Interpretations of Data in Ethical vs. Unethical Data Visualizations

by

Shaun O'Brien

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2017 by the
Graduate Supervisory Committee:

Claire Lauer, Chair
Eva Brumberger
Barbara D'Angelo

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

This paper presents the results of an empirical analysis of deceptive data visualizations paired with explanatory text. Data visualizations are used to communicate information about important social issues to large audiences and are found in the news, social media, and the Internet (Kirk, 2012). Modern technology and software allow people and organizations to easily produce and publish data visualizations, contributing to data visualizations becoming more prevalent as a means of communicating important information (Sue & Griffin, 2016). Ethical transgressions in data visualizations are the intentional or unintentional use of deceptive techniques with the potential of altering the audience's understanding of the information being presented (Pandey et al., 2015). While many have discussed the importance of ethics in data visualization, scientists have only recently started to look at how deceptive data visualizations affect the reader. This study was administered as an on-line user survey and was designed to test the deceptive potential of data visualizations when they are accompanied by a paragraph of text. The study consisted of a demographic questionnaire, chart familiarity assessment, and data visualization survey. A total of 256 participants completed the survey and were evenly distributed between a control (non-deceptive) survey or a test (deceptive) survey in which participant were asked to observe a paragraph of text and data visualization paired together. Participants then answered a question relevant to the observed information to measure how they perceived the information to be. The individual differences between demographic groups and their responses were analyzed to understand how these groups reacted to deceptive data visualizations compared to the control group. The results of the study confirmed that deceptive techniques in data visualizations caused participants to misinterpret the information in the deceptive data visualizations even when they were accompanied by a paragraph of explanatory text. Furthermore, certain demographics and comfort levels with chart types were more susceptible to certain types of deceptive techniques. These results highlight the importance of education and practice in the area of data visualizations to ensure deceptive practices are not utilized and to avoid potential misinformation, especially when information can be called into question.

TABLE OF CONTENTS

Page

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Looking at the news, marketing campaigns, election advertisements, social media channels, or any number of current forms of print or digital communications and it becomes apparent that data visualizations are everywhere. Whether it's a bar chart to represent the difference between two or more values or a line chart to show trends over time – data visualizations are used to communicate quantifiable information about important social issues (i.e. politics, environment, health) to large audiences through various communication channels (Kirk, 2012). Furthermore, novice and expert communicators create data visualizations as part of their communication packages as a way of quickly delivering information. With the increased use of data visualizations by so many, it is important to understand how people use and understand data visualizations.

Data visualizations are defined as "the mapping between discrete data and a visual representation (Manovich, 2011, p. 2)." Others have defined data visualization similarly as communication of abstract data through the use of interactive visual interfaces (Keim et al., 2006) or "[c]omputer graphics and interaction to assist humans in solving problems" (Kerren et al., 2008, p. 58). While the definitions differ slightly, they all seem to repeat a similar message – data visualizations are used to provide information to people with a visual representation.

Although it might seem like data visualizations are relatively new, humans have used visuals to communicate information for centuries (Kirk, 2012). Academics and practitioners within industries like engineering and statistics were the original developers and designers of data visualizations (Sue & Griffin, 2016). Today, anyone with access to a computer has the potential of developing or designing data visualizations for any number of communication platforms.

Modern technology and software now allow people and organizations to more easily produce and publish data visualizations (Chen, Hardle, & Unwin, 2008), which has contributed to making data visualizations more prevalent as a means of communicating important information (Sue & Griffin, 2016). Increased amounts of data that people and organizations now collect about all areas of our lives has necessitated an increase in our use of data visualizations to communicate large amounts of information quickly; however, an increased use of data visualizations to communicate important information has led to an increase in ethical transgressions with data visualizations (Sue & Griffin, 2016).

Because anyone can create data visualizations, and the information being communicated can be used to influence opinions about important issues, it is imperative that we study how using deceptive techniques might alter how readers interpret the information in the data visualization. Ethical transgressions in data visualizations are the unintentional or intentional use of deceptive techniques with the potential of altering the audience's understanding of the information being presented (Monmonier, 1991; Tufte, 1983). Increased pressure to turn around quick materials, the strong desire to mislead the audience, inexperience in creating data visualizations, and lack of familiarity with statistics are some of the reasons ethical transgressions occur with data visualizations (Pandey et al., 2015).

While these are some of the reasons ethical transgressions or deceptive practices might occur with data visualization, best practices and procedures for creating sound data visualization has been around for some time (Huff, 1954; Jones, 2011; Monmonier, 1991; Tufte, 1983). Scholars have long discussed the importance of adhering to ethical standards when developing data visualizations, but research about the effects of not adhering to these standards require further research and understanding.

While these practices and procedures have been around for some time, scientists have only recently that started to study how deceptive practices might influence the reader. A recent study by Pandey et al. tested the deceptive nature of visualizations by looking solely at how people interpret information presented to them through a variety of data visualizations, and the results determined participants were more likely to be misled in their interpretations of deceptive data visualizations (2015).

The recent Pandey et al. (2015) study was important as it provided insight into the deceptive potential of data visualizations when they contained some element of deception; however, the study focused on the data visualizations as a stand-alone component. Data visualizations are usually not stand-alone components, but typically presented across a variety of communication media often accompanied by text.

This study was designed to test and answer the question: In what ways does accompanying data visualizations with explanatory text change users' interpretations of the visualizations? By adding a paragraph of accurate text, the study attempts to mimic the way data visualizations are typically presented across various communication media (i.e. as both text and visualization in a newspaper, magazine, report, or advertisement). By incorporating these changes into the study, we can further measure the extent or severity of different distortion techniques in potentially deceiving the reader.

CHAPTER 2

BACKGROUND LITERATURE

What is Data Visualization?

We can see data visualizations everywhere we look – from television, news, internet, magazines, reports, or journals (Kirk, 2012). Some are using data visualizations as a complementary component to written text, while others are using data visualizations independently as a way to reduce the amount of written text (Pasternick & Utt, 1989). In addition to traditional data visualization types (i.e. charts and graphs), new methods of data visualization are being created through the use of new computing software (Kirk, 2012).

Chen, Härdle, and Unwin state that data visualizations or graphic displays are viewed as a great way of communicating information (2008). Data visualizations provide the viewer with a way of viewing trends, patterns and anomalies in quantitative and qualitative data that is not possible with text alone (Friendly, 2008). Data visualizations are an attempt by the communicator to display information in a way that is easier or quicker to understand than text-based methods by leveraging our ability to interpret data visually (Sue & Griffin, 2016).

With the increased popularity and use of data visualization as a means of communicating large amounts of information, it might seem like data visualizations are relatively new; however, data visualizations have been around for a long time and used for centuries (Sue & Griffin 2016).

In fact, the earliest examples of data visualizations were geometric diagrams illustrating the positioning of stars in the sky (Chen, Härdle, & Unwin, 2008). Many have discussed the theories and best practices for the creation and development of data visualizations (Huff, 1954; Jones, 2011; Monmonier, 1991; Tufte, 1983). Chen, Härdle, and Unwin's milestone project highlights the significant milestones achieved in data visualization and describes how data visualizations span as far back as pre-17^{th} century (2008). According to Friendly, the golden age of data visualization and statistical graphics was roughly between 1850 and 1900 with the invention of our most commonly used chart and graph types (2008). Some of the earliest forms of our most commonly used data visualization (i.e. charts and graphs) were visual representations of statistical data (Chen, Härdle, & Unwin, 2008) generally created by those within academics or practitioners within industries like engineering and statistics (Sue & Griffin 2016).

Karen Schriver discusses the importance of overall document design and the proper incorporation of both text and image on the page for the purpose of aiding the reader (1997). This work not only describes the proper application of document design, but it also highlights that data visualizations are not typically standalone components and are often combined with explanatory text.

While data visualization itself is not new, powerful new technology and systems capable of producing stunning visuals have made it easier for people to create data visualizations for inclusion by those from the fringe and mainstream over the past decade (Kirk, 2012). Today, anyone with access to a personal computer has the ability to create data visualizations (Sue & Griffin 2016). According to Sue and Griffin, computer processing of statistical information and the rapid adaptation of the personal computers in the 1980s has provided people with a new instrument for easily producing graphs (2016).

Similarly, Chen, Härdle, and Unwin continue to discuss how computers have been a great benefit to the increased use and production of data visualizations (2008).

Between 1984 and 1988, newspapers with graphics capabilities increased from 40 to 90 percent, and it was predicted that graphics would overtake photographic images within newspapers (Pasternick & Utt, 1989). Furthermore, data visualizations and graphics were starting to be viewed as a basic communication tool and could potentially replace large amounts of text in newspapers (Pasternick & Utt, 1989).

The increase in new display techniques for data visualizations requires good standards and practice to ensure that data visualizations are providing the reader with information that can be easily understood (Wainer, 1984).

Ethical Standards in the Communication Profession

The role of the professional communicator is to communicate truthful information to readers or consumers so that they can easily understand and interpret the data to make the reader informed (Skau, 2012). Data visualizations as a communicative tool allow for large amounts of information to be presented rather concisely in a visual representation (Kirk, 2012). What is the ethical obligation of the professional communicator in the development of data visualizations?

The Society of Professional Journalism publishes a list of ethical issues and guidelines for those within the field of journalism to follow (2014). Within each of their main ethical codes are some additional descriptions and guidelines that help journalists clarify their roles and responsibilities when creating content. According to the guideline, communicators have a responsibility to the audience to be truthful, minimize harm, act independently, and be accountable (Society of Professional Journalism, 2014). Based on these rules, the role of the communicator is to ensure that the audience receives truthful information including information found in data visualizations.

Additionally, Skau points out that it is the role of the information developer to represent the information and present it to the reader truthfully (2012). Data visualizations carry the same ethical importance as other forms of communication, and "working with data raises important ethical questions" (Ethics of Data, 2015).

Similar to journalists, technical communicators must follow a set code of ethics. According to the Society for Technical Communication (STC), "as technical communicators, we observe the following ethical principles in our professional activities" listing legality, honesty, confidentiality, quality, fairness, and professionalism as the main ethical categories for technical communicators (1998, p. 1).

While accurately displaying information is important, it is also important to keep in mind that information design utilizes rhetoric (Kinross, 1985). Information in itself is subject to rhetorical decisions on behalf of the reader and comes with rhetorical infiltration as soon as the designer gives it shape (Kinross, 1985). For these reasons, designing information applies the same rhetoric infiltration that photographers use as they crop images or photos – exposing the reader only to certain parts of information and not the whole picture. Additionally, our own perspectives change the meaning of images or information, which means that the reader's interpretation of the information is subject to their own perspectives.

While it is important to adhere to guidelines and best practice, it is also important to understand that data visualizations are still abstract forms of communication that the designer is using to relay information.

Ethical Issues with Data Visualization

Ethics in data visualization is not a new concept and the desire to alter statistics to shape the message has been in discussion for some time (Sue & Griffin 2016). Many have discussed the importance of standards and best practices for the development of data visualizations (Schriver, 1997). Scholars have long talked about the importance of ethically representing data.

In the 1950s, "How to Lie with Statistics" discussed various ways in which statistical information could be misinterpreted (Huff, 1954). In the 1980s, Eduard Tufte discussed the concept of graphical integrity and the lie factor to described ways that visual information could alter the reader's perception of information (1983). Similarly, two other publications expanded on the same concepts: "How to Lie with Charts" (Jones 2011), and "How to Lie with Maps" (Monmonier, 1991).

Researchers have studied the distortion of information caused by visual encoding. Such research has looked at how visual encoding of data is perceived and compared when the position, size, color, and angle of the data visual were represented differently (Bertin, 1983; Cleveland & McGill, 1984; Rogowitz, Treinish, & Bryson, 1996).

While these pivotal works all provide valuable information on the proper application and practice for the creation of data visualizations, researchers have only recently started to study to what extent would someone's perception of the information be affected when distortion techniques were utilized in a data visualization.

Linda, Martin, Cantor, and Rubenstein conducted a research project that looked at how physician's decisions regarding continuation of clinical trials could be influenced depending on the type of data visualization presented (1999). The results determined that the type of data display affected the physicians' decisions regarding the continuation of clinic trials. Additionally, the study found that more accurate recommendations were made when physicians observed icon displays versus traditional chart types like tables, pie, and bar charts (Linda, Martin, Cantor, & Rubenstein, 1999).

A recent study tested the deceptive nature of visualizations by looking solely at the way people interpreted information presented to them through a variety of data visualizations (Pandey et al., 2015). The study consisted of a user survey in which participants were randomly selected to analyze "control" and "test" data visualizations. The "test" data visualizations incorporated deceptive practices commonly utilized in data visualizations, such as Message Exaggeration/Understatements and Message Reversal. The results of the study determined that participants were more likely to be misled in their interpretations and incorrectly answer questions pertaining to the visualization when deceptive practices were used.

It has been pointed out that data is a representation of real life (Yau, 2013). This statement implies that data or the representation of data is a mirror of fact-based events within real life. One definition of data visualization states "[t]he representation and presentation of data that exploits our visual perception abilities in order to amplify cognition" (Kirk, 2012, p. 17). Analyzing these two statements together, the role of the information developer is to ensure data visualizations are both truthful to real life and represented in a way that it is easy to comprehend.

Knowing that the goal of information design is to represent understood data for the purpose of amplifying cognition, information developers should be mindful of ethical issues surrounding data visualization to ensure they are not exploiting the reader's perceptions of fact by injecting fiction or representing the data incorrectly. Furthermore, our ethical codes of conduct as information developers mandates that we provide truthful information to our audience.

Deceptive Data Visualization Techniques

Pandey et al. point out that visualization deception occurs at two levels – the chart level where the chart is interpreted incorrectly, and/or the message level where the message is interpreted incorrectly (2015). Additionally, Pandey et al. have categorized deceptive practices in data visualizations into two main categories, Message Exaggeration/Understatement and Message Reversal (2015).

Message Exaggeration/Understatement

Message Exaggeration/Understatement occurs when the facts are not distorted, but the way the information is presented is altered to intentionally or unintentionally exaggerate the facts, see Figure 1 and 2 for examples of Message Exaggeration/ Understatement (Pandey et al., 2015). Types of ethical transgressions or deceptive techniques used in Message Exaggeration/Understatement include a truncated axis, area as quantity, and aspect ratio. Each of these examples employs altering some element of the visual to exaggerate or reduce the appearance of visual for the desired appearance.

