely than others to have basic insights such as *Identify Extrema* and *Characterize Distribution*, indicating its poverty in generating richer insights.

4.3.3. How Good Are the Crowdsourced Visual Insights? Answering this research question helps us calibrate the quality of the user-derived visual insights, which we intend to use to guide the automated visualization compositions. To assess the quality of our crowdsourced insights, we examined the results in two ways: (1) examining the



(a) Line:Line ↦ Crossed Line (b) $\alpha = .38$ (c) Bar:Line ↦ Crossed Line (d) $\alpha = .26$

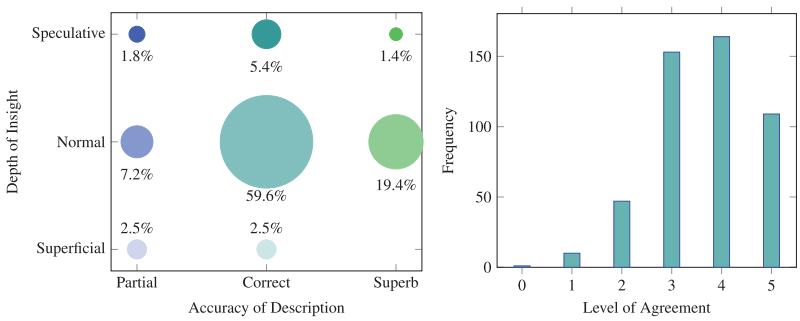
(e) Line:Line ↦ Nested Line (f) $\alpha = .28$ (g) Bar:Line ↦ Nested Line (h) $\alpha = .18$

(i) Bar:Bar ↦ Crossed Bar (j) $\alpha = .13$ (k) Bar:Line ↦ Crossed Bar (l) $\alpha = .18$

(m) Bar:Bar ↦ Nested Bar (n) $\alpha = .32$ (o) Bar:Line ↦ Nested Bar (p) $\alpha = .4$

(q) Bar:Bar ↦ Stacked Bar (r) $\alpha = .2$ (s) Bar:Line ↦ Stacked Bar (t) $\alpha = .51$

Fig. 6. Composite graphics used in Study 1 and the corresponding insight distributions.



(a) Percentage of Coded Quality Ratings (b) Histogram of Agreement with Experts

Fig. 7. Quality of comprehension.

accuracy and depth of the participants' insight descriptions and (2) comparing participants' choices of the insight statements about the graphics with those of the investigators.

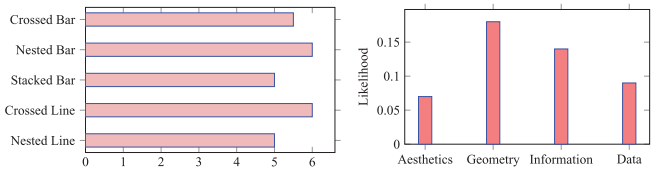
Quality of Insight Descriptions. All participants' insight descriptions were coded individually in terms of their *accuracy* and *depth* by the three investigators using the coding scheme shown in Appendix B.1. Intercoder reliability measured among the investigators was adequate for the *accuracy* of description ratings ($\alpha = 0.8$) and acceptable for the *depth* of insight ratings ($\alpha = 0.67$). When all three investigators gave a description a zero accuracy rating, the description was excluded from further analysis. A total of 86 descriptions (5.6%) were removed. For the remaining descriptions, the median of three investigators' ratings was used in the analysis.

The overall frequency of different quality ratings for all descriptions are shown in Figure 7(a).[10] The three levels of accuracy are coded in increasing saturation of green color, and the depth is similarly coded in blue color. Most descriptions were rated as accurate and with normal depth. Only about 5% of descriptions were wrong or superficial. This result suggests that people can derive quality visual insights from information graphics, including the composite ones.

Agreement with Experts. During the design of the study, three investigators also discussed and agreed upon the truth values of a set of five statements about each composite graphic. We call them the *experts' choices*. In our study, we asked each participant to rate the truth value of each statement. We then calculated the level of agreement between participants' choices and that of the experts, each represented by a five-element binary vector. We then calculate the degree of agreement between two vectors by computing their Hamming distance [Hamming 1950] and subtracting the distance from five (the maximum possible distance). Figure 7(b) shows the histogram of levels of agreement obtained. As can be seen, the level of overall agreement is high ($median = 4$). This result indicates that the participants recruited from the Internet acquired similar understanding of the graphics as people who do research on visual analytics, again confirming the quality of the crowdsourced visual insights.

4.3.4. How Easy Is It for a User to Comprehend a Graphic? Since we would like to generate graphics that are easy to understand, we asked each participant to (1) rate the *degree* of difficulty in comprehending a given graphic on a Likert scale and (2) answer an open-ended question to substantiate his or her subjective rating of difficulty.

[10]The category of being both Wrong and Superficial is not shown, which accounts for 0.3% of responses. All of the remaining situations had zero counts.



(a) Median Ratings by Graphics (with 7 Being Very Easy)

(b) Likelihood of Difficulty by Type

Fig. 8. Difficulty of comprehension.

Degree of Difficulty. Each participant rated the degree of difficulty in using a graphic to answer questions on a 7-point Likert scale, with 1 being very difficult and 7 being very easy. Overall, the level of difficulty was rated as being "somewhat easy"($median = 5$, see also Figure 8(a)). However, the type of graphics had a significant impact on the rating, (Kruskal-Wallis $H_4 = 0.25$, $p < 0.001$). Pairwise comparisons showed that Stacked Bar graphs were significantly more difficult to comprehend than Nested Bar and Crossed Line graphs. Other pairwise differences were not significant.

Types of Difficulty. The participants were also asked to explain the rationale for giving their subjective ratings of difficulty. We coded the answers based on the scheme shown in Appendix B.2. The intercoder reliability α test scores were 0.95, 0.96, 0.94, and 0.94, respectively, for four identified types of difficulty: aesthetics, geometry, information, and data (Table V in Appendix B).

