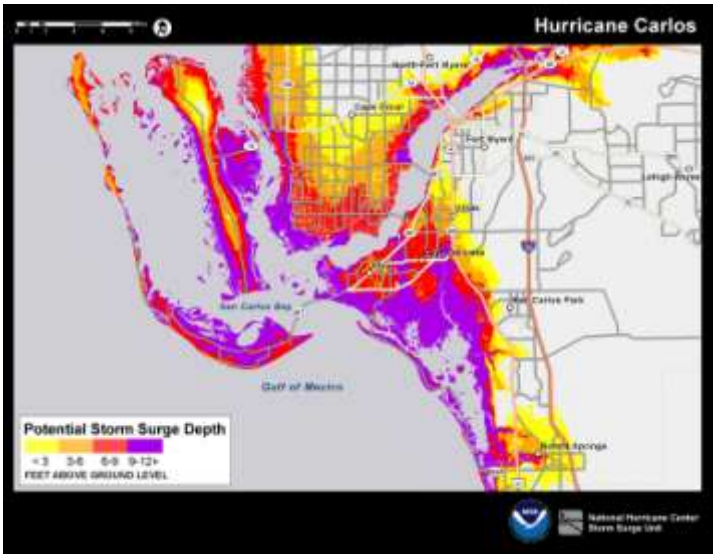
Effective Communication of Storm Surge Risk



Prepared for NOAA by:

Kathleen Sherman-Morris
Assistant Professor
Department of Geosciences
kms5@geosci.msstate.edu

Karla Antonelli
Graduate Student
Department of Psychology

Amanda Lea
Graduate Student
Department of Geosciences

Carrick Williams
Associate Professor
Department of Psychology

(all Mississippi State University)

September 28, 2012

Table of Contents

Effective Communication of Storm Surge Risk
Storm Surge Mapping Project, MSU Subaward

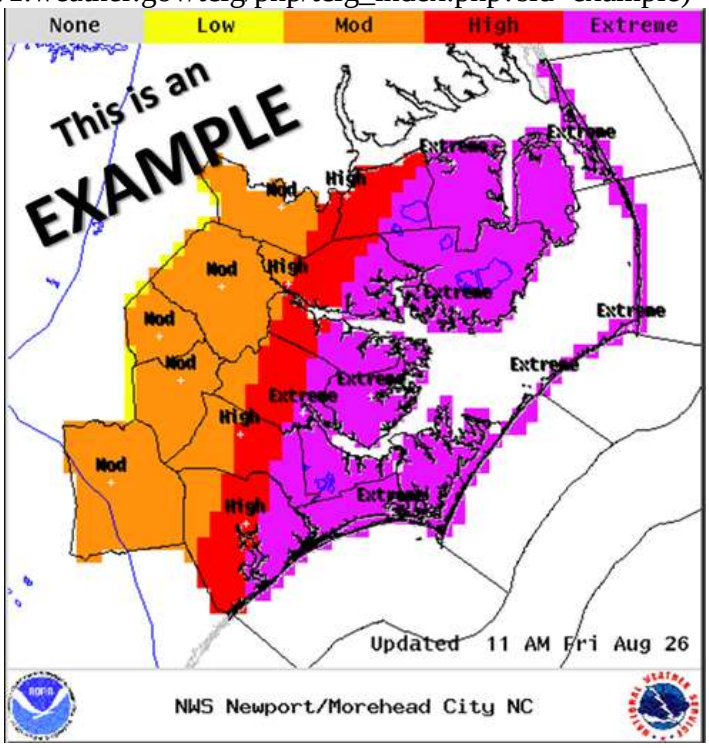
Summary

The goal of this portion of the storm surge mapping project was to test the graphics' effectiveness as measured by efficiency and accuracy in an eye-tracking experiment and in an online survey. Risk perception was also analyzed with different audiences and different graphic styles. At least one consistent pattern was detected in both methods (eye tracking experiment and online survey) and across in multiple hurricane scenarios. The green to red scale led to improved accuracy values among participants in both methods. The differences, although consistent across different tests, were not statistically significant in most cases. Qualitative categories (i.e. low, med, high and extreme) led to greater accuracy than quantitative categories measured in feet (i.e. <3, 3-6, 6-9, and 12+) in the online survey. This pattern was not repeated across all conditions in the eye tracking experiment. In that method, the scale in feet led to greater accuracy in the green-to-red scale but lower accuracy in the yellow-to-purple scale. Participants in the eye tracking experiment spent more time examining the maps when the legend was presented in feet rather than in text categories. The blue colored maps also led to increased examination times for all questions except one comparing two points to determine which was the most severe. The accuracy results from eye tracking were not significant for either color or legend type.

Background

The Alabama and Mississippi Gulf Coast is a region that experiences severe tropical weather regularly. Tropical cyclones bringing wind, rain, and surge and the associated forecasts issued should be occurrences to which residents of the area are accustomed. NOAA posts several types of forecasts for wind probabilities and severity, rain amounts, and surge potential and height. The current method for communicating the risk of storm surge is done with probabilities for certain heights. An interactive map is available on the National Hurricane Center's forecast website for every foot of storm surge probability from 2 to 25 feet. One storm surge height can be viewed at a time. On this map, only one color scheme (10 categories) is used ranging from green to yellow to red to purple; green at the lowest probability and purple at the highest probability. The color-coded probabilities are then placed along the coast line to indicate beaches with the highest probabilities of storm surge. Another color palette which is being used experimentally by National Weather Service offices is the following (Figure 1), which utilizes the yellow to purple scale. That color palette provided the baseline for testing the influence of changing colors on risk perception, efficiency and accuracy.

Figure 1. Sample wind impact product (Downloaded from http://w1.weather.gov/tcig/php/tcig_index.php?sid=example)



When designing a map, the author should consider that certain strong conventions exist in cartography related to color. One convention is the use of different hues for qualitative differences and varying a single hue's luminance, or lightness/darkness for quantitative differences (Brewer, 1994). Related to this is the common notion that "light is less—dark is more" (Garlandini and Fabrikant, 2009, p. 195; Mersey, 1990). The purpose for which the map is designed should be the determining factor in which type of color scales used. Breslow et al (2009), also confirmed research in this area such that multicolored scales are best used for identification tasks while brightness scales are more useful for comparison tasks. When used to communicate risk, other mapping conventions are important. Red is linked with warnings in Western culture (Hoffman et al., 1993; Monmonier, 1991). Other colors such as orange or blue may not have strong connotations one way or another. For example, 57.9% made at least one error in an experiment in ranking the colors in the Department of Homeland Security advisory system. The most frequent error made was in ranking blue/green and yellow/orange (Mayhorn et al., 2004). One's ability to obtain the correct information from a map also varies by individual factors. For example, making accurate inferences from a map is highly dependent upon a viewer's knowledge about the task at hand (Allen et al, 2006, Canham & Hegarty, 2010, Hegarty et al, 2010).

Online survey of MS/AL coastal zip codes--Procedure

To aide in the improvement of storm surge risk communication by the NHC and NWS to the public, an online survey was developed to assess public perceptions and their congruence with intended levels of risk for different mapping methods of storm surge

potential and intensity. Surveys were linked through Survey Monkey and tested the following question: Does one style graphic perform better in the following characteristics: 1) comprehension (accuracy of map-reading), 2) high(low) risk areas are perceived as high(low) risk, 3) higher risk areas are associated with higher perceived risk and intention to evacuate or some other relevant behavioral intent and 4) neither graphic should show a bias in the direction of errors made (i.e. does one graphic lead to an under or over estimation of risk?). Residents were presented with one graphic along with a scenario to estimate their risk, the risk to others and a behavioral intent. The graphics presented the same information (within a storm scenario), but varied in color. Two multicolored scales and one brightness scale were chosen: 1) green (lowest), yellow, orange, red (highest), 2) light blue (lowest) to dark blue (highest), 3) yellow (lowest), orange, red, purple (highest). Each group had four categories. The second manipulation was to the way surge height probability was displayed, either in text descriptions (low, med, high, extreme) or in feet (<3, 3-6, 6-9, and 12+). The legend labels were associated with the colors described above.

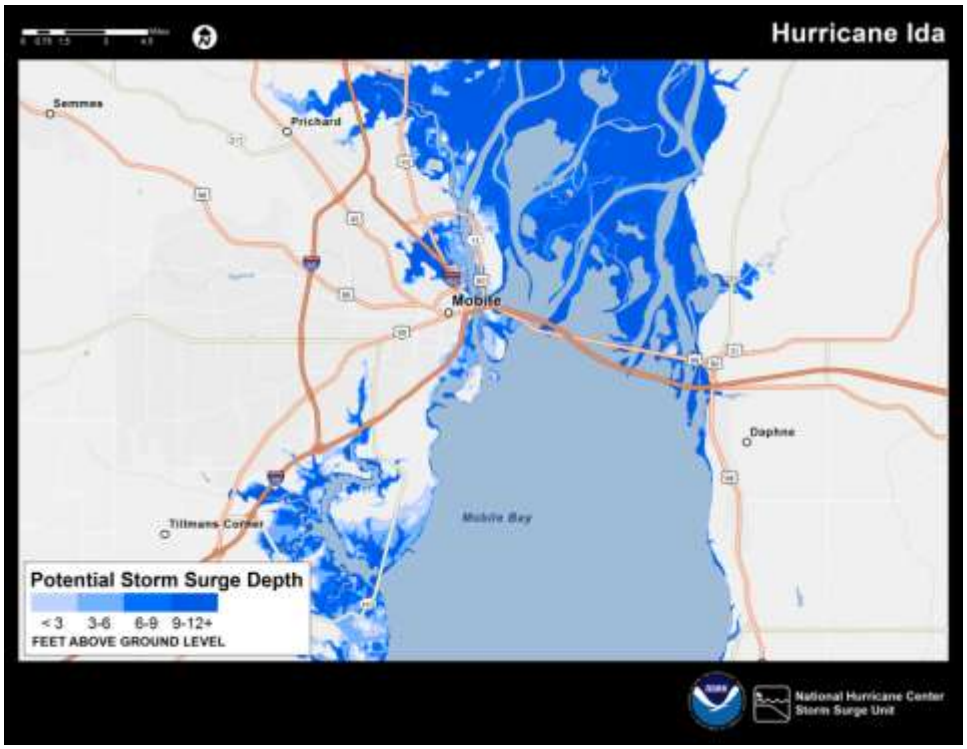
A total of 30 images were created, 2 hurricane strengths at 3 locations with 5 images apiece. Storm surge potential was depicted for a strong hurricane and for a weaker hurricane. Then the sets were divided among three areas along the Gulf Coast. Harrison County, MS, Jackson County, MS, and Baldwin and Mobile Counties, AL were the focus areas of the images. Each set graphics for a location had 5 images. The 5 images were green to red scale with text categories, green to red with numeric categories, yellow to purple with text categories, yellow to purple with numeric categories, and a blue monochromatic scale with numeric categories. Due to too small sample sizes, the Mississippi coastal counties will not be included in this report. Figure 2 a through e are examples of the "weaker" hurricane scenario, Hurricane Ida and Figure 3 a through e represent the "stronger" Hurricane Irma. While the hurricanes used to generate the images were different strengths, the resultant images were not very much different for the Alabama survey. Because there were no statistically significant differences between the two hurricane scenarios, responses for the two were collapsed in the analysis.

An email list, with 14000 names, from infoUSA was purchased to obtain a random sample of Alabama and Mississippi gulf coast residents in selected coastal counties to survey. Carrie Duncan of WLOX in Biloxi, MS and John Nodar of WKRG in Mobile, AL were asked to help recruit respondents by publicizing the survey through any of their means of communication which include, but were not limited to, television, radio, twitter, facebook, a blog, and the station website. The recruitment email and weather broadcasters directed respondents to a university domain website where residents of the gulf coast could consent to take the survey. Once there, respondents could choose one of three surveys depending on where they lived. Anyone living outside of four counties was advised to choose the county closest to where they live if they still chose to participate. The target response rate for the online survey was 750, 250 in each sample.

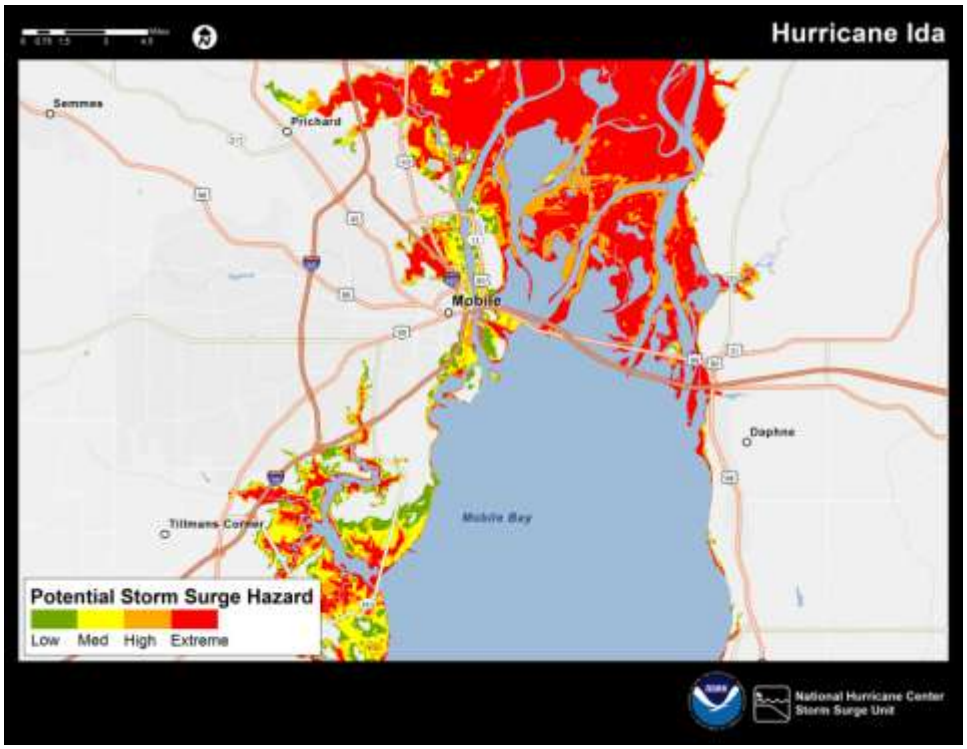
The survey itself focused on two scenarios. In scenario 1, the weaker hurricane, one of five images was shown to the respondent. Then a series of questions followed that evaluated the respondent's perceived risk from and comprehension of the provided graphic. Scenario 2, the stronger hurricane, with questions immediately followed in similar style. Previous experience with hurricanes, use of media for weather information, and demographic information were also requested of respondents.

