Building Better Graphs for Climate Change Communication: Evidence from Eye-tracking

by

Stephanie Courtney

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 5, 2019

Keywords: climate change, communication, eye-tracking,
geocognition, global warming, education

Approved by

Dr. Karen McNeal, Chair, Associate Professor of Geology
Dr. Martín Medina-Elizalde, Associate Professor of Geology
Dr. Christine Schnittka, Associate Professor of Science Education

Abstract

Reducing harm from global climate change will require public participation and therefore public education. Visual data representations such as graphs are often used for communication with public audiences but they are rarely designed using available evidence from cognitive science research. In the current study, undergraduate students took a pre-survey measuring climate change knowledge, climate scientist credibility perception, and perception of risk associated with climate change. Participants with low risk perception and low knowledge of climate change were then invited to the lab to view and answer questions about climate change graphs. Students either viewed original graphs from IPCC Summaries for Policymakers or new versions of the same graphs re-designed to fit evidence-based guidelines from Harold et al. (2016). Eye-tracking technology, which measures viewers' eye movements as a proxy for attention to different parts of the graphs, was used to evaluate usability. Overall participation in the activity increased participant risk perception, climate scientist consensus estimate, and perception of credibility of climate scientists. Results indicate many similarities in participant use of original and redesigned graphs, but slightly improved performance with original graphs primarily due to familiar formatting. Participants perceived the redesigned graphs as more credible and more satisfactory and one of the redesigned graphs as more worrying than the equivalent IPCC originals. Participant feedback was used to redesign the graphs again for use in a small third condition with improved results.

Acknowledgments

I am ever so grateful to all the folks near and far that have supported me and hopefully will continue to do so as I persist on the path to knowledge. I would love to specifically thank my parents (Chris and Mike), my brother (Josh), and my parents' cats (Francesca and, to a lesser extent, Mew). I have been sustained by electronic and cellular encouragement from Molly, Rachel, Jane, Gillian, Sarah, Willa, Stacey, Clee, and many others. I would also like to thank the mentors who guided me through my undergraduate progress that, after an important break, made this accomplishment possible – Drs. Marcia Björnerud, Andrew Knudsen, and Jim Moran. In this past not-quite-two-years, I have been fortunate to enjoy the companionship and colleagueship of the Auburn Geosciences Department and Geocognition lab group members: Eli, Nick, Lindsay, and Akilah. Thank you also to Drs. Jordan Harold and Tim Shipley for their crucial feedback during the design stage of this project. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1414475. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Lastly, I thank my wise mentors and committee members, Drs. Chris Schnittka and Martín Medina, and my dedicated advisor Dr. Karen McNeal.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AOI	Area of Interest
IPCC	Intergovernmental Panel on Climate Change
SPM	Summary for Policymakers
SYR	IPCC Synthesis Report
TtFF	Time to First Fixation
WG1	IPCC Working Group 1

INTRODUCTION

Climate change is among the greatest threats to human lives in the near future, and reducing harm requires public education about the facts and risks associated with global climate change (IPCC, 2013). In recent years, there has been a flood of available information on climate change science and impacts, especially via visual data representations, which are important tools for communicating with public audiences. Unfortunately, visual communication tools such as graphs are not usually designed using evidence from cognitive science. However, there exists a foundation of cognitive research that can provide guidance for creating data visualizations that are effective in aiding comprehension for various audiences (Harold, Lorenzoni, Shipley, & Coventry, 2016), and methods such as eye-tracking can provide novel insights to viewers' experiences with such visualizations.

Few studies have considered climate visualization design as it relates to viewers' prior knowledge (Harold et al., 2016; Atkins & McNeal, 2018). Prior knowledge and perceptions of the audience are particularly important to consider for the design of climate change-related communication tools. This is because public perceptions of climate change correlate strongly with cultural worldviews but do not always coincide with greater knowledge of the scientific facts of climate change (Kahan, Jenkins-Smith, & Braman, 2011; Leiserowitz & Smith, 2010). However, among undergraduate students, research has shown that knowledge-based approaches (i.e. content instruction) are more effective, possibly due to the malleability of belief systems of this age group (Aksit, McNeal, Libarkin, Gold, & Harris, 2017).

Climate Change Perceptions

In a spring 2018 Pew survey of over 27,000 people across 26 countries, 67% of respondents said that global climate change was a "major threat to [their] country," higher than any other threat, including ISIS (Poushter & Huang, 2019). The United States trailed slightly, with only 59% feeling majorly threatened by climate change. In fact, 15% Americans still do not believe global warming is happening at all and 24% believe it is due mostly to natural causes (Leiserowitz et al., 2019). However, climate literacy in the US is more complicated than belief or disbelief. In the same survey, 90% of self-identified liberal democrats agreed that global warming is due mostly to human activities compared to 28% of conservative republicans. The current study will take place in a region of the US with lower-than-average climate awareness; in a 2018 update of a model by the same authors, only 63% of Alabamians responded that global warming was happening and 37% said that warming was due to mostly natural causes (compared to 70% and 32% national averages) (Howe, Mildenberger, Marlon, & Leiserowitz, 2015).

Previous research strongly supports the relationship between climate change beliefs and cultural worldviews (Aksit et al., 2017; Kahan et al., 2011; Leiserowitz & Smith, 2010; Leiserowitz et al., 2019; McCright, Dunlap, & Xiao, 2013; van der Linden, Leiserowitz, Feinberg, & Maibach, 2015; van der Linden, Leiserowitz, & Maibach, 2017). However, instruction or communication of evidence supporting the existence of anthropogenic climate change can impact people’s beliefs and perception of risk associated with climate change (Aksit et al., 2017; McCright et al., 2013; van der Linden et al., 2017)

One indicator of overall climate literacy is the awareness of the consensus of around 97% of climate scientists that human activity is causing current global warming (Cook et al., 2016). Several studies have pointed to knowledge of the high scientific consensus as a reliable predictor of climate beliefs and support for relevant policy (McCright et al., 2013). This effect, called the “gateway belief model,” has been empirically validated. In addition, exposure to the fact of scientific consensus around climate change increases laypeople's belief that climate change is happening and is anthropogenic and increases their support for public action (van der Linden et al., 2015). In another study, exposure to the 97% scientific consensus reduced the polarization in beliefs between conservatives and liberals by 50%, showing the power of fact-based communication to overcome ideological differences (van der Linden, et al., 2017).

Because the gateway belief model is powerful, these authors also tested the messaging method for communication of the scientific consensus (van der Linden, Leiserowitz, Feinberg, & Maibach, 2014). Participants who viewed a pie chart displaying the 97% consensus, for example, estimated that the scientific consensus of climate scientists was 15% higher than they had before viewing the pie chart. In the same study, these results were compared to participants exposed to the same information in various text and image layouts and showed the pie charts were most effective in communicating this consensus. This work was primarily exploratory but represents a unique and crucial examination of the effectiveness of visuals in communicating climate change.

Several authors have created guidelines for creating more effective graphs and data visualizations; however, few of these are based on empirical evidence (Harold et al., 2016). Existing research on graph effectiveness often relies on survey-based methods, such as recording accuracy on data extraction tasks (e.g. Canham and Hegarty, 2010). Cognitive processes during tasks are then examined through interviews, which can distract the participants from task performance and thus influence results. Eye-tracking analysis, in contrast, can be conducted simultaneously with other tasks without interference, and can detect changes and features of cognition that may be too minute to detect with other methods (Bojko, 2013). For example, Atkins and McNeal (2018) examined expert-novice differences when viewing climate graphs and found that novice viewers spent proportionally more time viewing the axes and title than the data of the graphs than expert viewers.

Otherwise, climate graph research has so far focused on communication with policy- and decision-makers, including graphics used in the IPCC Summaries for Policymakers. Harold et al. (2016) reviewed climate communication and cognitive science literature to create evidence-based guidelines for designing more accessible visualizations and applied the guidelines to modify a major IPCC graphic (Harold et al., 2016). McMahon, Stauffacher, and Knutti (2016) explored participant affective reactions to different designs of visuals including four IPCC figures and two infographics and found that participants had less confidence in the infographics. This may be related to the authors' previous work that suggested that participants were highly confident in the data presented to them due to the complexity of the figure (McMahon, Stauffacher, & Knutti, 2015).

Graph Reading

Visual processing of graphs consists primarily of the interaction of top-down and bottom-up cognitive processing, according to the construction-integration model by Freedman and Shah (2002). This model has been used for text and graph comprehension and consists of two stages. The first stage represents the construction phase in which prior knowledge is activated based on the viewer's expectations and observation of the visual stimulus presented. In the second stage, the viewer integrates discrete observations and judgements of the stimulus into a single understandable message. In both stages, bottom-up processes describe the attention to visually salient features, i.e., what the eye is drawn to when viewing a stimulus based on aesthetic design. Top-down processes depend on the prior knowledge and expectations of the viewer, either concerning graphs in general or specific domain knowledge (Freedman & Shah, 2002).