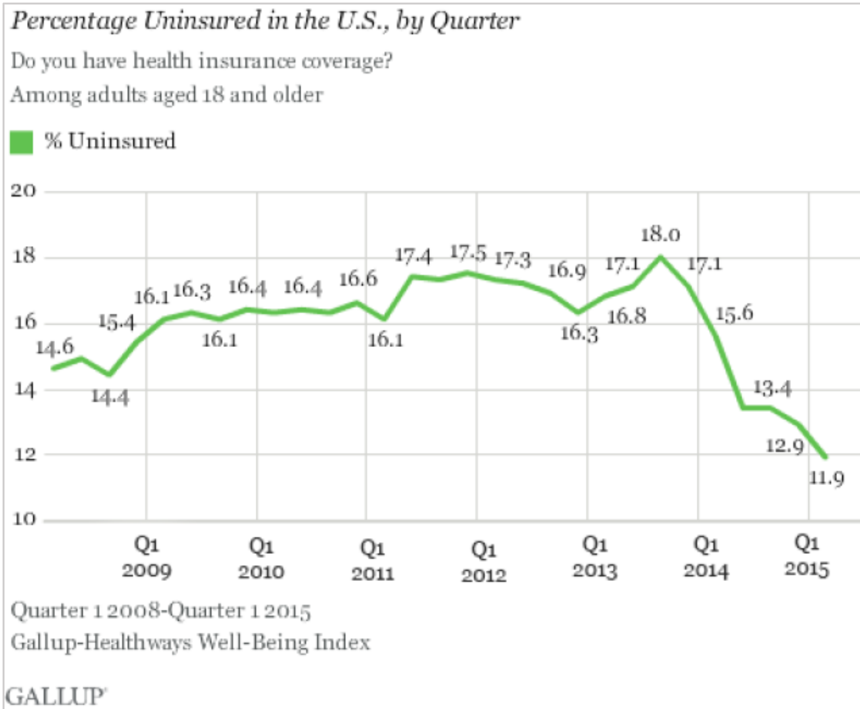


Figure 1 – Message Exaggeration (Truncated Axis)

Figure 2 - Message Exaggeration (Area as Quantity)

Message Reversal

Message Reversal encourages the user to interpret the fact in the message incorrectly, see Figure 3 for an example of message reversal (Pandey et al., 2015). The most common ethical transgression or deceptive technique used for Message Reversal is when the axis of a chart is inverted or flipped.

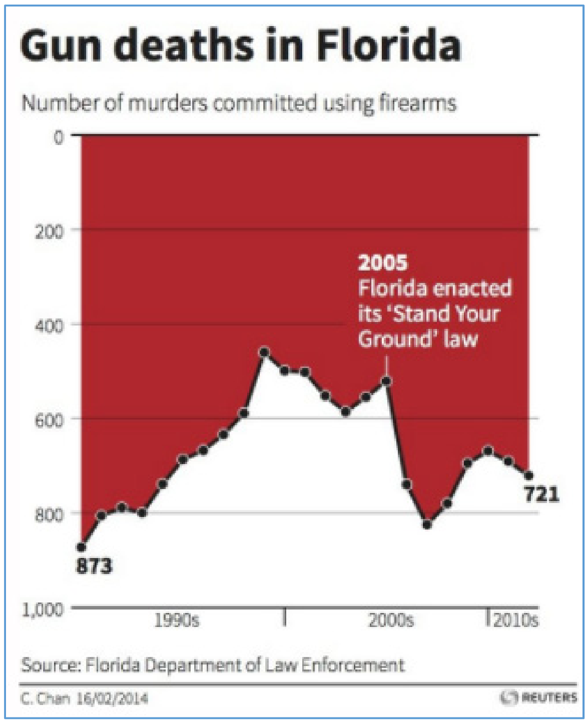


Figure 3 - Message Reversal (Inverted Axis)

These ethical transgressions occur both intentionally and unintentionally requiring extra attention to detail when developing data visualizations. For novice and experienced information developers, the desire to mislead, inexperience with statistics, emphasis toward aesthetics and graphics, and oversimplification are some potential reasons for ethical transgressions. With the increased use of data visualization across multiple disciplines, the need for ethical standards with data visualizations is strong (Bowen, 2013).

The literature review highlights both the importance of ethical practices with data visualizations and previous scholarly attempts to address this importance. However, it was only recently that researchers have started to look at how unethical practices with data visualizations might affect the reader. As important and informative as the Pandey et al. (2015) study was in looking at the deceptive potential of data visualizations, it also leads to some additional questions. How deceptive would the data visualization be when also accompanied by a paragraph of text that reports the same data accurately?

CHAPTER 3

METHODOLOGY

With the increase of data visualizations as a means of presenting important information quickly, this study was designed to understand more about how people interpreted the information presented in data visualizations. While scholars have long discussed theory and best practices for data visualization creation and usage, it was only recently that a study was conducted to understand truly how people read data visualizations.

The study conducted by Pandey et al. (2015) helped inform on the deceptive potential of data visualizations as standalone components and consisted of participants taking an online survey in which they analyzed data visualization and then answered a “how much” question to measure the individuals understanding of the data presented in the data visualization. The Pandey et al. (2015) study was designed to measure the difference between participant responses to survey questions for data visualizations without deceptive practice compared to data visualizations with some element of distortion. Additionally, the Pandey et al. (2015) measured participants chart familiarity and demographics to see if there were differences between certain types of participants. Participants were randomly assigned either a non-deceptive data visualization or deceptive data visualization and the results were compared statistically to measure the difference.

While the Pandey et al. (2015) study helped inform on the deceptive potential of data visualizations as standalone components, this study differs and was designed to understand in what ways accompanying data visualizations paired with explanatory text changes users' interpretations of the visualizations.

Like the Pandey et al. (2015) study, this study was designed to measure the difference in responses to deceptive versus non deceptive data visualization, but it was also designed to analyze the role of explanatory text with the data visualization to see if it would have any impact on the deceptiveness of the data visualization. Because data visualizations are rarely standalone, the purpose of adding the paragraph of text was to mimic the way data visualizations are typically presented across various communication media (i.e. as both text and visualization in a newspaper, magazine, report, or advertisement).

The approach and methodology of this study were to conduct an online survey of randomly sampled participants. The participants would complete a chart familiarity assessment, demographic questionnaire, and be randomly assigned either a test or control survey similar to the Pandey et al. (2015) study. Additionally, this study builds on the previous Pandey et al. (2015) study by asking several follow up questions upon completion of the control and test survey to measure where the participants received their information during the observation period of the survey.

This methodology section will describe the steps and process involved throughout the study and will provide detail on decisions made and procedures used. The methodology section will describe the creation of each section of the survey, participant recruitment, and how the survey was administered throughout the study. The control and test survey is included in Appendix A.

This study was designed similar to the previous Pandey et al. (2015) study. It incorporated a consent form, demographic questionnaire, chart familiarity assessment, and survey. All study components were designed and administered using an online survey tool provided by Qualtrics. Prior to disseminating the study to any participants, the entirety of the study and protocol was reviewed and approved by Arizona State University's Institutional Review Board. This section will describe each of the sections created for this study. See Appendix B for a copy of the Consent Form.

An Informed Consent Form was provided to all potential participants at the beginning of the online survey. The consent form provided participants with the title of the study, study details, potential benefits associated with their participation in the study, potential compensation (drawing for gift card), contact information for the PI and Arizona State University's Office of Research Integrity and Assurances (ORIA), and an option to continue the study by acknowledging consent or declining. Institutional Review Board approval was obtained through Arizona State University's ORIA prior to starting the study.

Demographic Questionnaire

Similar to the Pandey et al. (2015), a demographic questionnaire was designed and administered with the study. The original demographic questionnaire form the Pandey et al. (2015) was not accessible and the only demographic information reported was the education level of the participants. The demographic questionnaire for this study asked participants to provide their age and education level.

The information obtained from the demographic questionnaire helped to measure if age and education correlated to participant likelihood of being deceived by data visualizations. The questionnaire also asked participants if they had completed coursework related to data visualization. This question was designed to measure the participant's chart literacy and potential to identify potential deceptive practices utilized in the study.

The age demographic question was important to help understand the differences between age groups and their likelihood of being deceived by deceptive data visualizations. No participants under the age of 18 years old were recruited or allowed to take this survey. Participants were asked to select one of four possible responses to the following question: what is your age? The four possible four possible responses included:

- 18-29 years old,
- 30-49 years old,
- 50-64 years old, and
- 65 years and older.

The education demographic question, similar to the age question, was important to help understand the differences between education level groups and their tendencies in reading or interpreting information in data visualizations and accompanying text. Participants were asked the following question for the education demographic: What is the highest level of school you have completed or the highest degree you have received? Participants could respond by selecting one of eight possible options, which included:

- Less than high school degree,
- High school graduate or Equivalent GED,

- Some college but no degree,
- Associate’s degree,
- Bachelor’s degree,
- Master’s degree,
- Doctoral degree, and
- Professional degree (MD).

In order to measure how many participants had taken a course in data visualization, the following survey question was added as a yes or no question: Have you taken any courses in creating charts and graphs with data or visualizing data? This question was important in understanding the percentage of participants that would be familiar with potential deceptive techniques because they were discussed in a data visualization course. Even though the test was designed in a way that participants were not aware of the deceptive techniques in test treatments, participants familiar with data visualizations and chart literacy might cause the participant to find the distortion techniques and affect the overall study results. For this reason, it was important to measure how many participants could be familiar with data visualization or have this type of visual literacy by asking them if they had taken a course before.

Chart Familiarity Assessment

Similar to the Pandey et al. (2015), the chart familiarity assessment asked participants to rate their familiarity with certain chart types. It was not clear how this was measured in the Pandey et al. (2015) study. This information was important to help understand if comfort levels had any correlation to how someone might answer the survey. Would participants that said they were familiar or comfortable with a certain chart type spend less time analyzing the chart versus those with less familiarity? Additionally, these questions helped gauge the overall comfort level of the participants on certain chart types. Figures 4 through 7 show the charts created for the chart familiarity assessment.

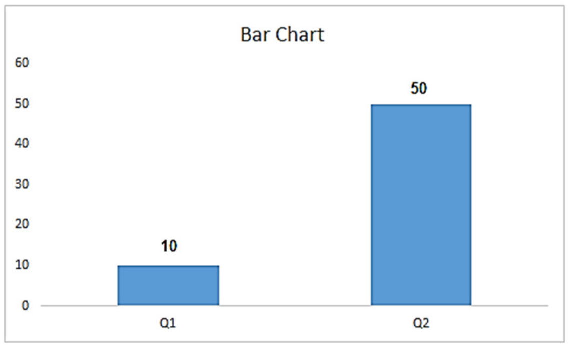


Figure 4 – Bar Chart Familiarity

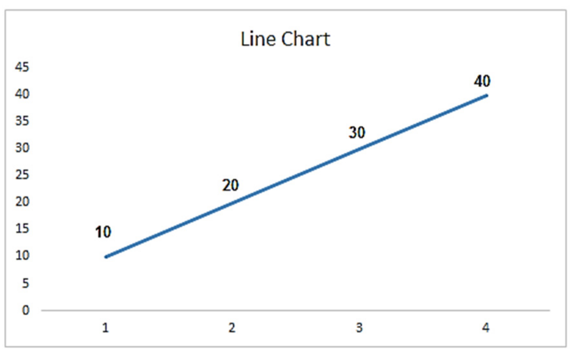


Figure 5 – Line Chart Familiarity

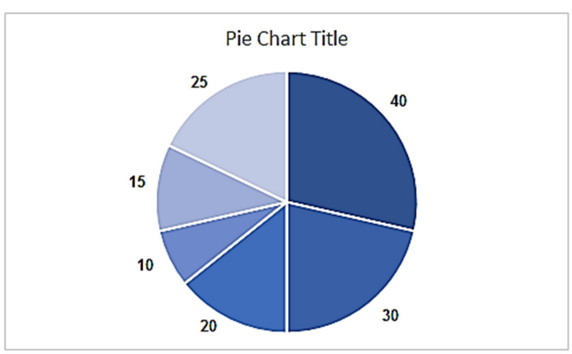


Figure 6 – Pie Chart Familiarity

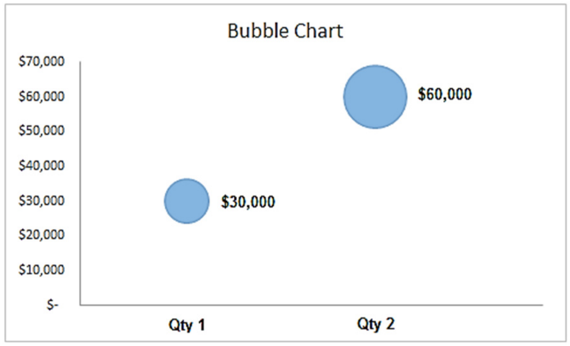


Figure 7 – Bubble Chart Familiarity

The design of the data visualizations for the chart familiarity utilized the same colors for all chart components (blue) and text (black). The font used throughout all of the charts shared the same font type (Ariel). Color in charts can be used as an aesthetic component, but it can also be utilized as a way of introducing new data or information. By controlling the color and only utilizing one color scheme the test reduced the amount of visual information the participant was being asked to process. Similarly, the use of only one text font reduced the amount of visual information or noise for the participant. The data visualizations all contained chart titles, data labels, and chart titles that were all created using the same font type.

Participants were asked to rate their comfort level with Bar, Line, Pie, and Bubble charts. These are four commonly used chart types and three of the four were also chart types utilized in the test and control survey questions. Participants were shown an image of the chart type and were asked how comfortable they were with that chart type. Participants could answer the question by selecting select only one of the following four options: Uncomfortable, Slightly Uncomfortable, Slightly Comfortable, and Comfortable. Figure 8 shows an example of the chart with the chart familiarity question.

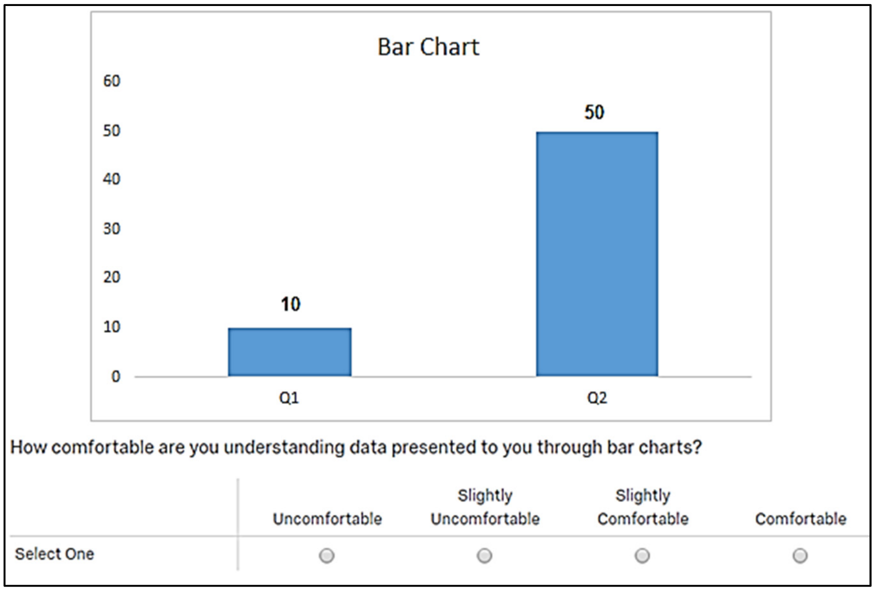


Figure 8. Chart Familiarity Question Example

Data Visualization Survey

Similar to the Pandey et al. (2015) study, this study required two treatments. The control treatments consisted of data visualizations free from distortion or deceptive practices. The test treatments consisted of the same data visualization types used in the control treatment but incorporated a single element of deception. Additionally, this study would include explanatory text that would accompany both the control and test data visualization to test how this might change how participants would respond to survey questions when provided both text and the data visualization. In order to develop the necessary components for the proposed study, the initial steps consisted of:

- Selecting the types of data visualization to test,
- Creating the data for the treatments,
- Creating control and test data visualizations, and
- Writing paragraphs of text.

