VISUAL-SPATIAL BIASES IN ENSEMBLE COGNITION

by

Lace M. K. Padilla

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychology

The University of Utah

December 2018

Copyright © Lace M. K. Padilla 2018

All Rights Reserved

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of Lace M. K. Padilla
has been approved by the following supervisory committee members:

<u>Sarah Creem-Regehr</u>	, Chair	<u>May 9th, 2018</u> Date Approved
<u>Jeanine Stefanucci</u>	, Member	<u>May 9th, 2018</u> Date Approved
<u>Trafton Drew</u>	, Member	<u>May 9th, 2018</u> Date Approved
<u>Miriah Meyer</u>	, Member	<u>May 17th, 2018</u> Date Approved
<u>William Thompson</u>	, Member	<u>May 10th, 2018</u> Date Approved

and by Lisa Aspinwall, Chair/Dean of
the Department/College/School of Psychology

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Given the widespread use of visualizations and their impact on health and safety, it is important to ensure that viewers interpret visualizations as accurately as possible. Ensemble visualizations are an increasingly popular method for visualizing data, as emerging research demonstrates that ensembles can effectively and intuitively communicate traditionally difficult statistical concepts. While a few studies have identified drawbacks to ensemble visualizations, no studies have identified the sources of reasoning biases that could occur with ensemble visualizations. Our previous work with hurricane forecast simulation ensemble visualizations identified a misunderstanding that could have resulted from the visual features of the display. The current study tested the hypothesis that visual-spatial biases, which are biases that are a direct result of the visualization technique, provide a cognitive mechanism to explain this misunderstanding. In three experiments, we tested the role of the visual elements of ensemble visualizations as well as knowledge about the visualization with novice participants ($n = 303$). The results suggest that previously documented reasoning errors with ensemble displays can be influenced both by changes to the visualization technique and by top-down knowledge-driven processing.

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF TABLES	vi
LIST OF FIGURES.....	vii
Chapters	
1. SIGNIFICANCE	1
1.1 Introduction.....	1
1.2 Ensemble Visualizations.....	5
1.3 Visual-Spatial Biases	7
1.4 Overview of Experiments	11
2. EXPERIMENT 1	13
2.1 Methods.....	14
2.1.1 Participants	14
2.1.2 Stimuli	15
2.1.3 Design.....	17
2.1.4 Procedure.....	17
2.2 Results.....	18
2.3 Discussion	22
3. EXPERIMENT 2	27
3.1 Methods.....	27
3.1.1 Participants	27
3.1.2 Stimuli and Design.....	28
3.1.3 Procedure.....	28
3.1.4 Coding	29
3.2 Results.....	30
3.3 Discussion	31
4. EXPERIMENT 3	33
4.1 Methods.....	34

4.1.1 Participants	34
4.1.2 Stimuli and Design	34
4.1.3 Procedure	35
4.2 Results	35
4.3 Discussion	37
5. GENERAL DISCUSSION	41
5.1 Conclusions	46
REFERENCES	49

LIST OF TABLES

Tables

2.1. List of fixed effects with coefficients, standard errors, <i>t-values</i> , <i>p-values</i> , and 95% confidence intervals from the statistical model predicting damage ratings. Collocation was coded such that the effects indicate a change from off-line to on-line.....	26
2.2 Proportion of correct responses for each visualization condition. *** Indicates <i>p-values</i> < .000, ** <i>p-values</i> < .005, and * <i>p-values</i> < .05 with the 9-track display as the referent.....	26
3.1. Coded strategies	32
4.1 List of fixed effects with coefficients, standard errors, <i>t-values</i> , <i>p-values</i> , and 95% confidence intervals from the statistical model predicting damage ratings.....	40
4.2 Proportion of correct responses for each visualization condition. *** Indicates <i>p-values</i> = .000, ** <i>p-values</i> < .005, and * <i>p-values</i> < .05.....	40

LIST OF FIGURES

Figures

1.1 Example of hurricane forecast display stimuli used in Padilla et al. (2017), where the red dots indicate the location of off shore oil platforms. In Figure 1.1 A, location A is collocated. In Figure 1.1 B, location B is collocated.	13
2.1. Example stimuli, showing the 9-track (A), 17-track (B), 33-track (C), and 65-track (D) displays. The black dot indicates the location of the offshore oil rig.	24
2.2. Example of the 33-track display, where one line was reflected over the mean line of the distribution of simulated ensemble members to make a collocation condition (A) and a noncollocation condition (B). C and D represent the mirror images of A and B, where the underlying map remained constant.	24
2.3. Damage change scores for the 9-, 17-, 33-, and 65-track displays. Error bars represent 95% confidence intervals around the mean.	25
4.1. Damage change scores for the 9-track display conditions with no instructions, general instructions, and task-specific instructions. Error bars represent 95% confidence intervals around the mean.	39
5.1. Icon arrays used in Stone et al. (1997) to illustrate the risk of standard or improved tires (reprinted with permission from Padilla et al., 2018).	48

CHAPTER 1

SIGNIFICANCE

We use data visualizations to make large-scale policy decisions such as where to allocate resources before a natural disaster and more personal life and death decision such as whether to evacuate before a forecasted hurricane. As visualizations have large-scale implications for the health and safety of our global community, we must have a clear understanding of how visualizations influence our decisions. The current work is a case study in evaluating how visualizations may produce errors in reasoning and provides practical recommendations for how to help viewers make their best possible decisions with visualizations of data.

1.1 Introduction

We use data visualizations—visual representations of data—to make common decisions such as selecting the fastest driving route as well as relatively infrequent decisions such as whether to evacuate before a forecasted hurricane. Given their widespread use and social impact, it is essential to understand how visualizations influence our judgments and actions. To make accurate predictions about how visualizations influence decisions, it is necessary to understand the underlying cognitive processes of the viewer. Some types of visualization decision processes are not obvious

and empirical studies have shown that viewers may interpret (even simple) data visualizations in unintended ways (Belia, Fidler, Williams, & Cumming, 2005; Correll & Gleicher, 2014; Newman & Scholl, 2012; Sanyal, Zhang, Bhattacharya, Amburn, & Moorhead, 2009; Scown, Bartlett, & McCarley, 2014). For example, Belia et al. (2015) found that experts in psychology, behavioral neuroscience, and medicine misunderstood how visually presented error bars, depicting 95% confidence intervals and standard errors, relate to statistical significance. These results are concerning because the majority of scientific publications use error bars to illustrate statistical significance. If expert academic authors have difficulty interpreting error bars, the untrained public has little chance of accurately interpreting these visualizations. Numerous other studies also find that visualizations can produce interpretations of the data that the visualization designers did not intend (for review see Padilla, Creem-Regehr, Hegarty, & Stefanucci, 2018). However, few studies have identified the cognitive processes that are responsible for misinterpretations of data visualizations. In the current study, we utilized one type of previously documented visualization misinterpretation as a case study in evaluating the nature of visualization decision-making processes. Specifically, we investigate the influence of *visual-spatial biases*, which are decision-making biases elicited by the visual elements of the display. Visual-spatial biases are produced by heuristics that may be either beneficial or detrimental to decision-making. Additionally, this work proposes practical recommendations for visualization designers, including methods for improving visualization techniques and the types of instructions that are helpful for users.

One context where prior work has documented multiple misinterpretations of data visualizations is in hurricane forecasting (Padilla, Ruginski, & Creem-Regehr, 2017;

Ruginski et al., 2016). Comparing multiple techniques for representing the uncertainty in hurricane path forecasts, our work has found that the current technique (*cone of uncertainty*) used by the National Hurricane Center produced more misinterpretations than an alternative approach called a *simulation ensemble* visualization, in which forecast data are represented as individual lines (see Figure 1.1) (Padilla et al., 2017; Ruginski et al., 2016). The term *ensemble visualization* has been used in various domains to describe different types of ensemble visualization techniques. In simulation science, simulation ensemble visualizations refer to visualizations of sets of forecast ensemble simulations or *ensemble members*. Scientists generate ensemble members by changing the initial conditions or parameters of a simulation and by utilizing different simulation models (for a more in-depth description of ensemble forecasts see Hamill, 2001; Potter et al., 2009). Others use the term ensemble more liberally to refer to situations where sets of visual information are displayed together, such as scatterplots (Rensink, 2014, 2016; Szafir, Haroz, Gleicher, & Franconeri, 2016), illustrations of objects (Sweeny, Wurnitsch, Gopnik, & Whitney, 2015), photographs of crowds, and faces (Leib et al., 2014), referred to here more broadly as *visual ensembles*. The simulation ensemble hurricane forecast tested in Padilla et al. (2017) was created using a mathematical model based on 5-year historical hurricane data. Each of the lines (seen in Figure 1.1) represents one run of the model, with perturbations to speed and bearing.