We calculated the likelihood for a type of difficulty to occur by summing up its number of occurrences divided by the total number of answers. Overall, the likelihood for the participants to complain about any type of difficulty was 0.37. Figure 8(b) shows the likelihood for each difficulty type to occur, where geometric difficulty occurred most often, followed by informational difficulty.

It is interesting to note that the aesthetics of a graphic was of the least concern for the participants. Therefore, it might not be a good investment to test "40 shades of blue" when designing a visualization. Our participants indicated that they had the most difficulty in comprehending the geometry of a graphic. Considering that our study used only bar and line graphs, this result strongly suggests that visualization should be geometrically simple in order for people to understand.

4.3.5. How Useful Is a Graphic for Deriving Insights? In our study, each participant was asked to rate the usefulness of each given graphic on a 7-point Likert scale, with 1 being not useful at all and 7 being very useful. Overall, participants rated the graphics in our study as "somewhat useful" ($median = 5$). This rating suggests that people can generally appreciate the value of information graphics.

A Kruskal-Wallis one-way ANOVA test showed that this perceived usefulness varied across different types of graphics, $H_6 = 53.1$, $p < 0.001$. Pairwise comparisons indicated that simple Line Graph graphics were rated significantly less useful than all composite graphics except for Stacked Bar; simple Bar Graph graphics were rated significantly less useful than Nested Bar and Crossed Line. All other differences were not significant.

To help validate the hypothesis of composite graphics being perceived as more useful than simple ones combined,[11] due to the additional insights offered in the composite

[11]Ideally, one would like to compare the usefulness of a pair of simple graphics with that of their composite graphics. As a limitation of the study, the surveys did not include such questions.

graphics, we calculated the correlation between the number of insights perceived in a graphic and its usefulness rating. A modest but significant correlation supported the hypothesis, $r = 0.19$, $p < 0.001$.

4.3.6. Summary of Key Findings. We briefly summarize the key findings in term of addressing the first set of research questions set out in the Introduction.

Insights Types and Distribution. Eight types of visual insights were identified (Table III). The insight types are orthogonal to one another. Several of them have not previously appeared in the literature. The overall distribution of perceived insight types follows Zipf's law, which implies that people are more likely to derive certain types of insights than others. Composite graphics appear to be associated with distinct patterns of insights (Figure 6).

Quality of Insights. The majority of user descriptions adequately captured the insights exhibited in the shown graphics. Our participants' judgment about the graphics were also in agreement with that of experts.

Difficulty of Comprehension. Participants rated the task of using information graphics to derive insights as "somewhat easy." However, more than one third of participants expressed certain difficulties. In particular, we found four main causes of the comprehension difficulty (Table V in Appendix B), with geometric type being the most frequent.

Usefulness of Graphics. Overall, the graphics were perceived as useful. Graphics providing richer insights were perceived as being more useful than others.

Our findings provide some jump-start knowledge for building automated graphics composition systems:

(1) The derived taxonomy of visual insights provides us the starting point that allows a user to express the type of visual tasks to be accomplished when requesting the creation of graphics.
(2) Our understanding of the distributional properties of visual insights could help the system optimize the limited visualization resources when achieving an intended visual task. For example, if a user's goal is to acquire multiple types of insights, the system may then choose to compose *one* graphic capable of achieving multiple insights *at once* instead of creating *multiple* graphics best for achieving each type of insight. Here, the distributional properties of the visual insights can be used to define the "weights" of a type of graphic for acquiring a particular type of insight.
(3) Those relatively rare types of insights are in the long tail portion of the Zipf's curve and require special care in graphics design to support them (Figure 5).
(4) Three highly popular insights occur at the head end of the curve, appearing in almost any type of graphics: *Compare Values*, *Identify Extrema*, and *Describe Correlation*. When the set of visual tasks to support is not known a priori, it would be prudent to choose the kind of graphics that support these three insights very well, maximizing the chances of serving a user's basic needs.

Our findings also suggest the effectiveness of different types of graphics when encoding the same amount of information. For example, our study showed that people consistently have more difficulty with *Stacked Bar* graphics than with any other graphics used in the study.

In our study, there was about 40% chance for people to express certain kind of difficulties with commonly used graphics presented in the study. In addition, most difficulties stemmed specifically from the geometric layout of the graphics. Thus, supporting the effective design of visualization clearly matters.

5. STUDY 2: RANKING COMPOSITE GRAPHICS

In Study 1, we focused on examining a person's comprehension of information graphics, especially the types of insights that one can derive from different information graphics. The results show the usefulness of composite graphics in helping people derive different types of visual insights. However, we did not study users' *explicit preferences* for a particular type of composite graphic in acquiring a specific type of insight. We thus designed and conducted Study 2 to elicit *explicit* user input on which composite graphics were preferred for deriving a specific type of insight.

5.1. Method

To elicit a user's preference for composite graphics, we designed test sets targeting each of the eight types of insights identified in Study 1. Specifically, for each type of insight, we designed two test sets using two different datasets from Study 1. As described later, we used the two sets of results to assess the result reliability. Each test set consisted of three parts.

The first part displayed five composite graphics side by side, each of which was uniquely labeled. The five composite graphics used in this study were the same five types of composite graphics used in Study 1 (Figure 6). In this study, they encoded the same dataset with different composition operators. The second part was a yes-no question, designed to test whether a participant could derive the intended insight from the five graphics displayed earlier. Here is a sample question on the insight type *Compare Values*: *"At four to seven years of age, do sedan and luxury sedan have similar claim cost?"* The question statement was selected from the turker's input in Study 1 but rephrased as a yes-no question.[12] If needed, minor edits were made to improve the readability of the statement. The third part asked the users to rank the five composite graphics in the order of their suitability to answer the question.