Selecting the types of data visualization to test

Similar to the Pandey et al. (2015) study, the type of data visualization used in this study consisted of a bar, line, and bubble charts. By keeping the same chart types as the original study, we are able to compare the original results with the results of our study to measure whether or not the addition of explanatory text changed the results.

Creating the data for the treatments

To differentiate slightly from the previous Pandey et al. (2015) study, the data to be used in the paragraphs and data visualizations would be new and different than the treatments provided in the previous study. The Pandey et al. (2015) treatments were all created using similar data; for example, the bar chart and bubble chart were both using the same percentages and data. The participants of the previous study only received one treatment as opposed to this study where the participants would receive all three control or test data visualizations.

Additionally, the information in the paragraph and data visualization were designed to be relevant to topics or scenarios that would typically pair both explanatory text with a data visualization. The bar chart data was designed to highlight information typically found in local real estate magazines or advertisements. The line chart is data was designed around the popular topic of health care coverage that can be found on the internet or television. Lastly, the bubble chart data was designed based on a scenario of information commonly found in advertisements or marketing materials for universities in an attempt to attract potential students.

Writing Paragraphs of Text

Prior to creating the actual data visualization for the control and test treatments, a paragraph of text was created for each of the scenarios to be used for the bar, line and bubble charts. The same paragraph of text accompanied both the control and test treatments as a way of limiting variables. Additionally, by having the same paragraph of text with both the control and test data visualization, the difference between the control and test data visualization could still be measured and compared against the original study since the only new variable was the addition of a paragraph of text. Figures 9 through 11 show the paragraphs for the scenarios to be included with the bar, line, and bubble charts.

> ABC Homes recently compared home prices for the Phoenix metropolitan area. Based on statistics, we saw 16,303 single-family homes sold in Phoenix, AZ in 2016 compared to 15,509 in 2015, an increase of roughly 5%. In 2016, the median sold price for a single-family detached home in Phoenix, AZ was $230,000 compared to $210,900 in 2015.

Figure 9 – Bar Chart Paragraph

> The percentage of uninsured Americans has seen some fluctuation over the course of the past four years. From January 2009 to July 2014, the highest percentage of uninsured reached 18.4 percent in 2011. The lowest percentage of uninsured comes in at 12.9 percent in 2014. That constitutes a 5.3 percent drop in the percentage of uninsured Americans.

Figure 10 – Line Chart Paragraph

> ABC University strives for excellence and innovation by offering in-demand degree programs and enriched learning opportunities to our students setting them up for success in their careers. Compared to the national average of 64.4 percent, 72.8 percent of ABC University students earn full-time employment upon graduation. That makes ABC University students 8.4% more likely to earn a full-time position after graduating.

Figure 11 – Bubble Chart Paragraph

Creating Control and Test Treatments

Following a similar methodology of Pandey et al. (2015), examples of ethical and unethical data visualizations were created and used during the control (ethical) and test (unethical) treatments portion of the online survey.

Similar to the Pandey et al. (2015) study, the data visualizations were created from common data visualization types (bar, line, and bubble). The information for both the control and test data visualization remained consistent with the information contained in the paragraph description. The difference between the control and test treatments was the use of common deceptive practices found with data visualizations.

The design of the data visualizations for the control and test treatments utilized the same color hue (blue) for all chart types and font style (Ariel). These design choices were utilized as a way of controlling the amount of visual information the participants would need to process. Font and color are both rhetorical design elements that can be used for aesthetic purposes but they also can be used to introduce new information and data for data visualizations. Figure 12 shows the difference between a test and control treatment.

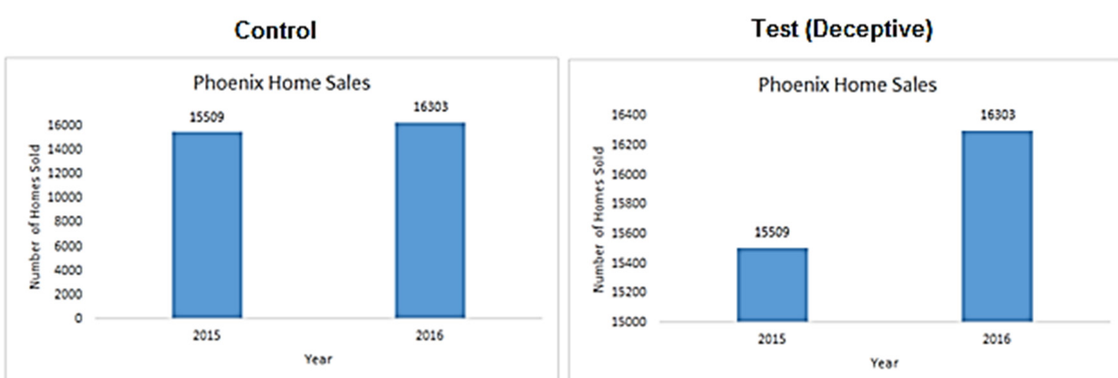


Figure 12. Control and Test Treatment Comparison

Control Treatments

The control treatments were created to look similar to the original Pandey et al. (2015) study but differed by using the information that was created for each scenario and paragraph of text.

The bar chart is used to illustrate visually the differences between certain values. For the control treatments, the x- and y-axis of the chart were set to zero. This allows the chart to show the true difference between the different bar values. If altered, the difference between the two bars in the chart would appear shortened or elongated resulting in the chart having an exaggerated or understated appearance. The data labels were left in to show the actual numbers for each data point. The color of the two bars in the bar chart was the same. By keeping the color the same, the participant would have less visual information to process and would focus more on the visual difference between the two bar lengths, the data labels, and other chart elements. The chart received a title of “Phoenix Home Sales” and the x-axis was labeled by “Years” and the y-axis was labeled by “Number of Homes Sold.” Figure 13 shows the bar chart control data visualization.

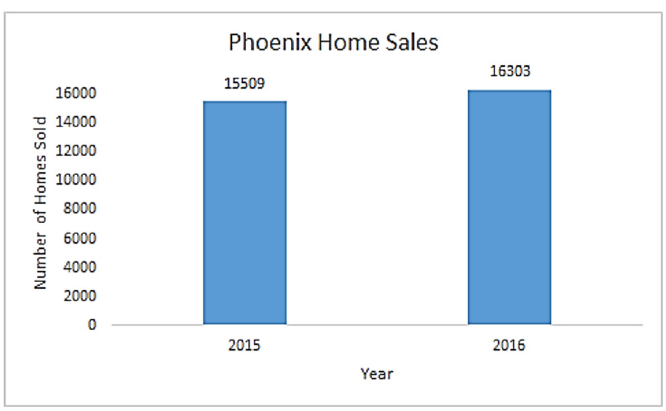


Figure 13 – Bar Chart Control Treatment

The line chart is a visual representation of data over a timeline to show a trend. The x- and y-axis of the chart were set to zero. This allows the trend line to accurately show the peaks and valleys of a trend line. If altered, the peaks and valleys of a trend line would become stretched or flattened and the values would be exaggerated or understated. Additionally, the x-axis was at the bottom of the chart. By having the x-axis inverted or at the top of the chart, the trend line would appear to have an opposite trend. For the line chart, several data points were used to create a trend line over the course of time. A semi-annual representation was used over the course of four years for a total of 12 data points. The trend line was designed to show a gradual upward trend but a sudden downward trend at the end of the chart to show a decrease in value. Only one color was used in the chart. This was done to allow the participant to focus just on the trend line in the chart, the data labels, and other chart elements. The chart received a title of "Percentage of Uninsured in America" and the x-axis was labeled by the semiannual date and the y-axis was labeled by "Percentage." Figure 14 shows the line chart control data visualization.

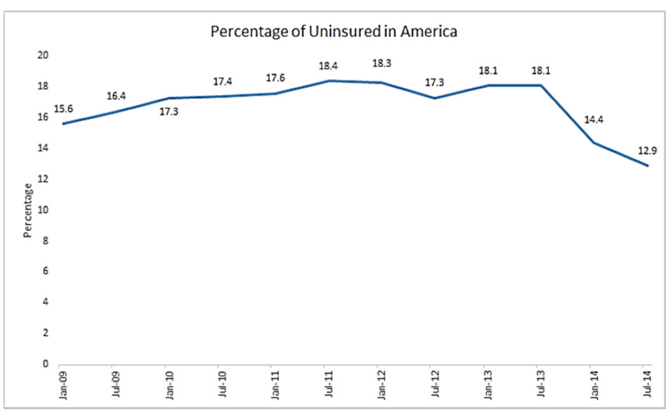


Figure 14 - Line Chart Control Treatment

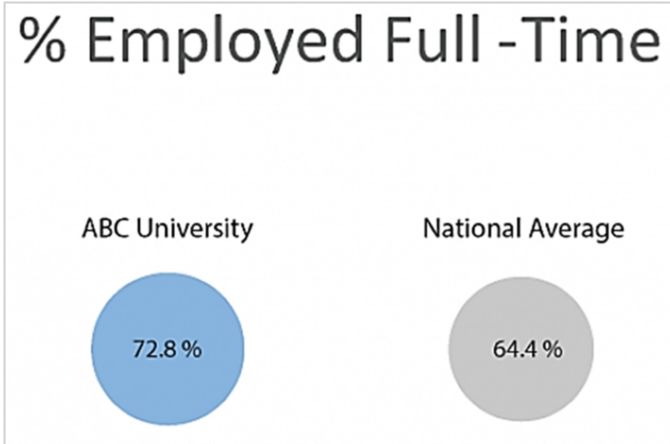


Figure 15 – Bubble Chart Control Treatment

The bubble chart is similar to the bar chart in that it is to show a visual representation of the difference between different data points. The bar chart does this by showing the height of a bar as the value of something, but the bubble chart does this by showing the size of a circle as the value of something. In this study, only two data points were used to create two bubbles for comparison. Two colors were used in the chart, one for each bubble. Data labels were included with each bubble to show the actual value the bubble was attempting to represent. The outer diameter of the bubbles or size of the bubble was created with the area of the bubble as the data point value. If the radius of the bubble was based on the data point value the bubble would have an exaggerated appearance. In addition to the data labels the chart received a title of "% Employed Full-Time." Figure 15 shows the bubble chart control data visualization.

Test Treatments

To limit variables between the previous Pandey et al (2015) study and this study, the chart types would receive the same method of deception as was tested in the Pandey et al. (2015) study. The bar chart received the Message Exaggeration/ Understatement deceptive technique known as “Truncated Axis.” The bubble chart received the Message Exaggeration/Understatement deceptive technique known as “Radius as Quantity.” The line chart received the Message Reversal deceptive technique known as “Inverted Axis.” The test treatments were created by taking the control treatments and employing the specific deceptive technique intended for each chart type.

The bar chart test treatment was altered by truncating the y-axis. The practice of truncating the y-axis can alter the appearance of the bars in the bar chart. To truncate the y-axis, the lower left corner was set to a value of 15,000. This made the lower valued bar appear shorter in comparison to the larger valued bar. All other elements of the chart were left unchanged to avoid additional variables and to test the effectiveness of the deceptive technique. Figure 16 shows the bar chart test data visualization.

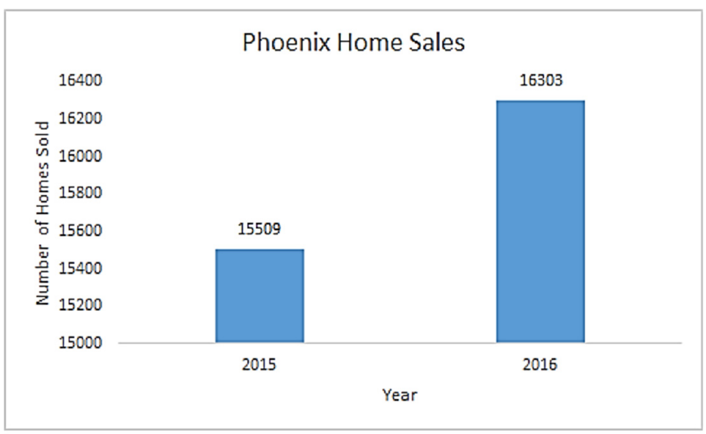


Figure 16 – Bar Chart Test Treatment

The line chart test treatment was altered by inverting the x-axis. The practice of inverting the x-axis alters the orientation of the trend line of a line chart. In this study, the x- and y-axis of the chart were set to zero. If the axis was set to anything other than zero the deceptiveness of the inverted axis would not be tested, but rather it would test the combination of both deceptive techniques when paired. To invert the x-axis the axis was positioned at the top of the chart. By doing this, two things happen to the chart. First, the values for the y-axis no longer increase upward but increase downward. Secondly, the trend line of the line chart is flipped having a reverse appearance. If the data illustrates an increase or decreased trend than the flipped chart could falsely indicate the opposite information. All other elements of the chart were left unchanged to avoid additional variables and to test the effectiveness of the deceptive technique. Figure 17 shows the line chart test data visualization.