Top-down processes are often analyzed with expert-novice studies, in which researchers compare performance across knowledge gaps either in graph reading or in domain content

(Atkins & McNeal, 2018; Stofer & Che, 2014). Experts have more resources for integrating discrete observations of a stimulus into a coherent representation (Freedman and Shah, 2002). Because bottom-up processes depend on visual salience, these processes are often examined in studies that compare design features of graphs. Many studies have shown that graph design can significantly alter comprehension and task performance of participants (Hegarty, 2011). Shah and Carpenter (1995) found that participants interpreted even very simple graphs differently depending on which data was shown on the x, y, or z axis, including graduate students experienced in graph comprehension and construction. Renshaw, Finlay, Tyfa, and Ward (2003) compared graphs of identical data designed either in accordance with or opposed to best-practice design guidelines. The graphs designed in accordance with the guidelines scored significantly higher in usability. However, evidence must show whether this kind of work extends to the unique domain of climate change, where knowledge and beliefs are particularly mediated by social and cultural contexts (Kahan et al., 2011).

Eye-Tracking

Eye-tracking can also be used to examine real-time visual and attentional processes, i.e., the construction and integration related to the visual features of the stimulus (Duchowski, 2007). Eye movements have been tracked and studied in some form since the 1800s, but in recent decades, improved technology has made eye-tracking much more precise and less invasive. Modern eye-tracking uses infrared light shone toward the participant and tracks the reflection back from the cornea and the retina. The corneal reflection only moves with the participant's head and is compared by the eye-tracker to the reflection from the retina, which indicates the location of the pupil. Humans have a total visual field around 180° but can see in greatest detail and brightest color in the center ~5° of the visual field, called the fovea (Duchowski, 2007). The fovea in visual processing is one illustration of the fact that humans, though presented with endless stimuli, have limited attention and processing capacity. Therefore, vision is “a piecemeal process” relying on the integration of small, discretely perceived areas into one bigger picture (Duchowski, 2007).

Most eye movement consists of only two actions, fixations and saccades, so modern eye-tracking measures both. Fixations are defined as a period of at least 70ms where the eye is stationary, and saccades are the extremely fast movements between fixation points (Bojko, 2013). Since the fovea is quite small, the area of each fixation is small as well, so fixations can be measured precisely. People are effectively blind during saccade movements but view and process stimuli during fixations. For this reason, most eye-tracking analysis centers around total duration of fixations, number and order of fixations, and fixations in particular areas of interest. Eye-tracking is used primarily to understand and compare viewers’ attention to particular visual features. When using eye-tracking, researchers can set an Area of Interest (AOI) corresponding with features particularly relevant to top-down or bottom-up processing to examine the AOIs influence on the user’s attention and task performance. Attention plays a significant role in thinking and processing, but eye-tracking is best used in tandem with other research methods which provide context for eye movement results, such as usability tasks or interviews (Bergstrom & Schall, 2014; Duchowski, 2007; Bojko, 2013).

OBJECTIVES

The goals of this study are to compare and evaluate the effectiveness of data visualizations of scientific evidence of climate change for communication with undergraduate students. The focus of the study is the comparison of original visualizations presented in Intergovernmental Panel on Climate Change (IPCC) reports with modifications of those visualizations designed to reflect findings from cognitive science research.

The primary research questions addressed in this study are: (1) How does graph design affect usability of the graph, where usability is characterized by efficiency, effectiveness, and satisfaction? (2) How does graph design affect visual attention to the graphs, as shown by eye-tracking? (3) How does graph design affect participant perceptions of credibility of the graphs and of climate scientists overall? (4) How does graph design affect participants' perception of risk associated with climate change? These questions were addressed using mixed qualitative and quantitative methods featuring eye-tracking, data extraction tasks, surveys, ranking exercises, and interviews, summarized in Table 1.

Table 1

Alignment of Research Questions, Metrics, and Analysis

Research Question	Dependent Var.	Instruments	Analysis
1. How does graph design influence usability of graphs?			
1a. Usability: Effectiveness	Participant accuracy	Data extraction questions	t-test (A vs. B) ANOVA (A, B, C)
1b. Usability: Efficiency	Time to answer question	Eye-tracking times	t-test (A vs. B) ANOVA (A, B, C)
1c. Usability: Satisfaction	Participant satisfaction rating	Satisfaction Q's, Rankings	t-test (A vs. B) ANOVA (A, B, C)
2. How does graph design influence attention to graphs?	Fixation metrics	Eye-tracking	t-test (A vs. B) ANOVA (A, B, C)
3a. How does graph design influence perception of credibility of graphs?	Participant credibility rating	Credibility Q's, Rankings	t-test (A vs. B) ANOVA (A, B, C)
3b. How does graph design influence participant perception of credibility of climate scientists?	Credibility rating, change from pre-survey	Pre- and Post-survey credibility instrument	Mixed ANOVA
4. How does graph design influence climate change risk assessment?	Risk assessment items, change from pre-survey	Pre- and Post-survey risk assessment items	Mixed ANOVA

Note. Non-parametric equivalents may be used (Mann-Whitney in place of t-test, Wilcoxon signed rank in place of paired t-test, and Kruskal-Wallis in place of one-way ANOVA)

MANUSCRIPT FOR SUBMISSION

BACKGROUND

The United States may bear a higher short-term cost from climate change than almost any other country in the world but its citizens are less concerned about climate change than those of many other countries (Leiserowitz et al., 2019; Poushter & Huang, 2019). This discrepancy can be explained somewhat by American unawareness of climate change; most recently, the Yale Program on Climate Change Communication found that, on average, only 74% of registered American voters think global warming is happening and 62% think that it is caused mostly by human activities (Leiserowitz et al., 2019). However, there are 56% and 62% differences in these statistics, respectively, between Americans from opposite ends of our political spectrum which illustrates one of the several complexities of climate change literacy in America.

This variation can be explained, at least partially, by the cultural cognition thesis presented by Kahan et al. (2011). The authors found that awareness around scientific consensus on a number of scientific issues, especially climate change, were tightly coupled with participants' worldviews as categorized along spectra of hierarchy and individualism. Awareness of the scientific consensus around climate change in particular is known to be important, and in one study, educating participants about the consensus led to significantly higher worry around the issue and higher beliefs that climate change is happening and is anthropogenic (van der Linden et al., 2015). However, educating the public about these issues also requires cultural consideration; the participants in the work of Kahan et al. (2011) were more likely to assess a fictional expert as knowledgeable and trustworthy if the expert is expressing a view that aligns with the participants' pre-conceived risk assessment of global warming.

There is a growing body of research concerning climate change communication with graphs and other data visualizations. Much of this inquiry was reviewed by Harold et al. (2016) to compile a set of guidelines to facilitate the design of more accessible graphics to communicate climate change (see Table 2). Graph use and comprehension can be assessed through the use of eye-tracking technology and the construct of usability (Goldberg & Kotval, 1999; Goldberg & Wichansky, 2003; Renshaw et al. 2003). Usability is task-focused and is characterized by effectiveness in completing the task, efficiency in completing the task, and user satisfaction with the product or interface. These metrics don't require eye-tracking but the additional instrument can enhance usability evaluation.

Graph design can also affect perceptions of risk, though this topic has been researched in far more depth as it applies to medicine than climate change (Ancker, Senathirajah, Kukafka, & Starren, 2006; Okan, Stone, & Bruine de Bruin, 2018). Additionally, there is not yet much research exploring the relationship between graph design and perceptions of credibility in climate change, but there is strong evidence that they may be related (McMahon et al. 2016). This study was designed to address each of these topics and is focused on asking how graph design affects (1) visual attention to the graphs, (2) usability of the graphs, (3) perceptions of credibility of the graphs and of climate scientists, and (4) climate change risk assessment. We have employed eye-tracking, survey, and interview methods to answer these questions in an explanatory mixed-methods study (Creswell & Clark, 2018).

METHODS

Participants first took an online pre-survey via emails to large introductory undergraduate classes. The survey included a 21-item climate knowledge inventory, composed of questions concerning the facts of Earth's climate and climate change from Libarkin, Gold, Harris, McNeal, & Bowles (2018), questions to gauge participant's perceptions of risk associated with climate change drawn from Libarkin et al. (2018) and Leiserowitz et al. (2019), an instrument on participants' perception of credibility of climate scientists adapted from McCroskey and Teven (1999), items concerning graph literacy and frequency of use from Atkins & McNeal (2018), and other demographic and background information items. The climate knowledge inventory from Libarkin et al. (2018) was selected for use in this study because it was developed with thorough validity and reliability methodology (including Rasch analysis) and has been used with similar undergraduate populations (Aksit et al., 2017; Libarkin et al., 2018). The instrument from McCroskey and Teven (1999) has also been highly reliable in other uses with undergraduates and measures the sub-factors of credibility used in source trust literature without applying the construct explicitly to internet use as many other instruments do (Connolly & Bannister, 2007). The risk items from Leiserowitz et al. (2019) had not been validated but have been used regularly by those authors and provide for comparison with many previous samples.