Figure 2A-E. Hurricane “Ida” images used to represent the weaker condition.

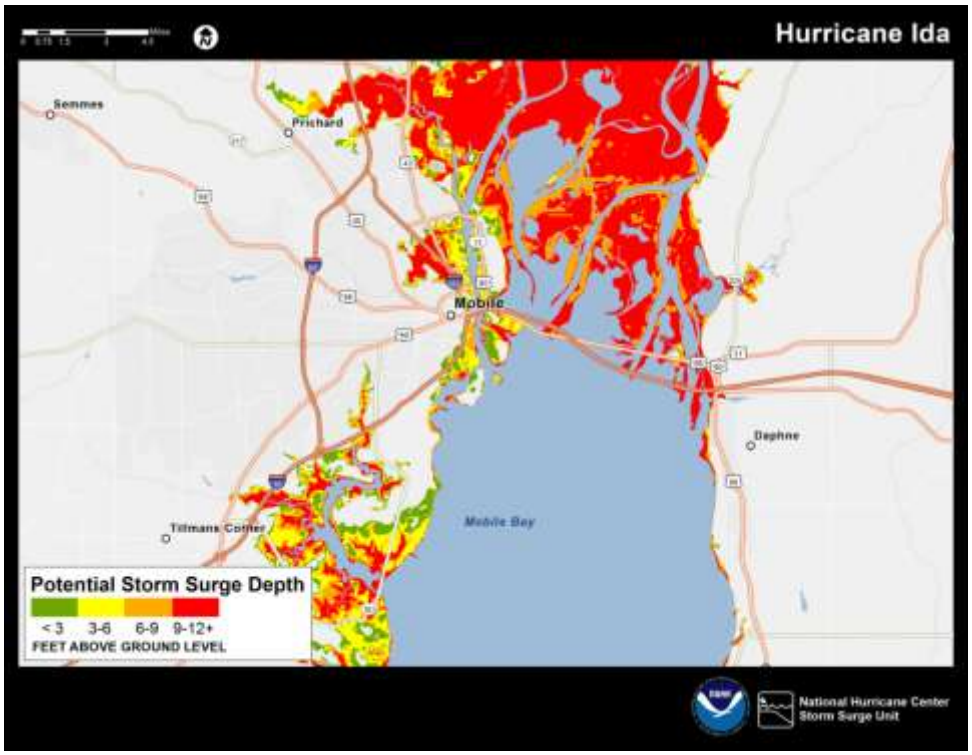
A.



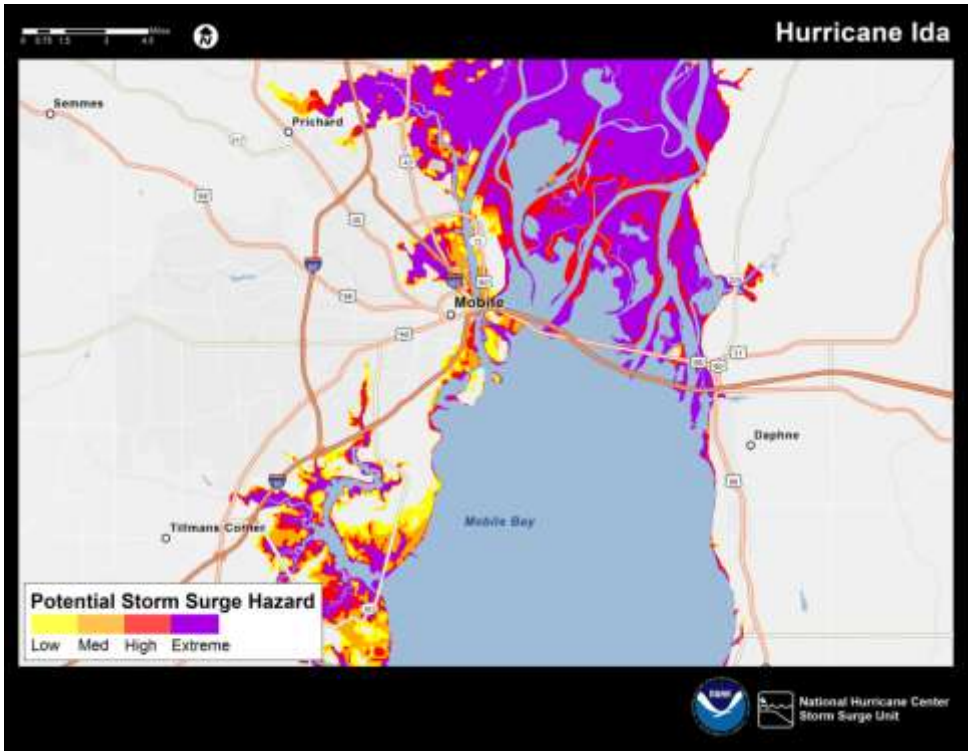
B.



C.


Hurricane Ida
Semmes
Prichard
Mobile
Daphne
Mobile Bay
Tillmans Corner
Potential Storm Surge Depth
< 3 3-6 6-9 9-12+
FEET ABOVE GROUND LEVEL
National Hurricane Center
Storm Surge Unit

D.


Hurricane Ida
Semmes
Prichard
Mobile
Daphne
Mobile Bay
Tillmans Corner
Potential Storm Surge Hazard
Low Med High Extreme
National Hurricane Center
Storm Surge Unit

E.

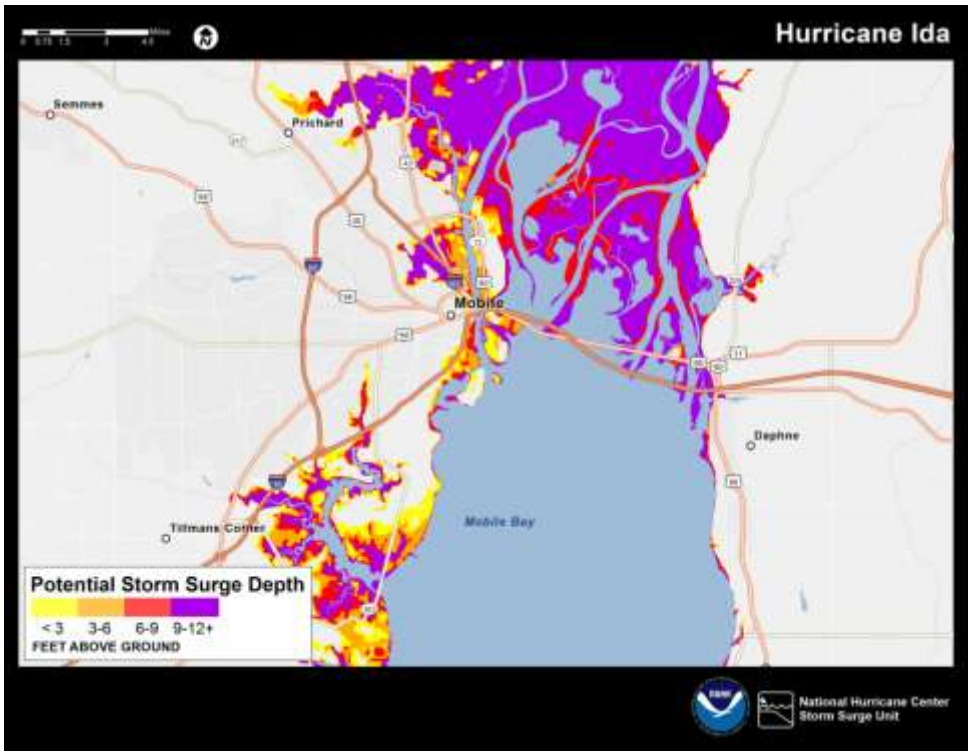
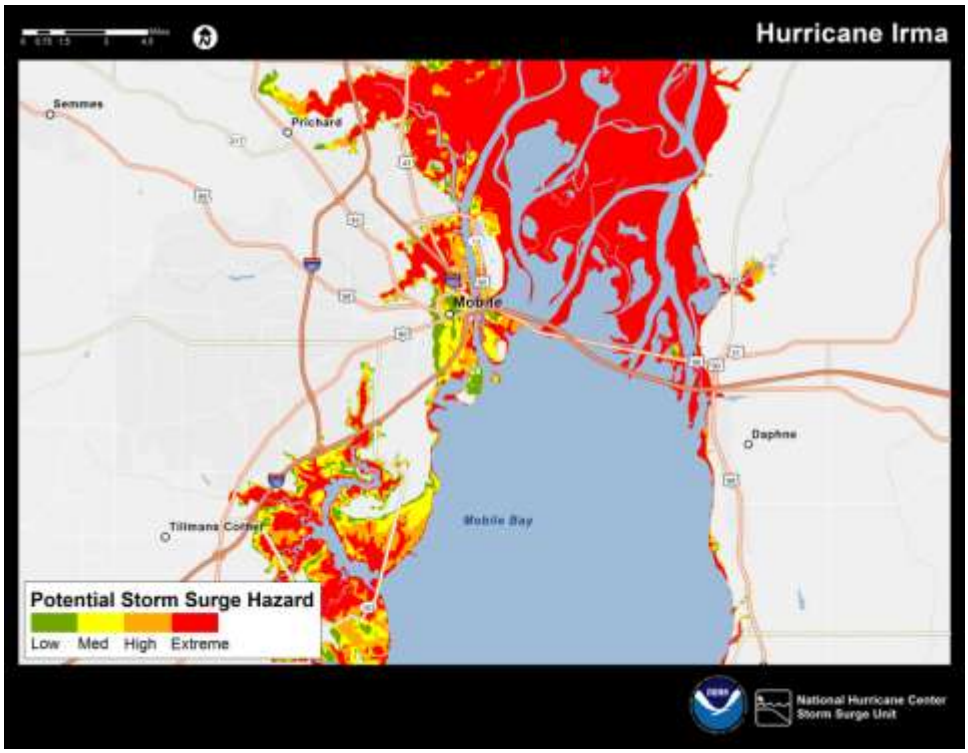


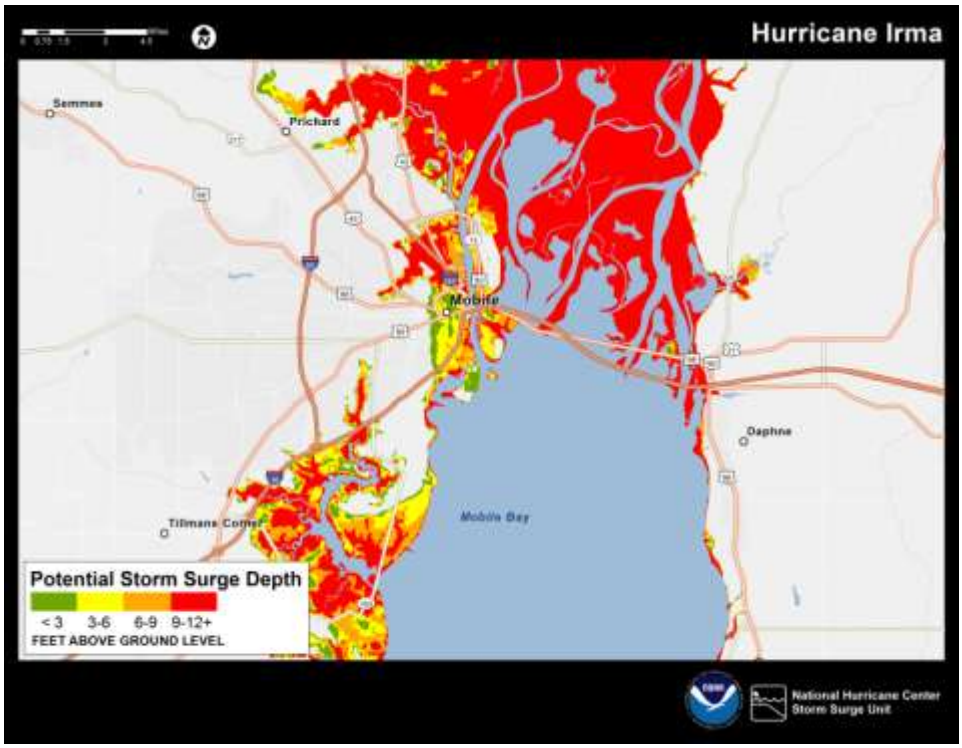
Figure 3A-E. Hurricane “Irma” images used to represent the stronger condition.

A.

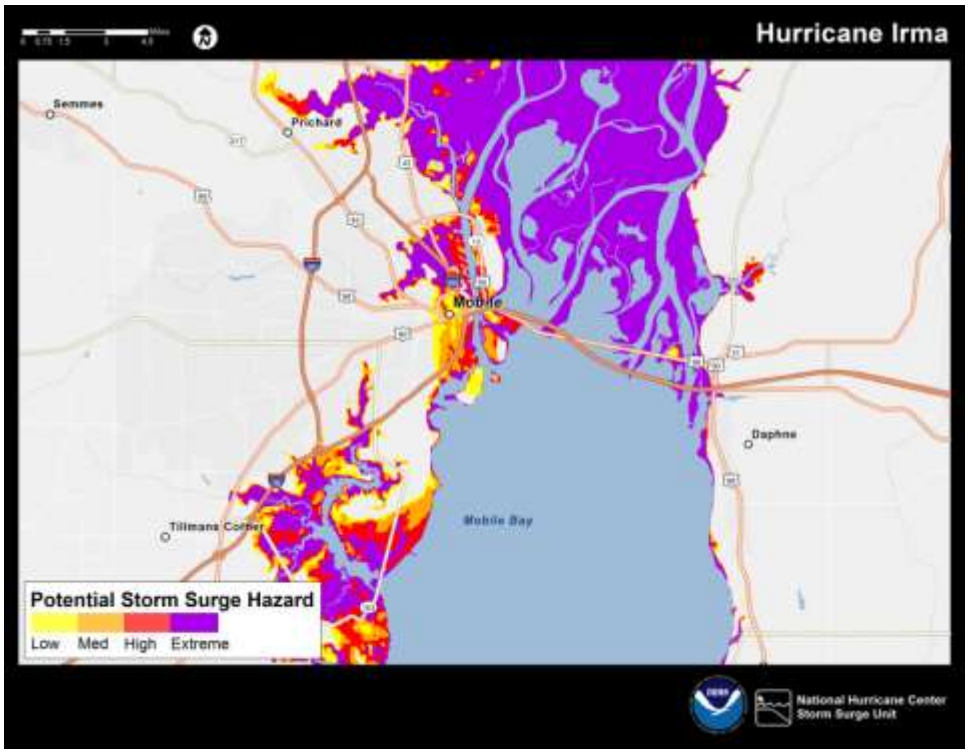
B.