While multiple studies have found that simulation ensemble visualizations intuitively communicate the probability of hurricane paths more effectively than other techniques (Liu et al., 2016; Padilla et al., 2017; Ruginski et al., 2016), one study showed that these ensemble visualizations can also bias viewers judgments for a specific task

(Padilla et al., 2017). In Padilla et al. (2017), we tasked participants with comparing potential damage to two oil platforms in the Gulf of Mexico, using simulated ensemble hurricane path visualizations. For each trial, one of the oil platforms was collocated with a forecasted hurricane path and the other was not (see Figure 1.1). This study examined whether viewers believed that oil platforms closer to the center of the storm (i.e., the area with the most densely populated grouping of lines) would receive more damage, as reported in Ruginski et al. (2016), or if participants' judgments would change when the farther oil rig was collocated with a hurricane track. Nonexpert viewers reported 99% of the time that the oil rig closer to the center of the storm would receive the most damage when it was collocated with a hurricane track (Figure 1.1 B). However, when the oil rig farther from the center was collocated with a hurricane track and the closer oil rig was not (Figure 1.1 A), participants reported that the closer oil rig to the center of the storm would receive more damage 54.59% of the time. In other words, placing a rig directly on a track led the viewer to interpret the likelihood of damage as greater. We call this the *collocation* effect and it is a misinterpretation of the type of simulation ensemble visualization used in the current study, as each of the lines is only a subset of the many possible outputs of the model. If one line intersects the oil rig, the potential of damage does not increase.

The goal of the current work was to identify the cognitive processes that underlie the collocation effect. To empirically examine the sources of the collocation effect, we draw on the integrative theory of visualization decision making by Padilla et al. (2018). This theory proposes that biases that are a direct result of the visualization technique may

be a unique category of biases termed visual-spatial biases.¹ These biases are directly driven by the visual stimulus and are different from those such as a familiarity bias that are driven by prior knowledge. Identifying whether the collocation effect is a visual-spatial bias provides a clue as to when the collocation effect may arise in the decision-making process. Padilla et al. (2018) suggest that visual-spatial biases likely originate early in the decision-making process, possibly during bottom-up attention. The early emergence of visual-spatial biases may be one reason why prior work has documented difficulty in helping viewers overcome these biases (Boone, Gunalp, & Hegarty, in press; Joslyn & LeClerc, 2013). If the collocation effect is a visual-spatial bias, then changing the visual features of the visualization technique should modify the bias.

1.2 Ensemble Visualizations

Both simulation ensembles and visual ensembles are increasingly popular methods for visualizing data, as emerging research demonstrates that they can efficiently and intuitively communicate difficult statistical concepts to novice viewers, such as probability distributions (Cox, House, & Lindell, 2013; Leib et al., 2014; Sweeny et al., 2015; Szafir et al., 2016), trends in central tendency (Szafir et al., 2016), and the slope, amplitude, and curvature of bivariate data (Correll & Heer, 2017) (for comprehensive reviews see Alvarez, 2011; Whitney, Haberman, & Sweeny, 2014). Sweeny et al. (2015) further showed that children as young as 4 could accurately judge the relative average size of visual ensembles. Ensemble visualizations are also desirable because they make a

¹ It should be noted that visual-spatial biases can be both beneficial and detrimental to decision making, even though this work only focuses on errors produced by visual-spatial biases.

representative portion of the data visually available, which preserves data for replication purposes (Liu et al., 2016) and maintain relevant outlier information (Szafir et al., 2016). Ensemble visualizations can also depict non-normal relationships in the data such as bimodal distributions, perceived as discrete clusters (Szafir et al., 2016). Several studies have shown that viewers can mentally summarize visual features of visual ensembles by perceiving the gist or integrating visual ensemble data into rich and quickly accessible information (Correll & Heer, 2017; Leib et al., 2014; Oliva & Torralba, 2006; Rousselet, Joubert, & Fabre-Thorpe, 2005). Szafir et al. (2016) have proposed that the ease by which visual ensembles communicate the gist of data produces efficient execution of four specific tasks (identification, summarization, segmentation, and structure estimation). In sum, there is evidence that adult novice viewers and children can, in some cases, intuitively derive complex statistical information from ensemble visualizations and that they can preserve potentially useful characteristics of the data.

While previous research indicates that there are various benefits to ensemble visualizations, there are also some drawbacks. The primary issue that has been proposed with ensemble visualizations is that they can appear overly complicated. Simulation ensembles have even been colloquially named “spaghetti plots” because they can look like a mess of intertwined lines (Diggle, 2002). For example, visual crowding may occur with simulation ensemble visualizations, which happens when ensemble members are plotted too closely together and cannot be easily differentiated. While some have developed algorithms to reduce visual crowding (Liu et al., 2016), visual crowding may still occur when all of the simulation ensemble members are plotted.

Beyond issues with the display technique, only a few studies have assessed if

ensemble visualizations produce adverse effects for decision making (Padilla et al., 2017; Rensink, 2014, 2016). Rensink (2014, 2016) found that viewers had a strong bias when estimating correlations from scatter plots but also demonstrated that the laws that viewers followed remained similar across variations of encoding techniques and data parameters, such as changes in density, aspect ratio, color, and the underlying data distribution. Rensink concluded that a linear (Weber) law for discrimination and a logarithmic (Fechner) law for perceived magnitude dictate perception correlation and are not affected by changes in the visual properties of visual ensembles. Together with the previously detailed collocation effect (Padilla et al., 2017), these studies document cases where simulation and visual ensembles bias judgments; however, no clear cognitive mechanism has been proposed that predicts how and why ensemble visualization biases occur.

1.3 Visual-Spatial Biases

The concept of visual-spatial biases was proposed by Padilla et al. (2018) in a review paper of decision making with static visualizations. In this theoretical review, we detailed some studies that reported inconsistent results when attempting to change participants' decisions with instructions or decision-making aids (Boone et al., 2018; Joslyn & LeClerc, 2013). For example, Boone, Gunalp, and Hegarty (2018) attempted to modify participants' judgments of hurricane forecasts by providing additional instructions about how the hurricane forecasts were generated. The type of hurricane forecast visualization tested was the *cone of uncertainty*, which is used by the National Weather Center to display hurricane track forecasts. The cone of uncertainty leads some people to believe that the hurricane is growing in physical size over time, which is a

misunderstanding of the increasing path uncertainty that the cone is intended to represent (Padilla et al., 2017; Ruginski et al., 2016). Boone et al. (2018) found that multiple types of instructions helped to reduce misconceptions about the cone of uncertainty, but instructions did not consistently influence participants' actual decisions about the storm. Additionally, Joslyn and LeClerc (2013) found that when temperature uncertainty was visualized as error bars around a mean temperature prediction, participants incorrectly believed that the error bars represented high and low temperatures. Participants maintained this belief despite a key, which detailed the correct way to interpret each temperature forecast (see also Grounds, Joslyn, & Otsuka, 2017). A key commonality between the two previously mentioned studies is that the visual elements of the displays were more influential to decisions than instructions or decision aids, which illustrates the powerful influence of visual elements on decision making. Biases that arise from the visual elements of the display have not been directly studied before (Padilla et al., 2018), which leaves open questions about why these biases may be particularly hard to overcome and if they constitute a unique category of decision biases.

One possible origin of visual-spatial biases is visual thought. Tversky (2011) proposed that visual thought is the way that we mentally organize information into spatial relationships and, like language, has been developed over centuries and is continuously being refined. Humans use visual thought to communicate concrete concepts such as relative locations between two towns and abstract concepts such as a number line. We communicate visual thought with all of our modalities from gestures to cave paintings. There are numerous examples of how visual thought dictates how we conceptualize information (Gentner, 2001; Lakoff & Johnson, 2008; Tversky, 2001, 2011). It is

possible that visual-spatial biases arise from viewers interpreting a visualization based on standardized visual thought when the visualization was not intended to be interpreted in this manner. For example, Padilla et al. (2018) observed a containment visual-spatial bias in several studies, where viewers interpret elements within a boundary as more similar than elements outside of a boundary (Belia et al., 2005; Boone et al., 2018; McKenzie, Hegarty, Barrett, & Goodchild, 2016). In one study, McKenzie et al. (2016) showed that participants who viewed Google Map's blue dot visualization with a hard boundary were more likely to use a containment heuristic than those who saw the same data but represented with a blurred edge created by a Gaussian fade (see also Newman & Scholl, 2012; Ruginski et al., 2016). Tversky (2011) illustrates the way that we conceptualize spatial boundaries with the analogy, "Framing a picture is a way of saying that what is inside the picture has a different status from what is outside the picture" (p. 522) (see also Fabrikant & Skupin, 2005). It is possible that visual thought, as described by Tversky (2011), is highly related to *perceptual grouping*, which is the process by which the visual-system separates objects from a background (Grossberg, Mingolla, & Ross, 1997). If visual-spatial biases are produced by perceptual grouping, this would suggest that the biases occur relatively early in the decision-making process. However, directed examination is needed to determine if visual-spatial biases exist, and their relationship to visual thought and perceptual grouping.