Three graphs, each with an A and B version, were initially selected for the study. The A version of each graph was originally published in either the Intergovernmental Panel on Climate Change (IPCC) Working Group 1 (WG1) Summary for Policymakers (SPM) or the IPCC Synthesis Report (SYR) SPM. Each B version of the graphs used the same data as the A version but was re-designed to adhere to guidelines for graph accessibility based on cognitive science research compiled in Harold et al. (2016), shown in Table 2. A rubric was created to assess application of the Harold et al. (2016) guidelines to graph design and the rubric was reviewed by the lead author of that publication. After the first re-designs of potential graphs for this study were created, the graphs and rubric were distributed to several experts including the lead and third authors of Harold et al. (2016) for review, after which the graphs were altered to reflect reviewer feedback. The B and C versions of these graphs were designed in Adobe Illustrator.

Table 2

Evidence-informed Guidelines to Improve Accessibility of Scientific Graphics of Climate Science from Harold et al. (2016)

	Psychological Insights	Associated guidelines to improve accessibility
	1. Intuitions about effective graphics do not always correspond to evidence-informed best practice for increasing accessibility	Use cognitive and psychological principles to inform the design of graphics; test graphics during their development to understand viewers' comprehension of them
Direct Visual Attention	2. Visual attention is limited and selective -- visual information in a graphic may or may not be looked at and/or processed by viewers	Present only the visual information that is required for the communication goal at hand. Direct viewers' visual attention to visual features of the graphic that support inferences about the data
	3. Salient visual features (where there is contrast in size, shape, color, or motion) can attract visual attention	Make important visual features of the graphic perceptually salient so that they 'capture' the attention of the viewer
	4. Prior experiences and knowledge can direct visual attention	Choose and design graphics informed by viewers' familiarity and knowledge of using graphics and their knowledge of the domain, that is, knowledge about what the data represents. Provide knowledge to viewers about which features of the graphic are important to look at, for example, in text positioned close to the graphic.
Reduce complexity	5. An excess of visual information can create visual clutter and impair comprehension	Only include information that is needed for the intended purpose of the graphic; break down the graphic into visual 'chunks', each of which should contain enough information for the intended task or message
Support inference-making	6. Some inferences may require mental spatial transformations of the data; experts may have strong spatial reasoning skills, non-experts may not.	Remove or reduce the need for spatial reasoning skills by showing inferences directly in the graphic. Support viewers in spatial reasoning, for example, by providing guidance in text.
	7. The visual structure and layout of the data influences inferences drawn about the data.	Identify the most important relationships in the data that are to be communicated; consider different ways of structuring the data that enable the viewer to quickly identify these relationships.
	8. Animating a graphic may help or hinder comprehension.	Decisions to create animated graphics should be informed by cognitive principles; consider providing user control over the playback and speed of the animation.
	9. Conceptual thought often makes use of cultural metaphors.	Match the visual representation of data to metaphors that aid conceptual thinking, for example, ‘up’ is associated with ‘good’ and ‘down’ is associated with ‘bad’; data with negative connotations may be easiest to understand if presented in a downwards direction.
Integrate text with graphics	10. When the graphic and the associated text are spatially distant, attention is split.	Keep the graphic and accompanying text close together, for example, use text within a graphic and locate the graphic next to the accompanying body text.
	11. Language can influence thought about the graphic.	Use text to help direct viewers’ comprehension of the graphic, that is, by providing key knowledge needed to interpret the graphic.

Graph 1 (SYR SPM.3) was selected in this study for its relative simplicity and the original (Figure 1) and redesign (Figure 2) are presented below.

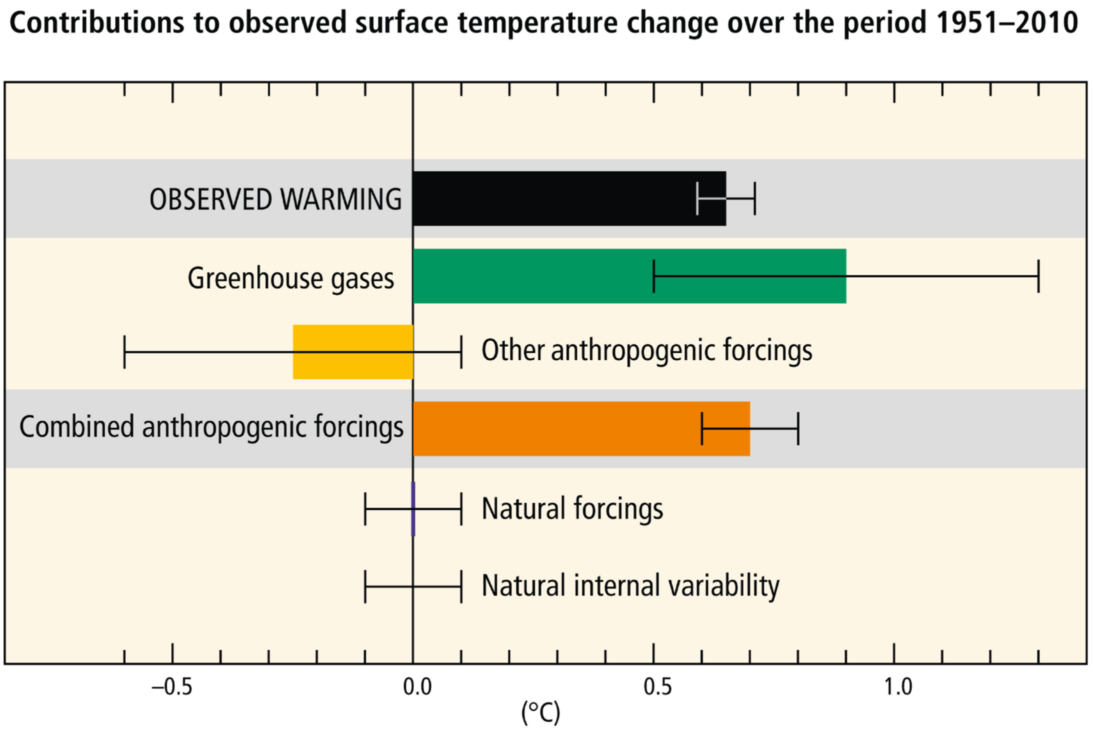

Contributions to observed surface temperature change over the period 1951–2010

Figure 1. Original IPCC graphic SYR SPM.3 (Graph 1A)

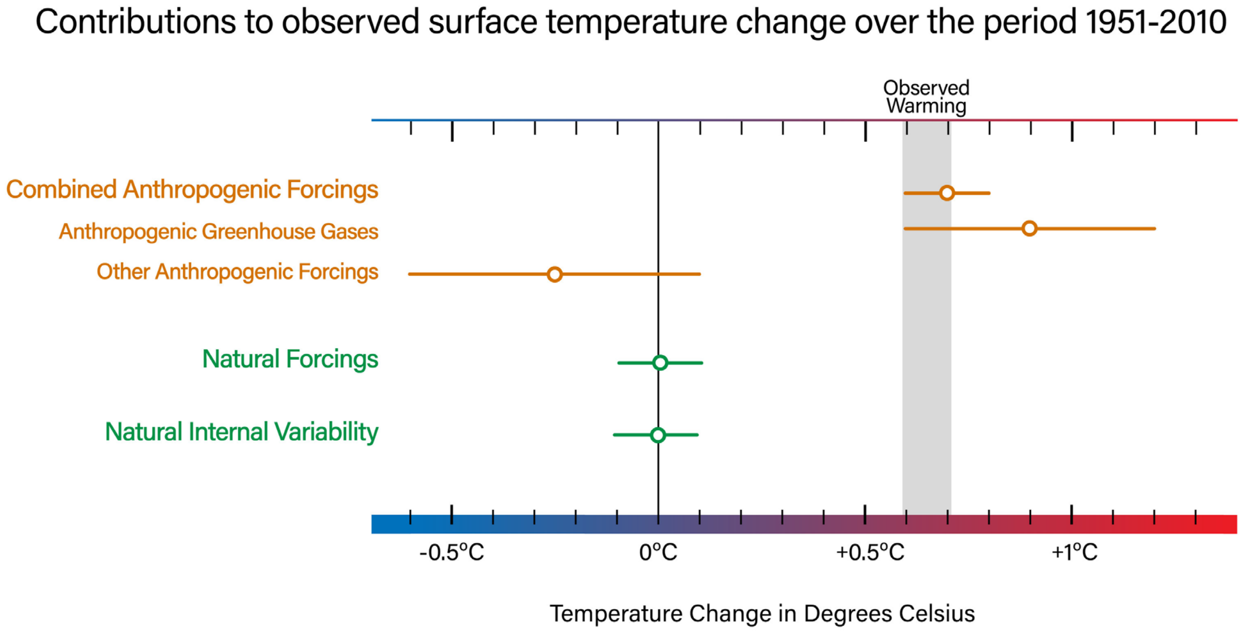

Contributions to observed surface temperature change over the period 1951-2010

Figure 2. Redesign of IPCC graphic SYR SPM.3 by author (Graph 1B)

Graph 2 (WG1 SPM.1) was cropped from its original publication version to reduce complexity for both the original (Figure 3) and redesign (Figure 4).