Hurricane Irma
Semmes
Prichard
Mobile
Daphne
Tillmans Corner
Mobile Bay
Potential Storm Surge Hazard
Low Med High Extreme
National Hurricane Center
Storm Surge Unit

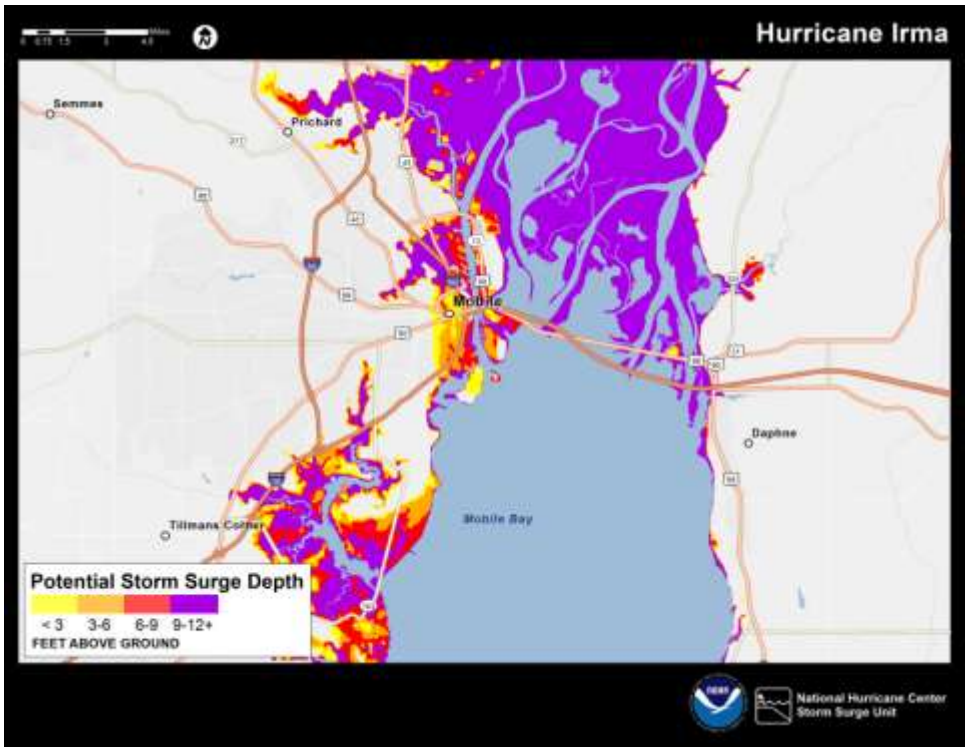
C.


Hurricane Irma
Semmes
Prichard
Mobile
Daphne
Tillmans Corner
Mobile Bay
Potential Storm Surge Depth
< 3 3-6 6-9 9-12+
FEET ABOVE GROUND LEVEL
National Hurricane Center
Storm Surge Unit

D.


Hurricane Irma
Semmes
Prichard
Mobile
Daphne
Tillmans Corner
Mobile Bay
Potential Storm Surge Hazard
Low Med High Extreme
National Hurricane Center
Storm Surge Unit

E.


Hurricane Irma
Semmes
Prichard
Mobile
Daphne
Tillmans Corner
Mobile Bay
Potential Storm Surge Depth
< 3 3-6 6-9 9-12+
FEET ABOVE GROUND
National Hurricane Center
Storm Surge Unit

Usability analysis of selected maps in MSU eye-tracking lab—Procedure

The eye-tracking study was predominately intended to evaluate comprehension and usability of the images, but also included a significant portion of questions about risk. In the eye tracking portion, participants saw all tested map presentations, whereas the online survey only showed each participant one map per scenario. Each participant viewed all five image conditions presented in varying order.

Participants. For the eye-tracking portion of the research, 40 people were selected to participate. The initial goal was 10 meteorology students, 10 non-meteorology students, and 20 members of the public. The final breakdown of the subjects was 9 meteorology experts (junior/senior students or faculty), 11 students from departments other than meteorology, and 20 members of the Starkville community. Due to a computer malfunction, one community participant's reaction time and accuracy data were lost. Flyers were posted in several university buildings and around town briefly explaining the research and need for participants. For their participation, participants were given a gift card valued at $20. The experiment lasted approximately 1 hour. Each participant was monitored during 6 map tasks in which they evaluated the relative risk of different areas in the map. Accuracy (how well they do on a task) and efficiency (how quickly they complete a task) was compared among different students and different images. The eye tracker showed which areas were attended to the longest and the pattern of movement around the map. Participants were 19-55 years of age; 21 male, 19 female.

Design and materials. The experiment conducted was a within-subjects design with five levels. Participants viewed images of storm surge predictions using three color presentation modes (blue, green-red, or yellow-purple) and two legend types (storm surge values in feet or warning categories). The blue color condition was only presented with storm surge values in feet. The resulting five conditions are presented in Figure 4.

A total of five different hurricane prediction graphs were provided by the National Hurricane Center and were used as stimuli. Each storm was associated with a different area of coastline, and thus there was no overlap in the geography of the maps that participants could use to aid in their performance across trials. Each participant saw each of the five storms in one of the five color-legend conditions. The combination of storm and condition along with the order of presentation was counterbalanced across participants.

Images of the storm were displayed on a 19 inch CRT monitor. At the distance of 71cm, the monitor subtended 28.5° x 21°, and the map of the storm subtended 23.3° x 14.9°. For each image a question/statement was displayed at the bottom of the screen that queried a specific element about the map. A total of 8 questions (see Table 1 for text of the questions) were presented for each storm to the participants for a total of 40 questions. Prior to each series of questions, a map containing only the coastline of the area was displayed to familiarize the participant with the area of coastline that the storm questions would address.

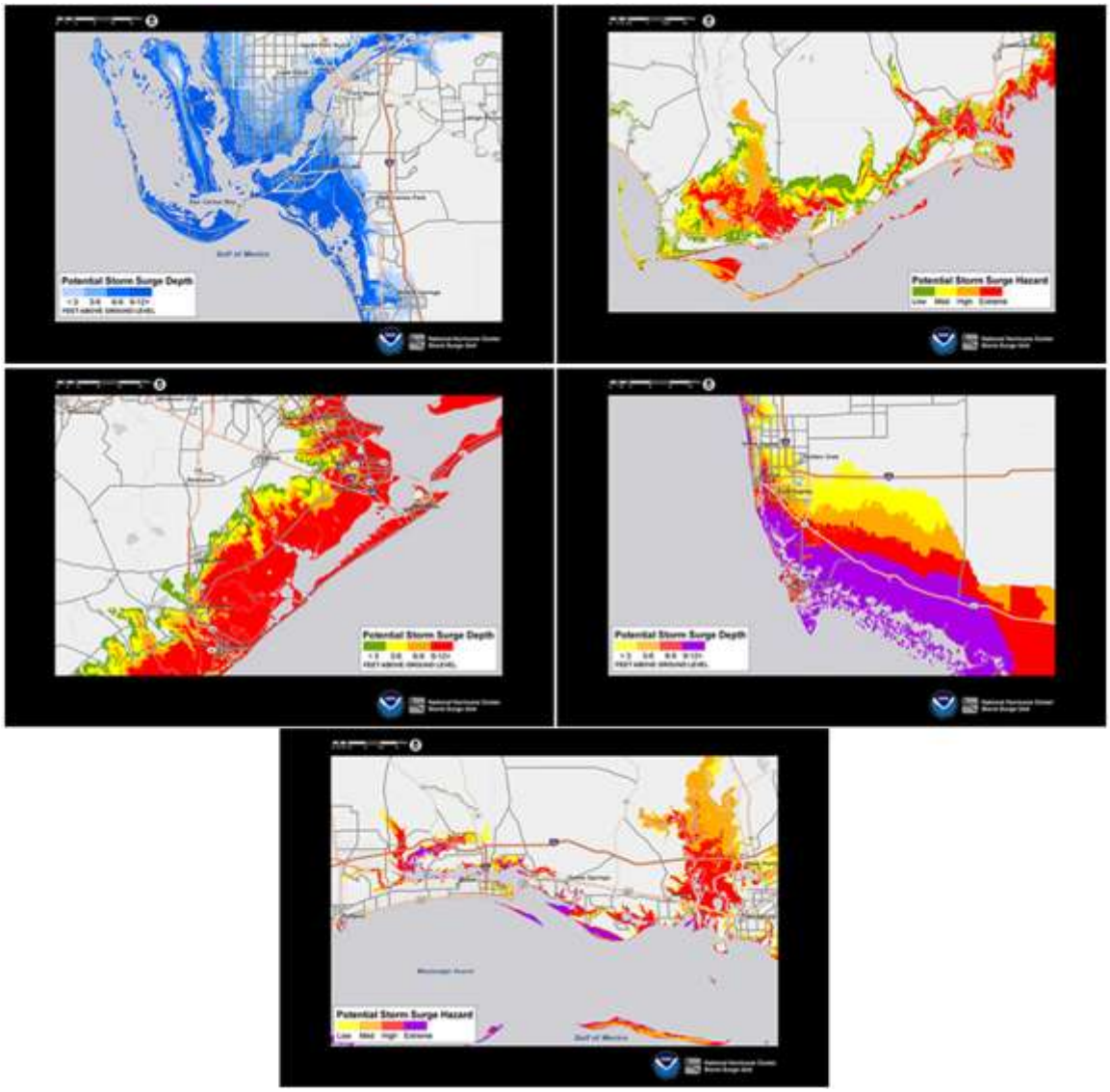


Figure 4 Each of 5 storms in each of 5 color and legend conditions.

Question 1	Which color do you believe is associated with the highest and worst storm surge?
Question 2	In which location is the storm surge forecast to be higher?
Question 3	What is the storm surge at point A forecast to be?
Question 4	Do you think a property located right at point A would experience storm surge flooding?
Question 5	If you lived in a single level house or on the ground level of an apartment building at point A, would you take any precautions to prevent damage to your home or belongings?
Question 6	If you lived in a single level house or on the ground level of an apartment building at point B, would you take any precautions to prevent damage to your home or belongings?

Question 7	On a scale from 1 to 8, where 1 is not bad at all and 8 is very bad, how would you rate this hurricane based on its storm surge potential?
Question 8	On a scale from 1 to 8, where 1 is not helpful at all and 8 is very helpful, how helpful do you think this image is in your ability to judge the storm surge risk associated with this hurricane?

Based on the question, a location was identified on the map via a letter (A or B) and a black arrow. These questions had participants make judgments about the severity of the storm surge at that location.

In addition to the main sequence of storm maps, participants were presented with two graphs that separately depicted each of the three color and two legend conditions (see Figure X) and asked participants to choose which they felt did the best job of providing storm surge prediction information. The presentation arrangement of condition images within the overall image (e.g., right or left, top or bottom row) was counterbalanced across participants.

Finally, participants were presented with graphs depicting hurricane wind damage predictions in one of three conditions based on where the legend was placed on the figure (top of graph, bottom of graph, or within the map area). The same image was presented four times, each with a different question that queried specific information from the map for a total of 4 questions.

Procedure. Prior to the experiment, participants completed an informed consent form and a general demographic form. Following that portion, participants were seated in front of the computer/eye tracker. To minimize head movements and maintain a consistent distance from the monitor, a chin rest was used. The participant's chair was adjusted so that he or she would be comfortable. The experimenter explained the procedure again and then began the calibration sequence for the eye tracker. Participants were instructed to fixate 5 locations on the monitor while the eye tracker was calibrated. Calibration was monitored throughout the experiment and checked between trials. If judged necessary, calibration was performed again.

Questions about individual storms were presented in a sequence of 9 images (the coastline only, followed by the eight questions about the storm surge predictions). The questions were presented in the same order for all participants and for each storm. Participants input their response via a standard keyboard. All of the answers were coded as a number and the participant was instructed to rest his or her hand on the number pad of the keyboard to facilitate input.

Following all of the storm questions, an additional sequence of the 4 images and questions on wind damage predictions and 2 images and questions about best depiction of storm surge risk were presented to participants.

Apparatus and analysis. Eye tracking was performed with an ISCAN ETL-400 eye tracker operating at 120 Hz. The system is accurate to less than 1° of visual angle. Eye tracking data were fed into a computer running E-prime 1.2 software (Schneider, Eschman, & Zuccolotto, 2002), which interpreted the data in coordinates consistent with the display.