In contrast to visual-spatial biases, Padilla et al. (2018) documented biases that are not influenced by the characteristics of the visual display. For example, Shen, Carswell, Santhanam, and Bailey (2012) were able to help viewers overcome a *familiarity bias* during a complex geospatial task with instructions. Familiarity biases have been widely

studied in behavioral economics and decision science, and illustrate how we believe that items and situations that are familiar to us are more important, probable, and preferable (Kahneman & Tversky, 1977; Park & Lessig, 1981). In the context of visualizations, prior work found that participants were more likely to select familiar map-like visualizations to make a judgment about terrorism threats rather than a visualization that would be more optimal for the task (Bailey, Carswell, Grant, & Basham, 2007). Shen et al. (2012) demonstrated that users were more likely to choose an efficacious map when given training concerning the importance of effective visualization techniques. In this case, viewers were able to use knowledge-driven processing to overcome the familiarity bias. A key difference between these studies is that Shen et al. (2012) used instructions to correct a familiarity bias, which is a cognitive bias that is not a result of the visual elements in the display. In contrast, we propose that the biases in Boone et al. (2018) and Joslyn and LeClerc (2013) were likely visual-spatial biases. It is possible that the reason visual spatial biases are so persistent is that they are driven by attention to salient features early in the decision-making process, thus influencing the entire downstream process. In a review of attention and decision making, Orquin and Loose (2013) detailed numerous studies that also find attention has a profound influence on downstream processes. For example, Shimojo, Simion, Shimojo, and Scheier (2003) found that manipulating the length of time that viewers gaze at a face increased their preference for the face. They propose that gaze duration has a “cascade effect” whereby attention has a compelling influence on many of the downstream decision-making processes (see also Richardson et al., 2009).

While numerous studies provide circumstantial evidence for the notion of visual-

spatial biases by proposing that the visual elements in the display are responsible for viewers interpretations (Belia et al., 2005; Joslyn & LeClerc, 2013; Liu et al., 2016; McKenzie et al., 2016; Newman & Scholl, 2012; Padilla et al., 2017; Ruginski et al., 2016), no work has directly tested the theory of visual-spatial biases. We propose that because visual-spatial biases are driven by the visual elements inherent to the display, they are unique to visualization decision making and may be especially robust, even when additional knowledge-driven information is provided.

1.4 Overview of Experiments

The goal of this work is to test the hypothesis that the collocation effect demonstrated in Padilla et al. (2017) is a visual-spatial bias (i.e., influenced by the visual elements of the visualization technique) and also test whether it can be influenced by top-down processing. To achieve this goal, we conducted a series of three experiments. In Experiment 1, we manipulated the number of total simulated ensemble members displayed. We predicted that viewers would report more damage for locations that are collocated with a simulated ensemble member when fewer overall ensemble members are plotted. This finding would suggest that the collocation effect is influenced by the visual elements of the display and provides evidence for a visual-spatial bias.

To test if the collocation effect can be influenced by top-down knowledge-driven processing, we conducted two additional experiments. In Experiment 2, we used think-aloud protocols to examine conscious strategies that might contribute to the collocation effect, using the simulated ensemble visualization from Experiment 1 with the fewest members. The goals of this experiment were to identify both conscious strategies and

errors in reasoning that may have produced the collocation effect. In Experiment 3, we utilized the findings of Experiment 2 to create two different types of instructions—one specific to the task and the other about the visualization technique. Then we tested if participants could incorporate top-down knowledge via instructions to overcome the collocation effect by comparing their responses to those of Experiment 1 where no instructions were provided. Importantly, it may be that manipulations of both the visualization (Experiment 1) and instructions (Experiment 3) would reduce the collocation effect.

Figure 1.1. Example of hurricane forecast display stimuli used in Padilla et al. (2017), where the red dots indicate the location of off shore oil platforms. In Figure 1.1 A, location A is collocated. In Figure 1.1 B, location B is collocated.

CHAPTER 2

EXPERIMENT 1

To test if the collocation effect is a visual-spatial bias, we conducted an experiment where the number of hurricane simulated ensemble members (9, 17, 33, and 65) was manipulated. Participants viewed a hurricane track visualization with either 9, 17, 33, or 65 tracks and estimated the level of damage that would incur to an off-shore oil rig at a specified location (see Figure 2.1 for example stimuli).

2.1 Methods

2.1.1 Participants

Based on the effect size in Padilla et al. (2017), a power analysis was conducted using G*Power (Buchner, Erdfelder, Faul, & Lang, 2017) to determine an adequate sample size. At an alpha of level 0.05, power of 0.80, and an effect size of $f^2 = 0.11$, the minimum number of participants needed is 54. Participants were 200 undergraduate students currently attending the University of Utah who completed the study for course credit. Seventy-three were male, and 127 were female, with a mean age of 22 ($SD = 5.58$). Each participant completed the task with visualizations that had one quantity of simulated ensemble members (9 $n = 52$, 17 $n = 50$, 33 $n = 50$, and 65 $n = 48$).

2.1.2 Stimuli

We used stimuli that were designed to mimic properties of simulation ensembles from Liu et al. (2016), while being simple enough to maintain experimental controls and test the collocation effect. Stimuli were presented online using the Qualtrics web application (Qualtrics, 2005). On each trial, participants were shown a display depicting a hurricane path visualization. To examine the collocation effect, custom code was generated to create artificial hurricane forecast images in such a way where one of the simulated ensemble members overlapped an “oil rig” depicted as a black dot (see Figure 2.2, A). In the custom code, a dot angle was specified that indicated the angle away from the midline that the oil rig would be placed. $N-1$ hurricane track lines were sampled from a clipped normal distribution of 40 degrees and a standard deviation of 5 degrees for each line, N being the number of desired lines. The lines could not overlap and two gaps in the distributions were specified. The first gap was placed around the oil rig, so that only one ensemble member intersected the oil rig. The second gap was equidistant from the center, which allowed the distribution of paths to be flipped over the midline to create additional stimuli and account for any skewing that may occur from random sampling. One line was then specified to intersect with the oil rig, producing one stimuli image (see Figure 2.2, A). A second image was also generated where the additional line was plotted through the gap that did not contain the oil rig (see Figure 2.2, B). The midline was accidentally plotted and did not adhere to the minimum distance constraint. The two resulting images were then flipped over the midline to create a total of four mirrored images, two collocated and two none-collocated (see Figure 2.2, C and D). Thin line widths for the hurricane tracks and a small diameter for the oil rig were selected to increase the precision of the dot

overlap with the line.

We chose the following distances to place the oil platforms relative to the mean of the distributions, 14 degrees and 12 degrees. Because the buffer creates a small gap in the distributions, the distances were chosen to be on the outer portions of the distributions. This reduced the salience of the gap by insuring it was located in less densely populated regions of the distributions. Each simulated ensemble member was a straight line of fixed length, characterized by its slope represented as an angle. Four quantities (9, 17, 33, and 65) of angles were randomly sampled from a clipped normal distribution with a maximum spread of 40 degrees, a standard deviation of 5 degrees, and a line thickness of 1 pixel. These quantities were selected to represent a wide range of values, which were created starting with a base of 8 and using a logarithmic scale to select, 16, 32, and 64. Each quantity had an additional ensemble member, which was the transient ensemble member that was either collocated with the oil platform or moved to the other side of the distribution. Sixty-five simulated ensemble members were subjectively the upper bounds of reasonable ensemble members to represent, given the standard deviation, max-spread, and thickness of lines that we specified. In our case, more than 65 simulated ensemble members would have resulted in even more overplotting, meaning that the distribution would no longer be perceivable. Given that we aimed to test if we can reduce the collocation effect by increasing the number of simulated ensemble members, we felt it was essential to test a wide range of quantities of ensemble members even if all the versions did not adhere to visualization design recommendations. Finally, to increase the number of trials, each of the permutations was seeded four times (i.e., randomly sample four times), and each of these was displayed with the midline oriented at three different

angles (-30, 0, -30), resulting in a total of 96 trials. All simulated ensemble distributions were digitally composited over a map of the U.S. Gulf Coast that had been edited to minimize distracting labeling. These images were displayed to the subjects at a pixel resolution of 960 x 640 pixels. Underneath the forecast, a scale ranging from 1 (no damage) to 7 (severe damage) was displayed.