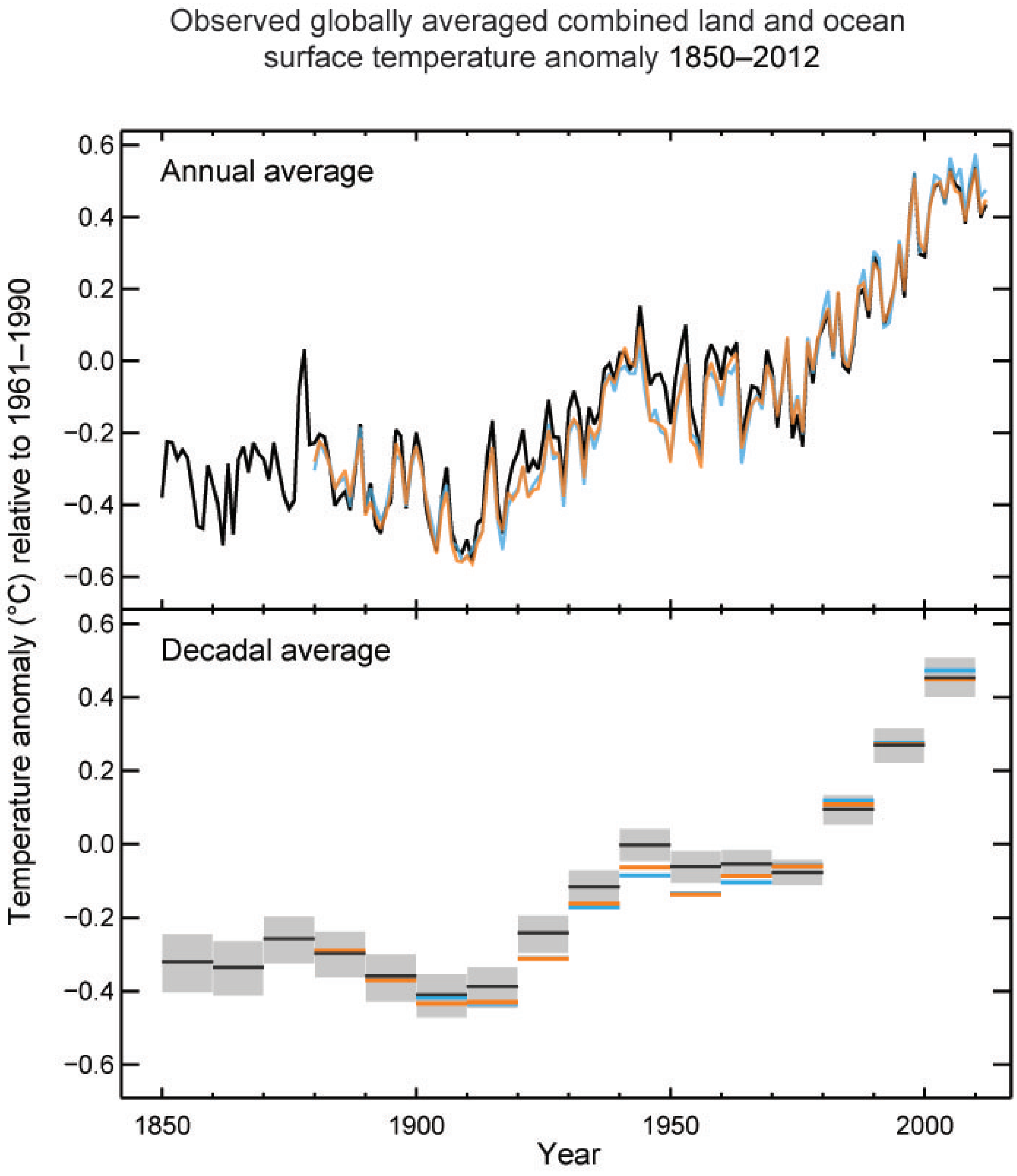


Figure 3. Original IPCC graphic WG1 SPM.1 (Graph 2A)

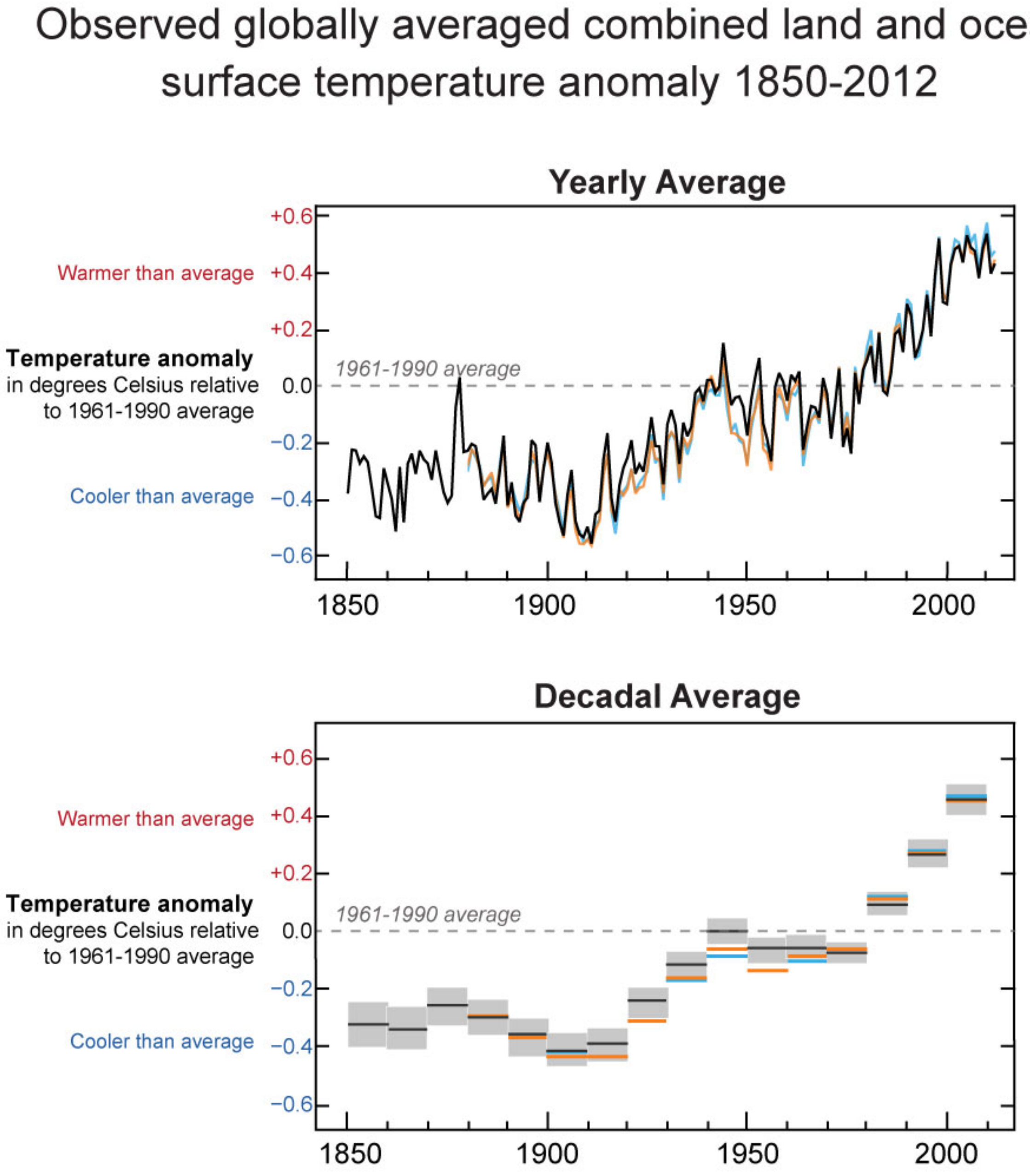


Figure 4. Redesign of IPCC graphic WG1 SPM.1 by author (Graph 2B)

The original publication version of graph 3 (WG1 SPM.5) and a redesigned version were both featured in Harold et al. as an example of the application of the guidelines compiled by those authors. Both the original (Figure 5) and redesign (Figure 6) were also cropped for size and total content for use in this study.