For all of the eye movement analyses described here, fixations were evaluated off line and were defined as a series of consecutive samples that were all within 8 pixels (as in Williams, Henderson, & Zacks, 2005). The minimum and maximum fixation durations

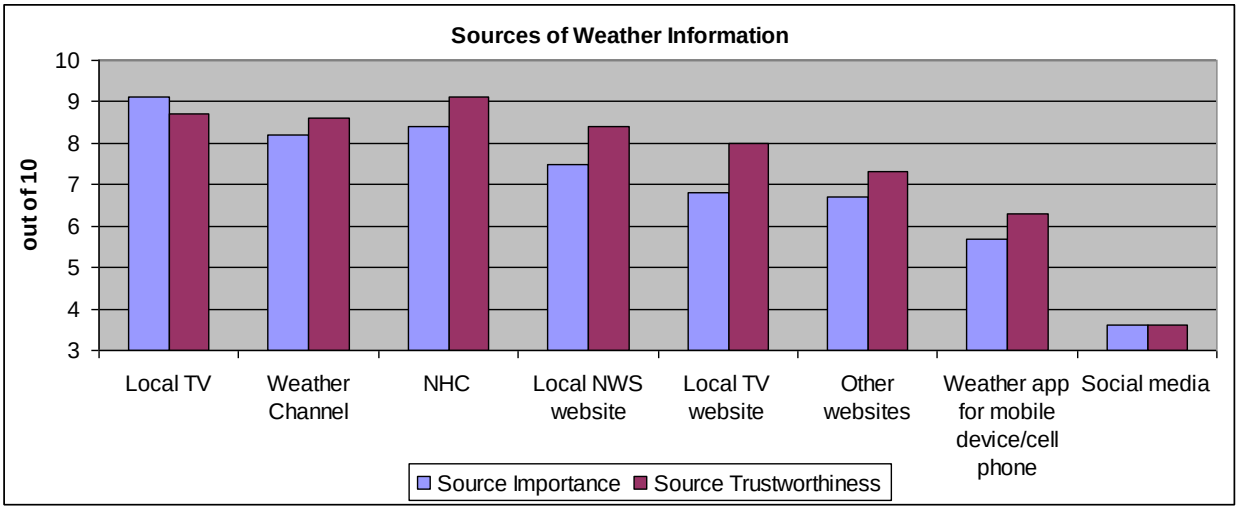
were set to 90 ms and 4000 ms respectively. Regions of interest were defined as the smallest rectangle that could enclose the critical area (Map, Scale, Logo, Question) or the smallest rectangle plus 10 pixels for the smaller regions of interest (Legend and Letter Markers). A fixation was assigned to a region if it fell within the boundary of the region.

Online Survey Results

A total of 129 surveys were submitted from Mobile and Baldwin counties. Fifty-two (40%) women and 55 (43%) men responded to the survey. (The rest did not identify their gender.) The sample consisted of people who identified predominately with the white race (76%). Most also were highly educated with 72% having at least some college experience. The average age of respondents was just above 55 and ranged from 19 to 84 years old. Single family, site built homes were the most common type of living arrangement (n=99, 76%). Most respondents also owned their home (n=100, 78%). Of the 129, 72 (56%) learned of the survey through the email sent out via infoUSA, 30 (23%) from a local TV meteorologist, and 4 (3%) from another person.

It seems that respondents are relatively weather-aware in that 74% obtain a forecast at least once per day, 3/4 of those multiple times per day; and 68% "pay very good attention" to forecasts when a hurricane threatens their area. Local television, the Weather Channel, the NHC, and the local NWS website were ranked the top four most important and trustworthy sources for hurricane forecasts (Figure 5). The most importance source was not the same as the most trustworthy source, however. The NHC was deemed more trustworthy than local TV, even though the local station was most important to respondents. Weather applications for mobile devices, while arguably grouped with the other sources by score, was the second least important and trustworthy source. Social media were deemed neither important nor trustworthy for hurricane forecasts with both scores below 4 on the 10 point scale.

Figure 5. Sources of weather information



When asked about previous preparatory actions taken for a hurricane, checking forecasts more often was the most common response (Figure 6). The more inconvenient the preparatory action, the less it was taken among respondents. This could be because respondents believe that they will be safe in their own home for the most likely of hurricane situations. Over 80% of respondents indicated that they would feel safe in their own home in wind speeds of 90 mph, the high end of category 1, occurred. As wind speed increased, that feeling of safety quickly faded and respondents became increasingly less confident in their home's structural integrity (Figure 7). This confidence, or lack of confidence depending on the strength of the hurricane, may be a result of just over half of the sample having experienced moderate to severe damage to their home.

Few of the respondents' homes (34, 26%) are elevated for potential flooding. This may be expected in that only 19 (15%) respondents had ever experienced flooding, seven (5%) of which had experienced damage in the same house in which they currently reside. Four (3%) reported minor damage, seven (5%) reported moderate damage, and eight (6%) reported severe damage.

Figure 6. Hurricane preparations made in the past.

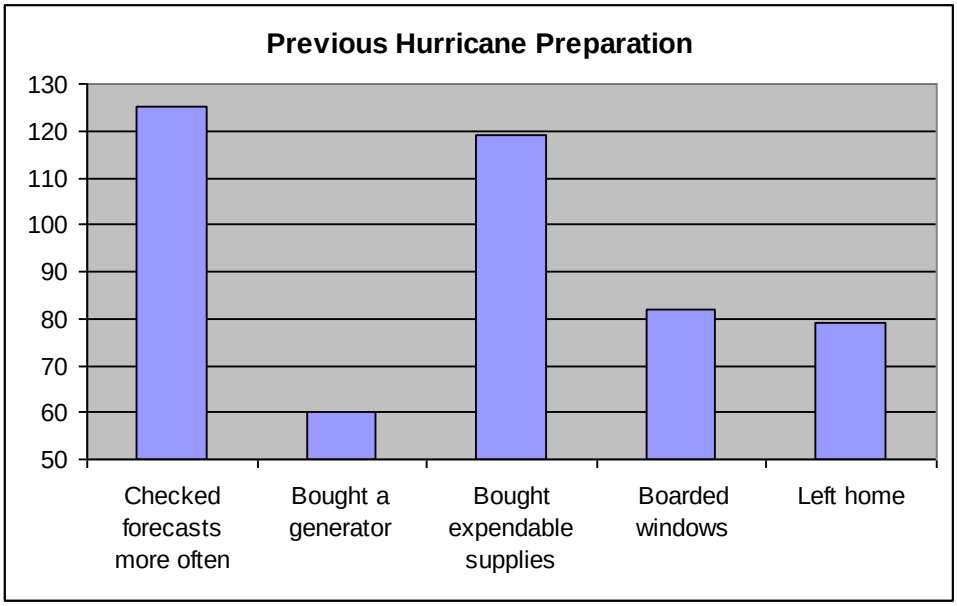
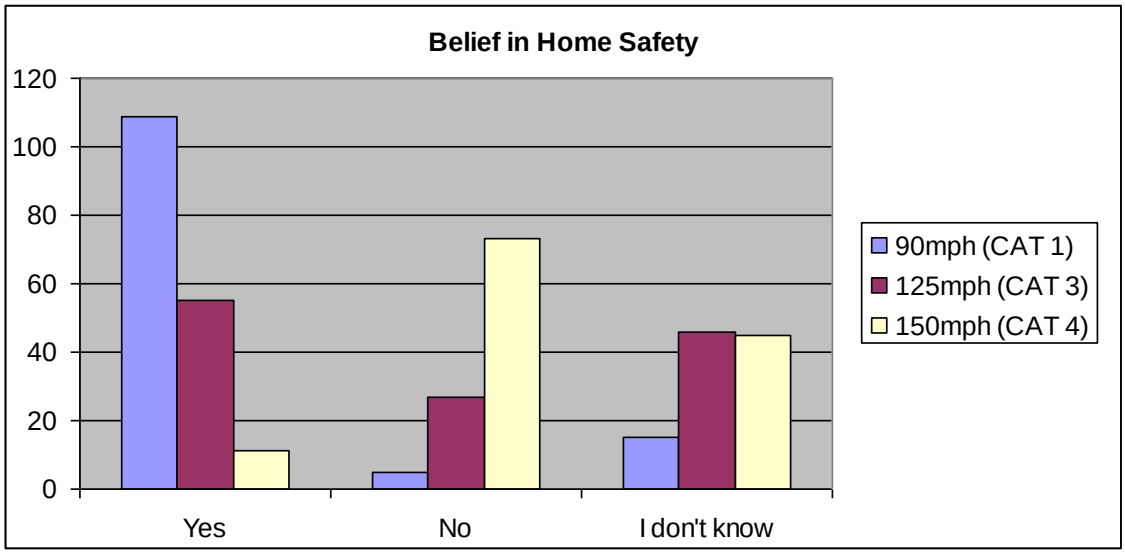


Figure 7. Perception of home safety as a function of hurricane wind strength.



Two images of storm surge forecasts, a weaker hurricane (Ida) and a stronger hurricane (Irma), were shown to respondents. The distribution of the hurricane images seen among respondents (Table 1 & 2) was relatively even, although fewer people continued participating through the survey from the first image to the second image. Data from the online survey did not indicate significant differences among the five image types, although the sample size (129) was too small to expect anything but large effects.

Table 1 Weak Hurricane Image Distribution

		Legend		
		Category	Feet	Total
Color	Red	*18*	*24*	**42**
	Blue	*0*	*31*	**31**
	Purple	*28*	*16*	**44**
	Total	**46**	**71**	**117**

Table 2 Strong Hurricane Image Distribution

		Legend		
		Category	Feet	Total
Color	Red	*25*	*17*	**42**
	Blue	*0*	*19*	**19**
	Purple	*28*	*21*	**49**
	Total	**53**	**57**	**110**

Risk perception scores fairly closely followed the statistical normal curve for both hurricane images (Figure 8). Yellow to purple images were rated more 'bad' on that rating scale than the other categories and green to red scales were rated slightly more helpful. Overall, very few differences in risk perception were found to be significant. Figures 9-12 show differences in a combined risk perception variable among the five images and among the two storm scenarios (Hurricanes Ida and Irma). While not significant, the risk perception values were higher for the yellow to purple scale with the descriptive text legend and green to red scale with legend in feet.

Figure 8. Risk perception scores based on 5 variables.

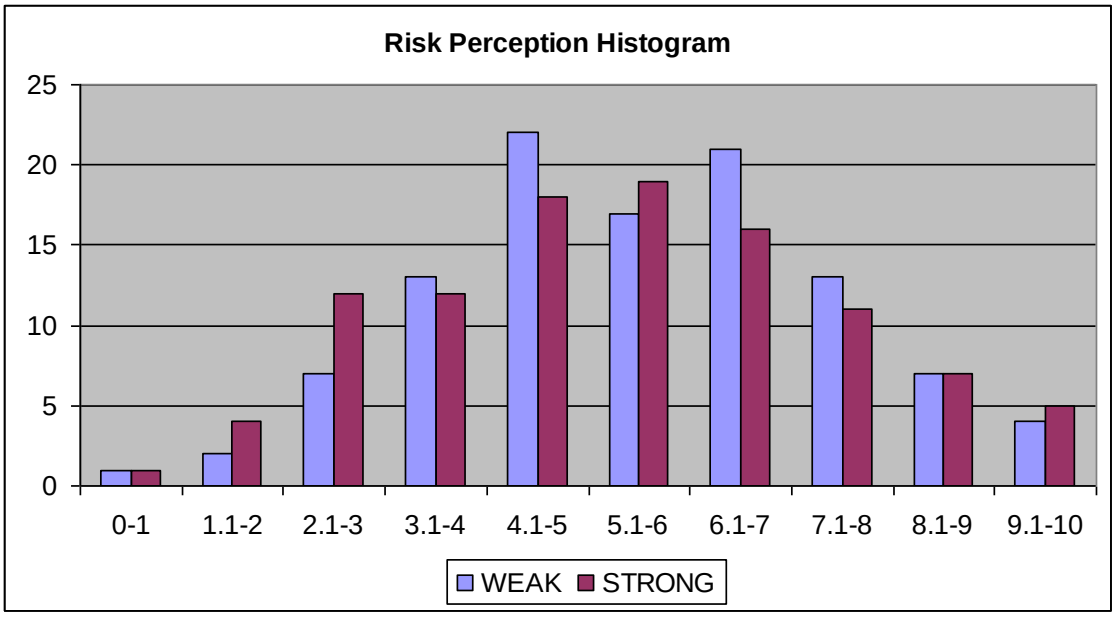


Figure 9. Risk perception scores by image type and storm category

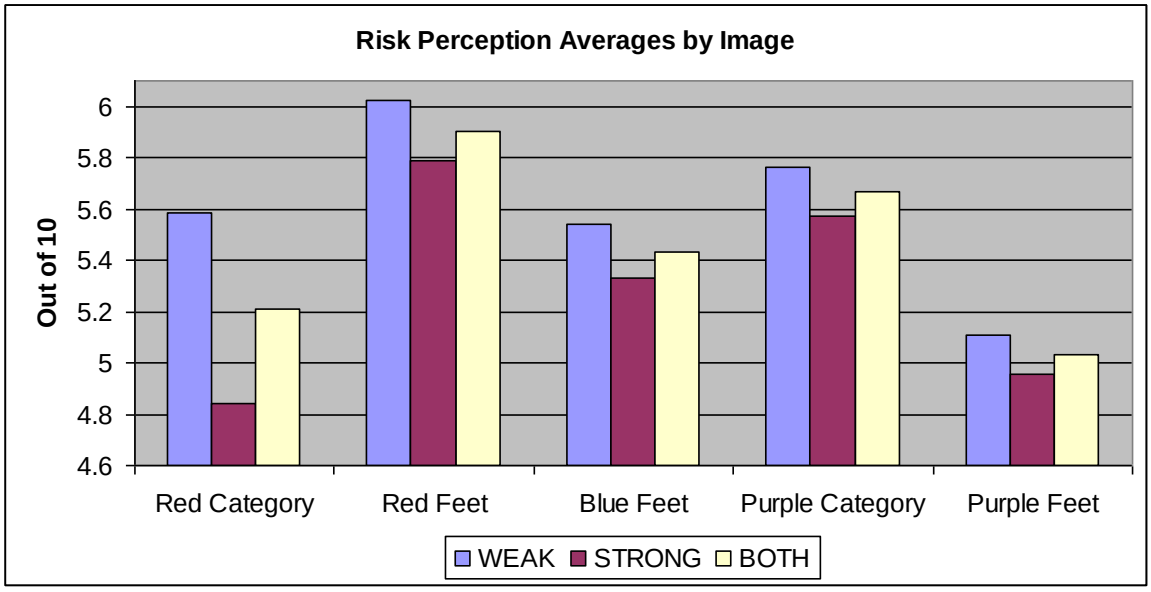


Figure 10. Risk perception scores separated by storm strength. (Hurricane Ida was supposed to be the weaker storm).