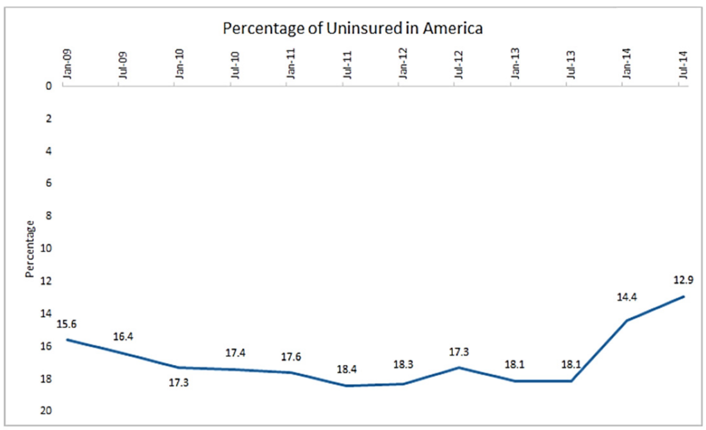


Figure 17 - Line Chart Test Treatment

The bubble chart test treatment was altered by changing altering the size of the bubbles through a deceptive technique known as “Radius as Quantity.” An unaltered bubble would have been created by having the area of the bubble be the value or quantity of the data point being represented. When the “radius as quantity” deceptive technique is used the outer radius of the bubble is based on the value or quantity of the data point being represented. To create the treatment for this study, both bubbles needed to be altered to have the radius of the bubbles by equal to the value being represented by the data label. Doing this made the larger valued bubble appear even larger than the smaller valued bubble when compared to the control bubble chart. All other elements of the chart were left unchanged to avoid additional variables and to test the effectiveness of the deceptive technique. Figure 18 shows the bubble chart test data visualization.

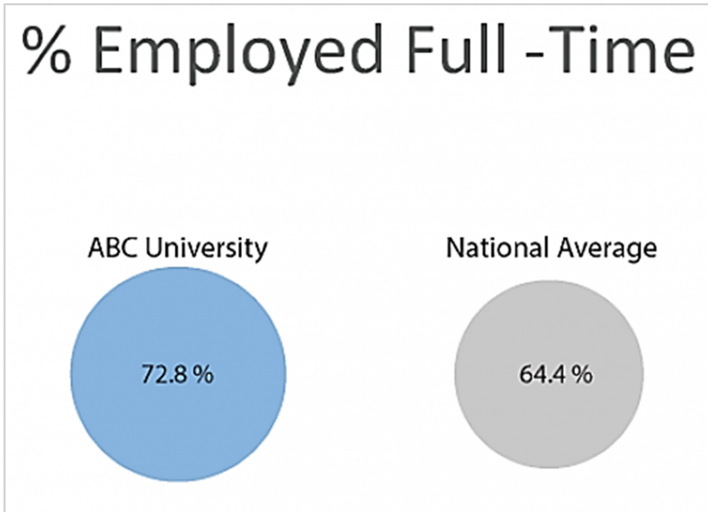


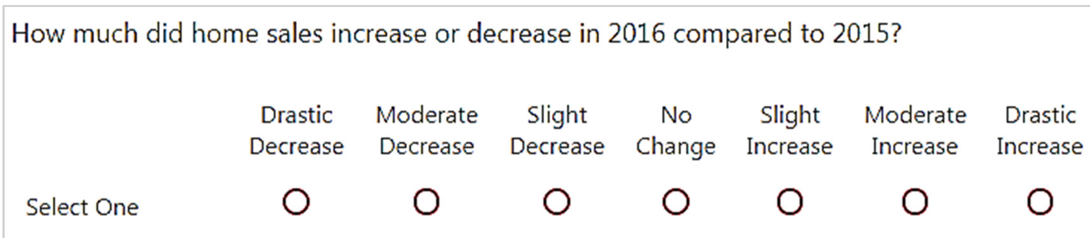
Figure 18 – Bubble Chart Test Treatment

Creating Survey Questions

Upon completion of the Demographic Questionnaire and Chart Familiarity assessment, the survey randomly assigned one of two treatments to participants. Participants were evenly distributed either the control survey or test (deceptive) survey. Each treatment took participants approximately 5-7 minutes to complete.

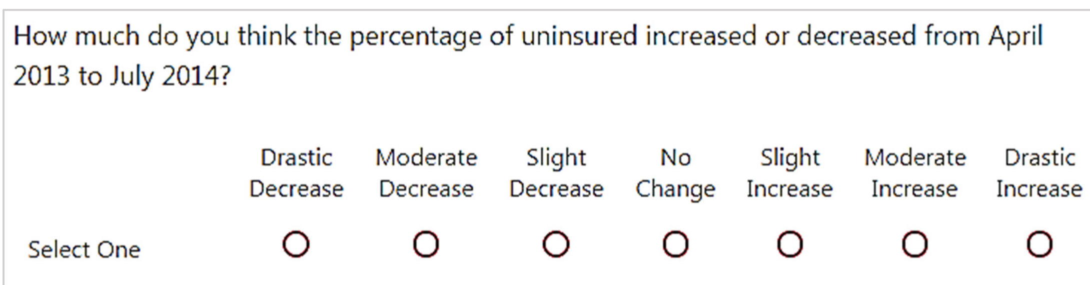
Participants were asked to examine both the paragraph of text and data visualization. Once the participant proceeded to the next screen, participants were then asked a question regarding the information from the previous screen. The survey questions for each scenario and chart type were designed as a “How much” question to measure the participant’s interpretation of the information. The question was phrased in a way that it left the answer up to the participant and that no one answer would be correct or incorrect but purely opinion. By doing this, we would see if the chart or data visualization influenced the participants understanding of the information. This was especially important for the deceptive data visualizations because it would help determine if the exaggerations or reversal of information influenced the participants' answers.

Participants were given seven possible responses to the survey questions. Figures 19 through 21 show the survey questions for each chart type.


How much did home sales increase or decrease in 2016 compared to 2015?

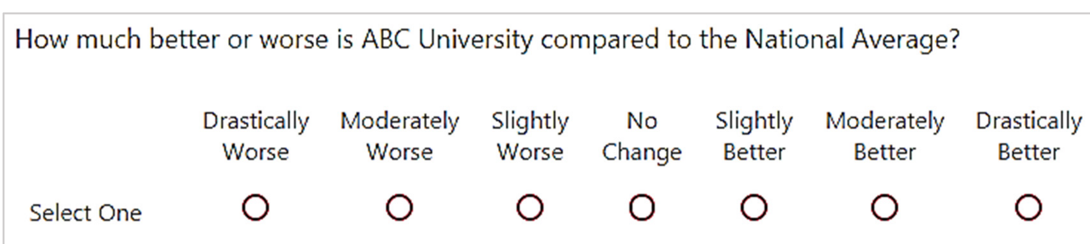
	Drastic Decrease	Moderate Decrease	Slight Decrease	No Change	Slight Increase	Moderate Increase	Drastic Increase
Select One	○	○	○	○	○	○	○

Figure 19 – Bar Chart Survey Question


How much do you think the percentage of uninsured increased or decreased from April 2013 to July 2014?

	Drastic Decrease	Moderate Decrease	Slight Decrease	No Change	Slight Increase	Moderate Increase	Drastic Increase
Select One	○	○	○	○	○	○	○

Figure 20 – Line Chart Survey Question


How much better or worse is ABC University compared to the National Average?

	Drastically Worse	Moderately Worse	Slightly Worse	No Change	Slightly Better	Moderately Better	Drastically Better
Select One	○	○	○	○	○	○	○

Figure 21 – Bubble Chart Survey Question

After completing an initial round of the survey, some modifications and changes were made to the survey to obtain some additional data.

The first modification was to enable a timing feature in Qualtrics to measure how long participants observed the paragraph of text and data visualization. This information would help understand how people took the survey, and help measure whether or not they spent enough time with each scenario prior to answering the questions. Lower times could indicate that participants only observed either the paragraph of text or data visualization but not both. Higher times could indicate that participants observed both the paragraph of text and data visualization prior to answering the survey question.

The second modification was the inclusion of a second set of questions to measure where participants were getting their information when answering the data visualization survey. Because the data visualization surveys included both a paragraph of text and data visualization, it was important to understand which combination or preferred data source they utilized to answer the subsequent question. For each chart type, the question "Where did you find your answer to the question?" was asked. The participant was given four possible options for a response, which included:

- In the paragraph of text
- In the chart
- Both in the paragraph of text and in the chart
- I'm not sure

In addition to selecting one of these responses, the participant could elaborate further and type a lengthier response in a freeform text box. This would provide a deeper and richer understanding of why participants chose to use one source versus the other.

In order to extrapolate the data obtained from the freeform responses, a qualitative analysis was performed by coding the participant responses. First, participant responses were looked at for any specific reasoning for why they chose to use one type of data source over another. Secondly, all responses were coded to ascertain a number of similar responses and to measure what data source participants utilized to answer the survey question. Participants that mentioned both the chart and text in their response were coded with a "B," participants that stated they only utilized the chart to answer their question were coded with a "C," and participant that said they only utilized the paragraph of text were coded with a "T."

Participant Recruitment

Participants were recruited from a pool of Psychology 101 students, as well as faculty and staff at Arizona State University's Polytechnic campus and other ASU students, faculty and staff where available. Additionally, participants were recruited from a national listserv after a message about the survey was posted to the WPA-l, which is a listserv of writing program administrators. The goal of the study was to recruit as many participants as possible within the timeframe of the project period with an original goal of approximately 50 to 100 total participants. All participants needed to be 18 years old or older to participate in the study and no parent permission was necessary. Of the recruited participants, approximately half were exposed to the control treatment and half were exposed to the test (deceptive) treatment.

For participation in the study, participants were given the option of providing a contact e-mail address at the end of the study and entered into a drawing. The drawing was a chance to win one $25 gift card. The $25 gift card provided some incentive to participants for completing the survey while not creating undue pressure. Information regarding the gift card was provided in the Consent Form at the being of the survey.

Participants were convenience sampled largely from students and faculty who were affiliated with Arizona State and similar universities. As one of the largest universities in the United States, however, Arizona State University offers a unique opportunity to sample from a large, diverse group of individuals with varying cultures, perspectives, and experiences.

CHAPTER 4

RESULTS

This study consisted of a user survey designed to understand in what ways does accompanying data visualizations with explanatory text change users' interpretations of the data visualizations when they contain deceptive techniques. The study examined the use of message exaggeration and message reversal techniques on data visualizations and pairing them with explanatory text. As discussed in the Methodology section, the study consisted of a demographic questionnaire, chart familiarity assessment, and data visualization survey. This section will discuss the participant breakdown and study results.

Study Participation

A total of 305 participants were recruited; 256 participants completed the informed consent form and answered the survey. The control and test surveys were evenly distributed with 128 receiving the control treatment and 128 receiving the test treatment.

The original survey was designed to end after the chart survey, but initial results and further analysis revealed that including additional questions that asked participants where they derived their answers from would help shed light on why people chose the answers they did (e.g., if they were deceived by a visualization, was it because they only looked at the visual and did not read the text?). Of the 305 recruited participants, 114 recruited participants received the updated survey, with 100 of those participants acknowledging consent and answering the survey questions. These 100 participants are included in the total 256 participants that also took the initial study.

Table 1 shows the total distribution of participants between the control and test treatments for each chart type, including those who had the additional qualitative questions and those who did not.

Table 1 – Participant Distribution

	Control	Test
Bar	128	128
Line	126	124
Bubble	121	122

Quantitative Results

In order to determine whether there was a significant difference between the control and test results, the results of the data visualization study were analyzed using the Mann-Whitney U-Test. Because the recruited participants were randomly selected and assumptions about the population cannot be made, a non-parametric test was required to conduct the analysis. The Mann-Whitney U-Test is designed to measure the statistical significance or differences between two samples and is comparable to the parametric students' T-test. Furthermore, the original Pandey et al. (2015) study also utilized the Mann-Whitney U-Test to determine the statistical significance of the control and test data.

The U-test produces a statistical p-value used to determine if differences between two groups are statistically significant. When the p-value measures $< .05$ the difference between the two groups is considered to be significant, while p-value $>.05$ are not considered significant. For the purposes of this study, the measurable difference or p-

value indicates that participants were susceptible to the deceptive technique. For example in the case of a bar chart (truncated axis) survey, the significance of the results means that the values recorded by the test group were larger than the values collected for the control group and the difference was determined to be statistically significant. Table 2 shows the *p*-values for the Mann-Whitney U-Test for all responses for each chart type.

Table 2 – Overall Mann-Whitney U-Test p-values

	***p*-values**
Bar	0.00008
Line	0.01046
Bubble	0.13622

Statistically significant when p < .05

Based on the data obtained from 256 participants, the results of the Mann-Whitney U-Test determined that participant responses to the test survey compared to the control survey were considered statistically significant for the bar chart (*p*-value 0.0008) and line chart (*p*-value 0.01046). This means that participants who received the test survey were susceptible to the deceptive techniques presented in the bar (truncated axis) and line (inverted axis), and participant responses for the test survey differed from those of the control group to such an extent that they were considered to be statistically significant. In contrast, the bubble chart (*p*-value 0.13622) indicates that the difference in values between the test survey and control survey was not enough to be considered statistically significant and participants were not susceptible to the bubble chart deceptive technique (area as quantity).

Age

Based on the age demographic, the results of the Mann-Whitney U-Test determined participant responses to the test survey compared to the control survey for the bar chart were considered statistically significant for demographics **18-29** years old (*p*-value 0.00104) and **30-49** years old (*p*-value 0.01552). This means that these demographics were susceptible to the bar chart (truncated axis) deception technique because participant responses to the test survey differed from the control group to such an extent that it was considered statistically significant.

Additionally, participants' responses to line chart test surveys compared to control surveys were considered statistically significant for demographics **18-39** years old (*p*-value 0.03846). This demographic was susceptible to the line chart (inverted axis) deception technique as the responses to the test survey differed from the control group to such an extent to be considered statistically significant. Table 3 shows the *p*-values for the Mann-Whitney U-Test based on age demographics.

Table 3 – Age Mann-Whitney U-Test p-values

	18-29 yrs	**30-49 yrs**	**50+ yrs**
Bar	0.00104	0.01552	0.29372
Line	.03846	0.36282	0.23404
Bubble	0.37346	0.39532	0.4965

Statistically significant when p < .05

The age demographic questionnaire received 256 responses and the results of the questionnaire are shown in figure 22. These results show that participants were predominantly 18-29 years old, but 46% of the results came from individuals that were 30 years old or older. The results of the age demographic are constant with the location of recruitment efforts with the vast majority of participants coming from Arizona State University.

Figure 22 – Age Demographic Results

Education

Based on the education demographic, the results of the Mann-Whitney U-Test determined participant responses to the test survey compared to the control survey for the bar chart were considered statistically significant for demographics with **some college but no degree** (*p*-value 0.00278) and **Master's degree** (*p*-value 0.0394). This means that these demographics were susceptible to the bar chart (truncated axis) deception technique because participant responses to the test survey differed from the control group to such an extent that it was considered statistically significant. Table 4 shows the *p*-values for the Mann-Whitney U-Test based on education demographics.