2.1.3 Design

We utilized a 4 (number of simulated ensemble members: 9, 17, 33, and 65) x 2 (collocation: on- and off-line) x 2 (oil rig locations: 12° and 14°) x 2 (side of the distribution: left and right) x 3 (angle of storm: -30, 0, and 30) x 4 (seeds) mixed factorial design. Collocation, oil rig locations, side, angle of storm, and seeding were within-participant variables, resulting in a total of 96 trials per participant. Participants were randomly assigned to one of four visualization conditions (9, 17, 33, and 65 simulated ensemble members) as a between-participants factor.

2.1.4 Procedure

Individuals were first given the following instructions of the task and visualization.

In the following experiment, you will view maps showing the forecast path of different hurricanes as they travel over the Gulf of Mexico, towards land. The maps will also show the location of one offshore oil platform in the Gulf. Oil platforms are large structures on the surface of the water with components that extend to the ocean floor for drilling and storing oil.

See the sample map below. A set of potential forecast paths of where the hurricane will move in the next three days are shown in red and the location of the oil platform is shown by a small black circle. Your task is to estimate the level of damage that the platform will incur based on the depicted forecast of the hurricane path on a scale of 1 to 7 where 1 is no damage and 7 is severe damage.

You will make your judgments of potential damage to the oil platform using the damage scale provided below the map, which will be presented to you along with the forecast maps on each trial. To respond you should check the box (1 through 7) associated with the level of damage that you believe will occur to the oil platform as a result of the forecasted hurricane. The hurricane forecasts and the locations of the oil platforms will vary across trials.

Additionally, each trial included the text as a reminder of the task, “What is the level of damage that the oil platform will incur?” Following the instructions, participants completed all of the trials presented in a different random order for each participant. Lastly, participants answered questions related to comprehension of the hurricane forecasts.

2.2 Results

Multilevel models (MLM) were fit to the data using Hierarchical Linear Modeling 7.0 software and restricted maximum likelihood estimation procedures (Raudenbush & Bryk, 2002). Multilevel modeling is a generalized form of linear regression that is used to analyze variance in experimental outcomes predicted by both individual (within-participants) and group (between-participants) variables. The package lme4 in R (Bates, Maechler, & Bolker, 2012) was used to calculate the regression weights, and the code and data for this analysis can be found in the supplemental material. Visualization was dummy coded such that the 9-track visualization was the referent. Collocation was coded such that the coefficients indicate a change from off-line trials to on-line trials, meaning that a significant positive slope reveals a collocation effect. Collocation (off and on), Visualization (9-track, 17-track, 33-track, and 65-track), Distance (12° and 14°), and the interaction between Collocation and Visualization were entered as fixed-effects. Participants were entered as random effects. Self-report measures of experience with

hurricanes and hurricane prone regions were also collected. The results of this analysis can be seen in Table 2.1. As the participants were students at the University of Utah, so few students had experienced a hurricane (3%) or had lived in hurricane-affected regions (7%) that we did not include these measures as covariates.

Our primary hypothesis was that we would see less of a collocation effect for hurricane track visualizations with more simulated ensemble members. To start, there was a main effect of Collocation, meaning that for the 9-track display (the referent) and at the 12° distance (also the referent), damage ratings increased by 1.7 points (on the 7 point Likert scale) when the oil rig was intersected by one of the lines compared to when it was not. This finding is consistent with the results of Padilla et al., (2017), which revealed that when participants were asked to compare two locations, their judgments were biased when one of the locations was collocated with a simulated ensemble member. Further, the current results generalize the collocation effect from a comparison of potential damage between two locations (used in Padilla et al., 2017), to a judgment of the extent of damage in a single location. The general public commonly makes estimates of the risk of a hurricane hitting a specific location, such as their town.

Consistent with our predictions, we also found a significant interaction between collocation and each of the visualizations compared to the 9-track display. The negative coefficient for each of these interactions indicates that the difference between the on-line and off-line trials is significantly smaller for the 17-, 33-, and 65-track displays compared to the 8-track display at the closest distance. The 9 track on-line trials ($M = 4.55$, $SD = 1.56$) elicited 1.71 more damage than the off-line trials ($M = 2.84$, $SD = 1.47$). The difference between the on-line and off-line trials is 0.42 smaller for the 17-track display

(on-line: $M = 4.27$, $SD = 2.99$, off-line: $M = 2.99$, $SD = 1.56$), 0.52 smaller for the 33-track display (on-line: $M = 3.98$, $SD = 1.36$, off-line: $M = 2.80$, $SD = 1.28$), and 0.15 smaller for the 65-track display (on-line: $M = 4.24$, $SD = 1.53$, off-line: $M = 2.68$, $SD = 1.47$) compared to the 9-track display.

To visualize the reduction in the collocation effect, we transformed the dependent variable by calculating the difference between the on-line damage ratings and off-line damage ratings at the same oil platform location, seed, and storm angle. This resulted in a damage change score where zero indicates no collocation effect, positive values indicate an increase in reported damage for on-line trials compare to off-line trials, and negative values would indicate a decrease in reported damage for on-line trials compare to off-line trials. These data are displayed in Figure 2.3.

As illustrated in Figure 2.3, the 17-, 33-, and 65-track visualizations show significantly less of a collocation effect. However, unexpectedly, the 65-track visualization shows more of a collocation effect than the 17- and 33-track displays. A post-hoc analysis confirmed that after setting the 65-track visualization as the referent and running the same model as previously described, there were significant interactions between Collocation and each of the visualizations compared to the 65-track display. The negative coefficients for the interactions between Collocation and the 17-track display ($b = -0.27$), $t(191) = -6.46$, $p < 0.00$, 95% $CI [-0.35, -0.19]$ and the 33-track display ($b = -0.37$), $t(191) = -8.74$, $p < 0.000$, 95% $CI [-0.45, -0.28]$ indicated that the collocation effect was significantly smaller for the 17- and 33- track displays compared to the 65-track display at the 12° distance.

There was also significant main effect of distance, which revealed that at the

distance closer to the center of the distribution (12° , $M = 3.63$, $SD = 1.06$), participants believed that the oil rig would receive more damage compared to the farther oil rig location (14° , $M = 3.47$, $SD = 1.09$). While significant, this is a small change on a Likert scale from 1-7. In our prior work, we utilized a wide range of distances from the center of the distribution and found correspondingly larger differences in damage ratings (Padilla et al., 2017; Ruginski et al., 2016). Our intention here was not to add to the distance finding, but to conceptually replicate it to ensure that the changes we made to the visualization technique did not result in unintended consequences to the viewers' perception of the distribution. In sum, this finding is in line with past work that suggests viewers effectively perceived the probability distribution that the hurricane track simulated ensemble visualization is intended to represent.

Participants also reported their confidence in their judgments for each trial using a Likert scale ranging from 1 (not at all confident) to 7 (very confident), along with follow-up questions about the visualizations. Using a multilevel model, we evaluated the impact of Visualization and Collocation (fixed effects) on confidence ratings with participants as random effects. This analysis revealed no significant change in confidence from the 9-track display ($M = 4.56$, $SD = 1.63$) compared to the 17- ($M = 4.65$, $SD = 1.51$), 33- ($M = 4.39$, $SD = 1.54$), and 63-track displays ($M = 4.87$, $SD = 1.44$). However, there was a main effect of Collocation, which showed the participants were more confident about their judgments for the on-line ($M = 4.68$, $SD = 1.50$) trials than the off line trials ($M = 4.56$, $SD = 1.58$), ($b = .117$), $t(198) = 9.03$, $p < 0.000$, 95% CI [0.09, 0.14]. However, the increased confidence for the online trials was quite small, 1.71%.

The results of the survey questions can be found in Table 2.2. Using a general

linear model, Visualization was used to predict question accuracy with the 9-track display as the referent. Full output of the models can be found in the supplementary materials. For Q1, which references uncertainty in the visualization, there were no significant differences between the 9-track display and other visualization techniques, with participants at chance. For Q2, which references the collocation effect, participants viewing the 17- and 33-track displays showed lower correct responses than those viewing the 9-track display. This is surprising because participants' behavioral judgments were in the opposite direction where the 17- and 33-track displays show the least collocation effect. For Q3, which references the number of simulated ensemble members, viewers of the 65-track display were significantly worse at this question compared to the 9-track display. This result suggests that when too many simulated ensembles are plotted, one unintended effect may be that viewers may believe that they represent all of the possible outcomes. To follow-up on this finding, a linear model was conducted with Q3 predicting the damage change score (collocation effect) of the 64-track display. This analysis revealed that participants who answered Q3 correctly showed significantly less of a collocation effect ($M = 1.38$, $SD = 1.36$) compared to those who answered the question incorrectly ($M = 1.68$, $SD = 1.56$), ($b = -0.30$), $t(46) = -6.82$, $p < 0.00$, 95% $CI [-0.38, -0.21]$.