This experiment is a between-subjects design, where eight groups of participants were recruited, with each group working on one of the eight types of insights. After a small pilot study, we recruited 30 turkers for each group, with a total of 240 participants for eight groups. Before the actual task started, each participant was asked to perform a pretest, where he or she was asked to rank five dummy graphics by a simple criterion. This qualification test ensured that participants know how to rank graphics and helped filter out inattentive participants who produced wrong answers.

5.2. Results and Analysis

The goal of Study 2 was to discover which type(s) of composite graphics are preferred for deriving a specific type of insight. It was achieved by obtaining the rankings of the five types of composite graphics for each of the eight types of insights. For each type of insights, the ranks of five graphics were compared using Friedman tests [Friedman 1937]. The effects of graphic type are statistically significant for all insight types. Figures 9 and 10 summarize the results. For each insight type, we list the graphics by their median ranks in ascending order from left to right. By this representation, the graphic at the left-most position is the most preferred choice for the given insight type. When several graphics do not differ significantly in the pairwise comparison, they are grouped and underlined together.

To assess the reliability of the derived rankings, we compared the rankings derived from the two test sets in each type of insight. Using the Mann-Whitney test, we found the rankings reliable, since there were no significant differences between two rankings

[12]The statement testing Va differed from others. We used a text box to solicit a numeric value. The answer was validated to contain at least a number, or the system would not go forward. The goal here was to encourage participants to engage in creating conceptual structure of Va.

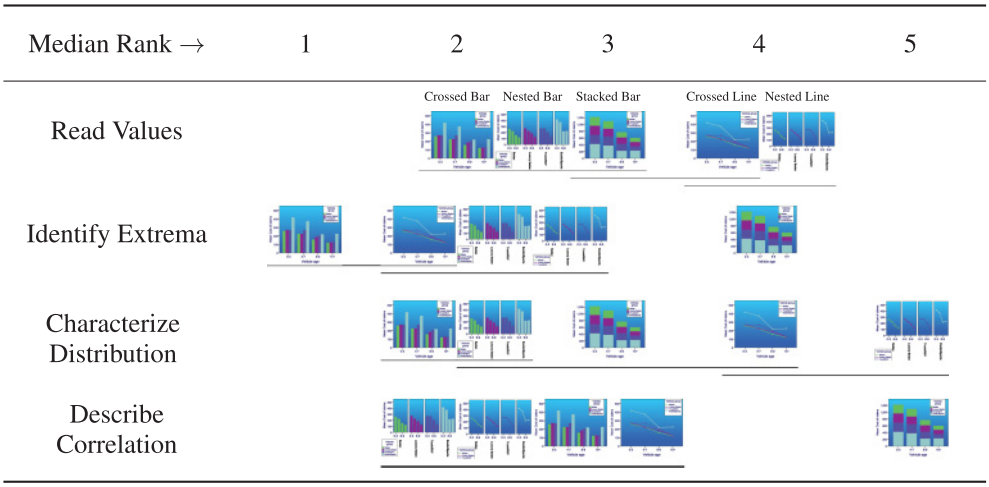


Fig. 9. Preference of composite graphics for basic insights (rank 1 is the most preferred).

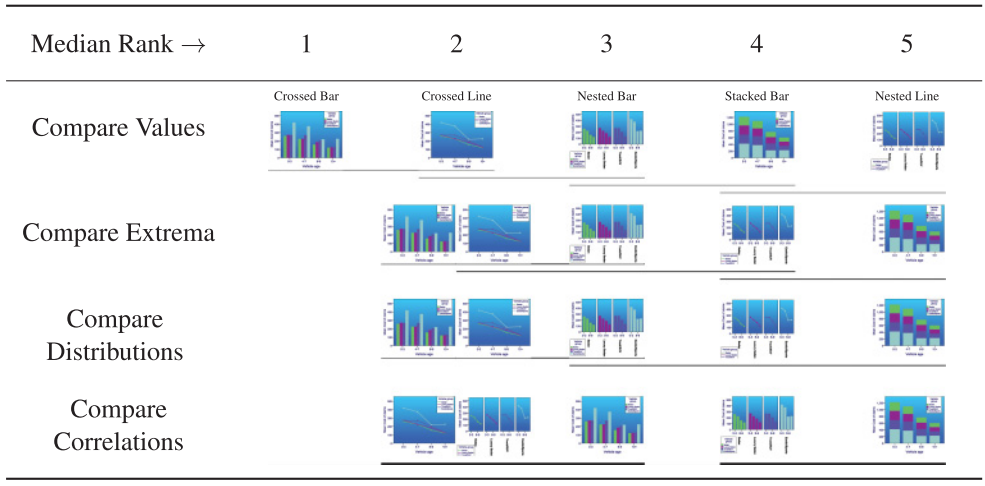


Fig. 10. Preference of composite graphics for comparative insights (rank 1 is the most preferred).

for all cases, with two exceptions: the ranks of Nested Line graphics (4 vs. 5) in Characterize Distribution and the ranks of Stacked Bar (3 vs. 4) and Nested Bar (2 vs. 3) in Read Value were slightly different.

It is interesting to note that *Cross Bar* was consistently ranked among the most preferred choice, whereas *Stacked Bar* was consistently ranked among the least preferred except for *Read Value* (Figures 9 and 10). For comparative visual tasks, the *Nested Line* chart was also consistently ranked among the most preferred ones.

Based on our findings, we can build a system that will recommend appropriate composition operators for a given visual task (e.g., reading a value or making a comparison). Let us take the car-buying scenario introduced earlier as an example (Figure 2). If the couple want to compare how the claim cost changes with the age for two different vehicles, a *Compare Correlation* task, *Crossed Line* (Figure 2(c)) would be their top choice. In contrast, if they simply look for the least expensive vehicle with the youngest age, an *Identify Extrema* task, a *Crossed Bar* (Figure 3(a)) would be the best option.