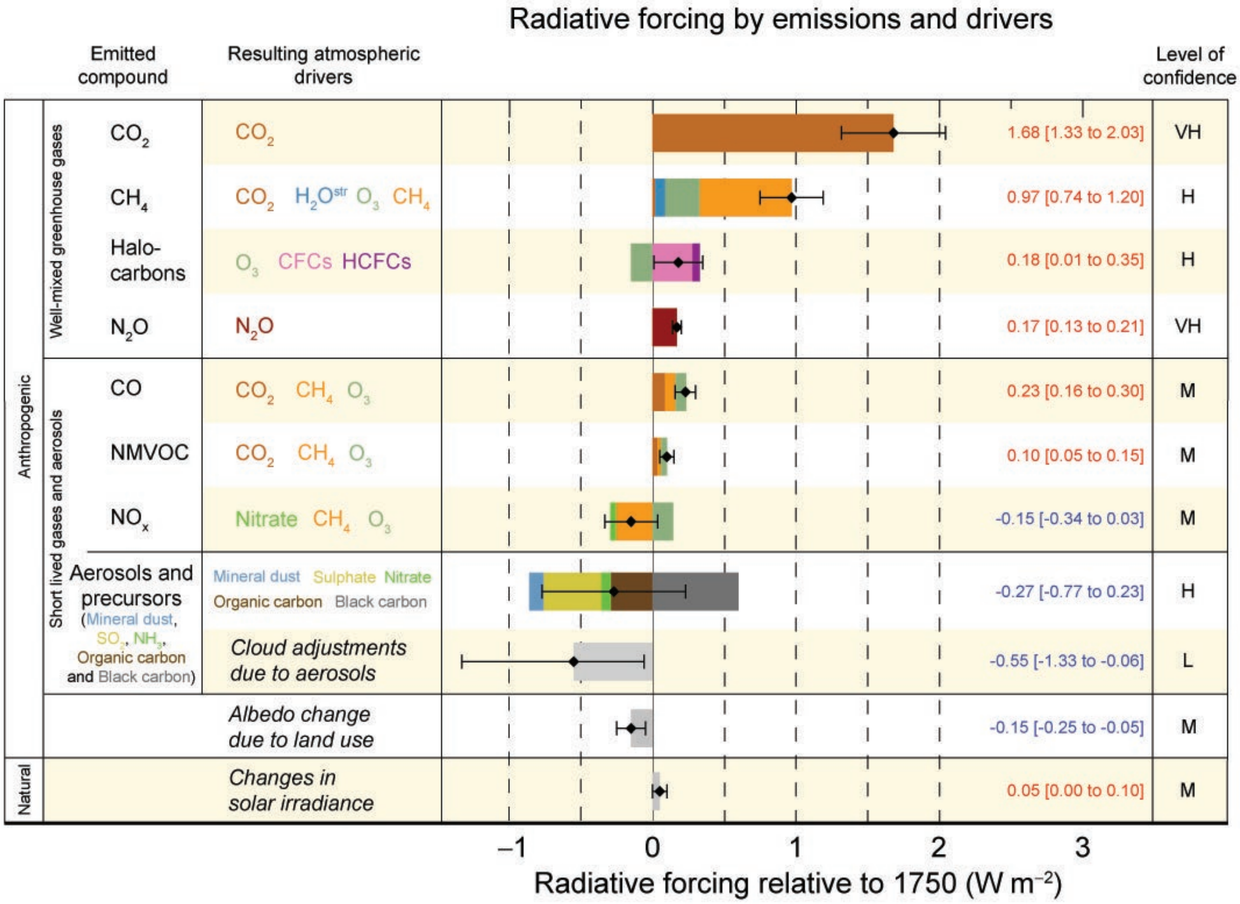


Figure 5. Cropped IPCC graphic WG1 SPM.5 (Graph 3A)

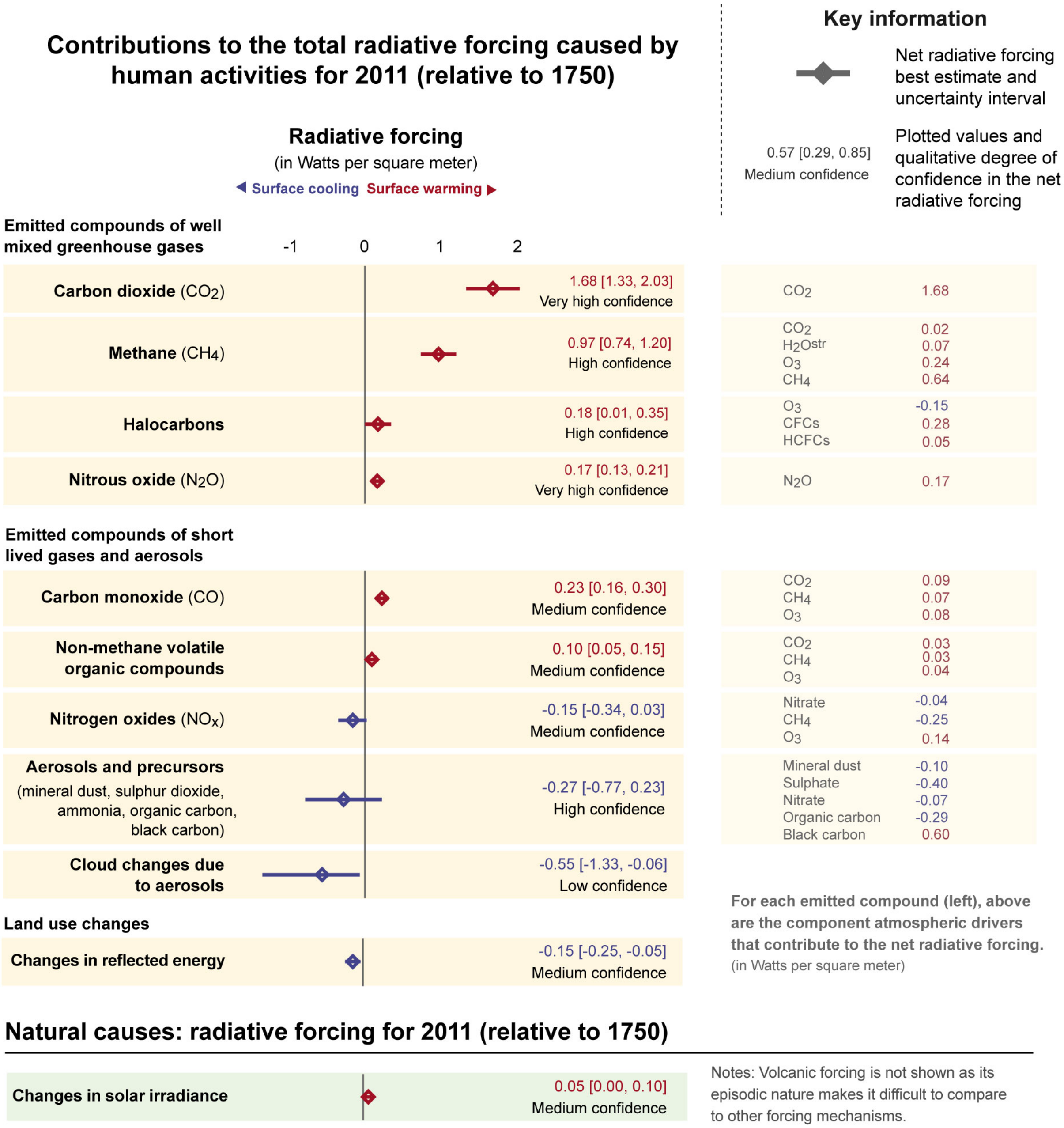


Figure 6. Cropped and redesigned IPCC graphic WG1 SPM.5, modified from Harold et al. (2016) (Graph 3B)

After initial data collection, participant feedback was compiled and a third C version of graphs 1 and 2 were created by the author for additional data collection (Figures 7 and 8).