2.3 Discussion

The results of this experiment showed that novice users are less biased by the impact of a single simulated ensemble member when more ensemble members are represented. This finding supports our prediction that the collocation effect can be

influenced by changes to the visualization technique (without any changes to the instructions), which suggests that it is in part a visual-spatial bias. Additionally, there were several effects relating to visualizing the largest number of simulated ensemble members (65-track). The primary unpredicted finding is that for the 65-track display, the collocation effect was greater than for the 17- and 33-track displays (although still less than the 9-track display). Additionally, the postsurvey Q3 suggested that viewers were more likely to believe that the 65-track display represented all of the possible paths the hurricane could take. A follow-up analysis confirmed that, for the 65-track display, incorrect beliefs about the visualization representing all of the possible paths increased the collocation effect. In sum, while increasing the number of simulated ensemble members can reduce the collocation effect, there is evidence that when many simulated ensemble members are represented, more viewers believe that all of the possible outcomes are shown.

Importantly, it should be noted that the collocation effect was never completely ameliorated. The 34-track visualization showed the largest (30%) reduction of the collocation effect compared to the 9-track display. However, participants still reported that oil-platforms that were directly hit by one of the simulated ensemble members would receive 1.28 units of more damage on the Likert scale than oil-rig locations that were not directly hit. This finding suggests that the collocation effect is not solely a visual-spatial bias. Nonvisual factors such as knowledge-driven processing may also contribute to the collocation effect. To examine conscious knowledge-driven strategies that participants may have used, Experiment 2 was conducted, which utilized think-aloud protocols to evaluate how participants believed they completed the task.

Figure 2.1. Example stimuli, showing the 9-track (A), 17-track (B), 33-track (C), and 65-track (D) displays. The black dot indicates the location of the offshore oil rig.

Figure 2.2. Example of the 33-track display, where one line was reflected over the mean line of the distribution of simulated ensemble members to make a collocation condition (A) and a noncollocation condition (B). C and D represent the mirror images of A and B, where the underlying map remained constant.

Figure 2.3. Damage change scores for the 9-, 17-, 33-, and 65-track displays. Error bars represent 95% confidence intervals around the mean.

Table 2.1. List of fixed effects with coefficients, standard errors, *t-values*, *p-values*, and 95% confidence intervals from the statistical model predicting damage ratings. Collocation was coded such that the effects indicate a change from off-line to on-line.

Fixed Effects	<i>Coeff</i>	<i>Std. Error</i>	<i>t-value</i>	<i>p-value</i>	95% <i>CI</i>
(Intercept)	3.87	0.17	21.83	0.00	(3.52, 4.21)
Collocation	1.70	0.02	58.40	0.00	(1.65, 1.76)
17-track	0.14	0.21	0.67	0.50	(-0.27, 0.55)
33-track	-0.05	0.21	-0.23	0.81	(-0.46, 0.36)
65-track	-0.16	0.21	-0.75	0.45	(-0.58, 0.25)
Distance	-0.07	0.007	-10.54	0.00	(-0.09, -0.06)
Collocation*16-	-0.42	0.04	-10.21	0.00	(-0.50, -0.34)
Collocation*32-	-0.52	0.04	-12.53	0.00	(-0.60, -0.44)
Collocation*64-	-0.15	0.04	-3.58	0.0003	(-0.23, -0.06)

Table 2.2. Proportion of correct responses for each visualization condition. *** Indicates *p-values* < .000, ** *p-values* < .005, and * *p-values* < .05 with the 9-track display as the referent.

Questions	9-track % Correct	17-track % Correct	33-track % Correct	65-track % Correct
Q1. The display indicates that the forecasters are less certain about the path of the hurricane as time passes.	50%	46%	58%	37.5%
Q2. Locations that are touching a hurricane track are more likely to be hit by the storm than locations equidistant from the center of the forecast but not touching a hurricane track.	34.6%	16%*	12%*	25%
Q3. The hurricane forecast shows all possible paths the hurricane could take.	63%	54%	54%	42%*

CHAPTER 3

EXPERIMENT 2

The second aim of this work is to test if top-down knowledge-driven processing can influence the collocation effect. The first step in achieving this goal is to identify what types of conscious decision-making strategies participants are aware of using, in order to determine how top-down knowledge may be able to influence the collocation effect. Based on prior work that examined the use of strategies in a mental rotation task that included visual information (Hegarty, 2017), we used a concurrent verbal protocol and a retrospective protocol. Work by Ericsson and Simon (1992) suggests that verbal protocols do not effectively capture low-level processing mechanisms because participants may be unaware of components of visual processing, such as saccades. The objective of this experiment was to study the processes that participants were aware of, in the case that they may be adopting deliberate cognitive strategies that may contribute to the collocation effect.

3.1 Methods

3.1.1 Participants

Participants were 20 undergraduate and graduate students currently attending the University of Utah who received \$10 for participation. Eight participants were male, and

12 were female, with a mean age of 25.75 ($SD = 4.3$).

3.1.2 Stimuli and Design

The same 9-track stimuli were used as in Experiment 1. Participants completed concurrent verbal protocols during 10 randomly sampled trials from Experiment 1. Participants viewed only the 9-track displays for this experiment.

3.1.3 Procedure

Participants began by receiving instructions on how to complete concurrent verbal protocols, which involve instructing participants to verbalize their thoughts as they complete each stage of the study, including the practice trials (Ericsson & Simon, 1992). In line with recommendations from Ericsson and Simon (1992), three practice trials were used to help participants become comfortable with verbalizing their thoughts while completing the task. Participants were then given 10 think-aloud trials in which they were instructed to verbalize everything that came to mind as they completed all steps of the task. Following recommendations from Ericsson and Simon (1992), participants were encouraged to “keep talking” rather than a social communication request, such as “tell me what you think.” Finally, participants completed retrospective protocols where they reported what they thought while they completed the think-aloud protocols. The entire process was video recorded and transcribed.

3.1.4 Coding

Three distinct strategies and three combinations of strategies were defined a priori. The strategies were motivated by the behavioral data and informal self-reports in Padilla et al. (2017). *Distance strategy* was coded when participants reported determining their damage rating based on how far the oil rig was from the center of the distribution of simulated ensemble members. *Collocation* was coded when participant specifically commented on rating oil rig locations that are collocated with a simulated ensemble member as receiving more damage. *Surrounding ensemble members* was coded when participants reported making their damage judgments based on the distance of the oil rig to the surrounding simulated ensemble members. For example, a participant responded, “I looked at how close the oil (rig) was to the red travel lines and guessed how much damage would occur by the location and proximity of space.” Combinations of these strategies were also observed. Two raters independently coded each of the 10 concurrent verbal protocols, for all 20 participants, along with all of the retrospective protocols for evidence of these three strategies and their combinations. Trials that did not meet any of the coding criteria were coded as *No discernable strategy* and included in the analysis. For two think-aloud trials, one person spoke too quietly for the audio recording device to detect and one other person did not speak aloud for one of the trials. For the retrospective protocols, 20% of people did not remember using the strategies that they reported while completing the task and instead remembered anecdotal information about hurricanes or how they would have liked to be given more instructions about what the visualizations represent.

3.2 Results

Intraclass correlations were computed using SPSS (IBM, 2013) to determine interrater reliability (Landers, 2015). The results showed that the agreement between the raters was reasonable for the concurrent verbal protocols ($ICC .817$, 95% $CI [.759 - .862]$, $F(199, 199) = 5.478$, $p < .001$) and retrospective verbal protocols ($ICC .809$, 95% $CI [.759 - .862]$, $F(199, 199) = 5.478$, $p < .001$). Table 3.1 shows the frequency of strategies. Consistent with previous behavioral data (Liu et al., 2016; Padilla et al., 2017; Ruginski et al., 2016), the majority of people reported using the distance strategy (40%) and a combined strategy of distance and collocation (16.7%). The dominance of the distance strategy is an additional indicator that this type of simulated ensemble visualization does effectively communicate a distribution of uncertainty in the hurricane path. The next most common strategy was to base judgments on surrounding members + Collocation (12.8%). While few participants reported only using the collocation strategy (5.2%), strategies that included collocation collectively represented 29.5% of trials. Interestingly, the concurrent and retrospective protocols did not fully agree. Notably, 20% of participants did not recall thinking about any of the coded strategies, even though when they were completing the trials they reported using them. The instructions for the retrospective protocols were, “Please describe in as much detail as possible everything you remember thinking about during the experiment, including during the instructions.” This experiment might be unique compared to other work that has used retrospective protocols (Ericsson & Simon, 1992; Hegarty, 2017), in that the task may have elicited prior knowledge about hurricanes, probabilities (both correct and incorrect), and historic events. Prior knowledge may have been easier to recall compare to specifics about how

participants completed each trial. For example, one participant commented in the retrospective protocol:

I remember thinking about Hurricane Ike and Hurricane Katrina. Those are probably the most famous hurricanes I remember in my lifetime. I remember thinking about earthquake scales, and wondering if hurricanes were graded the same way. Especially with the scale of damage of the rig from 1-7.