Table 4 – Education Mann-Whitney p-values

	High School	**Some College**	**Associates**	**Bachelors**	**Masters**	**Doctoral/ Professional**
Bar	0.25428	0.00278	0.88866	0.27572	0.0394	0.4902
Line	0.32708	0.14986	0.11184	0.93624	0.25848	0.33706
Bubble	0.93624	0.56868	0.24604	0.97606	0.71884	0.1141

Statistically significant when p < .05

The education demographic question received 256 responses and the results of the questionnaire are shown in figure 23. According to the results of the education demographic question, the largest percentage of participants only had some college with no degree obtained. Of the 256 participants, fifty-one percent of the participants had less than a Bachelor's degree, while 49% of participants had at least a Bachelor's degree or higher. The second largest group or participants at 19% stated they had a Master's degree. The smallest group of participants at 2% were those with a Professional degree (i.e. MD). For the purposes of running statistical testing, the Professional degree results were included with the Doctoral degree results.

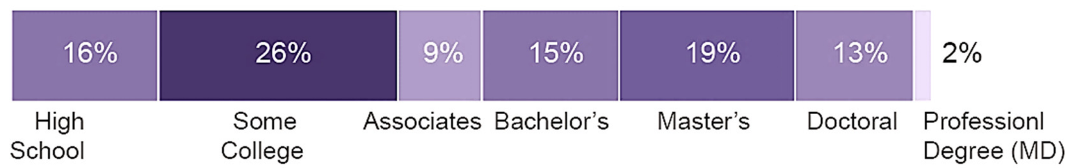


Figure 23 – Education Demographic Results

Course Taken

In order to understand how many participants might be informed or educated about potential ethical transgressions with data visualization or have a higher level of chart literacy, we asked participants if they had taken a data visualization course previously.

261 participants responded to the "course taken" question and 160 participants stated that they had not taken a course previously with 101 participants saying that they had taken a course. This would indicate that over half of the participants would not be aware of potential ethical transgressions with data visualization and would provide a clear example of how average consumers of data visualizations would respond given the study scenarios. Figure 24 shows the percentage breakdown for the course taken question.

Figure 24 – Data Visualization Course Results

Based on the course-taken demographic, the results of the Mann-Whitney U-Test determined participant responses to the test survey compared to the control survey for the bar chart were considered statistically significant for participants that said they **had taken a course** (*p*-value 0.02852) as well as those that said they **had not taken a course** (*p*-value 0.0008). This means that each of these demographics were susceptible to the bar chart (truncated axis) deception technique because participant responses to the test survey differed from the control group to such an extent that it was considered statistically significant. This basically means that regardless of whether a respondent had taken a course in data visualization or not, they were equally susceptible to the bar-chart deception in test survey.

Additionally, for the line chart, participant responses to the test survey compared to the control survey were considered statistically significant for participants that said they **had taken a course** (*p*-value 02642). This demographic was susceptible to the line chart (inverted axis) deception technique as the responses to the test survey differed from the control group to such an extent to be considered statistically significant. Table 5 shows the *p*-values for the Mann-Whitney U-Test based on the course taken question.

Table 5 – Course Taken Mann-Whitney U-Test p-values

	Data Visualization Course Taken?	
	Yes	**No**
Bar	0.02852	0.0008
Line	0.02642	0.18352
Bubble	1.33533	0.4902

Statistically significant when p < .05

Chart Familiarity

The chart familiarity assessment results determined that participants were mostly comfortable with bar, line and bubble charts. Only 2 percent of participants stated that they were "slightly uncomfortable" for both the bar and line charts with no participants stating they were "uncomfortable." Ninety percent of participants stated that they were comfortable with the bar chart and 82% stated they were comfortable with the line chart, while 8% and 16% stated they were slightly comfortable, respectively. Figure 25 shows the percentage breakdown for the chart familiarity assessment.

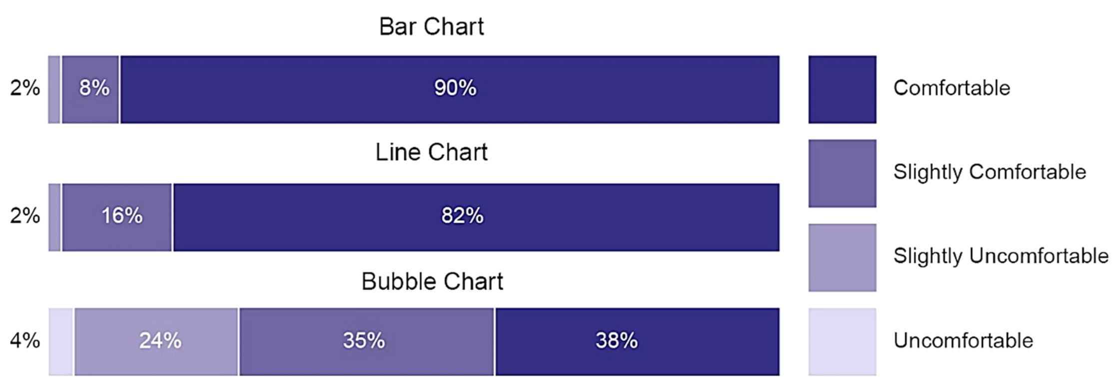


Figure 25 – Chart Familiarity Results

Majority of the participants stated they were both comfortable and slightly comfortable with bubble charts, at 38% and 35% respectively; however, twenty-four percent of participants stated they were only slightly comfortable and 4% were uncomfortable with the bubble chart.

In addition to the Mann-Whitney U-Test, the second round of the survey included a feature built into Qualtrics to measure participant times spent observing the paragraph of text and data visualization prior to answering the question. This feature was added after the initial recruitment, so there were only 100 participant responses collected.

Based on those 100 participants responses, participants spent a combined average of 30.15 seconds observing both the paragraph of text and data visualization with a combined minimum average of 2.37 seconds and a maximum average of 82.9 seconds. Figure 26 shows the participant observation times for all chart and treatment types.

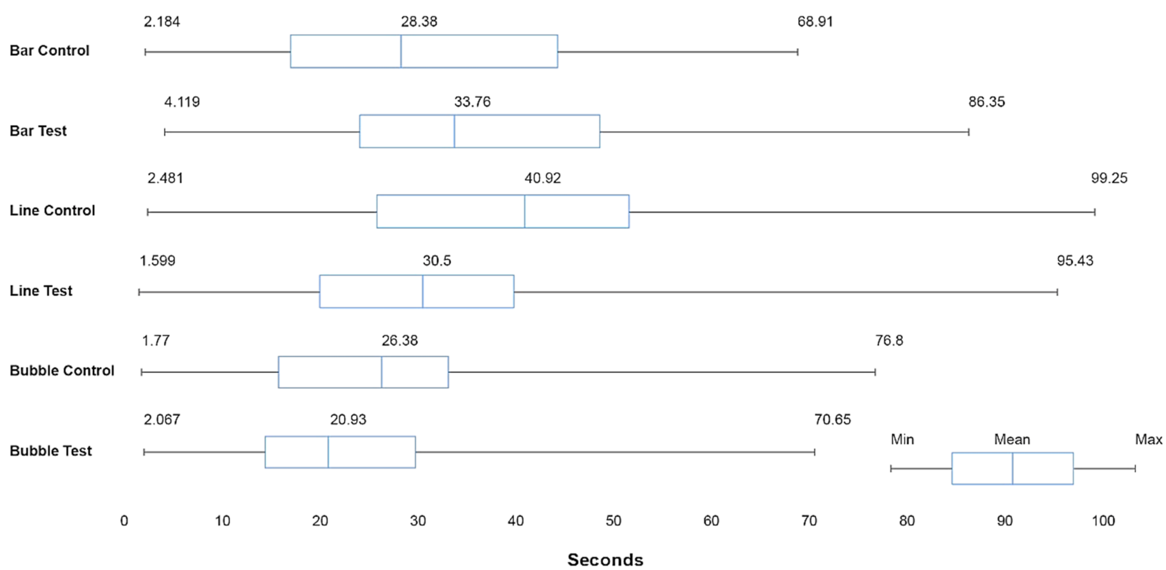


Figure 26 – Participant Survey Observation Times

Participants spent more time observing the line control treatment on average at 40.92 seconds compared to the other chart types, and participants spent less time with the bubble test treatment on average at 20.93 seconds. For the bar chart, participants observed the control treatment for an average of 28.38 seconds compared to 33.76 second on average for the test treatment, which is a difference of 5.38 seconds. Participants spent only 30.5 seconds on average for the line test treatment compared to 40.92 second for the control treatment, which is a difference of 10.42 seconds. Participants spent 26.38 seconds on average observing the bubble control treatment compared to 20.93 seconds, which is a difference of 5.45 seconds.

Qualitative Results

In order to understand where participants drew information from when answering the survey question, the second round of the survey also included qualitative questions after the chart questions. Ninety participants responded to the new survey questions out of 100 who completed the second round of the survey. Information about this part of the survey is detailed in the Methods section and the participant recruitment for this modified survey is described earlier in the results section.

The questions for the new section of the survey were designed to add a qualitative component to the data by allowing the participants to provide a direct response to a question about each chart, but participants were also encouraged to provide a freeform response for the participant to elaborate. I used both the direct response and answers provided in the free-response boxes to better understand where participants got their answer for the survey questions.

Based on the data from roughly 90 responses, participants utilized the chart to answer the survey question more for the line chart than the bar or bubble charts, with 41% percent of participants stating they used the line chart compared to 29% for the bar and 26% for the bubble. Thirty-seven percent of participants stated that they utilized both the paragraph of text and the chart when answering the survey question for both the bar and bubble chart with 34% saying the same for the line chart. Only 21% of participants used the paragraph of text when answering the line chart survey question compared to

32% for the bar chart and 34% for the bubble chart. Figure 27 shows the percentage breakdown for participant responses to what they utilized to answer the survey question.

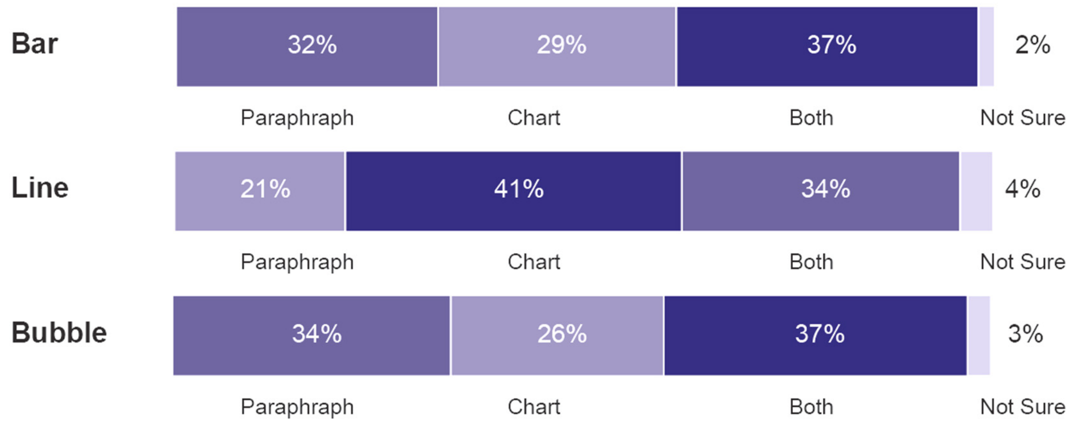


Figure 27 – Overall Qualitative Answers

Based on the data from roughly 45 responses who received the control survey, participants utilized the chart to answer the survey question more for the line chart than the bar or bubble charts with 46% of participants stating they used the line chart compared to 24% and 27%, respectively. Forty-two percent of participants stated that they utilized both the paragraph of text and the chart when answering the survey question for both the bar chart, while 33% said the same for the bubble chart and only 24% said so for the line chart. Only 24% of participants used the paragraph of text when answering the line chart survey question compared to 31% for the bar chart and 36% for the bubble chart. Figure 28 shows the percentage breakdown for control group participant responses to what they utilized to answer the survey question.

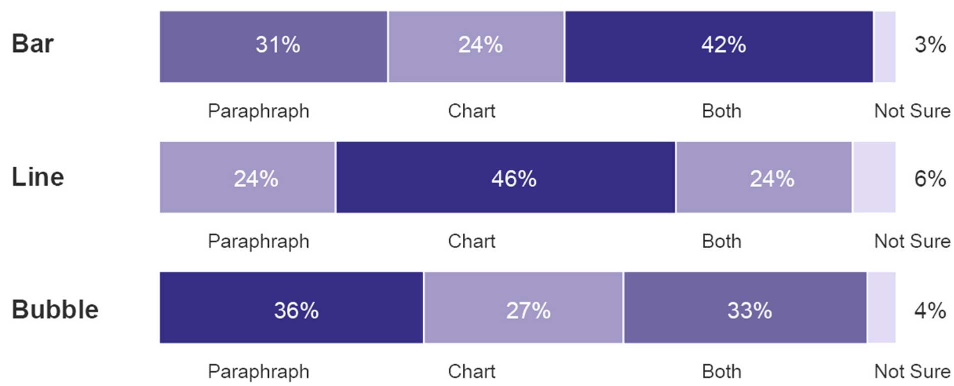


Figure 28 – Control Qualitative Answers

Based on the data from roughly 45 responses who received the test survey, participants utilized the chart to answer the survey question more for the line chart than the bar or bubble charts with 36% of participants stating they used the line chart compared to 33% and 24%, respectively. Forty-four percent of participants stated that they utilized both the paragraph of text and the chart when answering the survey question for the line chart, while 40% said the same for the bubble chart and only 31% said so for the bubble chart. Only 18% of participants used *only* the paragraph of text when answering the line chart survey question compared to 33% for both the bar and bubble charts. Figure 29 shows the percentage breakdown for test group participant responses to what they utilized to answer the survey question.

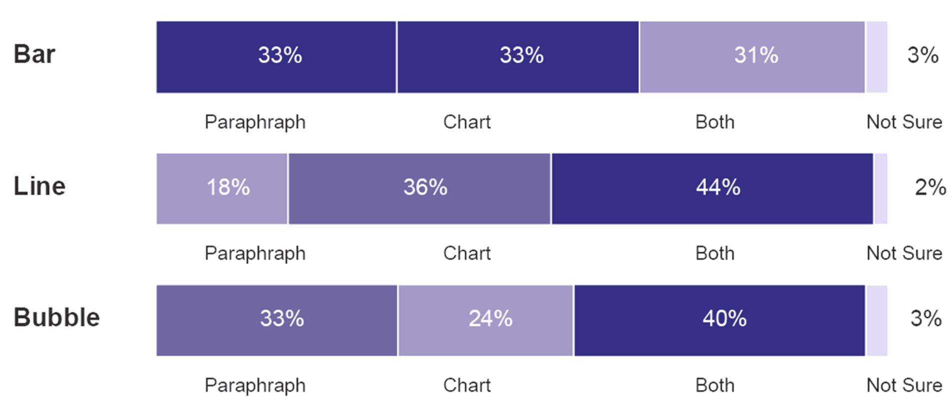


Figure 29 – Test Qualitative Answers

Only 58 participants chose to elaborate and provide a further response when asked whether they utilized the paragraph of text, data visualization, or both to answer the survey question.