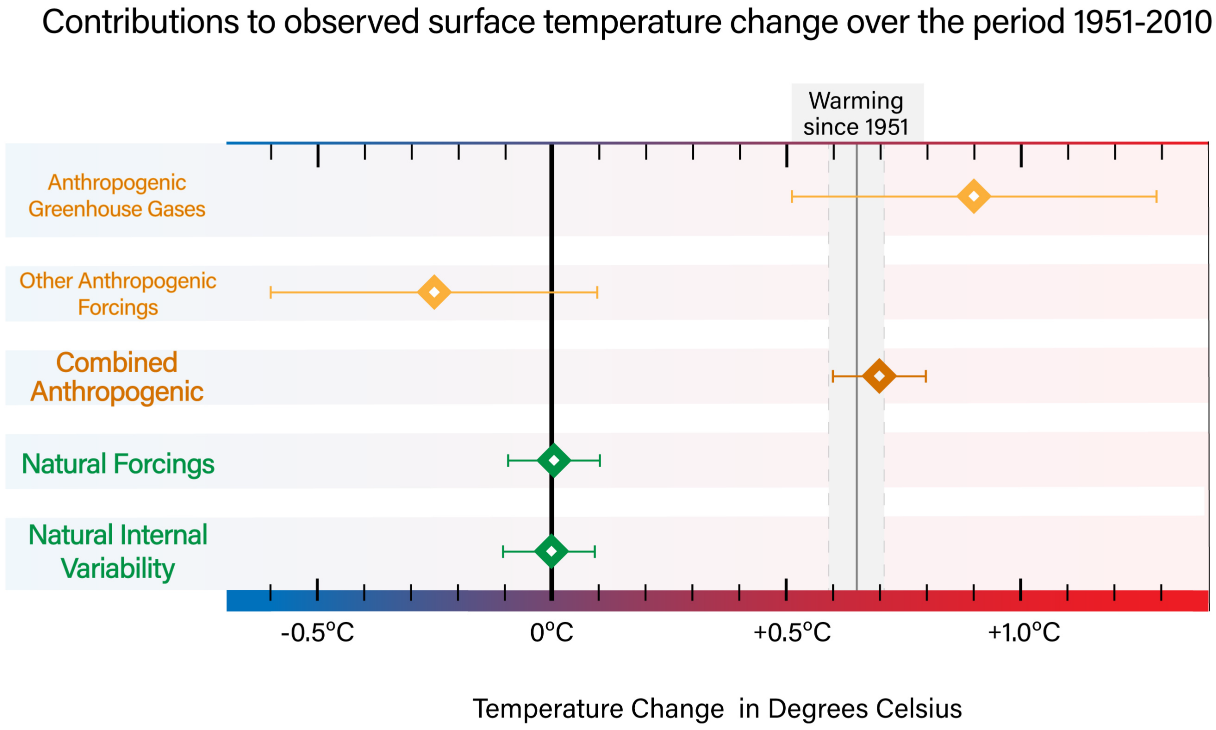


Figure 7. Redesign of IPCC graphic SYR SPM.3 by author (Graph 1C)

Observed globally averaged combined land and ocean surface temperature anomaly 1850-2012

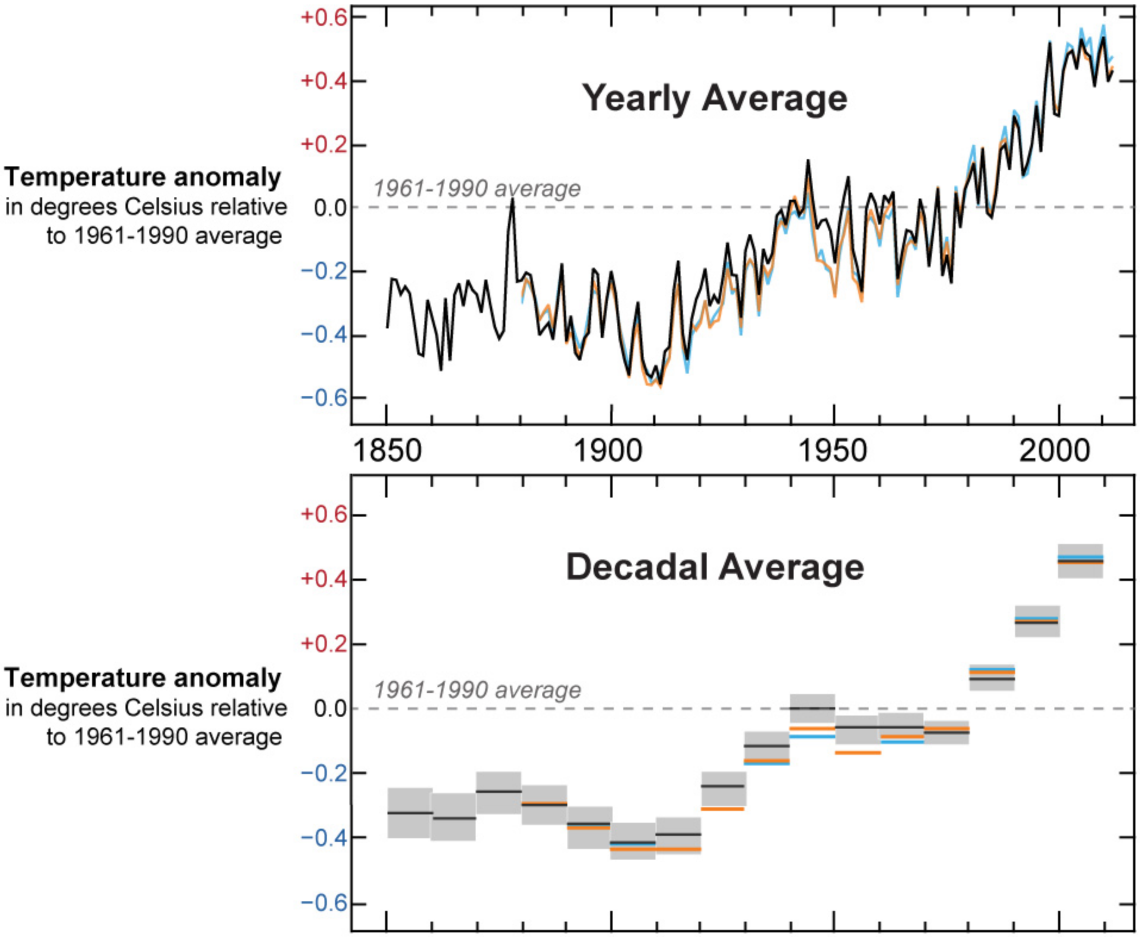


Figure 8. Redesign of IPCC graphic WG1 SPM.1 by author (Graph 2C)

Main Study

From pre-survey data, students with below-median scores in both climate change knowledge and climate change risk perception were invited to the lab to participate in the study. In the main phase of the study, participants were randomly assigned to one of two conditions for an A/B between-subjects eye-tracking study. The computer activity part of the experiment was designed and performed in Tobii Studio software and recorded with a Tobii TX300 eye-tracker. The eye-tracker is attached to the bottom of the monitor, requires no chin-rest or other physical restrictions, and samples at a frequency of 300 Hertz. Before the activity, each participant's eyes were calibrated to the eye-tracker which requires that participants sit 50-73cm away and directly in front of the screen but allows for shifting and adjustments within that range during the activity. Data from the two participants with less than 70% gaze samples (weighted to include presence or absence of each eye) were not used for any subsequent analysis.

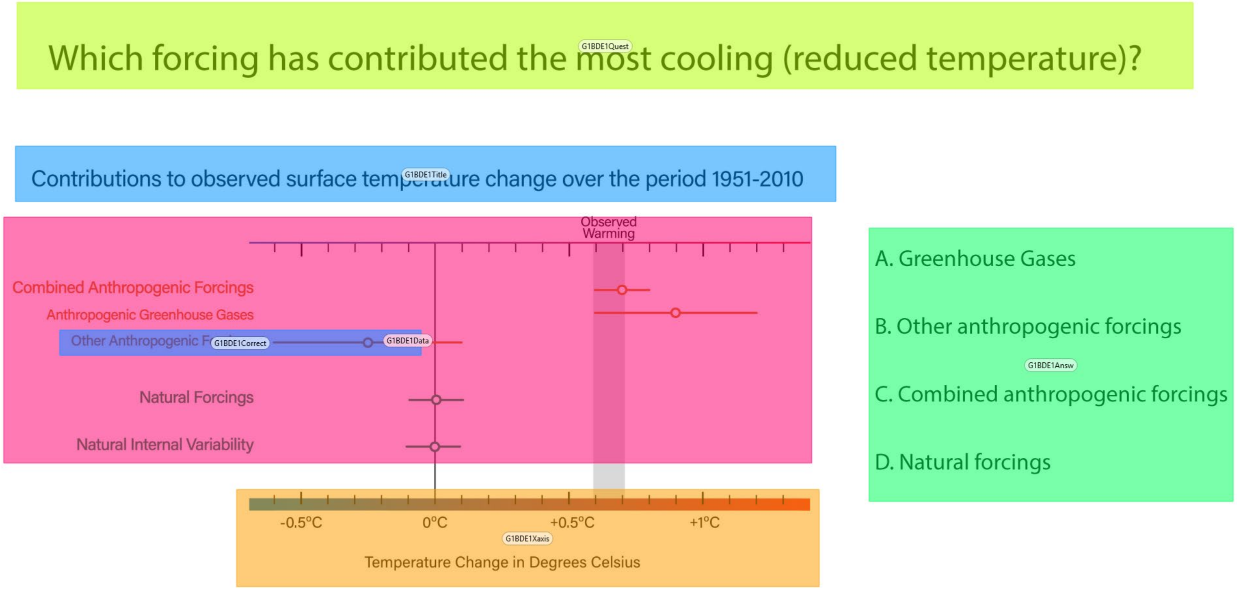
Each participant first had unlimited time to read definitions of three key terms used in the graphs (anthropogenic, forcing, and anomaly) to provide some context to those who had never seen the terms before. Participants then completed a practice question unrelated to the subject matter to become acquainted with the system, and then began the experiment that consisted of answering questions about three graphs of either the A condition designs or the B condition redesigns. The A condition participants viewed three graphs (Figures 1, 3, & 5) from the most recent IPCC Assessment Report Summaries for Policymakers. The SPMs are shorter, more concise reports and represent the materials intended for a non-expert audience. The B condition participants viewed three re-designs of the same IPCC graphs (Figures 2, 4, & 6). After initial A and B group data collection, a smaller C condition was also run (n = 9). The C condition consisted of new redesigns of the two graphs created by the researcher based on preliminary results from participants (Figures 7 & 8).