Further, the majority of participants (17 of 20) reported that they wanted more information about how to interpret the hurricane forecasts, including if it mattered that one of the simulated ensemble members was collocated with the oil rig.

3.3 Discussion

We found that the majority of participants were consciously aware of using a distance strategy and many people reported factoring in the collocation of the oil rig and a simulated ensemble member. The results of this study provide evidence that some participants strategically increased their damage ratings when the oil rig was collocated with an ensemble member. Many participants also reported basing their judgments on the surrounding simulated ensemble members to the oil rig, rather than on the distribution as a whole. Given that some people were aware of the influence of collocation and the surrounding simulated ensemble members, it is possible that if they are given instructions about how to overcome the collocation effect and interpret the visualizations correctly, they will be able to incorporate this information into their decisions.

Table 3.1. Coded strategies

Strategy	Concurrent Protocols: Proportion of trials code in each strategy by both raters		Retrospective Protocols: Number of Participants Reporting Strategy N (%)	
	Mean	# of trials (Rater 1, 2)	Mean	# of participants (Rater 1, 2)
Distance	40%	(76, 84)	22.5%	(5, 4)
Surrounding members	12%	(33, 14)	7.5%	(1, 2)
Collocation	5.2%	(12, 9)	2.5%	(1, 0)
Distance + Surrounding members	12.5%	(23, 27)	5%	(1, 1)
Distance + Collocation	16.7%	(28, 39)	25%	(5, 5)
Surrounding members + Collocation	12.8%	(25, 26)	17.5%	(3, 4)
No discernable strategy	1%	(3, 2)	20%	(4, 4)

* Indicates that only one of the raters specified this code.

CHAPTER 4

EXPERIMENT 3

The second step in determining if top-down knowledge can influence the collocation effect is to test if participants can utilize top-down knowledge to overcome the collocation effect. Prior work is inconclusive as to whether viewers can incorporate additional information to interpret a visual display (Shen et al., 2012) (c.f., Boone et al., 2018; Joslyn & LeClerc, 2013). To test if knowledge-driven processing, in the form of instructions, can influence the collocation effect, in Experiment 3, we tested if either specific information about how to overcome the collocation effect (task instructions) or more general information about how simulated ensemble hurricane forecast tracks are generated (general instructions) can reduce the collocation effect. We predicted that specific instructions about how to overcome the collocation effect would reduce the effect significantly but not completely. This finding would support the insights from Experiment 2 and suggest that visual-spatial biases can be influenced by top-down knowledge. Additionally, we predicted that the more general instructions would reduce the collocation effect but not to the degree of the task-specific instructions.

4.1 Methods

4.1.1 Participants

Participants were 83 undergraduate students currently attending the University of Utah who received course credit for participation. Three participants were disqualified for not following instructions. Of the 80 (40 in each instructions group) that were included in the analysis, 23 participants were male, and 57 were female, with a mean age of 21 ($SD = 3.7$).

4.1.2 Stimuli and Design

The same 9-track display stimuli were utilized along with the same study design as in Experiment 1. However, before receiving the experiment instructions, participants viewed one of two videos. The task-specific video included narrated instructions about the collocation effect and information about how the simulated ensemble hurricane forecasts were generated along with visual examples (length of 3.13 minutes). The sequence of the task instructions video was as follows:

1. Overview of the functions of hurricane forecasts
2. Description of how the type of hurricane forecast used in this study was generated
3. Information about uncertainty in hurricane forecasts
4. Instructions on how to identify the center of the distribution of paths and that the center represents the most likely path of the hurricane will take.
5. Illustration of how static simulated ensemble hurricane forecasts represent a subset of the many possible paths generated by the forecast models

6. Description of the collocation effect

7. Practice overcoming the collocation effect with example questions

The general instructions video was an edited version of the collocation effect video (1.37 minutes) and included elements 1-5 of the list above. Specific information about the collocation effect and practice overcoming the effect was cut out. The videos are available in the supplemental materials.

4.1.3 Procedure

Participants were randomly assigned to one of two groups (task instructions or general instructions). After consent was obtained, participants viewed the relevant video then completed the same procedure detailed in Experiment 1 but with only the 9-track visualization. As the 9-track visualization exhibited the largest collocation effect, we utilized it as a baseline to try to reduce the collocation effect with the instruction videos.

4.2 Results

As in Experiment 1, we used a multilevel logistic regression model to determine the influence of instructions on the damage ratings. We compared the 9-track display results from Experiment 1 to new data from participants that received the additional instructions. Instructions (none, task, and general), Collocation (off and on), Distance (12° and 14°), and the interaction between Collocation and Instructions were entered as fixed-effects (see Table 4.1). Participants were entered as random effects. Collocation was coded such that effects indicate a change from off-line to on-line trials and the no-instructions condition was specified as the referent.

As illustrated in Figure 4.1, the participants who viewed the general ($M = .96$, $SD = 1.22$) and task-specific instructions ($M = .66$, $SD = 1.01$) demonstrated significantly less of a collocation effect compare to those that received no instructions ($M = 1.70$, $SD = 1.56$). The coefficients for the interactions indicate that the task-specific instructions reduced the bias by 1.04 on the Likert scale. For the general instructions, the bias was reduced by .74 on the Likert scale or a 44% reduction of the collocation effect observed with the 9-track display. To test if the task-specific instructions reduced the collocation effect more than the general instructions, the same analysis was conducted, but the general instructions were specified as the referent. This analysis revealed that the task-specific instructions reduced the collocation effect significantly more than the general instructions, ($b = -0.30$), $t(125) = -6.12$, $p < 0.00$, 95% $CI [-0.39, -0.20]$.

For confidence, a multilevel model was used to evaluate the impact of Instructions and Collocation (fixed effects) on confidence ratings with participants as random effects. This analysis revealed that participants who viewed both the general ($M = 4.6$, $SD = 1.4$) ($b = .89$) $t(128) = 3.83$, $p < 0.000$, 95% $CI [0.43, 1.35]$ and the task-specific instructions ($M = 4.93$, $SD = 1.48$) ($b = 1.23$) $t(128) = 5.26$, $p < 0.000$, 95% $CI [0.77, 1.69]$ had significantly more confidence in their damage ratings compared to those that received no instructions ($M = 3.7$, $SD = 1.74$). To test if participants with task-specific instructions were more confident in their ratings compared to those who received the more general instructions, the same analysis was conducted but the general instructions were specified as the referent. This analysis revealed participants with the task-specific instructions were not more confident in their responses compared to those with general instructions, ($b = 0.33$), $t(125) = 1.35$, $p = 0.17$, 95% $CI [-.15, .82]$.

The results of the survey questions can be found in Table 4.2. Using a general linear model, Visualization was used to predict question accuracy with the no-instructions condition as the referent. Full output of the models can be found in the supplementary materials. For Q1, there were no significant differences between the 9-track display with no instructions and either of the instruction conditions. For Q2 (collocation), participants with the task-specific instructions were more likely to answer the question correctly compared to those that received no instructions. For Q3 (all possible paths), participants who received either instruction condition answered the question more correctly than those without instructions.

4.3 Discussion

In Experiment 3, we found that instructions were able to significantly, but not fully, reduce the collocation effect. The task-specific instructions attenuated the bias to a greater degree than general instructions, as evidenced by both the objective behavioral measures and the participants' self-report measures of their understanding. This work suggests that providing viewers with specific instructions about how to understand visualizations of data, including how to utilize the data for specific tasks, is vital for insuring that they fully understand and can use the visualization. Further, providing viewers with instructions improves how they feel about their judgments.

In sum, we find that the collocation effect is in part a visual-spatial bias and can be influenced by knowledge-driven processing. However, it is noteworthy that the task-specific instructions did not fully reduce the bias. The inability of the task-specific instructions to fully reduce the bias provides strong evidence that the visual

characteristics of the simulated ensemble visualization are difficult to fully override with knowledge-driven processing, making the collocation effect a compelling example of a persistent visual-spatial bias.