In the case of not knowing users' potential tasks a priori, a default approach is to weigh all of the tasks by their expected likelihood of occurrences. By doing so, the most preferred composition is a *Crossed Bar* graphic, since it is the top-ranked graphic for all of the three most frequently perceived insights for composite graphics. In other words, the *Crossed Bar* graphic can be considered a universal default composition.

6. IMPLICATIONS

To control the scope of our study, we focused on examining two most commonly used business graphics and their composites: bar and line graphs. Prior to our study, we were worried about the significance and applicability of our results due to the limited study scope. As summarized earlier, our study results not only help answer our two original sets of research questions but also bear two important implications on user-centered visualization research. First, our study results help develop advanced, user-centered visualization systems that are traditionally difficult to build due to a lack of computational foundations built on top of rigorous empirical evidences. Second, our methodology of crowdsourcing nontrivial users' visual cognitive tasks and rigorously analyzing the crowdsourced results can be applied to a wide range of empirical investigations in visualization research.

6.1. Applications of Crowdsourced Results

Not only do our study results provide us with guiding principles for visualization composition, but they also help us develop intent-driven information graphics generation and retrieval systems.

6.1.1. Intent-Driven Information Graphics Generation and Retrieval. The main motivation of our work is to lay foundations for building automated visual composition systems, where users can compose complex graphics from existing ones without having to deal with many challenges in the process, such as data composition and visual encoding of multidimensional data. Although our study results directly help achieve such a purpose, they also facilitate the development of intent-driven graphics generation systems. In particular, the insight taxonomy developed in the study (Table III) helps map a user's high-level intent (i.e., visual insights to be perceived and conceptual structure to be created) to one or more desired target graphics. This approach is similar to the use of user perceptual tasks to guide visual composition [Casner 1991]. Unlike the previous work, however, our taxonomy is backed up by solid empirical evidences that associate a user's intent (visual insight) with the underlying desired graphics. As a result, we can develop a high-level declarative visual language based on the insight taxonomy to support intent-driven visualization. Given a description of the visual insight that one is seeking, the system can automatically generate the appropriate visualizations that match the visual insight.

In addition, users can also use the intent to *search for* existing information graphics. Finding a target information graphic is challenging, as an information graphic is often complex and difficult to describe. Since our results record the connections between an intent (visual insight) and a graphic, we can also *automatically* annotate and index the similar type of information graphics by the matching intent. As a result, such information graphics can then be searchable by their semantics (intent). Such intelligent visualization generation and retrieval systems operate at a level much closer to users' native domains of knowledge and intuitions than to the often unfamiliar data and visualization details.

6.1.2. Natural Language–Driven Information Graphics Generation and Retrieval. Given an intent-driven visualization system, the next natural step would be to build a natural language (NL)-based user interface to data visualization. Since our studies collected

users' descriptions about their intent in the form of NL, such data can be used as a training corpus for such a system. For example, the users' NL input can be used as a training corpus to automatically generate textual captions for all similar information graphics. Such automatic caption generation also helps achieve better accessibility to the graphics (e.g., helping visually impaired people comprehend the graphics). For another example, our collected NL descriptions of graphics can also be used as a training corpus for a system to allow a user to express his or her intent in NL. Such a system will automatically map the user's NL input to one or more intent descriptions and then generate the target graphics that satisfy the intent. Note that our distributional analysis of the insights can also be helpful for the system to resolve ambiguities rising in NL interpretation. For instance, if a user's intent from his or her NL input is ambiguous, the system may ask for clarification based on the distribution of the insight types (e.g., asking for the more likely insight type first).

The general theme behind our ideas is to leverage crowdsourced empirical evidences to jump-start and help build more intelligent visualization systems that are traditionally very difficult to develop. As we leverage our study results toward this direction, we hope that both the HCI and information visualization communities work together to develop empirical approaches for advanced visualization systems, especially by leveraging the power of the crowd to develop people-centric visualization systems.

6.2. Crowdsourced Research Methodology

Previous work has demonstrated the effectiveness of crowdsourced approaches to understanding one's visual perception [Heer and Bostock 2010]. Given the crowd diversity (e.g., skills diversity in turkers), however, there is little evidence on the effectiveness of crowdsourced approaches to complex visual cognitive tasks or on the quality and reliability of the crowdsourced results for such tasks. Since our studies aimed at examining users' comprehension of information graphics in depth and at scale, our methods of systematically instrumenting crowdsourced studies and rigorously analyzing the quality and reliability of crowdsourced results can be used to investigate users' visual cognitive behavior in a way that has never been done before. Moreover, our content analysis method, including the coding principles based on cognitive semantics, is valuable to researchers who also wish to harvest crowdsourced rich content.

6.2.1. Crowdsourcing Users' Understanding of Graphics via Text Input. One way of understanding users' comprehension of graphics is their own verbal (textual) description of the graphics [Goldsmith 1984]. Although there are other works on studying users' language use in describing graphics [Metoyer et al. 2012], our work shows the feasibility of crowdsourcing users' input from a large number of people (i.e., more than 500 people) on the Internet. Our approach of combining introspection-based semantic analysis and corpus-based statistical analysis produces highly replicable results. We also show how to assess the quality and reliability, including consistency, of the crowdsourced results to determine the effectiveness of our approach. Since our results demonstrate a high level of quality and reliability among user input, we believe that our approach can be effectively applied to study users' comprehension of any other types of information graphics beyond relational graphics, such as network diagrams and text visualization.

In addition, the goal of our content analysis is to uncover conceptual structures induced by graphics. As suggested by the study results, simpler graphics facilitate the discovery and categorization of more generally applicable visual insights. According to the theories of cognitive semantics, such basic conceptual structures are the building blocks of more complex conceptualizations. As a first study of this kind, this article focuses on building a solid foundation for understanding complex and interactive

visualization. Therefore, the results of the study are very generalizable, because the understanding of complex and interaction visualization can only be based on them.