Utilizing qualitative analysis and coding of participant responses, twenty-four percent of participants stated they utilized the paragraph only when answering the bar chart question and 29% stated they used only the chart. Thirty-four percent of the participants stated they utilized both the bar chart and paragraph of text to answer the survey question. Thirteen percent of participants stated they were not sure if they used either the paragraph of text or the data visualization to answer the survey question. While these numbers differ from the pointed responses, the results from the freeform responses show that participants utilized a combination of the chart or char plus text more than just the paragraph of text for the bar chart.

Responses to the line chart showed that 39% of participants used only the chart to answer the survey question, while 20% of participants used only the paragraph of text. Thirty-seven percent of participants stated they used both the line chart and paragraph of text to answer the survey question. Only four percent of participants stated they were not sure if they used either the paragraph of text or the data visualization to answer the survey question. The results of the free-form responses for the line chart mirror the responses from the pointed responses with only some slight difference. The results show the participants reliance on the chart or combination of chart plus text to answer the survey question. This highlights further that participants utilized the paragraph of text as a standalone component was less than the chart. Two participants responded to the freeform response that they observed the inverted axis in the test survey.

For the bubble chart, forty-one percent of participant stated they used only the paragraph of text to answer the survey question, while only 23% stated they used only the data visualization. Thirty-two percent of participants stated they used both the data visualization and paragraph of text to answer the survey. Only four percent of participants stated they were not sure if they used either the paragraph of text or the data visualization to answer the survey question. In contrast to the bar and line charts, the bubble chart freeform responses show that participants utilized either the paragraph of text or combination of text and chart to answer the survey question. This was reinforced by some participants elaborating on their responses and stating their comfort level with the bubble chart caused them to use the paragraph over the chart when answering the question.

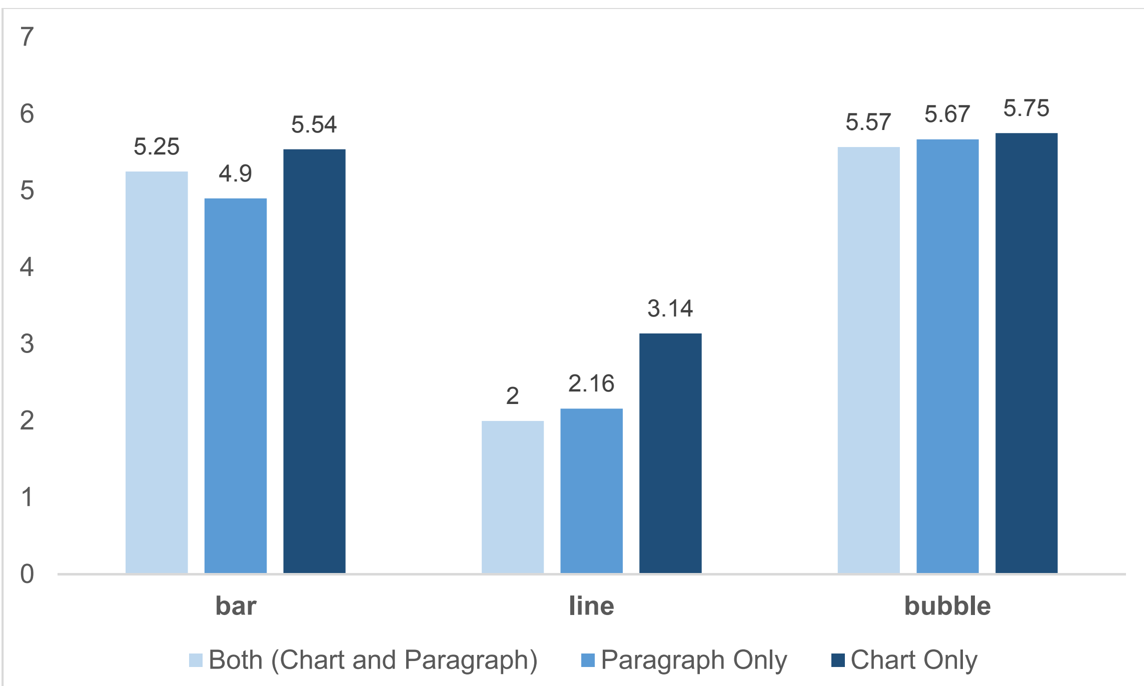


Figure 30 – Average Participant Response by Chart Type and Information Source

By analyzing the freeform responses and observing the average participant response to the “how much” question, the average participant responses were more exaggerated for all chart types when the participant only observed the chart. These responses were more exaggerated in comparison to those that read the paragraph or used both the paragraph and chart. The most severe difference between the average participant responses was for the line chart with those that only looked at the chart averaging 3.14 compared to an average of two for those that looked at both the paragraph of text and chart. Figure 30 shows the average participant responses by chart type and information source.

CHAPTER 5

DISCUSSION AND CONCLUSION

The goal of this study was to investigate in what ways accompanying data visualizations with explanatory text changes a user's interpretations of the visualization. Additionally, the study was designed to test and compare results to a previously conducted study by Pandey et al. (2015). The Pandey et al. (2015) study determined that participants were deceived by data visualizations when they contained deceptive techniques. The Pandey et al. (2015) study found that regardless of chart type or deceptive technique, participants were deceived and misinterpreted information in data visualizations that incorporated deceptive techniques. Unlike this study, they were unable to find any direct correlations to education level and participants potential to be deceived.

This study mirrored much of the Pandey et al. (2015) but differed by including a paragraph of explanatory text with the data visualization. The results of the study confirmed that even with accompanying text that reiterates the actual differences and trends in the data, people are susceptible to deceptive, unethical data visualizations and their perception of the information presented in data visualizations can be manipulated. The overall results confirmed that participants were susceptible to the bar (truncated axis) and line chart (inverted axis) deceptive techniques with the difference between participants responses to test (deceptive) charts compared to the control group were considered statistically significant. However, the study did not confirm the original results for the bubble chart (area as quantity) as the difference between the participant responses to test (deceptive) charts compared to the control group were not considered statistically significant.

Furthermore, the study found that certain demographics were susceptible to deceptive techniques, while other demographic groups were not. While the results show that only certain demographics and age groups fell more susceptible to the deceptive techniques compared to others, the overall results showed that the deceptive techniques still caused participants to perceive the information differently. Interestingly, the test results showed that individuals that stated they had taken a data visualization course before were more susceptible to deceptive techniques of the bar chart (truncated axis) and line chart (inverted axis), while those who stated they had not taken a course in data visualization were only susceptible to line chart (inverted axis) deceptive practices. This further highlights that regardless of the additional information (descriptive text) or level of perceived familiarity, participants were still susceptible to the deceptive techniques in the data visualizations and as a result may become misinformed.

The responses to the chart familiarity assessment showed a large majority of participants were comfortable with the bar and line chart types, and people found the bubble chart to be the least comfortable of the chart types. The qualitative analysis determined that participants were not comfortable with the bubble chart and resulted to using the information in the paragraph of text over the chart to answer the question. The change in results from the Pandey et al. (2015) to this study for bubble chart would indicate that the paragraph of text did alleviate the potential for the bubble chart to deceive the participants; however, participants resulted to using the paragraph of text over the chart to avoid trying to interpret the bubble chart.

Upon analyzing the freeform responses and observing the average participant responses, the study confirms that participants exaggerated their responses more when they only looked at the chart only compared to those who looked at the paragraph only or both. While this confirms that the inclusion of explanatory text with a data visualization resulted in less exaggerated responses, the overall results of the study still confirmed that regardless of including explanatory text participant responses to test treatments were exaggerated compared to control treatments.

The study conducted by Linda, Martin, Cantor, and Rubenstein looked at how decisions can be influenced by using different data visualization types over others and showed that better decisions were made when non-traditional data visualization types were used in making recommendations for patients (1999). While this study did not attempt to answer the same question, results also showed participants tendency to select certain visualizations over others. This was shown in the chart familiarity assessment as well as the free-form responses. During these two assessments, participants showed they favored the bar and line charts over the bubble chart with some even choosing not to use the bubble chart to answer the survey question. The study also confirmed that the more favorable chart types (bar and line) also resulted in more exaggerated responses from the participants compared to the unfavorable type (bubble).

Robin Kinross discusses how data visualizations or information design employs rhetoric and that ‘pure’ information only exists for the designer and not the reader (1985). While this implies that information design comes with it some design liberty and rhetoric, the potential for misleading the reader is enhanced when data visualizations are manipulated by message exaggeration/understatement, and message reversal techniques.

The information obtained from this study provides valuable insight to understand how people use data visualizations when paired with explanatory text. This research helps inform future practitioners and educators about the potential for reader misunderstanding of information when deceptive techniques are used in data visualizations. The intent of the data visualization is to inform the reader, but employing deceptive practices in the creation of data visualizations has the potential of causing the reader to become misinformed.

Some limitations of the study included the original determination for participant distribution. Originally, the target recruitment of 50 to 100 participants required the three visualizations types to be distributed to all participants who received the control and test treatments. This was done as a way of maximizing the number of responses for each chart and treatment type. The Pandey et al. (2015) study only distributed one chart type per participant. By providing the participants with all three chart types in the same order, the participants appeared to become familiar with the survey format and for that reason appeared to decrease the amount of time spent on each question as the survey progressed.

Further limitations include not randomly ordering and assigning chart types. The study could have avoided the previously mentioned limitation regarding the distribution of the chart types by distributing participants any of the three chart types in random order. This would have avoided the issues of participants spending less time at the end of the survey on the bubble chart.

Further research can be done to look at how people observe data visualization by tracking individual eye movements and observing what key features of the data visualization are being observed. Other chart types were not tested in this study to limit variables; however, it would be important to test how other types of data visualization are used. Additionally, further research could be done to examine the exact relationship of participant comfort with chart and their potential to be deceived.

In conclusion, deceptive practice with data visualizations – whether intentional or not – have the power to leave the reader misinformed and ultimately misunderstand the information being presented to them. With the increase use of data visualizations as a means of communicating large amounts of information, research such as this further advances the area of study to better understand how these practices impact readers and to a larger extent society as a whole. It is important that we continue to research this topic and educating those that produce and consume data visualizations about the potential for possible deception and misinformation, which seems appropriate considering constantly changing and evolving social and political climates.

REFERENCES

Bergstrom, C., & West, J. (2016). Visualizing, Misleading axes on graphs. *Calling Bullshit.* Retrieved from http://callingbullshit.org/tools/tools_misleading_axes.html

Bertin, J. (1983). *Semiology of graphics*. Madison, Wis.: University of Wisconsin Press.

Bowen, J. P., Keene, S., Ng, K., & MyiLibrary. (2013). *Electronic visualization in arts and culture*. London: Springer.

Cairo, A. (2014). *Ethical Infographics: In data visualization, journalism meets engineering.* Retrieved from https://www.scribd.com/document/230474170/Ethical-Info-Graphics-In-data-visualization-journalism-meets-engineering

Chen, C., Härdle, W., & Unwin, A. (2008). Handbook of Data Visualization (Springer Handbooks Comp.Statistics). Berlin, Heidelberg: Springer Berlin Heidelberg.

Cleveland, W., & Mcgill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association, 79*(387), 531-554.

Dick, M. (2015). Just fancy that: An analysis of infographic propaganda in The Daily Express, 1956–1959. *Journalism Studies*, 16(2), 152-174. doi:10.1080/1461670X.2013.87241

Dragga, S., & Voss, D. (2001). Cruel pies: The inhumanity of technical illustrations. *Technical Communication, 48*(3), 265-274.

Elting, L. S., Martin, C. G., Cantor, S. B., & Rubenstein, E. (1999). Influence of data display formats on physician investigators' decisions to stop clinical trials: Prospective trial with repeated measures. *British Medical Journal, 318*(7197), 1527-1531.

Ethics of Data. (2015, May 12). [Blog] *Berkman Klein Center for Internet & Society at Harvard University.* Retrieved from https://cyber.harvard.edu/getinvolved/studygroups/datastorytelling_ethics

Friendly, M. (2008). The Golden Age of Statistical Graphics. *Statistical Science, 23*(4), 502-535. Retrieved from http://www.jstor.org/stable/20697655

Huff, D. (1954). *How to lie with statistics*. London: WW Norton & Company.

Jones, G. E. (2011). *How to lie with charts*. LaPuerta Books and Media

Keim, D., Mansmann, F., Schneidewind, J., & Ziegler, H. (2006). Challenges in Visual Data Analysis. *Information Visualization, 2006. IV 2006. Tenth International Conference on,* 9-16.

Kerren, Andreas, Stasko, John T, Fekete, Jean-Daniel, North, Chris, Purchase, Helen C, Andrienko, Natalia, . . . Ward, Matthew. (n.d.). *Information visualization : Human-centered Issues and perspectives /*. Berlin ;: Springer.

Kirk, A. (2012). *Data Visualization : a successful design process*. Birmingham: Packt Publishing.

Kinross, R. (1985). The Rhetoric of Neutrality. *Design Issues, 2*(2), 18-30. doi:10.2307/1511415

Manovich, L. (2011). What is visualisation?. *Visual Studies*, 26(1), 36-49. doi:10.1080/1472586X.2011.548488

Monmonier, M. S. (1991). *How to lie with maps*. Chicago: University of Chicago Press.

Neuhaus, F., & Webmoor, T. (2012). Agile ethics for massified research and visualization.*Information, Communication & Society, 15*(1), 43-65. doi:10.1080/1369118X.2011.616519

Pandey, A. V., Rall, K., Satterthwaite, M. L., Nov, O., & Bertini, E. (2015, April). How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1469-1478). ACM.

Pasternack, S., & Utt, S. (1989). *Reader Use and Understanding of Newspaper Informational Graphics*. S.l.]: Distributed by ERIC Clearinghouse.

Representation. (n.d.). Retrieved November 16, 2016, from http://www.merriam-webster.com/dictionary/representation

Rogowitz, B., Treinish, & Bryson. (1996). How Not to Lie with Visualization. *Computers in Physics, 10*(3), 268-273.

Schriver, K. (1997). *Dynamics in document design : Creating text for readers*. New York: Wiley Computer Publishing.

Skau, D. (2012, February 07). A Code of Ethics for Data Visualization Professionals. Retrieved from http://www.scribblelive.com/blog/2012/02/07/a-code-of-ethics-for-data-visualization-professionals/

Society of Professional Journalists (2014, September 6). *Society of Professional Journalists Cost of Ethics.* [Flyer] Retrieved from https://www.spj.org/pdf/spj-code-of-ethics.pdf

Society of Technical Communication (1998) *As technical communicators, we observe the following ethical principles in our professional activities.* Retrieved from https://www.stc.org/about-stc/ethical-principles/

Sue, V and Griffin, M. (2016). *Data Visualization and Presentation with Microsoft Office*. Thousand Oaks, CA: Sage.