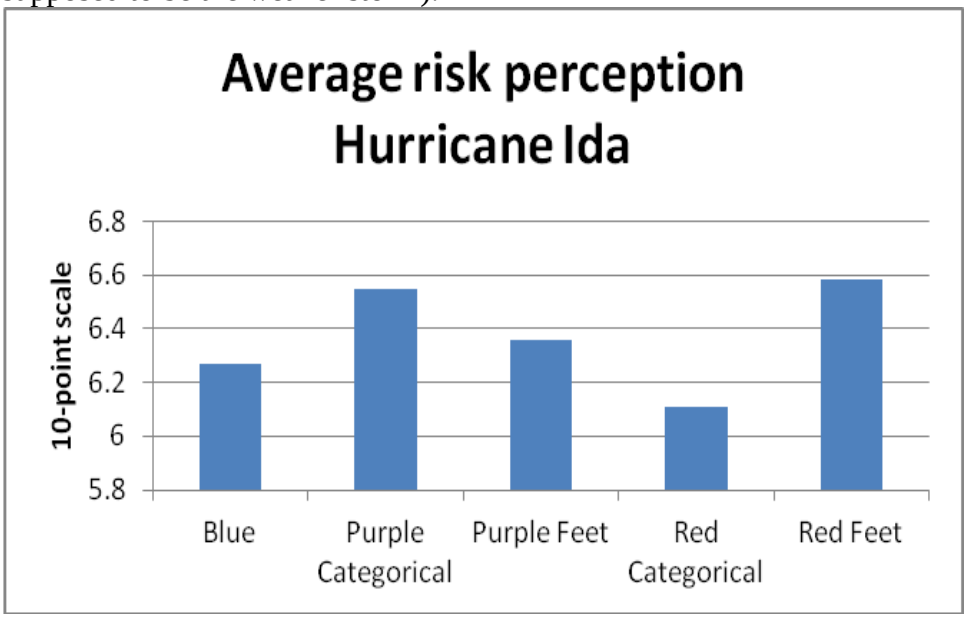


Figure 11. Risk perception scores separated by storm strength. (Hurricane Irma was supposed to be the stronger storm).

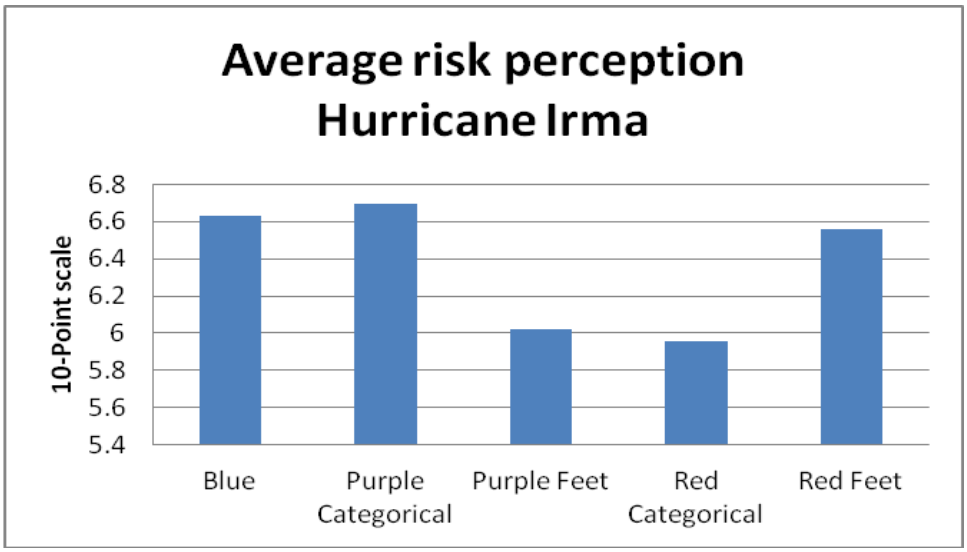
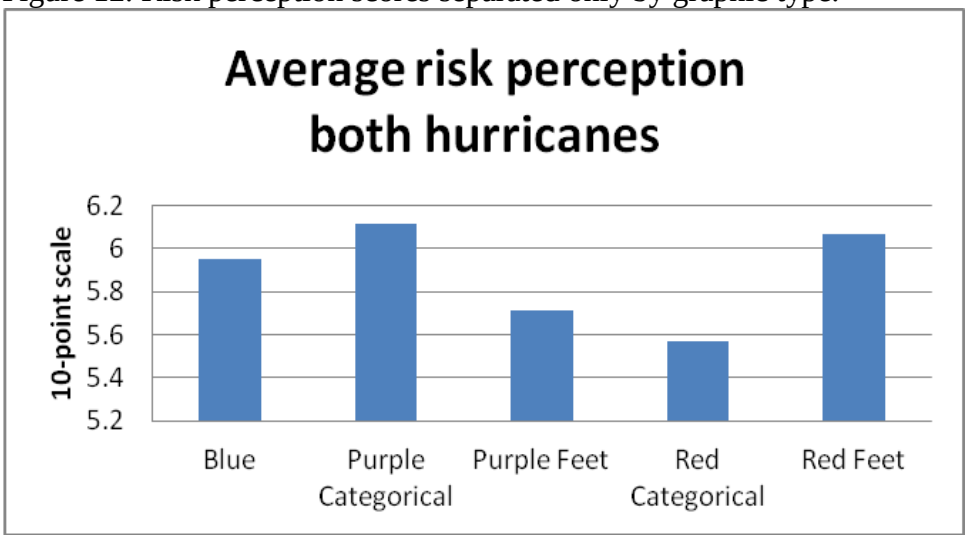


Figure 12. Risk perception scores separated only by graphic type.



Accuracy of worse storm surge potential between two locations did reveal a slightly different result. When accuracy from both hurricanes were combined, the difference in accuracy was marginally significant ($\chi^2 = 3.77$, $p = 0.052$) with a likelihood ratio of 3.87 ($p = .049$). Accuracy did not significantly differ among the three color palettes ($\chi^2 = 4.174$, $df = 2$, $p = 0.124$). However, the legend that displayed feet resulted in less accurate comparisons than the categorical legend (Table 4). The green to red color scheme resulted in more accurate comparisons than the other two palettes, thus the purple feet legend had the least accurate results and the red categorical legend had the most accurate results.

Table 3. Accuracy by hurricane and legend type

	Category	Feet
Ida	87.0%	75.0%
Irma	85.0%	75.0%
Combined	85.7%	75.2%

Table 4. Accuracy by color palette and legend type

	Category	Feet
Red	88.4%	85.0%
Blue		73.5%
Purple	83.6%	66.7%

Since the green-red spread had the best comparison results (86.7% accurate), the accuracy means from the other two palettes were combined (75.7% accurate) and tested against the former. This test yielded $\chi^2 = 3.94$, df = 1, p = 0.047, meaning that respondents viewing the green-red color palette performed significantly better when comparing storm surge potential between two locations than the other two color palettes combined. To check that this result was not an artifact of the sample size, the yellow-purple scheme accuracy (81.8%) was tested against the other two (76.9%). This test yielded $\chi^2 = 0.80$, p = 0.37. Therefore, better performance from the green-red color palette was not an artifact of the sample size and is, in fact, a real effect of the color palette chosen for the map display.

Results of the online survey are suggestive of differences in accuracy based on map color palette and type of scale depicted. That is, green-to-red images led to more accurate results than the other two images and qualitative, text-based scales resulted in more accurate responses. Few differences were statistically significant, however. There were no significant differences in risk perception among color palettes, between legends and between storm strengths. The survey may not be generalizable to the public, but may adequately measure a similar demographic as individuals who would actively seek information from the NHC online during a hurricane. A greater sample size is needed for analysis of the Mississippi coastal counties.

Eye Tracking Experiment Results

All analyses were performed using SPSS 20 software using repeated measures ANOVAs. All analyses were performed with the within subjects factor of color-legend condition (Blue feet-values, Green-Red category-text, Green-Red feet-values, Yellow-Purple category-text, Yellow-Purple feet-values) and the between subject factors of participant experience (Meteorological Faculty/ Student – Experts, Undergraduate Student, Community Member) and storm order list. Task irrelevant information on the

map distracts viewers from the task at hand regardless of a viewer's subject knowledge (Canham & Hegarty, 2010), however the influence of map design on one's ability to accurately interpret a map is mediated by better domain knowledge and understanding of the subject matter (Hegarty et al, 2010). That was why it was important to distinguish participants by their level of experience. For all analyses reported here a significance level of .05 was assumed except where marginal effects are described. Greenhouse-Geisser corrections were applied when violations of sphericity were found. Finally, partial eta squared is provided as a measure of effect size.

The dependent variables are discussed separately below. **Question accuracy** is an indication of the ability of the participants to accurately judge the answer to the question in the different color-legend conditions. **Response time** is a global measure of processing effort of a particular question or judgment. As will be seen, for some questions response times in the different color-legend combinations were elevated. The final section of dependent variables addresses the **eye movement** data. Fixation patterns on the images give an indication of what was attended in the image. Because the locations that are fixated have been actively attended, the use of eye movements can be informative about deployment of attention during processing an image (e.g., Deubel & Schneider, 1996; Henderson, 1993; Hoffman & Subramaniam, 1995; see Rayner, 2009 for a recent review of the use of eye movements). The measure of interest in the current study are the proportion of fixations during a trial that fell on a particular region of interest. In this case, proportions were used because the response times were highly variable. The proportion measure accommodates the differences in response time.

Question Accuracy and Response Times.

Participant accuracy averaged collapsed across Question 1-4 are presented in Table 5. Questions 1-4 all had objectively correct answers, whereas the other questions asked for a judgment. Overall, accuracy was fairly high for all groups (all conditions were greater than 84% accurate). There was no overall effect of color-legend condition, F (4, 68) = 1.60, p = .18, η_p^2 (effect size) = .086. In addition, there was no statistical difference between participant experience groups (Expertise), F (2, 17) = 2.145, p = .15, η_p^2 = .202, and expertise did not interact with color-legend condition, F (8, 68) = 0.99, p >.20, η_p^2 = .104.

Table 5. Mean accuracy for Expertise group by Color-Legend condition.

	Blue Values	Green-Red Text	Green-Red Values	Yellow-Purple Text	Yellow-Purple Values	Expertise Group Mean
Experts	94.6%	91.1%	96.4%	85.7%	94.6%	92.5%
Community	87.7%	88.9%	84.3%	84.7%	85.9%	86.3%
Undergraduates	91.0%	95.1%	94.4%	85.4%	88.2%	90.8%
Weighted Condition Means	90.8%	91.3%	90.9%	85.2%	89.3%	

Response times (see figure 13) were analyzed for Questions 1-6 separately because each question required the participant to find or address a different element of the image. Unless mentioned, the participant meteorological experience did not affect response times.

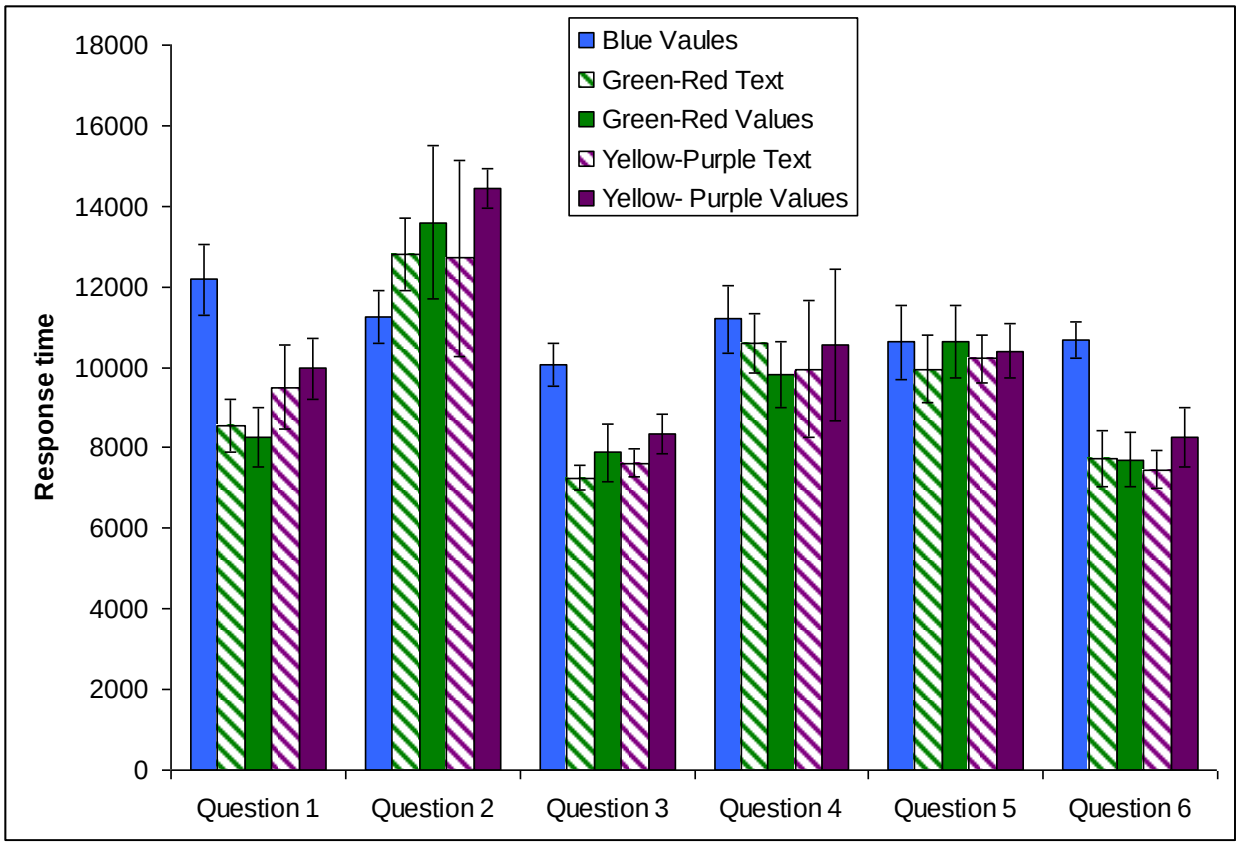


Figure 13: Mean response times for the 5 conditions for Questions 1 – 6. Error bars are the standard error of the mean.