Figure 4.1. Damage change scores for the 9-track display conditions with no instructions, general instructions, and task specific instructions. Error bars represent 95% confidence intervals around the mean.

Table 4.1. List of fixed effects with coefficients, standard errors, *t-values*, *p-values*, and 95% confidence intervals from the statistical model predicting damage ratings.

Fixed Effects	<i>Coeff.</i>	<i>Std. Error</i>	<i>t-value</i>	<i>p-value</i>	<i>95% CI</i>
(Intercept)	4.52	0.19	23.87	0.00	(4.15, 4.89)
Collocation	1.71	0.03	56.07	0.00	(1.64, 1.76)
Task Instruction	-0.18	0.22	-0.82	0.41	(-0.60, 0.24)
General Instructions	0.81	0.22	3.74	0.0001	(0.38, 1.23)
Distance	-0.13	0.009	-13.46	0.00	(-0.14, -0.10)
Collocation*Task Ins.	-1.04	0.05	-22.54	0.00	(-1.13, -0.95)
Collocation*General Ins.	-0.74	0.05	-16.02	0.00	(-0.83, -0.64)

Table 4.2. Proportion of correct responses for each visualization condition. *** Indicates *p-values* = .000, ** *p-values* < .005, and * *p-values* < .05.

Questions	None % Correct	General % Correct	Task % Correct
Q1. The display indicates that the forecasters are less certain about the path of the hurricane as time passes.	50%	35%	42%
Q2. Locations that are touching a hurricane track are more likely to be hit by the storm than locations equidistant from the center of the forecast but not touching a hurricane track.	34.6%	42%	90%***
Q3. The hurricane forecast shows all possible paths the hurricane could take.	63%	87%*	92%*

CHAPTER 5

GENERAL DISCUSSION

Together, the current experiments show that the collocation effect could be reduced by either manipulating the visualization technique or by instructions. However, in either case, the collocation effect was never fully eliminated. Our first study demonstrated that the collocation effect was in part a visual-spatial bias, by showing that the effect was influenced by the visual characteristics of the display. We found that the collocation effect could be reduced when more simulated ensemble members were visualized. However, we also found some negative effects of representing a large number of simulated ensemble members, including the belief that the many ensemble members represented all of the possible outcomes. In Experiment 2, we found that viewers were consciously aware of the strategies they used to complete the task, including the influence of a simulated ensemble member when it was collocated with their point of interest. In Experiment 3, we found that viewers could incorporate instructions about the visualization and task to partially overcome the collocation effect.

Notably, specific and admittedly frank instructions on how to overcome the collocation effect did not fully reduce the bias. Our work proposes that the visual elements of the display have a powerful influence on decision making and should be taken seriously. This finding supports the claim that visual-spatial biases are a unique

category of bias that may be driven by low-level visual information. It is important to examine the nature of visual-spatial biases as they can have large-scale impacts on global health and safety when they affect visualizations that people use to make decisions with risks, such as hurricane evacuation or global climate change. Thus, visual-spatial biases warrant directed study to evaluate their nature and implications.

While this work supports the proposal that visual-spatial biases are a unique type of decision making bias, many questions remain open as to the true source of these biases. As Padilla et al. (2017) proposes, biases could arise from viewers attending to salient information in a visualization. It is possible in the current study, for example, that viewers paid more attention to a specific simulated ensemble member when it affected their point of interest and that the additional attention produced an overweighting of the relevant ensemble member. Viewers' attention may have been drawn to the specific simulated ensemble member because it had a small buffer, which possibly made the collocated ensemble member more salient. Participants may have also utilized top-down attention to focus on the collocated simulated ensemble member because they believed it was relevant for the task. The results of Experiment 2 suggest that some participants believed that the collocated ensemble member was relevant for the task, and those viewers may have also paid more attention to it.

In a review of decision making and attention Orquin and Loose (2013) noted that, "...decision makers will fixate on salient stimuli with a higher likelihood, regardless of its importance to the decision" (p. 191). Many studies also find that salient information in visualizations draws viewers' attention (Fabrikant, Hespanha, & Hegarty, 2010; Hegarty, Canham, & Fabrikant, 2010; Hegarty, Friedman, Boone, & Barrett, 2016; Padilla et al.,

2017; Schirillo & Stone, 2005; Stone et al., 2003; Stone, Yates, & Parker, 1997). For example, Stone et al. (1997) propose that the biases they observed in visualizations of health risk were due to attention. In one study, viewers were shown pictograms with icons that indicated how many people out of 5,000,000 were injured when driving on standard tires (30 injured) compare to the 15 people who were injured with improved tires (see Figure 5.1). When asked how much they would pay for the improved tires, in one study, participants were willing to pay an extra \$102 for the improved tires, which is a 45% increase. Stone et al. (1997) propose that viewers focused on the change in number of icons (30 - 15) to make their judgments rather than comparing the real change given the large base rate of 5,000,000 (i.e., actual change of .0000003). It is possible that other visual-spatial biases, such as the collocation effect, may also be driven by attention. One way to test how hurricane forecasts direct viewers' attention is to utilize eye tracking measures. A gaze analysis could determine if participants are paying more attention to the simulated ensemble member that is collocated with the oil rig, which would suggest that the collocation effect is related to attention.

In addition to attention, visual-spatial biases may also be produced by incorrect application of *graph schemas* or prior knowledge about a visualization, such as graph conventions (Pinker, 1990). If a viewer has learned a schema that is relevant to the visualization, Pinker (1990) proposes that the viewer compares the schema to the visualization and updates the mental representation of the visualization to include this prior knowledge. For example, if a viewer sees a bar chart, she may remember the conventions for the X- and Y-axis and use that information to interpret the graph. However, issues arise when a viewer uses the wrong graph schema to interpret a

visualization. For example, Joslyn and LeClerc (2013) found that when participants viewed error bars around a mean temperature forecast, they incorrectly believed that the error bars represented high- and low-temperatures. Joslyn and LeClerc (2013) propose that viewers incorrectly utilized the mental schema for high- and low-temperature forecasts because they looked similar. Participants maintain this incorrect judgment despite a key detailing the correct way to interpret the forecasts. Additional empirical examination is needed to evaluate if other visual-spatial biases such as those in our current study are produced from graphic schema errors as well. Similar to Joslyn and LeClerc (2013), viewers of the simulated ensemble hurricane forecasts may have utilized the incorrect graphic schema. For example, it could be the case that viewers employed a graphic schema for driving routes. Applications, such as Google Maps and Waze, show users multiple finite paths that they could take to arrive at a destination. If viewers use a driving route graphic schema, they would incorrectly assume that each of the hurricane paths represents individual paths the hurricane could take rather than a distribution of possible paths. However, additional testing is needed to identify the specific graphic schema viewers use to interpret simulated ensemble hurricane forecasts.

This work also illustrated methods for developing both general and task-specific instructions and showed that instructions could be used to reduce biases. Prior work demonstrated inconsistent findings as to if people could utilize prior knowledge to change their judgments when viewing visualizations (Bailey et al., 2007; Boone et al., 2018; Joslyn & LeClerc, 2013; Shen et al., 2012). In line with the recommendations from Zapata-Rivera, Zwick, and Vezzu (2016), this work finds that instructions can be used to reduce a *specific error* in reasoning with visualizations. We believe that the task-specific

instructions were more successful than the general instructions because they targeted a specific visual bias rather than trying to broadly improve judgments. This is potentially why other work did not find consistent improvements in visualization decision making when providing viewers with more information (Boone et al., 2018; Joslyn & LeClerc, 2013). One future direction for this work is to combine the collocation reduction observed with the 33-track display and the instructions. If viewers of the 33-track display are given instructions about how to complete the task, it is possible that the collocation effect could be reduced entirely.

Our intention for not providing participants with detailed instructions on how to interpret hurricane forecasts in Experiment 1 was to first understand what type of biases were naturally elicited purely by the visualization technique. Further, in real-world contexts, such as in hurricane forecasts on the news, it is rare that viewers are given a full description of how the forecast visualizations were generated and how to effectively interpret the forecasts. We sought to examine how people make judgments about storm damage with limited background information, to better understand what elements of the visualization technique are eliciting biases that would likely be observed in the real world.