The three graphs in each condition were presented in alternating sequences. For each graph, the participants (1) viewed the graph alone, (2) answered three data extraction questions about the graph (simple graph-reading not requiring extrapolation or content knowledge), (3) answered a graph satisfaction question, and (4) answered a credibility perception question about the graph. Each question was multiple choice and the graphs were shown on screen when each question was presented. The participants had unlimited view time for each step until they advanced the activity with a mouse click. The satisfaction and credibility perception items were presented as statements with 4-point Likert scales answer options ranging from strongly disagree to strongly agree. The operative words of the credibility statements (misleading, accurate, credible) were drawn from an existing instrument (McCroskey & Teven, 1999). One of the many instruments designed to measure human-computer interface satisfaction, by Chin, Diehl, & Norman (1988), was sampled and adapted to create graph satisfaction items.

After the graph computer activity, participants completed a post-survey consisting of items repeated from the pre-survey concerning risk perception, credibility perception, and relevant climate change knowledge and new additional items concerning climate change and policy. After the post-survey, participants engaged in a recorded retrospective interview about their experience completing the graph activity and strategies used to answer the questions while watching the recording of their eyes. Retrospective interviews cued by gaze plot videos can lend more specific insight into the processes and difficulties that participants had during the activity (Olsen & Strandvall, 2010), and is often recommended over concurrent think-aloud methods because it is less likely to alter gaze patterns (Duchowski, 2007). After the eye-tracking retrospection, the recorded interview continued with questions about graphs and climate change perceptions and an activity in which graphs were ranked for ease of use, trustworthiness, and cause for worry about climate change. The graphs were compared for ease and trustworthiness within the participant's condition (all A, all B, or both C graphs), as data-equivalent pairs (e.g. 1A vs. 1B), and finally across all graphs presented to the participant.

Analysis

The primary analysis of eye-tracking data was conducted to assess usability of the graphs. Eye-tracking data were analyzed by *t*-test between A and B conditions and by Analysis of Variance (ANOVA) between all three conditions. The Tobii Studio software time until mouseclick data were used as the metric for total time spent on each graph because it includes time that fixations may have not been measured by the eye-tracker but during which participants may have still been viewing or deliberating on the question. Fixation count and duration for individual areas of interest were plotted and correlated to ensure that the metrics were highly related and that using only fixation duration in analyses would suffice (see Figure 11). The Areas of Interest (AOIs) used for duration and time to first fixation analyses were equal in size and placement for all questions asked for each graph (see Figure 9). Time to first fixation (or TtFF) is a metric in the eye-tracking software that counts the time from the introduction of the stimulus to the first fixation within an AOI, which can indicate which features are most salient to viewers. Fixation duration for each AOI was also normalized to the size of the AOI. This normalization is performed because participants would statistically spend more view time in larger AOIs so considering the impact of AOI size is valuable. However, because AOIs were data-equivalent, raw fixation duration data were interpreted as most meaningful and therefore used for final analyses.


Which forcing has contributed the most cooling (reduced temperature)?

A. Greenhouse Gases

B. Other anthropogenic forcings

C. Combined anthropogenic forcings

D. Natural forcings

G1BDE1Quest
Contributions to observed surface temperature change over the period 1951-2010
G1BDE1Title
Observed Warming
Combined Anthropogenic Forcings
Anthropogenic Greenhouse Gases
Other Anthropogenic Forcings
G1BDE1Correct
G1BDE1Data
Natural Forcings
Natural Internal Variability
-0.5°C
0°C
+0.5°C
+1°C
G1BDE1Xaxis
Temperature Change in Degrees Celsius
G1BDE1Answ

Figure 9. Example of graph question area of interest boundaries (graph 1B)

For comparisons between the participant pre- and post-tests, such as for change in credibility perception of climate scientists and risk assessment of climate change, data were analyzed with mixed ANOVA to detect effects from both participation (pre/post within-subjects) and group membership (A/B/C between-subjects).

Participant answers to data extraction, satisfaction, and credibility perception questions were extracted from Tobii Studio. Data extraction answers were coded as correct or incorrect and compared across groups. Individual answer distributions and composites of answers were highly non-normal and therefore analyzed with nonparametric tests such as the Mann-Whitney U test (between two groups only) and the Kruskal-Wallis H test (all three conditions). Participant ranks for ease, trustworthiness, and worry/risk were inverted such that higher values would correspond to higher assessment of the variable (i.e., 1st place in trustworthiness inverted to 6 trust “points”). Ranked data are ordinal and therefore also analyzed by nonparametric tests, specifically the related-samples Wilcoxon signed rank W test since analysis included multiple graph ranks for each individual. Additionally, since the C condition was ranking only four graphs rather than six, the rank values are not comparable to the A and B groups, but graph ranks can still be compared within the C condition. Non-parametric test results are reported with the standardized *Z*-statistic.

Recorded interviews were transcribed and data were coded primarily for specific features of each graph that contributed to participant perceptions of usability, credibility, and climate change risk. An Auburn Geocognition Lab member co-coded excerpts sampled from 25% of participants to ensure good inter-rater reliability (93% agreement, Cohen's Kappa = .69). Analysis was performed primarily by searching for co-occurrences between individual graph codes, value codes (praise/ease, criticism/difficulty, etc.) and codes for features of graphs (amount of information, use of color, organization/layout, etc.). Three a-priori codes were drawn from the three factors of McCroskey and Teven (1999) and additional codes for risk perception and credibility were used as they emerged from the data. Because the interview pivoted around comparing and ranking the graphs, most comments were in reference to specific graphs and features. However, participants also spoke generally about risk, credibility, and graphs, so some analysis separate from graph features was appropriate.

RESULTS

The 69 total participants in the study were, typically, 19-year-old (*M* = 19.17, *SD* = 1.465, range 18-24), white (n = 60), conservatives ("conservative" n = 35, "very conservative" n = 8), who attend church annualy or more often (n = 42). A little over half of the participants were women (n = 37). About half were in STEM majors (n = 33) and most were currently enrolled in a course that they identified as within STEM (n = 55). About half were "sympathetic toward the environmental movement" (n = 36), none identified as active with it, 26 participants said they were neutral toward it, and 7 were “unsympathetic toward the environmental movement”. Participants initially estimated that 72.16% of climate scientists “think that human-caused global warming is happening” on average (*SD* = 18.13). Participant scores on the initial climate knowledge inventory was significantly correlated with their self-reported frequency of graph use and creation (Pearson’s $r = .408$, $p = .001$) and their performance on a 4-item axis identification task (Spearman’s $\rho = .327$, $p = .006$). The graph measures were also correlated with each other ($\rho = .287$, $p = .017$), but not with any other variables most relevant to the research questions.

There was one participant with some kind of colorblindness in each of the A, B, and C groups. Participants in each condition were statistically equivalent except the tendency for C-group members to be in earlier school years (i.e., more freshmen) than the B group, $p = .003$. Each participant group is named for the graphs that they used during the computer activity, though by the end of the study the A and B groups saw all of the A and B graphs, and the C group saw both the A and C graphs.