In **Question 1**, participants determined the color of the highest storm surge of without a legend. In responses to this question, there was a significant difference, F (4, 68) = 5.94, $p < .001$, $\eta_p^2 = .259$, with the Blue color condition producing significantly longer response times than the green-red conditions (ps < .01 after Bonferroni correction), but not the yellow-purple conditions (ps > .12 after Bonferroni correction). The remaining conditions were all statistically equivalent.

For **Question 2**, participants were required to compare the severity of a storm surge at two different locations to determine which was more severe. In contrast to Question 1, there were no statistical differences in the response times based on color-legend condition, F (4, 68) = 1.35, $p > .20$, $\eta_p^2 = .073$. Although there was no overall statistical difference, it is interesting to note that for this question, the Blue values condition was responded to the quickest, likely because comparing two points on a single color dimension is easier than across color categories.

Question 3 asked participants to judge the level of storm surge at a single location. Similar to Question 1, participants did demonstrate a reliable difference in response times, F (4, 68) = 6.11, $p < .001$, $\eta_p^2 = .264$, with the Blue values condition having reliably longer response times than the Green-Red and Yellow-Purple category-text conditions (ps < .01 after Bonferroni correction), and marginally longer times than the Green-Red and Yellow-Purple feet-values conditions (.05 < ps < .10 after Bonferroni correction). The Green-Red and Yellow-Purple conditions were not statistically different.

To answer **Question 4**, participants had to determine if a property at a specified location would be flooded by the storm (yes or no). As can be seen, Figure 13, there was no difference in response times based on color-legend condition, F (4, 68) = .134, $p >$ .20, $\eta_p^2 = .008$.

With respect to the **Questions 5 and 6**, participants were asked to make a judgment about whether they personally would take precautions to prevent damage to a single level house or ground floor apartment at a specific location on the image. Although the questions were identical (only the location changed), the results were different. In Question 5, no difference was found based on the color-legend condition, F (4, 68) = .151, $p > .20$, $\eta_p^2 = .009$. However, in **Question 6,** there was not only a significant difference in the response times, F (4, 68) = 6.20, $p < .001$, $\eta_p^2 = .267$, with the Blue value condition being slower than the other conditions (ps < .015 after Bonferroni correction), but there was also an interaction between the color-legend condition and participant meteorological experience, F (8, 68) = 5.38, $p < .001$, $\eta_p^2 = .388$. This interaction (See Figure 14) resulted from the fact that the Expert group (faculty/students) spent on average 5 seconds more than the Community members and 6 seconds more than the Undergraduate students examining the Blue values image. In the remaining conditions, the Expert group's response times were similar to the other groups. Because the questions were similar, there is no readily available explanation for the differences in pattern on Questions 5 and 6.

Finally, in order to evaluate if any differences exist between the Green-Red and Yellow-Purple conditions and the different legend types (categorical text or feet values), a separate analysis was performed collapsing across Questions 2 – 6 (Question 1 had no legend). The results appear in are in Figure 15. Although the Green-Red conditions were numerically faster (by approximately 200 ms) than the Yellow-Purple conditions, they were not statistically different, F (1, 17) = .232, $p > .20$, $\eta_p^2 = .013$. With respect to the legend type, Category text maps were responded to approximately 500 ms faster than the Feet values condition, but this trend was only marginally significant, F (1, 17) = 3.10, $p =$ .096, $\eta_p^2 = .154$.

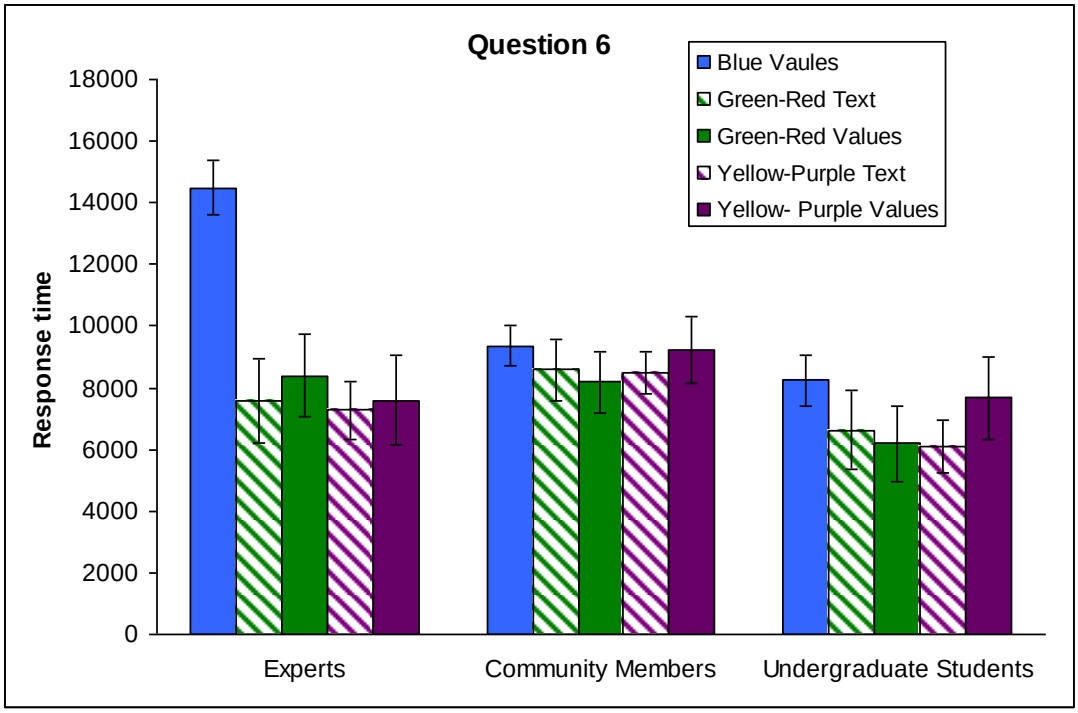


Figure 14: Mean response times for the 5 conditions for Question 6 by meteorological experience. Error bars are the standard error of the mean.

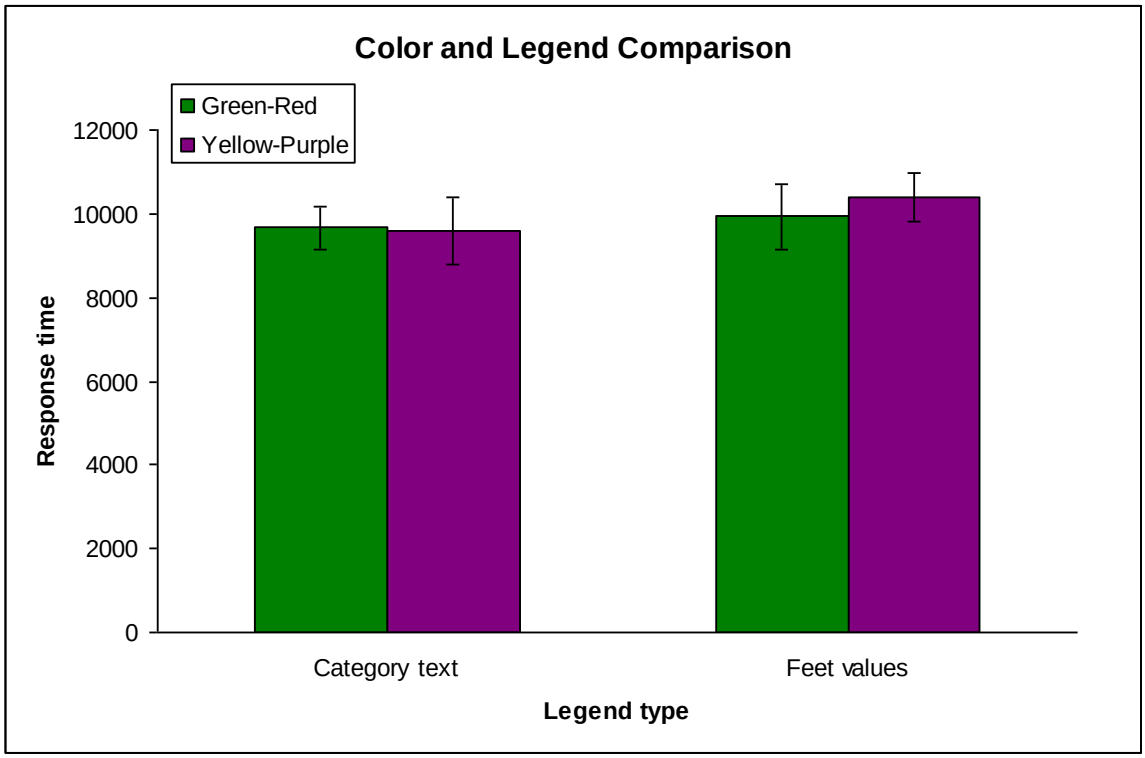


Figure 15: Mean response times for Green-Red and Yellow-Purple conditions only. Error bars are the standard error of the mean.

Response Time Discussion. Overall, the response time data indicate that Blue value condition is the most difficult to interpret in that it takes longer to answer most of the questions that were asked. The one exception (though only a numerical trend) was in the direct comparison of the severity of the surge at two separate locations in the same map. Because a single dimension was used in this judgment, it may have been easier to make the comparison directly on that dimension rather than across a color category boundary. Although it was not statistically significant and it may be of limited use in the general public who may not need to make this type of judgment often, it is an interesting trend that could point to future research questions. With respect to the other conditions, they were all similar in response times, though there was a trend that using categorical references like Extreme, High, Medium, and Low were easier to interpret than the same information in feet values. If the goal is to speed the interpretation of the danger in storm surges, maps that use color category distinctions and category warning labels appear to be better than ones that rely on actual predictions of the height of the surge at a location.

Eye tracking results. The response time data give a global measure of the processing during a trial. However, eye tracking allows for an examination of the processing of various elements during the trial. Humans make approximately three separate fixations each second. By analyzing the location of these fixations during trial, we can determine what elements of the stimuli were critical to answer the question asked. For the analyses described below, we calculated the proportion of the fixations within a single trial that were located on region of interest (the map, the legend, or the letter marker on the map). As described above, response times were variable between conditions, and thus to address the difference in time a proportion measure was created. In order to determine if fixation patterns were different for the color-legend conditions, we analyzed the proportion of fixation to the critical regions separately across Questions 2- 6 (the questions that had all 3 critical regions represented). Question was entered as an additional within subjects variable in the analysis. Because of the likelihood of violations of sphericity, Greenhouse-Geisser corrections were applied to all analyses described in this section. However, for clarity, the uncorrected degrees of freedom are listed.

In Figures 16-18, we have plotted the locations of all the fixations made on a particular image by all of the participants in each of the three meteorological experience groups. These figures depict Storm Charley in Blue Values condition, Question 4. Note that the number of participants shown in each figure is constrained by the number of participants from their defining group who saw that Storm/Condition combination. The green dots represent the location of the fixations.

Figure 16 Fixations to the map
Storm Charley in Blue Values condition, Question 4
Community, based on 5 participants

Figure 17 Fixations to the map
Storm Charley in Blue Values condition, Question 4
Meteorological, based on 1 participant

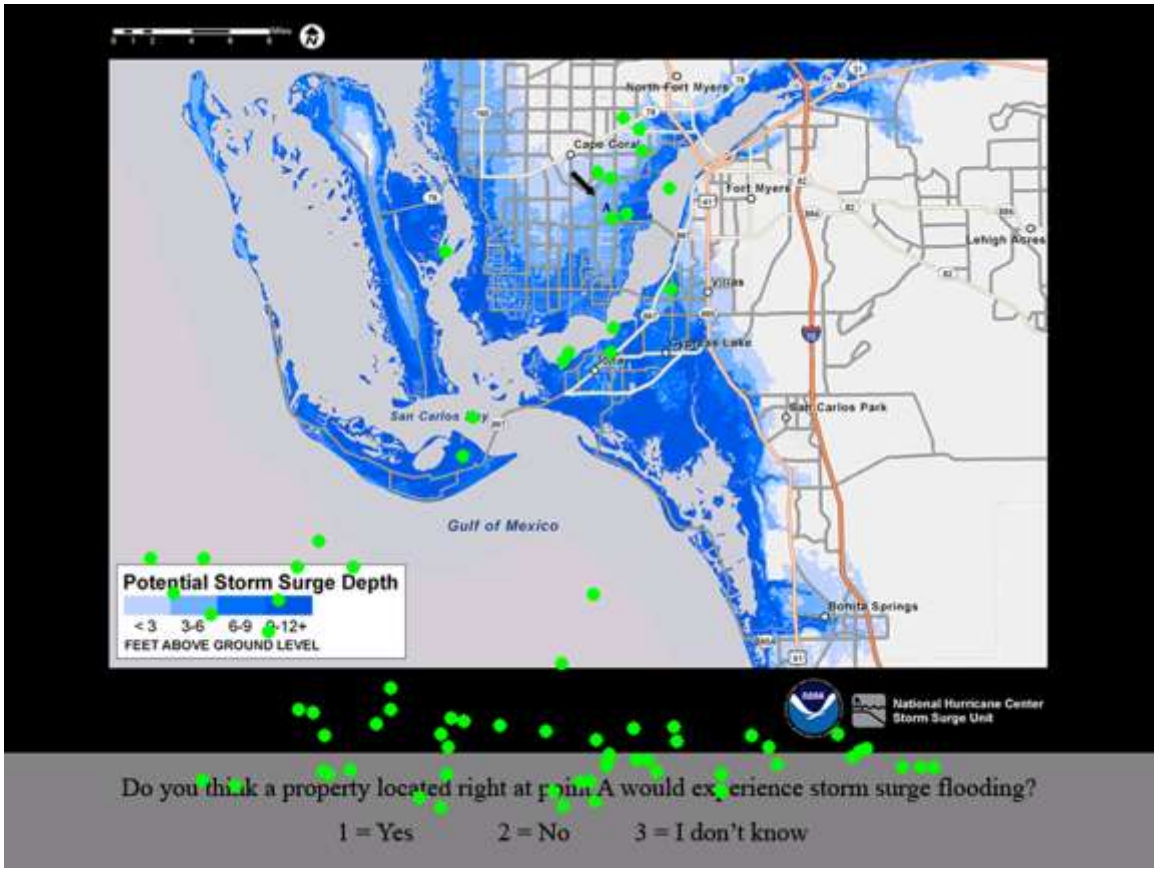


Figure 18 Fixations to the map
Storm Charley in Blue Values condition, Question 4
Student, based on 2 participants

Proportion of Fixation analyses. Figures 19-21 show the proportion of fixations that were located on the Legend, the Letter markers, and the Map area (excluding fixations on the legend and letter markers). Of critical interest is whether or not color-legend condition affected proportion of fixations to the region and if question interacted with the color-legend condition. For fixations to the legend, there was a significant effect of the color-legend condition, F (4, 72) =4.47, p = .012, η_p^2 = .199, but there was no interaction with question or with expertise. The Blue values condition had the greatest proportion of fixations to the legend (.120) followed by the Yellow-Purple values condition (.109), the Green-Red values condition (.098), and the Yellow-Purple and Green-Red category text conditions (both .095). Although there was an overall difference, when a Bonferroni correction was applied, no condition was statistical different from the other conditions.