The applied contributions of this work are to demonstrate that simulated ensemble hurricane forecasts are effective for intuitively communicating uncertainty in hurricane paths. Additional work is needed to test if these findings generalize to other contexts where ensemble visualizations are used. Further, if more simulated ensemble members are plotted, then the negative effects of this visualization technique are reduced. Our findings suggest that around 30 simulated ensemble members reduces the collocation

effect, while not suffering from cognitive effects of representing too many lines, for the max-spread and line width used in this study. Then, if additional instructions are included about how the hurricane forecast was generated and the collocation effect, the negative effects of this visualization technique can be further reduced. Considering newer visualization methods that allow scientists to represent the size and intensity of the storm in addition to the predicted storm path (Lui, Padilla, Creem-Regehr, & House, under review), we suggest that simulated hurricane ensemble forecasts are one of the most promising visualization techniques currently available.

4.1 Conclusions

Ensemble visualizations are an increasingly popular method for visualizing data, as emerging research demonstrates that ensemble visualizations can effectively and intuitively communicate traditionally difficult statistical concepts to novice viewers, such as probability. Simulated ensemble visualizations are now being used to help people make large-scale decisions such as whether to evacuate a town before a hurricane strike. Given their widespread use and social impact, it is essential to understand how ensemble visualizations influence our judgments and actions. We found that simulated ensemble visualizations that include a greater number of ensemble members (but not too many) are more appropriate for multiple types of decision-making tasks and that providing instructions about how the visualization is created can help people make more effective decisions. Further, this work demonstrates the importance of evaluating both the lower level perceptual and higher level cognitive processes at work when making decisions with visualizations. By understanding the cognitive processes associated with

visualization reasoning, we can make more effective predictions about viewers' judgments and create increasingly targeted visualization improvements and decision aids.

How much would you be willing to pay the IMPROVED tires?:
\$ _____ for 4 tires.

Figure 5.1. Icon arrays used in Stone et al. (1997) to illustrate the risk of standard or improved tires (reprinted with permission from Padilla et al., 2018).

REFERENCES

- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15. doi:10.1016/j.tics.2011.01.003
- Bailey, K., Carswell, C. M., Grant, R., & Basham, L. (2007). Geospatial perspective-taking: How well do decision makers choose their views?. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 51(18), 1246 - 1248.
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10(4), 389.
- Boone, A. P., Gunalp, P., & Hegarty, M. (2018). Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. *Journal of Experimental Psychology: Applied*, 24(3), 275.
- Buchner, A., Erdfelder, E., Faul, F., & Lang, A.-G. (2017). G*Power (Version 3.1.9.3) Retrieved from <https://stats.idre.ucla.edu/other/gpower/>
- Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 2142-2151.
- Correll, M., & Heer, J. (2017). Regression by eye: Estimating trends in bivariate visualizations. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3025453.3025922
- Cox, J., House, D., & Lindell, M. (2013). Visualizing uncertainty in predicted hurricane tracks. *International Journal for Uncertainty Quantification*, 3(2), 143-156.
- Diggle, P. (2002). *Analysis of longitudinal data*. Oxford, UK: Oxford University Press.
- Ericsson, K. A., & Simon, H. A. (1992). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Fabrikant, S. I., Hespanha, S. R., & Hegarty, M. (2010). Cognitively inspired and perceptually salient graphic displays for efficient spatial inference making. *Annals of the Association of American Geographers*, 100(1), 13-29.

- Gentner, D. (2001). Spatial metaphors in temporal reasoning. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 203-222). Cambridge, MA: MIT Press.
- Grossberg, S., Mingolla, E., & Ross, W. D. (1997). Visual brain and visual perception: How does the cortex do perceptual grouping? *Trends in Neurosciences*, 20(3), 106-111.
- Grounds, M. A., Joslyn, S., & Otsuka, K. (2017). Probabilistic interval forecasts: An individual differences approach to understanding forecast communication. *Advances in Meteorology*, Volume 2017, Article ID 3932565. 1-18. <https://doi.org/10.1155/2017/3932565>.
- Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3), 550-560.
- Hegarty, M. (2017). Ability and sex differences in spatial thinking: What does the mental rotation test really measure? *Psychonomic Bulletin & Review*, 25(3), 1212-1219.
- Hegarty, M., Canham, M. S., & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(1), 37.
- Hegarty, M., Friedman, A., Boone, A. P., & Barrett, T. J. (2016). Where are you? The effect of uncertainty and its visual representation on location judgments in GPS-like displays. *Journal of Experimental Psychology: Applied*, 22(4), 381.
- IBM. (2013). SPSS statistics for Windows (Version 21.0). Armonk, NY: IBM.
- Joslyn, S., & LeClerc, J. (2013). Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, 22(4), 308-315.
- Kahneman, D., & Tversky, A. (1977). *Intuitive prediction: Biases and corrective procedures*. Mclean, VA: Decisions and Designs Inc.
- Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. Chicago, IL: University of Chicago Press.
- Landers, R. N. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower*, 2, e143518.
- Leib, A. Y., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision*, 14(8), 26-26.

- Liu, L., Boone, A. P., Ruginski, I. T., Padilla, L., Hegarty, M., Creem-Regehr, S. H., ... House, D. H. (2017). Uncertainty visualization by representative sampling from prediction ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(9), 2165-2178.
- McKenzie, G., Hegarty, M., Barrett, T., & Goodchild, M. (2016). Assessing the effectiveness of different visualizations for judgments of positional uncertainty. *International Journal of Geographical Information Science*, 30(2), 221-239.
- Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, 19(4), 601-607. doi:10.3758/s13423-012-0247-5
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155, 23-36.
- Orquin, J. L., & Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, 144(1), 190-206.
- Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, 3(1), 29.
- Padilla, L., Ruginski, I. T., & Creem-Regehr, S. H. (2017). Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications*, 2(1), 40.
- Park, C. W., & Lessig, V. P. (1981). Familiarity and its impact on consumer decision biases and heuristics. *Journal of Consumer Research*, 8(2), 223-230.
- Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73-126). New York, NY: Psychology Press.
- Potter, K., Wilson, A., Bremer, P.-T., Williams, D., Doutriaux, C., Pascucci, V., & Johnson, C. R. (2009). *Ensemble-vis: A framework for the statistical visualization of ensemble data*. Paper presented at the Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference.
- Rensink, R. A. (2014). On the prospects for a science of visualization. In W. Huang (Ed.), *Handbook of human centric visualization* (pp. 147-175). New York, NY: Springer.
- Rensink, R. A. (2016). The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review*, 24(3), 776-797.

- Richardson, D. C., Spivey, M. J., Hoover, M. A., Taatgen, N., van Rijn, H., Nerbonne, J., & Schomaker, L. (2009). *How to influence choice by monitoring gaze*. Paper presented at the 31th Annual Conference of the Cognitive Science Society, Austin, TX.
- Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the “gist” of real-world natural scenes? *Visual Cognition*, 12(6), 852-877.
- Ruginski, I. T., Boone, A. P., Padilla, L., Liu, L., Heydari, N., Kramer, H. S., . . . Creem-Regehr, S. H. (2016). Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2), 154-172.
- Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., & Moorhead, R. J. (2009). A user study to compare four uncertainty visualization methods for 1D and 2D datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 1209-1218. doi:10.1109/TVCG.2009.114
- Schirillo, J. A., & Stone, E. R. (2005). The greater ability of graphical versus numerical displays to increase risk avoidance involves a common mechanism. *Risk Analysis*, 25(3), 555-566.
- Scown, H., Bartlett, M., & McCarley, J. S. (2014). *Statistically lay decision makers ignore error bars in two-point comparisons*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Shen, M., Carswell, M., Santhanam, R., & Bailey, K. (2012). Emergency management information systems: Could decision makers be supported in choosing display formats? *Decision Support Systems*, 52(2), 318-330.
- Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12), 1317.
- Stone, E. R., Sieck, W. R., Bull, B. E., Yates, J. F., Parks, S. C., & Rush, C. J. (2003). Foreground: Background salience: Explaining the effects of graphical displays on risk avoidance. *Organizational Behavior and Human Decision Processes*, 90(1), 19-36.
- Stone, E. R., Yates, J. F., & Parker, A. M. (1997). Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of Experimental Psychology: Applied*, 3(4), 243.
- Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental Science*, 18. doi:10.1111/desc.12239

- Szaifir, D. A., Haroz, S., Gleicher, M., & Franconeri, S. (2016). Four types of ensemble coding in data visualizations. *Journal of Vision, 16*. doi:10.1167/16.5.11
- Tversky, B. (2001). *Spatial schemas in depictions*. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 79-111). Cambridge, MA: MIT Press.
- Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science, 3*(3), 499-535.
- Whitney, D., Haberman, J., & Sweeny, T. D. (2014). From textures to crowds: Multiple levels of summary statistical perception. In J. S. Werner & L. M. Chalupa (Eds.), *The new visual neurosciences* (pp. 695-710). Boston, MA: MIT Press.
- Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment, 21*(3), 215-229.