Sula, C. (Presenter). (2012, December 12). Digital Library Seminar Series: 2012-12-12. *The Ethics of Visualization.* Video retrieved from http://chrisalensula.org/the-ethics-of-visualization/

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT.: Graphics Press.

Tukey, J. (1977). *Exploratory data analysis* (Addison-Wesley series in behavioral science: quantitative method). Reading, Ma.: Addison-Wesley.

Wainer, H. (1984). How to Display Data Badly. *The American Statistician, 38*(2), 137-147. doi:10.2307/2683253

Yau, N. (2017). How to Spot Visualization Lies. *Guides: Tips and suggestions for working and designing with data.* Retrieved from https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Yau, N. (2013). *Data points: Visualization that means something* (1. Aufl.;1; ed.). Indianapolis: John Wiley & Sons, Inc

APPENDIX A

CONTROL AND TEST SURVEY

Demographic Questionnaire

What is your age?

- ○ 18-29 years old (1)
- ○ 30-49 years old (2)
- ○ 50-64 years old (3)
- ○ 65 years and over (4)

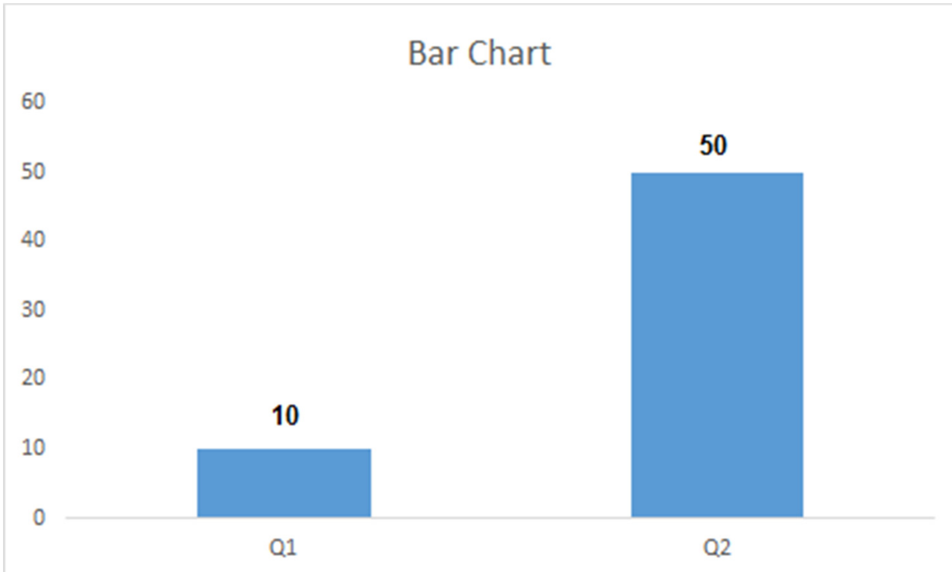
What is the highest level of school you have completed or the highest degree you have received?

- ○ Less than high school degree (1)
- ○ High school graduate (high school diploma or equivalent including GED) (2)
- ○ Some college but no degree (3)
- ○ Associate degree in college (2-year) (4)
- ○ Bachelor's degree in college (4-year) (5)
- ○ Master's degree (6)
- ○ Doctoral degree (7)
- ○ Professional degree (JD, MD) (8)

Have you taken any courses in creating charts and graphs with data or visualizing data?

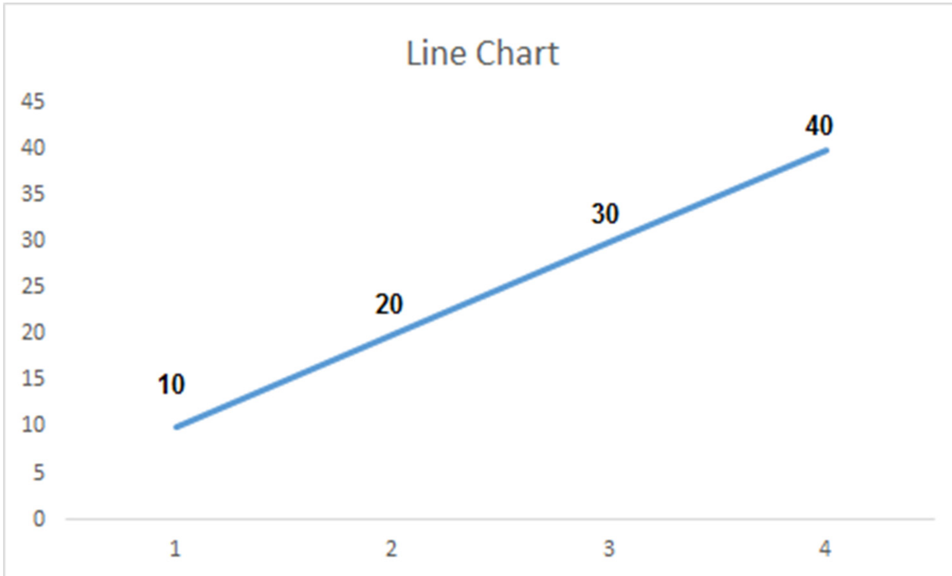
- ○ Yes (1)
- ○ No (2)

Chart Familiarity



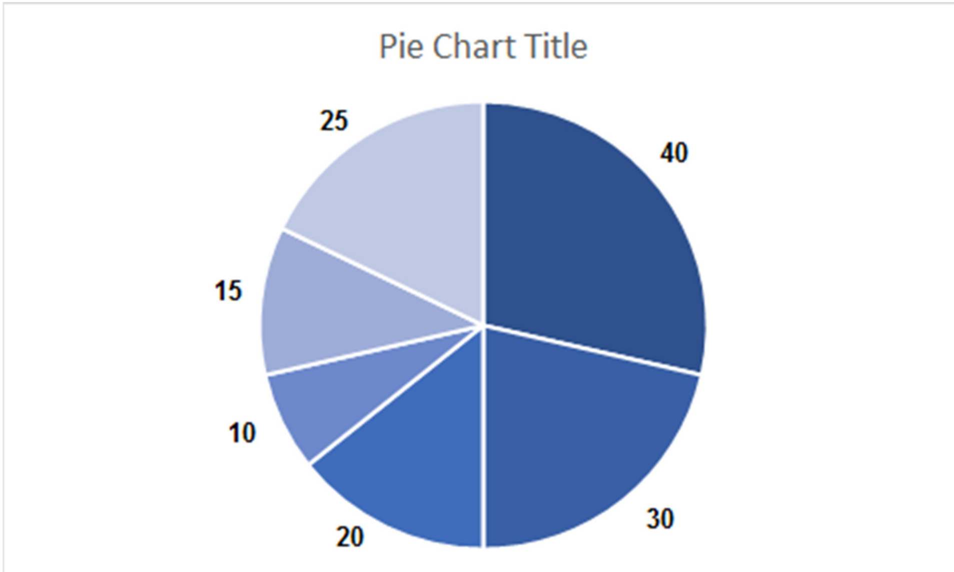
How comfortable are you understanding data presented to you through bar charts?

	Uncomfortable (1)	Slightly Uncomfortable (2)	Slightly Comfortable (3)	Comfortable (4)
Select One (1)	○	○	○	○



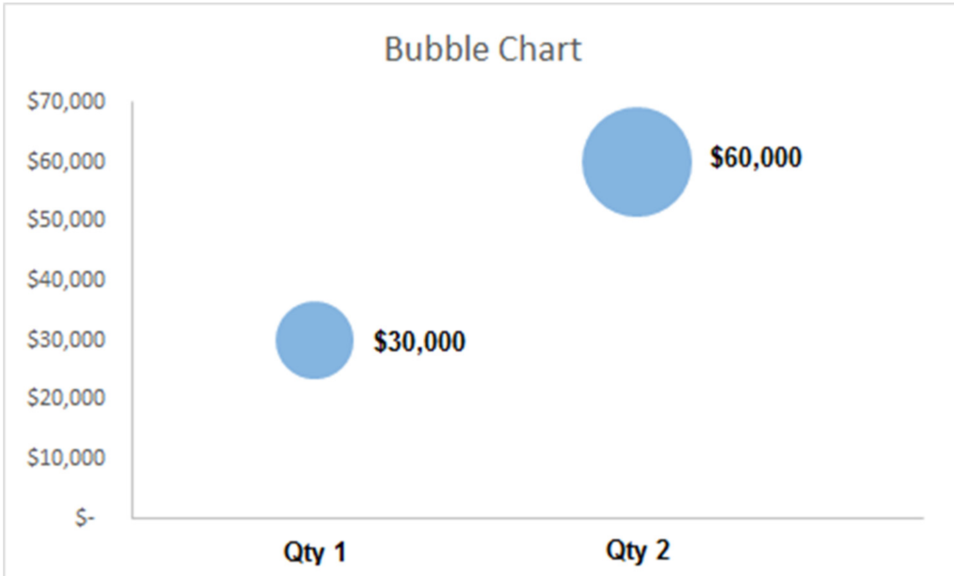
How comfortable are you understanding data presented to you through line charts?

	Uncomfortable (1)	Slightly Uncomfortable (2)	Slightly Comfortable (3)	Comfortable (4)
Select One (1)	○	○	○	○



How comfortable are you understanding data presented to you through pie charts?

	Uncomfortable (1)	Slightly Uncomfortable (2)	Slightly Comfortable (3)	Comfortable (4)
Select One (1)	○	○	○	○



How comfortable are you understanding data presented to you through bubble charts?

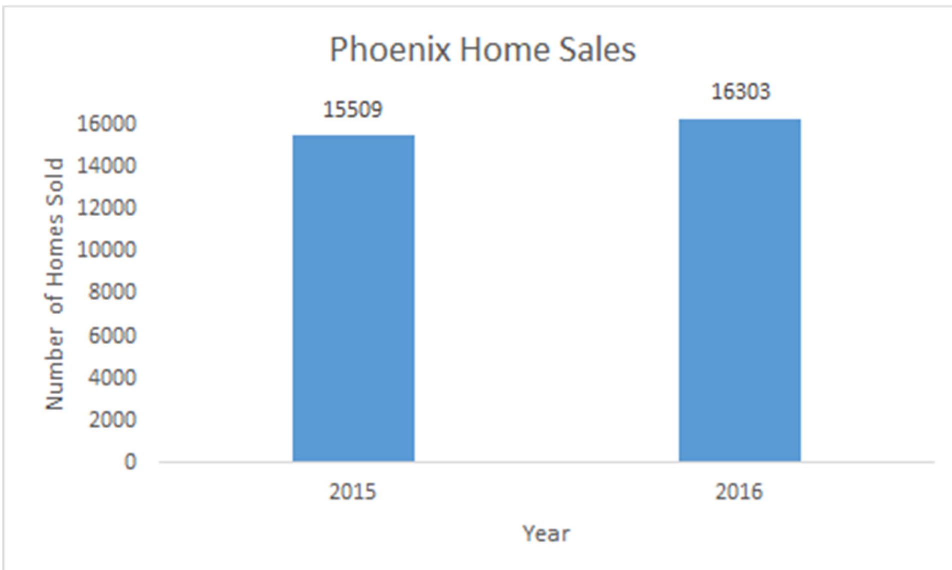
	Uncomfortable (1)	Slightly Uncomfortable (2)	Slightly Comfortable (3)	Comfortable (4)
Select One (1)	○	○	○	○

End of Block: Default Question Block

Control Treatment

Carefully look over and read the information below. The next screen will ask a question related to this information and you will not be able to go back.

ABC Homes recently compared home prices for the Phoenix metropolitan area. Based on statistics, we saw 16,303 single-family homes sold in Phoenix, AZ in 2016 compared to 15,509 in 2015, an increase of roughly 5%. In 2016, the median sold price for a single-family detached home in Phoenix, AZ was $230,000 compared to $210,900 in 2015.



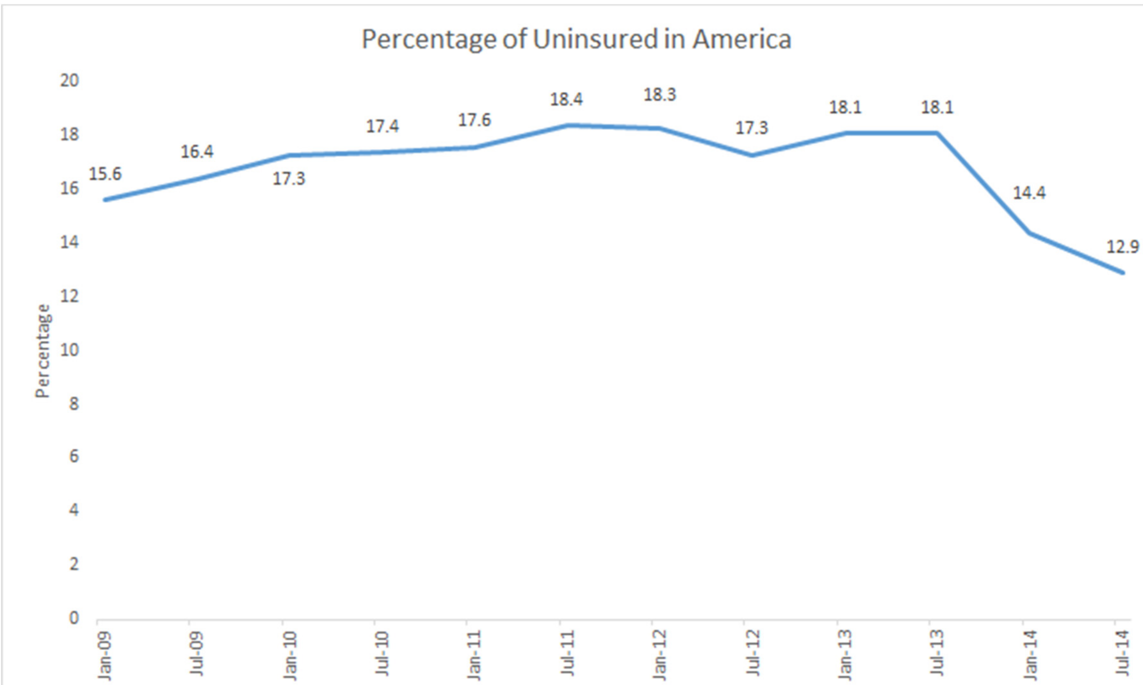
Page Break

How much did home sales increase or decrease in 2016 compared to 2015?