One key principle of the cognitive semantic approach is the so-called embodied cognition thesis, which states that our conceptual structures are originated from our bodily experience and the abstract thoughts are metaphorical projection thereof [Lakoff and Johnson 2008; Lakoff 1987]. For data visualization, what is the embodied experience that the people can drawn upon? From our content analysis of description of visualization, a viable hypothesis is that the measurement experience in the physical world provides the necessary foundation for creating the conceptual structures for visualization. All eight types of visual insights discovered have clear corresponding physical measuring processes. One implication for creating comprehensible visualization is to draw inspiration from physical activities that people perform in the physical world.

6.2.2. Crowdsourcing Information Visualization Taxonomy. As part of the study results, we obtained a taxonomy that characterizes the user-identified insights from the graphics under study (Table III).Previous efforts on constructing visualization taxonomies (e.g., Casner [1991], Amar et al. [2005], Chuah and Roth [1996], Kosslyn [1989], Shneiderman [1996], and Zhou and Feiner [1998]) were either based on visual cognitive theories or limited experimental results. In contrast, we used a crowdsourced approach to collect a large quantity of data (i.e., more than 1,500 insight descriptions) in a short period of time (i.e., within a couple of days). The collected data set also allowed us to discover several insight types not previously reported in the literature, such as *Compare Extrema*, *Compare Distributions*, and *Compare Correlations*. On the other hand, with only eight types of insights, the absence of more types seem to be more surprising. For example, the composite insights are all comparative types rather than of other possible combinations. The cognitively viable conceptual structures do not seem to occupy the full space of possible structures.

Furthermore, the collected large dataset helped us measure the properties of a taxonomy. For example, we were able to measure the distributional properties of the insights in our taxonomy and found the insight patterns to be distributed according to Zipf's law. In addition, three insight types, *Compare Values*, *Identify Extrema*, and *Describe Correlation*, exhibit a much higher frequency of occurrences than others. Such findings are useful especially when applying the taxonomy to guide the creation of information graphics. For example, if a user's goal is to explore a dataset from different aspects, the system should perhaps start with the information graphics that can help the user extract the insights that the user is most familiar with to reduce the barrier (cognitive load) to entry.

In short, our crowdsourced methodology can be used to collect data from a large number of users quickly and analyze the quality and reliability of the collected results rigorously. We believe that our methodology can be applicable to a wide array of HCI-driven information visualization research, ranging from understanding users' comprehension of a specific type of information graphics to building a comprehensive, empirically validated visualization taxonomy.

7. LIMITATIONS AND FUTURE WORK

Although our work offers unique value to both HCI and visualization communities, we have found several areas that warrant future work.

Coverage of Insights. From the insight descriptions produced by our participants, more than half of such descriptions contained only a single insight. This phenomenon may not signal people's inability to derive multiple insights from an information graphic. Instead, it could be an artifact of the experiment settings, where turkers just wanted to write down the first insight coming to mind to finish the task quickly.

Although our original goal was not intended to investigate the coverage of insights communicated by an information graphic, we found this to be an interesting research topic. For example, if an information graphic is capable of achieving multiple insights *simultaneously*, a system may choose to use such a graphic for accomplishing multiple visual tasks (e.g., reading value *and* making a comparison) instead of using multiple graphics. The research challenge here is how to motivate participants to extract as many insights as possible. Besides providing incentives for the participants (e.g., giving a monetary bonus for each insight extracted), alternative study approaches such as essay writings or interviews might be used.

Factors Affecting Insights. Our studies allowed us to characterize users' understanding of relational graphics and observe different insight patterns perceived across relational graphics. Although such patterns are useful for a system to recommend suitable relational graphics for deriving a particular insight, other factors that we have not systematically studied could also affect the choice of the graphics. Such factors include the characteristics of the dataset, the number of data points to be encoded, a user's visual preferences, and the interaction context. In our studies, we have already observed the influence of certain factors. For example, *Identify Extrema* needs obvious extreme data points, whereas *Describe Correlation* needs significant trending character in the dataset. Note that previous research efforts have investigated the influence of various factors on the use of information graphics, such as data characteristics [Zhou and Feiner 1996] and interaction context [Wen et al. 2005; Gotz and Wen 2009]. However, most of such work relies on a very limited number of case studies instead of building on collected significant empirical evidences as we do. Thus, it would be interesting to use the methodology similar to ours to collect empirical evidences and systematically investigate the influence of other factors on the generation and composition of relational graphics.

Understanding Other Visualizations. Our current studies focus on relational graphics; however, it would be interesting to apply our research methodology to study other types of visualizations. The purpose of this extension is twofold. First, it will help generate a more complete taxonomy of visual insights. Second, it will help us further refine and augment our current empirical methodology in crowdsourcing and analyzing users' comprehension of general information graphics. For example, if we ask people to comprehend information graphics with which they are unfamiliar, what type of results would we get? Would the people be unable to describe their comprehension at all or inaccurately describe their comprehension? As one of the most powerful features in information graphics is its interactivity, it is an extremely important research direction to investigate how to adopt our current methodology to study interactive visualizations. This extension will be especially challenging on an open crowdsourcing platform where the participants must be carefully guided to perform visual interaction tasks that few research efforts have addressed.