Because Questions 2-6 all had markers placed on the map in order to evaluate the participant's understanding of the storm surge information, fixations to the Letter Marker (including the arrow) were analyzed. Once again, there was a significant effect of the color-legend condition, F (4, 72) =6.13, p = .001, η_p^2 = .254. In addition, there was a significant interaction with question number, F (16, 288) = 3.52, p = .001, η_p^2 = .163, and a three way interaction with condition, question, and expertise, F (32, 288) = 1.69, p = .049, η_p^2 = .158. As in the legend analysis, the Blue values condition (.098) had a greatest

proportion of fixations to the letter markers. The remaining conditions were all similar with

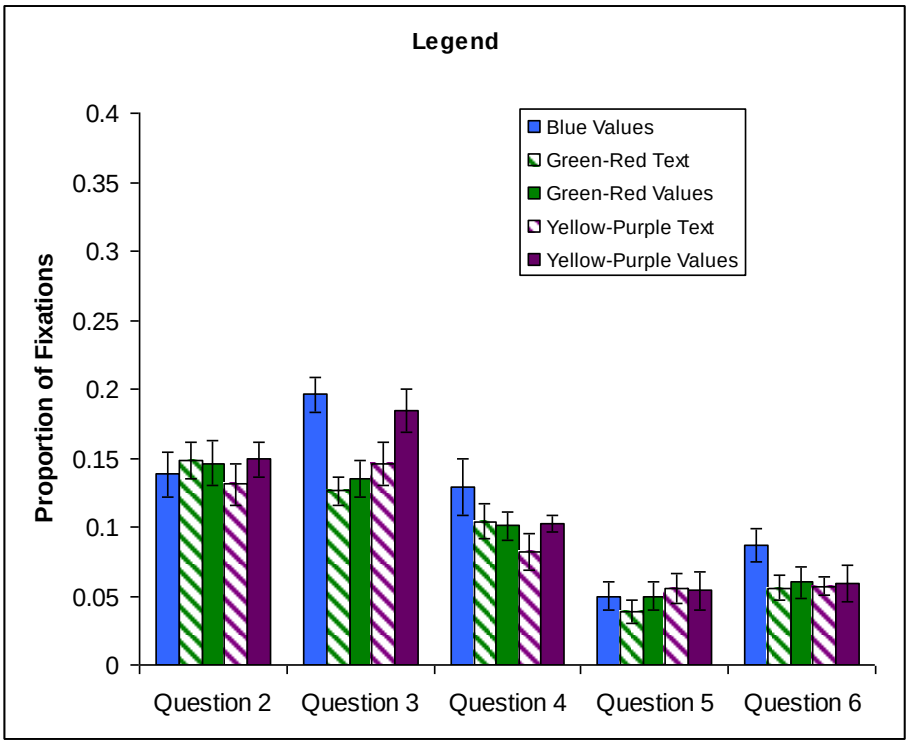


Figure 19. Proportion of Fixation to the Legend by question and color-legend condition. Error bars are the standard error of the mean.

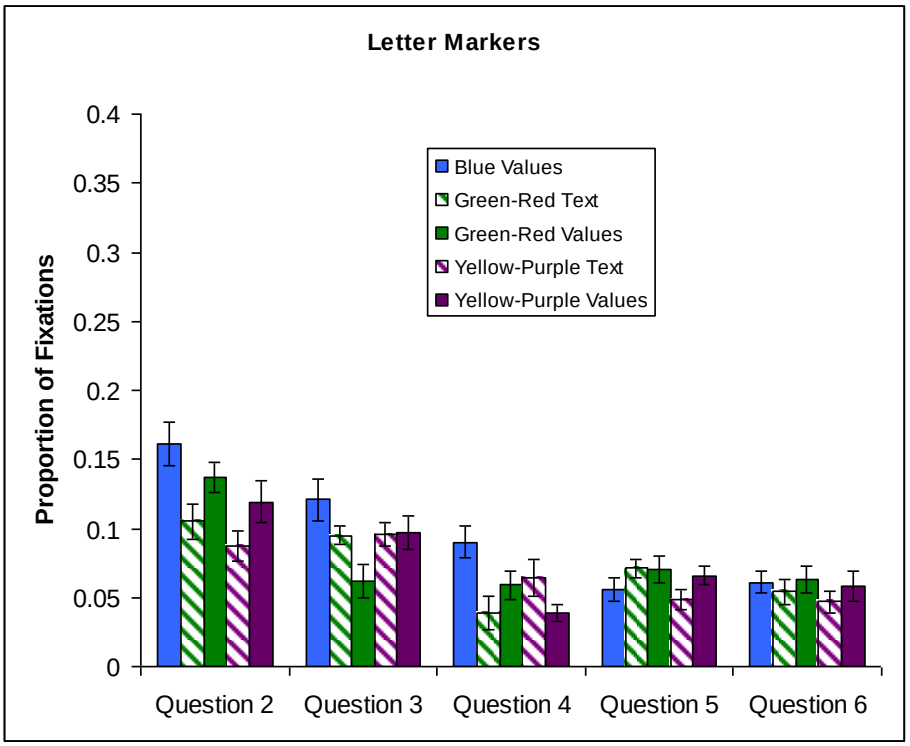


Figure 20. Proportion of Fixation to the Letter Markers on the map by question and color-legend condition. Error bars are the standard error of the mean.

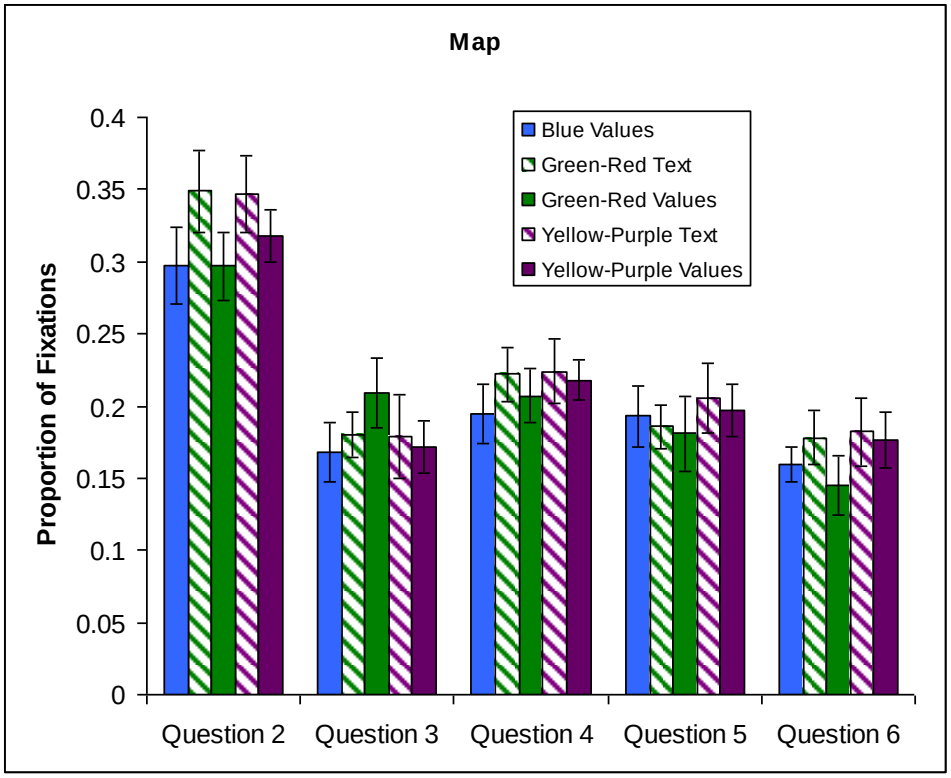


Figure 21. Proportion of Fixation to the Map area by question and color-legend condition. Error bars are the standard error of the mean.

Green-Red values (.078), Yellow-Purple values (.076), Green-Red category text (.073) and Yellow-Purple values (.068). In comparing the conditions to each other, after a Bonferroni correction, the only statistical difference was between Blue values and Green-Red category text (p = .03). The interaction between question and condition appears to result from the fact that for Questions 2-4, letter markers in the Blue values condition received a greater proportion of fixations compared to the other conditions. However in Questions 5 ad 6, this difference was not present. The three-way interaction is likely the result of the fact that community members fewer fixations to the letter markers especially in the Blue values condition compared to both the Experts and the Undergraduate students. In general, Experts (.097) had a greater proportion of fixations to the letter marker than the Undergraduate students (.073) and the Community members (.069), F (2, 18) = 4.22, p = .031, η_p^2 = .319.

The final comparison was the proportion of fixations to the map area excluding those fixations to the legend and the letter markers. These fixations are, in essence, fixations that are not directly related to answering the question. The proportions appear in Figure FF. Although there was an effect of question, F (4, 72) = 49.47, p < .001, η_p^2 = .733, color-legend condition and expertise had no effect or interaction. Thus, the non-question specific fixations were equally distributed across conditions.

Wind Map Preferences and Accuracy

Participants were asked their preferences of the map color and legend styles. Lastly, they were shown a wind potential map with legend moved in various locations on

the screen. The purpose of this test was to determine the best placement of the legend. The results show that participants overwhelmingly believed the green-to-red map did "the best job in informing the public about their storm surge risk" (Figure 22, Table 6). Preferences were nearly evenly divided between text values (Low, Med, High, Extreme) and numerical values (in feet). See Figure 23 and Table 7. The map with the legend at the top of the page led participants to perform the best across four accuracy questions. The difference was not significant.

Table 6. Number of participants responding for each color condition. (Total 39 participants).

	Blue	Green – Red	Yellow – Purple
Experts	2	5	2
Community	2	14	3
Undergraduates	1	7	3
Total	5	26	8

Figure 22. Maps for Question 1: Which of these maps do you think does the best job of informing the public about their storm surge risk?

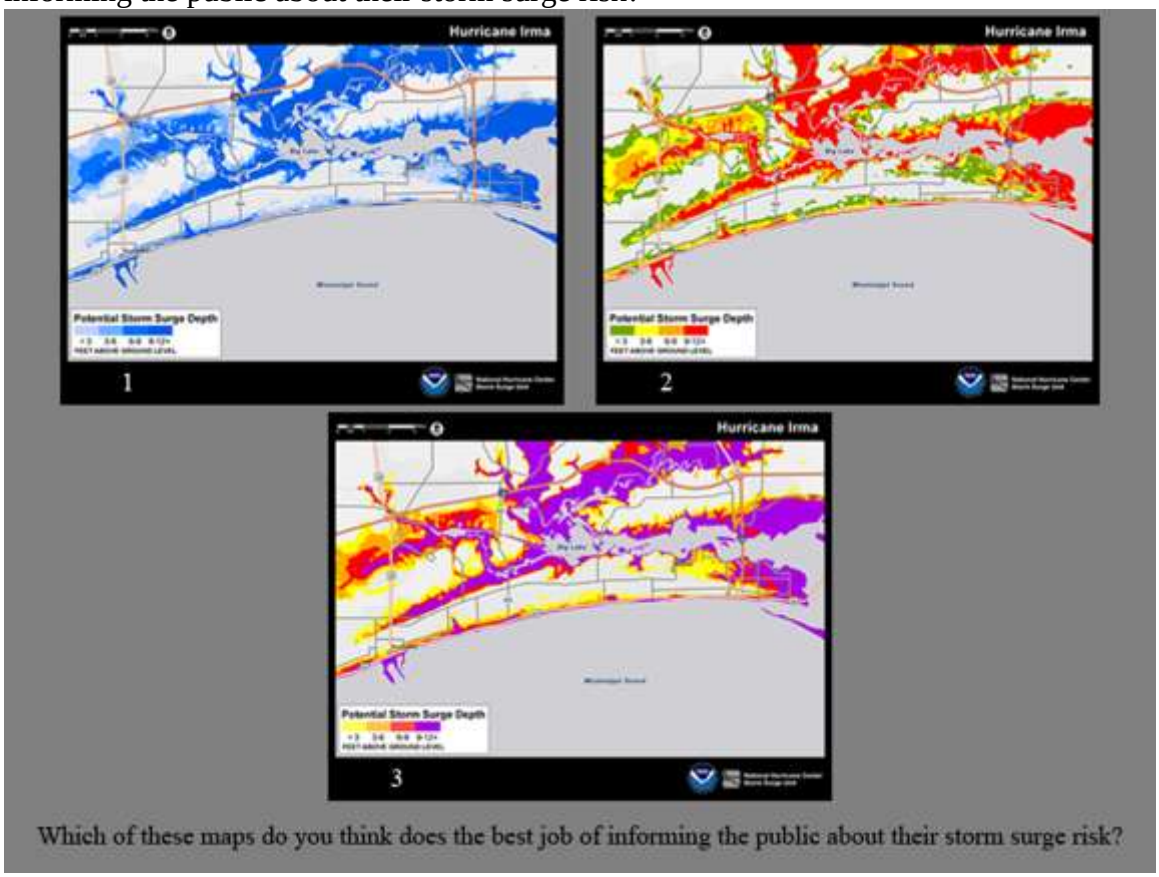


Table 7. Number of participants responding for each color condition. (Total 40 participants).