	Drastic Decrease (1)	Moderate Decrease (2)	Slight Decrease (3)	No Change (4)	Slight Increase (5)	Moderate Increase (6)	Drastic Increase (7)
Select One (3)	○	○	○	○	○	○	○

Carefully look over and read the information below. The next screen will ask a question related to this information and you will not be able to go back.

The percentage of uninsured Americans has seen some fluctuation over the course of the past four years. From January 2009 to July 2014, the highest percentage of uninsured reached 18.4 percent in 2011. The lowest percentage of uninsured comes in at 12.9 percent in 2014. That constitutes a 5.3 percent drop in the percentage of uninsured Americans.



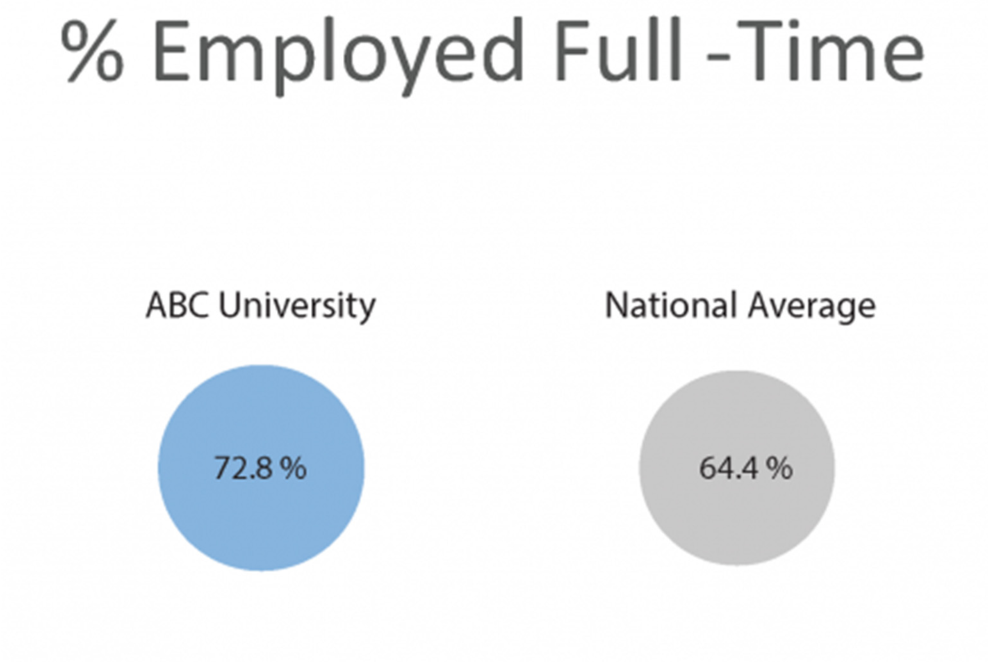
Page Break

How much do you think the percentage of uninsured increased or decreased from April 2013 to July 2014?

	Drastic Decrease (1)	Moderate Decrease (2)	Slight Decrease (3)	No Change (4)	Slight Increase (5)	Moderate Increase (6)	Drastic Increase (7)
Select One (2)	○	○	○	○	○	○	○

Carefully look over and read the information below. The next screen will ask a question related to this information and you will not be able to go back.

ABC University strives for excellence and innovation by offering in-demand degree programs and enriched learning opportunities to our students setting them up for success in their careers. Compared to the national average of 64.4 percent, 72.8 percent of ABC University students earn full-time employment upon graduation. That makes ABC University students 8.4% more likely to earn a full-time position after graduating.



Page Break

How much better or worse is ABC University compared to the National Average?

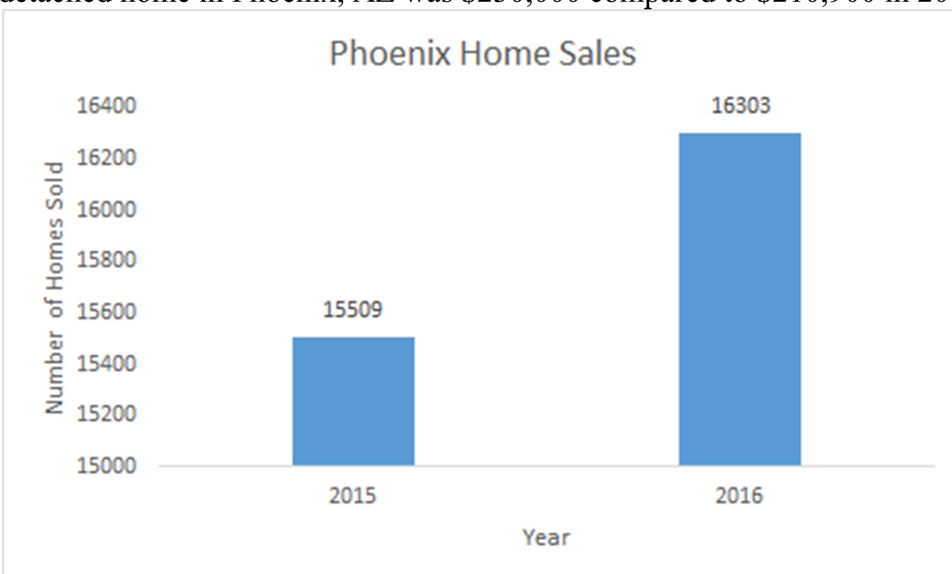
	Drastically Worse (1)	Moderately Worse (2)	Slightly Worse (3)	No Change (4)	Slightly Better (5)	Moderately Better (6)	Drastically Better (7)
Select One (1)	○	○	○	○	○	○	○

End of Block: Control Survey

Deceptive Treatment

Carefully look over and read the information below. The next screen will ask a question related to this information and you will not be able to go back.

ABC Homes recently compared home prices for the Phoenix metropolitan area. Based on statistics, we saw 16,303 single-family homes sold in Phoenix, AZ in 2016 compared to 15,509 in 2015, an increase of roughly 5%. In 2016, the median sold price for a single-family detached home in Phoenix, AZ was $230,000 compared to $210,900 in 2015.



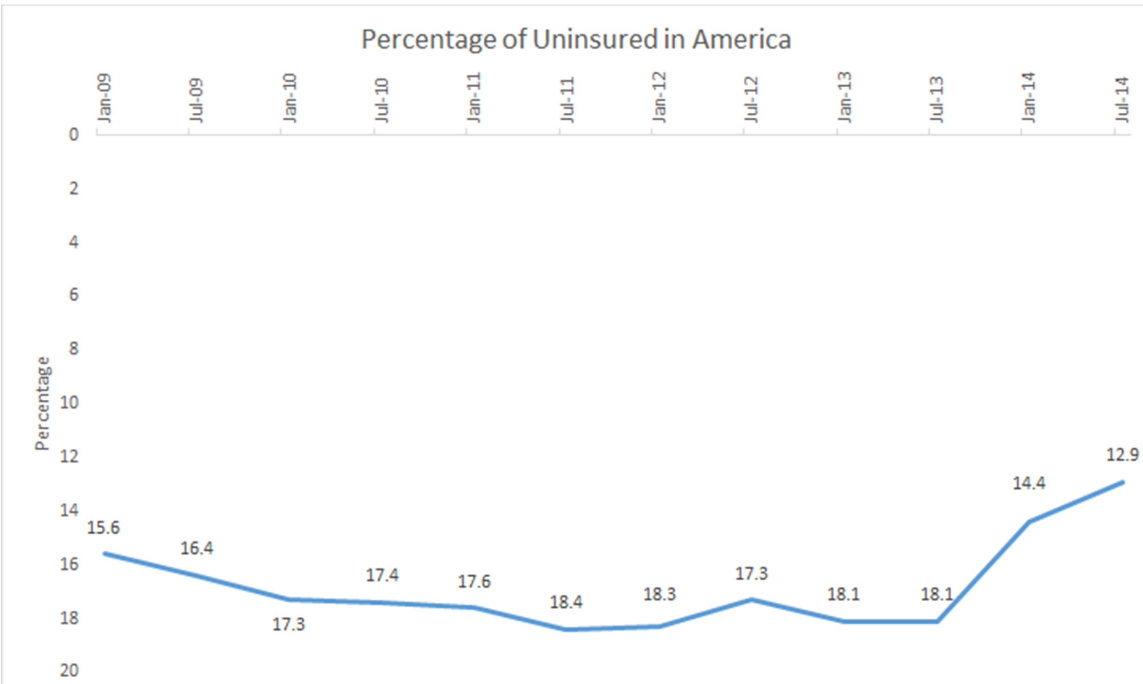
Page Break

How much did home sales increase or decrease in 2016 compared to 2015?

	Drastic Decrease (1)	Moderate Decrease (2)	Slight Decrease (3)	No Change (4)	Slight Increase (5)	Moderate Increase (6)	Drastic Increase (7)
Select One (1)	○	○	○	○	○	○	○

Carefully look over and read the information below. The next screen will ask a question related to this information and you will not be able to go back.

The percentage of uninsured Americans has seen some fluctuation over the course of the past four years. From January 2009 to July 2014, the highest percentage of uninsured reached 18.4 percent in 2011. The lowest percentage of uninsured comes in at 12.9 percent in 2014. That constitutes a 5.3 percent drop in the percentage of uninsured Americans.

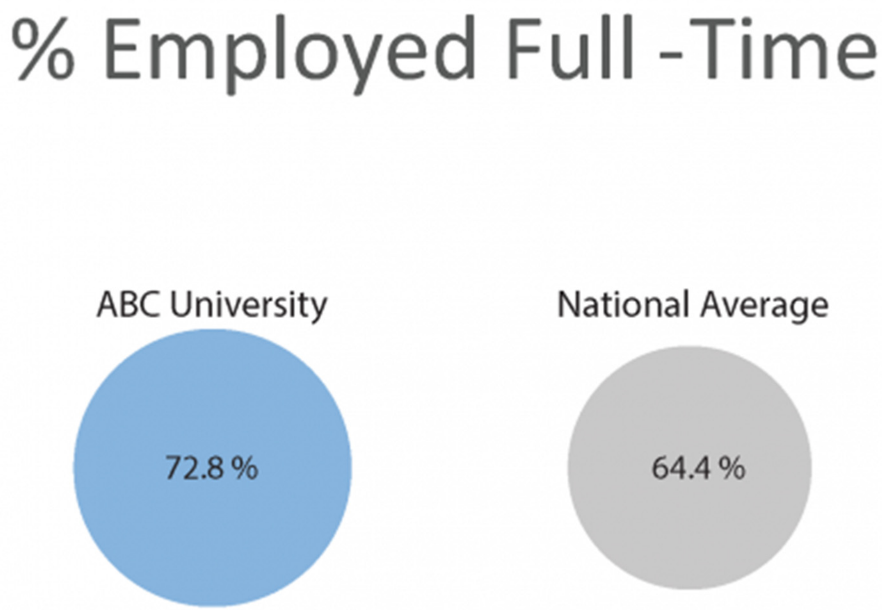


How much do you think the percentage of uninsured increased or decreased from April 2013 to July 2014?

	Drastic Decrease (1)	Moderate Decrease (2)	Slight Decrease (3)	No Change (4)	Slight Increase (5)	Moderate Increase (6)	Drastic Increase (7)
Select One (1)	○	○	○	○	○	○	○

Carefully look over and read the information below. The next screen will ask a question related to this information and you will not be able to go back.

ABC University strives for excellence and innovation by offering in-demand degree programs and enriched learning opportunities to our students setting them up for success in their careers. Compared to the national average of 64.4 percent, 72.8 percent of ABC University students earn full-time employment upon graduation. That makes ABC University students 8.4% more likely to earn a full-time position after graduating.



How much better or worse is ABC University compared to the National Average?

	Drastically Worse (1)	Moderately Worse (2)	Slightly Worse (3)	No Change (4)	Slightly Better (5)	Moderately Better (6)	Drastically Better (7)
Select One (1)	o	o	o	o	o	o	o

End of Block: Treatment Survey

Qualitative Questions -

Where did you find your answer to the question in the survey about housing prices?

- In the paragraph of text (1)
- In the bar chart (2)
- Both in the paragraph of text and in the bar chart (3)
- I'm not sure (4)

Please elaborate (no character limit)

____

____

____

____

____

Where did you find your answer to the question in the survey about uninsured Americans?

- In the paragraph of text (1)
- In the line chart (2)
- Both in the paragraph of text and in the line chart (3)
- I'm not sure (4)

Please elaborate (no character limit)

____

____

____

____

____

Where did you find your answer to the question in the survey about students earning full-time employment?

- ○ In the paragraph of text (1)
- ○ In the bubble chart (2)
- ○ Both in the paragraph of text and in the bubble chart (3)
- ○ I'm not sure (4)

Please elaborate (no character limit)

____

____

____

____

____

APPENDIX B

CONSENT FORM

Consent_Form

Welcome to the research study!

I am a graduate student under the direction of Professor Claire Lauer in the College of Integrative Sciences and Arts at Arizona State University. I am conducting a research study about data visualizations.

Your participation in this study is voluntary. If you choose not to participate or to withdraw from the study at any time, there will be no penalty, (for example, it will not affect your grade). Participation in this survey will make you eligible and enter you into a drawing for a $25 Amazon gift card. You must be 18 or older to participate in the study. This study looks to recruit roughly 50-100 participants, and participants will have roughly 1:100 odds of winning the gift card. Participant recruitment will last approximately 2 months. Upon completion of the recruitment period, participants and the winner will be notified via email that a winner has been selected.

Participants will complete a five to seven-minute survey in which they are asked to read a paragraph of text and data visualization and then answer a question related to both. Prior to the survey, you will be asked a series of basic demographic questions and questions about your familiarity with common graph types. You have the right to not to answer any questions and to stop participation at any time.

Although there is no benefit to you, your participation will aid and benefit developers of information by providing further insight into how we learn from data visualizations. There are no foreseeable risks or discomforts to your participation.

Your responses will be kept confidential and the results of this study may be used in reports, presentations, or publications but your name will not be used. Additionally, the results of the study will be shared in the aggregate form only.

If you have any questions concerning the research study, please contact the research team at ShaunObrien@asu.edu or 480-440-4004. Additionally, you may contact Dr. Claire Lauer at Claire.Lauer@asu.edu or 480-828-3881. If you have any questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk, you can contact the Chair of the Human Subjects Institutional Review Board, through the ASU Office of Research Integrity and Assurance, at (480) 965-6788. Please let me know if you wish to be part of the study.

By clicking the button below, you acknowledge that your participation in the study is voluntary, you are 18 years of age, and that you are aware that you may choose to terminate your participation in the study at any time and for any reason.

Please note that this survey will be best displayed on a laptop or desktop computer. Some features may be less compatible for use on a mobile device.

- I consent, begin the study (1)