The studying of simple graphics in the article is congruent with our research goal of deriving a set of cognitively solid conceptual structures to be used as building blocks to understand more complex visualizations and tasks. The simplest graphics allow us to uncover the most generally applicable visual insights. One challenge of more complex and unconventional visualizations is that they are often very removed from the bodily experience with which people are familiar. For these complex visualizations, the simple image schema–based insight types uncovered in this article may still serve as the basic building blocks for the corresponding conceptual structures, but understanding their dominant features may require higher-level machinery of cognitive semantic, such as frame [Fillmore 1985; Barsalou 1999], metaphor [Lakoff and Johnson 2008], metonymy [Kövecses and Radden 1998], mental spaces [Fauconnier 1994], and conceptual

blending [Fauconnier and Turner 2008]. It is our hope to see a fruitful marriage between visualization research and cognitive semantics.

8. CONCLUSIONS

To support the composition of two or more existing information graphics, we are building an automated graphic composition system. As an initial step, we would like to acquire a thorough understanding of people's comprehension and preferences of composite graphics under various conditions (e.g., data and tasks). Toward this goal, in this article, we have reported two crowdsourced studies conducted on Amazon Mechanical Turk involving more than 750 questionnaires. Our first study focused on examining users' comprehension of graphics to derive visual insights. The second study was on extracting users' explicit preferences of composite graphics for achieving various visual tasks. As a result, our studies present several key findings. First, we have identified eight orthogonal types of visual insights along with their distribution properties among different types of graphics. We have also found that the type of compositions significantly affects the kinds of insights to be acquired. For each type of insights, we have obtained explicit user preference orders for each type of composite graphics. Not only do our findings provide the foundation for building an intent-driven automated visual composition system, but our work also bears important implications to the HCI and visualization communities. In particular, our crowdsourced methodology for analyzing crowdsourced rich content and developing visualization taxonomy can be extended to conduct a wide range of empirical studies in visualization. Moreover, our crowdsourced results can be leveraged by the communities to build intent-driven, NL-based visualization generation and retrieval systems.

APPENDIX A: STUDY 1 SURVEY QUESTIONS

After introducing the dataset and the variables involved, the second page of the survey displayed a simple graphic followed by three questions. The first two were open-ended questions:

Please describe the graph in your own words. You may start your description like "This graph shows. . . ."".

What insight, observation, or information have you gained from this graph? You may start your description like "I learned that. . . ."

Turker's free-text responses to these questions were the raw materials of our content analysis.

The final question was a 7-point Likert scale question:

How useful is this graph for you in general (e.g., helping you learn about the topic or satisfy your curiosity about the topic)?

The third page displayed the second simple graphic encoding the same dataset but with a different categorical variable, followed by the same three questions as on Page 2.

After displaying two simple graphics, the fourth page, also the final page, presented a composite graphic that was one possible composition of the two simple graphics. It also had three associated questions. The first was a multiple choice question:

According to the graph, which of the following statements are true? Check all true statements.

Five choices were provided to describe the composite graphic. This answers were used to assess the level of agreement between the turker's understanding and our own understanding of the composite graphic.

Table IV. Coding Scheme: Quality of Descriptions

Code	Description
Accuracy	
0 Wrong	Incomprehensible or completely wrong/irrelevant description. `Example:` "I learned that there are a lot of people who need car insurance."
1 Partial	Part but not all of the description is correct. `Example:` "The graph shows that the average cost of claims decreases with vehicle age, with vehicles that are 10 years or older costing the least. I learned that the average cost of claims does not increase with the age of a vehicle." The first half of the description is correct but the second half is wrong.
2 Correct	The description is correct. `Example:` "I learned that the older the vehicle, the lower the mean average cost of claims is."
3 Superb	The description not only is correct but also contains fine details and/or qualifications. `Example:` "Vehicle age is an important factor because the difference of 7–8 years in a car's age can change the cost of claims by more than 50%"
Depth	
1 Superficial	The description contains no substantial information beyond the variable names and the type of the graphic. `Example:` "I learned that ratio of average weight over the desirable weight differs by the frequency of vigorous activities."
2 Normal	The description contains specific information directly perceivable in the graphic. `Example:` "I learned that those who exercise less than once a week have the highest ratio of weight over desirable weight."
3 Speculative	The description contains not only perceivable information but also the participant's own injected information, such as assumptions, hypotheses, rationale, and so on. `Example:` "Just from this graph, I learned that more exercise means being closer to the desirable weight. However, the graph does not take into account those who exercise more than once per day, who probably weigh more because of increased muscle mass."

The second question was a 7-point Likert scale question:

How easy was it to answer the previous question using the graph above?

The final question in our survey was an open-ended question:

For the previous question, explain the reasons for your rating.

Free-text input to this question was used to analyze the type of difficulties that people may have with the composite graphic.

APPENDIX B: STUDY 1 CODING SCHEME

We coded the written descriptions of participants for all of the open-ended questions according to a coding schema. The definitions and examples of the codes are described next. All of the examples are the actual text written by participants. The coding schema consists of three aspects, as follows:

B.1. Quality of Descriptions

We coded the quality of participants' descriptions of graphics along two dimensions, the *accuracy* of description and the *depth* of insights, as shown in Table IV.

The accuracy of description measures the degree of the text description matching the content of the corresponding graphic. If the text does not actually describe the meaning

Table V. Coding Scheme: Types of Difficulty in Comprehension

Code	Description
Aesthetics	Issues related to the aesthetics of a graphic (color, background, etc.). Example: "The graph bars and the background color do not have a very good contrast." "The three shades of purple are somewhat hard to tell apart."
Data	Issues related to understanding the nature and the ambiguity of data. Example: 'Isn't clear whether the holders of bachelor's degrees are being counted separately from those with bachelor's degrees and higher degrees (i.e., graduate)." "In the left-side scale, it is not shown the units, just the value. How to know what is it?"
Geometry	Issues related to the geometric aspects of a graphic (layout, shape, scale, size, etc.). Example: "The graph did not give a very good scale on the Y axis. As a result it was a little confusing." "Since the graphs aren't overlapping, it makes it difficult to tell whether some of the points were higher or lower than other ones"
Information	Issues related to the nature and the amount of information being represented. Example: "The graphs consisted of too much information, so it became a bit confusing." "As I do not properly understand the graph, I cannot say that I got it fully."

of the graphic, it is rated as *wrong*. There are also cases of *partial* accuracy, where some parts of the text depict situations contrary to the situation in the graphic. If all parts of the description match what are shown in the graphic, it is coded *correct*. In some cases, the descriptions are not only correct but also are very detailed (e.g. noticing subtle patterns); they are coded as *superb*.