	Values	Text
Experts	7	2
Community	9	11
Undergraduates	5	6
Total	21	19

Figure 23. Maps for Question 2: Which of these maps do you think does the best job of informing the public about their storm surge risk?

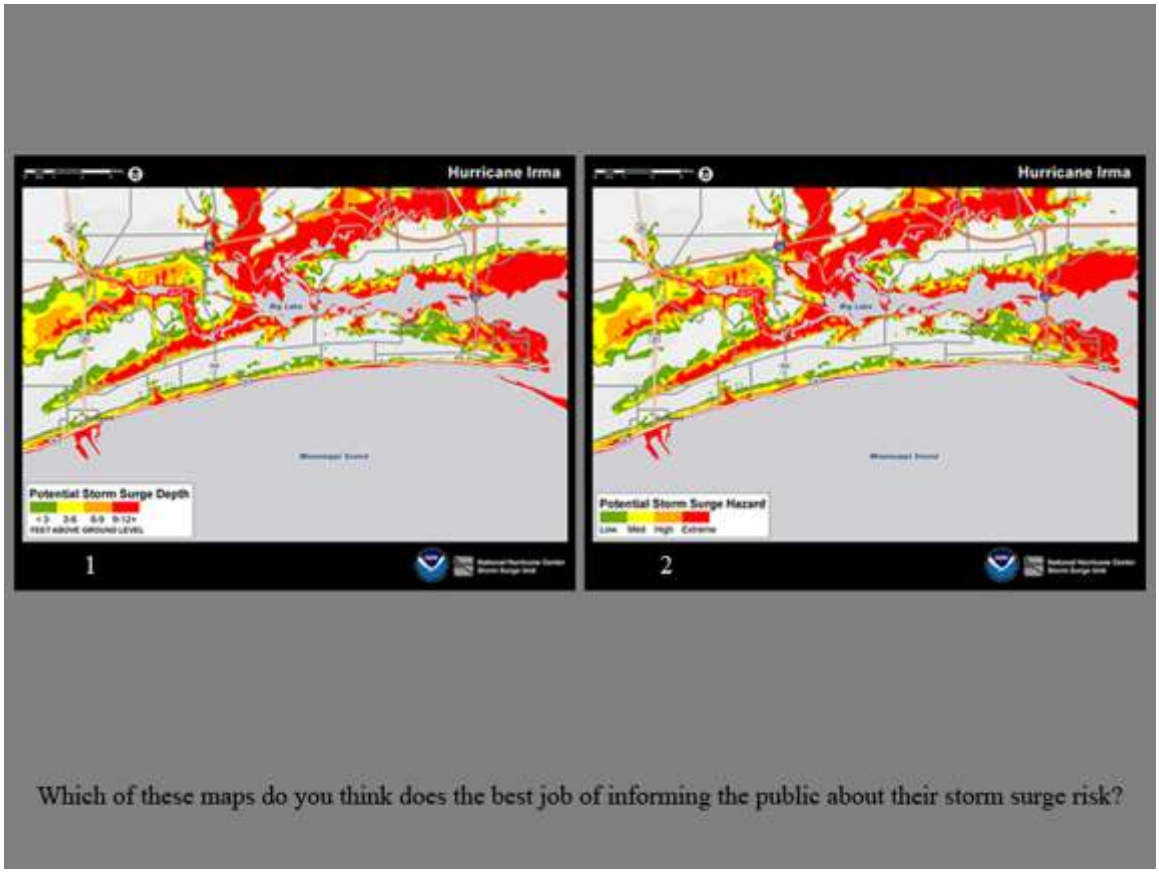
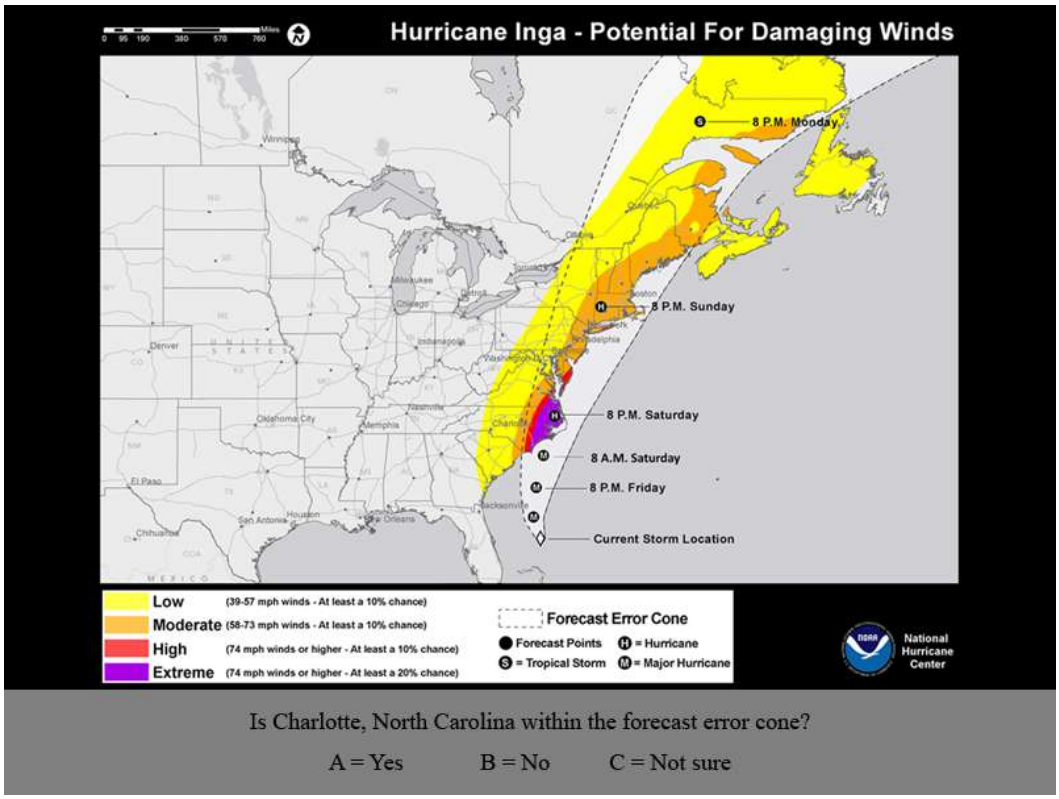


Table 8. Wind damage questions accuracy, collapsed across questions:

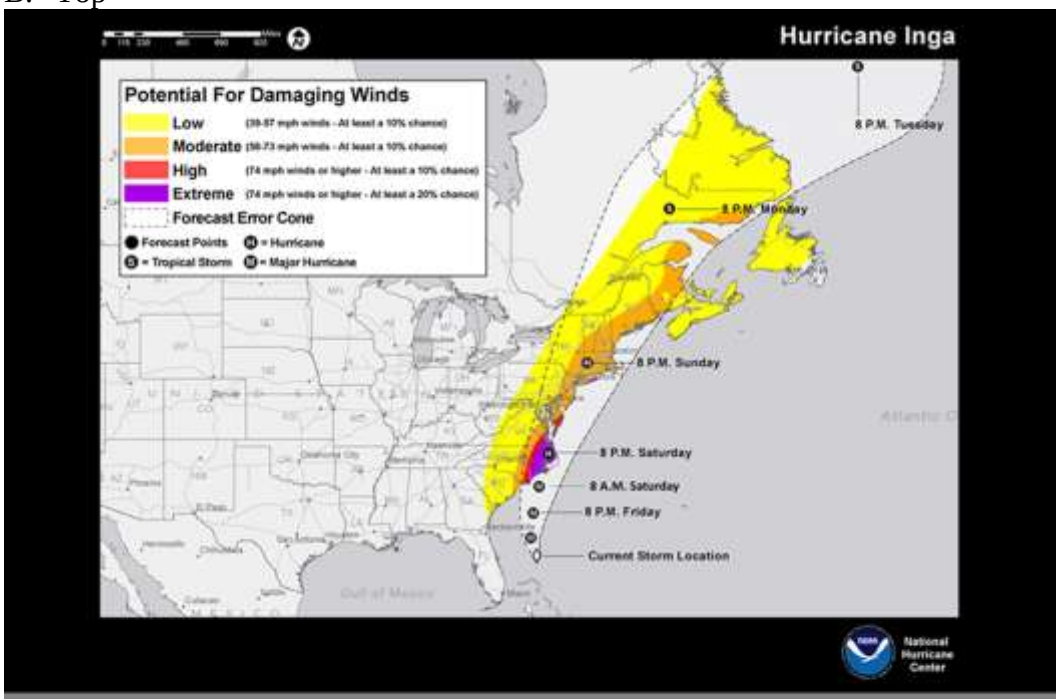
% Accurate	Bottom	Top	Split
Experts	75	83	75
Community	46	50	46
Undergraduates	45	58	48
Total Average	53	60	53

Figure 24 A-C. Images used to test accuracy of wind map use with three of four accuracy questions.

A. “Bottom”



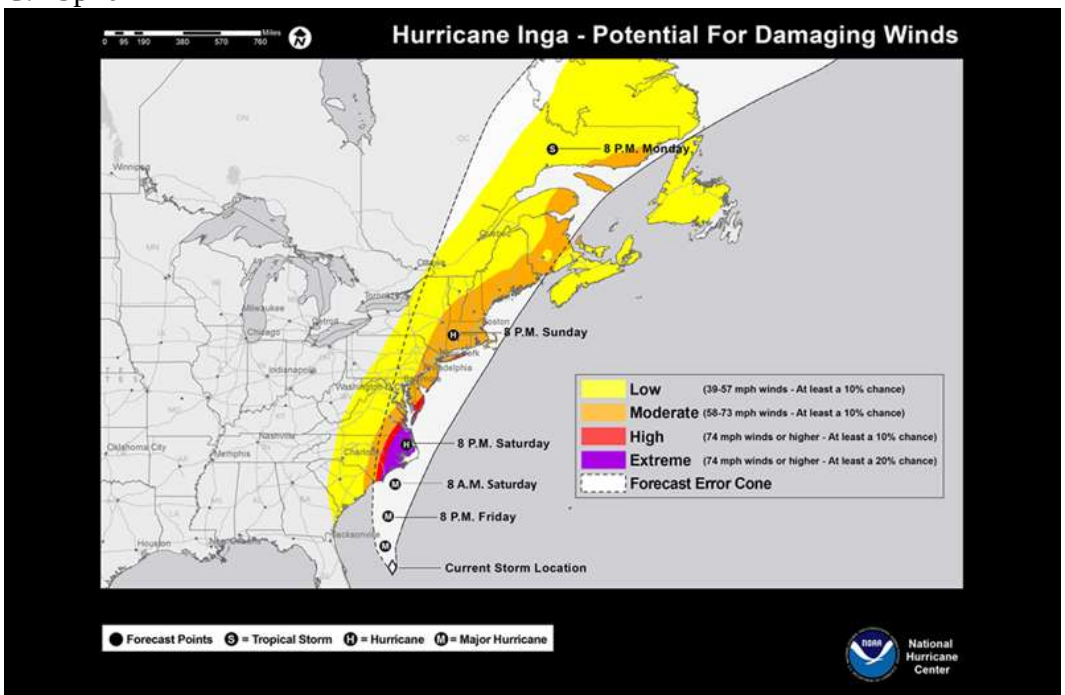
B. “Top”



What do you believe is the potential for damaging winds from Hurricane Inga at 8 A.M. Saturday?

1 = None 2 = Low 3 = Moderate 4 = High 5 = Extreme

C. “Split”



Between 8 P.M. Saturday and 8 P.M. Sunday, does the potential for damaging winds from Hurricane Inga increase, decrease, or stay the same?

A = Increase B = Decrease C = Stay the same

References

Allen, G.L., Miller Cowan, C. R., Power, H. 2006. Acquiring information from simple weather maps: Influences of domain-specific knowledge and general visual-spatial abilities. *Learning and Individual Differences.* 16: 337-349.

Canham, M. and Hegarty, M. 2010. Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*. 2: 155-166.

Brewer, C. A., 1994. Color use guidelines for mapping and visualization. In: MacEachren, A. M, Fraser Taylor, D. R. (Eds), *Visualization In Modern Cartography*. Elsevier Science Inc, New York 123-147.

Deubel, H. & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36,* 1827-1837.

Fabrikant, S., Hespanha, S., and Hegarty, M. 2010. Cognitively inspired and perceptually salient graphic displays for efficient spatial inference making. *Annals of the Association of American Geographers,* 100: 13-29.

Garlandini, S. and Fabrikant, S.I. 2009. Evaluating the effectiveness and efficiency of visual variables for geographic information visualization. In K. S. Hornsby et al., (eds). COSIT, Berlin: Springer-Verlag, 195-211.

Hegarty, M. Canham, M.S., and 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: 37-53.

Henderson, J. M. (1993). Eye movement control during visual object processing: Effects of initial fixation position and semantic constraint. *Canadian Journal of Experimental Psychology, 47*, 79–98.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception and Psychophysics, 57*, 787-795.

Hoffman, R.R., Detweiler, M., Conway, J. A., Lipton, K., 1993. Some considerations in using color in meteorological displays. *Weather and Forecasting* **8**, 505-518.

Mayhorn, C. B., Wolgarter, M. S., and Shaver, E. F. 2004. What does code red mean? Ergonomic in Design, (fall):12

Mersey, J. E., 1990. The role of colour scheme and map complexity in chloropleth map communication. *International Publications on Cartography*. University of Toronto Press, Toronto.

Monmonier, M., 1991. *How to Lie With Maps*. University of Chicago Press, Chicago.

Weinstein N.D., Sandman P.M. 1993. Some criteria for evaluating risk messages. *Risk Analysis* 13:103–14.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62,* 1457-1506.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime (Version 1.2) [Computer Software]. Pittsburgh, PA: Psychology Software Tools, Inc.