The depth of insights reflects the level of mental engagement of the participants with the graphics. A *superficial* description contains mere recitation of the variable names shown in the graph and the name of the graph type shown. It does not involve the pattern recognition, interpretive, and integrative processes necessary to decode the meaning of the graphics [Carpenter and Shah 1998]. A *normal* description contains the outcome of the processes shown, whereas a *speculative* description reflects not just those shown but also the outcome of some inference processes that involve the participants' own knowledge, beyond what the graph itself entails.

B.2. Types of Difficulty in Comprehension

We developed a taxonomy of participants' difficulties in understanding the graphics. These were coded using binary codes (presence or absence), according to the participants' written rationale on their easiness ratings for understanding each graphic. Table V shows the definitions and examples of the four types of difficulties: aesthetics, data, geometry, and information.

REFERENCES

J. S. Allwood and P. Gärdenfors. 1999. *Cognitive Semantics: Meaning and Cognition*. John Benjamins Publishing Company.

J. G. Altonji and R. M. Blank. 1999. Race and gender in the labor market. *Handbook of Labor Economics* 3, 3143–3259.

R. Amar, J. Eagan, and J. Stasko. 2005. Low-level components of analytic activity in information visualization. In *InfoVis*.

F. J. Anscombe. 1973. Graphs in statistical analysis. *American Statistician* 27, 17–21.

L. W. Barsalou. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22, 4, 577–660.

R. A. Becker, W. S. Cleveland, and M.-J. Shyu. 1996. The visual design and control of trellis display. *Journal of Computational and Graphical Statistics* 5, 2 (1996), 123–155.

A. J. Berinsky, G. A. Huber, and G. S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* (R. M. Alvarez, ed.). DOI:http://dx.doi.org/10.1093/pan/mpr057

J. Bertin. 1983. *Semiology of Graphics* (W. J. Berg, trans.). University of Wisconsin Press, Madison, WI.

T. J. Blocker and D. L. Eckberg. 1997. Gender and environmentalism: Results from the 1993 general social survey. *Social Science Quarterly* 78, 841–858.

M. Canham and M. Hegarty. 2010. Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction* 20, 2, 155–166. DOI:http://dx.doi.org/10.1016/j.learninstruc.2009.02.014

P. A. Carpenter and P. Shah. 1998. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied* 4, 2 (1998), 75.

S. M. Casner. 1991. Task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics* 10, 2 (1991), 111–151.

C. Chen and M. Czerwinski. 2000. Empirical evaluation of information visualization: An introduction. *International Journal of Human-Computer Studies* 53, 5 (2000), 631–635.

M. Chuah and S. Roth. 1996. On the semantics of interactive visualizations. In *IEEE InfoVis'96*. 29–36.

W. Cleveland and R. McGill. 1985. Graphical perception and graphical methods for analyzing scientific data. *Science* 229, 828–833.

C. Collins and S. Carpendale. 2007. VisLink: Revealing relationships amongst visualizations. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1192–1199.

W. Croft and D. A. Cruse. 2004. *Cognitive Linguistics*. Cambridge University Press.

H. M. Culbertson and R. D. Powers. 1959. A study of graph comprehension difficulties. *Educational Technology Research and Development* 7, 3 (1959), 97–110.

F. R. Curcio. 1989. *Developing Graph Comprehension. Elementary and Middle School Activities.* National Council of Teachers of Mathematics.

C. Danis, F. Viegas, M. Wattenberg, and J. Kris. 2008. Your place or mine: Visualization as a community component. In *CHI'08*. 275–284.

G. DeSanctis. 1984. Computer graphics as decision aids: Directions for research. *Decision Sciences* 15, 463–487.

M. Dörk, S. Carpendale, C. Collins, and C. Williamson. 2008. VisGets: Coordinated visualizations for Web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1205–1212.

D. Downey. 1995. When bigger is not better: Family size, parental resources, and children's educational performance. *American Sociological Review* 60, 746–761.

V. Evans and M. Green. 2006. *Cognitive Linguistics: An Introduction*. Lawrence Erlbaum Associates.

H. D. Fain, E. L. Kick, B. L. Davis, and T. J. Burns. 1997. World-system position, tropical climate, national development, and infant mortality: A cross-national analysis of 86 countries. *Human Ecology Review* 3, 197–203.

G. Fauconnier. 1994. *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge University Press.

G. Fauconnier and M. Turner. 2008. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

C. J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6, 2 (1985), 222–254.

L. Frank, M. Andresen, and T. Schmid. 2004. Obesity relationships with community design, physical activity, and time spent in cars. *American Journal of Preventive Medicine* 27, 87–96.

M. Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32, 675–701.

S. N. Friel, F. R. Curcio, and G. W. Bright. 2001. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education* 124–158.

D. Gillborn and H. S. Mirza. 2000. *Educational Inequality: Mapping Race, Class and Gender. A Synthesis of Research Evidence*. Office for Standards in Education, London.

B. Glaser and A. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction.

E. Goldsmith. 1984. *Research into Illustration: An Approach and a Review*. Cambridge University Press.

D. Gotz and Z. Wen. 2009. Behavior-driven visualization recommendation. In *IUI*.

R. W. Hamming. 1950. Error detecting and error correcting codes. *Bell System Technical Journal* 29, 2, 147–160.

A. F. Hayes and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1, 1 (2007), 77–89.

J. Heer, F. B. Viegas, and M. Wattenberg. 2007. Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. In *CHI*.

J. Heer and M. Bostock. 2010. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *CHI'10*. 203–212.

J. Heer, N. Kong, and M. Agrawala. 2009. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *CHI'09*. 1203–1312.

J. Heer and G. Robertson. 2008. Animated transitions in statistical data. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2008), 1240–1247.

A. Inselberg and B. Dimsdale. 1991. Parallel coordinates. In *Human-Machine Interactive Systems*. Springer, 199–233.

P. G. Ipeirotis. 2010. Demographics of Mechanical Turk. *CeDER-10-01 working paper*.

W. Javed and N. Elmqvist. 2012. Exploring the design space of composite visualization. In *Pacific Visualization Symposium (PacificVis)*. 1–8.

M. Johnson. 1990. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press.

S.-H. Kim, H. Yun, and J. S. Yi. 2012. How to filter out random clickers in a crowdsourcing-based study? In *2012 BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization*. 15.

O. Kitamura, T. Kuroiwa, Y. Kawamoto, and E. Kaneko. 1998. A study on the improved tanker structure against collision and grounding damage. In *PRADS*.

S. M. Kosslyn. 1989. Understanding charts and graphs. *Applied Cognitive Psychology* 3, 185–226.

Z. Kövecses and G. Radden. 1998. Metonymy: Developing a cognitive linguistic view. *Cognitive Linguistics* 9, 37–78.

G. Lakoff. 1987. *Women, Fire, and Dangerous Things*. University of Chicago Press.

G. Lakoff and M. Johnson. 2008. *Metaphors We Live By*. University of Chicago Press.

J. Lemaire. 1985. *Automobile Insurance: Actuarial Models*. Hingham, MA: Kluwer-Nijhof.

J. Mackinlay. 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics* 5, 2, 110–141.

J. Mackinlay, P. Hanrahan, and C. Stolte. 2007. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visual and Computer Graphics* 13, 6 (2007), 1137–1144.

P. McCullagh and J. A. Nelder. 1989. *Generalized Linear Models*. Chapman & Hall.

R. Metoyer, B. Lee, N. H. Riche, and M. Czerwinski. 2012. Understanding the verbal language and structure of end-user descriptions of data visualizations. In *CHI'12*.

National Center for Health Statistics. 2000. National Health Interview Survey.

National Opinion Research Center. 1991. General Social Survey 1991 [United States].

National Opinion Research Center. 1993. General Social Survey 1993 [United States].

C. North. 2006. Visualization viewpoints: Toward measuring visualization insight. *IEEE Computer Graphic and Application* 26, 3 (2006), 6–9.

S. Pinker. 1990. A theory of graph comprehension. *Artificial Intelligence and the Future of Testing* 73–126.

H. Rosling. 2009. Homepage. Retrieved from www.gapminder.org.

S. F. Roth, M. C. Chuah, S. Kerpedjiev, J. Kolojejchick, and P. Lucas. 1997. Toward an information visualization workspace: Combining multiple means of expression. *International Journal of Human-Computer Interaction* 12, 1 (1997), 131–185.

S. F. Roth and J. Mattis. 1991. Automating the presentation of information. In *AAAI*.

W. S. Sarle. 1990. *The VARCLUS Procedure*.

P. Shah and E. G. Freedman. 2011. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science* 3, 3 (2011), 560–578.

P. Shah, E. G. Freedman, and I. Vekiri. 2005. The comprehension of quantitative information in graphical displays. *The Cambridge Handbook of Visuospatial Thinking* 426–476.

P. Shah, R. E. Mayer, and M. 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.

B. Shneiderman. 1996. The eyes have it: A task by data type taxonomy of information visualizations. In *IEEE Visual Languages'96*. 336–343.

D. K. Simkin and R. Hastie. 1986. An information-processing analysis of graph perception. *Journal of the American Statistical Association* 82, 454–465.

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

L. Talmy. 2000. *Toward a Cognitive Semantics*. Vol. 2. MIT Press.

J. J. Thomas and K. A. Cook. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society.

E. R. Tufte and G. Howard. 1983. *The Visual Display of Quantitative Information*. Graphics Press.

United Nations. 1995. *World Economic and Social Survey 1995*.

F. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. 2007. ManyEyes: A site for visualization at Internet scale. *IEEE Trans. Vis. Comp. Graphics* 13, 6 (2007), 1121–1128.

Z. Wen and M. Zhou. 2008a. Evaluating the use of data transformation for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1309–1316.

Z. Wen and M. X. Zhou. 2008b. An optimization-based approach to dynamic data transformation for smart visualization. In *IUI'08*. 70–79.

Z. Wen, M. X. Zhou, and V. Aggarwal. 2005. An optimization-based approach to dynamic visual context management. In *InfoVis*. 25–32.

H. Wickham. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer.

H. Wickham and H. Hofmann. 2011. Product plots. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2223–2230.

L. Wilkinson. 2005. *The Grammer of Graphics* (2nd ed.). Springer.

L. Wilkinson. 2012. The grammar of graphics. In Handbook of Computation Statistics (J. E. Gentle, W. K. Härdle, and Y. Mori, eds.). Berlin/Heidelberg: Springer, 375–414.

D. Woods. 1984. Visual momentum: A concept to improve the cognitive coupling of person and computer. *International Journal of Man-Machine Studies* 21, 229–244.

M. Zhou and M. Chen. 2003. Automated generation of graphic sketches by examples. In *IJCAI*.

M. X. Zhou and S. K. Feiner. 1996. Data characterization for automatically visualizing heterogeneous information. In *IEEE Information Visualization '96*. 13–20.

M. X. Zhou and S. K. Feiner. 1998. Visual task characterization for automated visual discourse synthesis. In *CHI*.

G. K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Received September 2012; revised May 2013; accepted October 2013