Usability: Effectiveness, Efficiency, Satisfaction

Effectiveness, as measured by accuracy on data extraction questions, varied slightly by group and by question. On average between all three graphs, participants in condition A performed significantly better than participants in condition B, $Z = -3.10$, $p = .002$. However, when separated by graph, only performance on questions for graph 1 was significantly different, $Z = -2.82$, $p = .005$ (see Table 3). Further, participants did not perform consistently on items overall or by graph (Cronbach’s $\alpha = .296$ for all 9 items combined, graph 1 $\alpha = .434$, graph 2 $\alpha =$ negative, graph 3 $\alpha = .173$). Considered as individual questions, group A performed significantly better than B on one question for graph 1 and one question for graph 2 (see Figure 10). Group C performed statistically equally to group A on all individual questions and scales, and better than group B on the scale of all 6 applicable data extraction questions combined, $Z = 2.81$, $p = .005$.

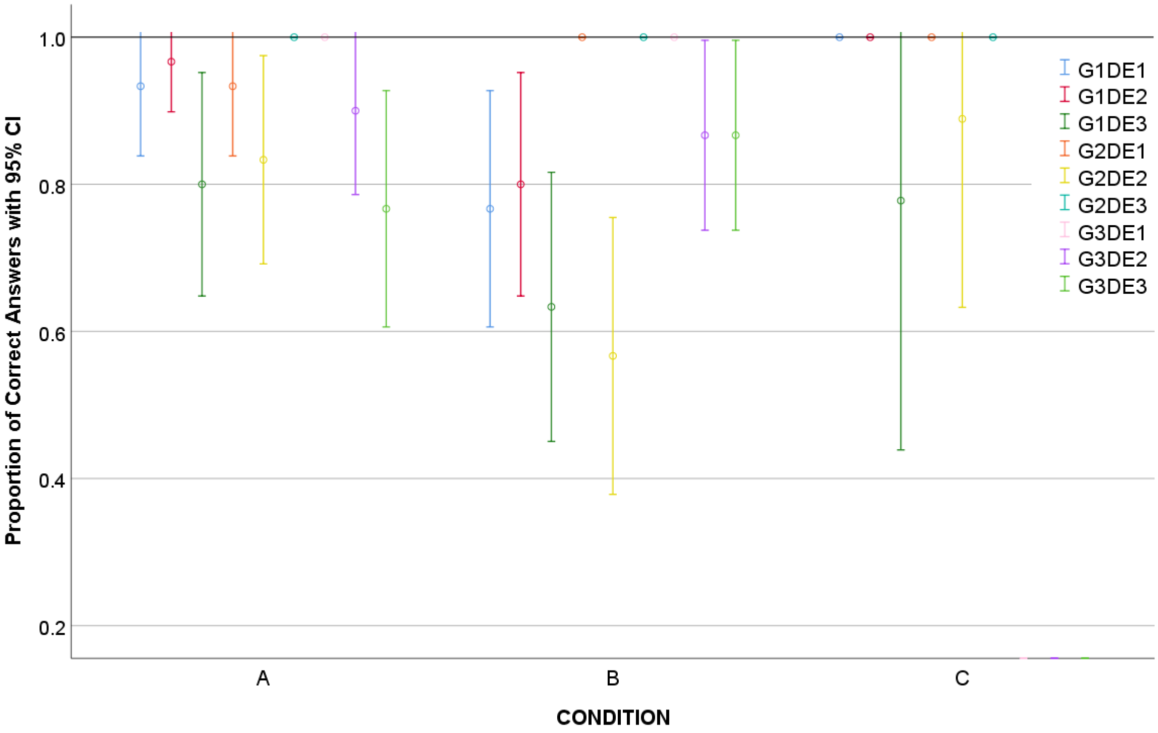


Figure 10. Accuracy by question by condition in proportion of correct answers with 95% confidence intervals shown. C condition did not include graph 3. Data is coded by graph number (G1) and data extraction question number (DE1).

Table 3

Mean Data Extraction Accuracy (out of 3 possible)

CONDITION		Graph 1	Graph 2	Graphs 1-2	Graph 3	Graphs 1-3
A	Mean	2.70*	2.77	5.47	2.60	8.07
	Std. Dev.	.702	.430	.860	.563	1.26
B	Mean	2.20	2.57	4.77	2.43	7.20
	Std. Dev.	.805	.504	.858	.504	1.06
C	Mean	2.78	2.89	5.67		
	Std. Dev.	.441	.333	.500		
Total	Mean	2.49	2.70	5.19	2.52	7.38
	Std. Dev.	.760	.464	.896	.537	1.34

Note. The Ccondition did not include graph 3.

*Indicates significant A/B group difference, $p < .05$.

In this study, efficiency was measured with total time spent by each participant on each question via the Time to First Mouseclick metric in the Tobii software. Participants voluntarily advanced through the questions, so efficiency was highly variable between individuals (see Table 4). In general, there were very few statistically significant differences between groups. Question-by-question, two questions took participants significantly more or less time by condition including the second question for graph 1, *F*(2, 66) = 10.477, *p* < .001. Tukey HSD post hoc comparisons revealed that participants in the B group spent significantly more time on that question (*M* = 37.9, *SD* = 24.1) than participants in either the A (*M* = 17.5, *SD* = 10.5) or C groups (*M* = 21.4, *SD* = 7.47). The B group participants also took more time than the A group participants on the first question for graph 3, *t* (36.75) = 2.43, *p* = .020. When questions are compiled into composite scores by graph, groups used statistically equal time aside from the B group spending more total time on graph 1, *F*(2, 66) = 4.21, *p* = .019 (see Table 4).

Table 4

Mean Time Spent on Questions by Graph by Group (seconds)

CONDITION		Graph 1	Graph 2	Graphs 1-2	Graph 3	Graphs 1-3
A	Mean	74.06	63.63	137.7	79.49	217.2
	Std. Dev.	31.10	24.71	41.22	27.63	54.73
B	Mean	99.10*	61.82	160.9	87.52	248.4
	Std. Dev.	40.97	17.88	46.71	54.01	82.62
C	Mean	75.78	54.06	129.8		
	Std. Dev.	22.79	18.86	24.85		
Total	Mean	85.17	61.60	146.8	83.51	232.8
	Std. Dev.	36.62	21.16	43.48	42.73	71.25

Note. The C condition did not include graph 3.

* Indicates significant A/B group difference, *p* < .05.

Lastly, usability is also characterized by user satisfaction with the product or interface. In this study, a combination of satisfaction questions during the computer activity and the ranking activities afterward were completed to measure participant satisfaction. These quantitative results are supplemented by qualitative data about participants' perceptions of the graphs and features of the graphs. From the computer activity Likert-style satisfaction questions (one per graph), there were no significant differences between composite scales or individual graphs (see Table 5). This stage of the activity is the only point at which A group members were rating only the A graphs, etc. Overall participant satisfaction at this stage was correlated with overall performance (effectiveness) (Spearman's ρ = .249, p = .039), but there was no significant relationship with overall time spent (efficiency) on the questions. For graph 1 alone, satisfaction was correlated to both effectiveness (ρ = .277, p = .021) and efficiency (ρ = .293, p = .015). Graph 3 satisfaction was related to efficiency (ρ = .249, p = .039) but not effectiveness. No such relationships exist for graph 2.

Table 5

Computer Activity Satisfaction Ratings of Graphs by Group (4-point Likert)

CONDITION		Graph 1	Graph 2	Graphs 1-2	Graph 3	Graphs 1-3
A	Mean	2.77	3.20	5.97	2.37	8.33
	Std. Dev.	.626	.664	.890	.765	.922
B	Mean	2.73	3.10	5.83	2.67	8.50
	Std. Dev.	.583	.885	1.12	.758	1.41
C	Mean	3.00	3.00	6.00		
	Std. Dev.	.866	1.12	1.12		
Total	Mean	2.78	3.13	5.91	2.52	8.10
	Std. Dev.	.639	.821	1.01	.770	1.43

Note. The C condition did not include graph 3. No significant differences.

In the ranking activity after the eye-tracking retrospection, each participant first compared the graphs pairwise for satisfaction (A/B or A/C). At this stage, the redesigned versions of graphs 2 and 3 (graph B for groups A and B, graph C for group C) were ranked higher for ease of use (satisfaction) by most participants (see Table 6).

Table 6

Pairwise Satisfaction Ranking of Graphs by Group (Proportion Redesigned Higher)

CONDITION		Graph 1 Ease	Graph 2 Ease	Graph 3 Ease
A	Proportion B	.50	.93	.64
B	Proportion B	.45	.80	.73
C	Proportion C	.56	.67	
Total	Proportion B + C	.49	.84	.69

Note. The C condition did not include graph 3.

Later in the interview, each participant also ranked all 6 graphs (A and B conditions) or 4 graphs (C condition) that they had been presented with. The results for all-graph ranking are shown in Table 7. Out of all A- and B-group participants, graph 2B was ranked significantly higher than 2A ($Z = 4.37$, $p <.001$) and 3B was ranked higher than 3A ($Z = 3.16$, $p = .002$). There were no significant differences between the A and B groups, i.e., participants who first saw the A graphs did not rank them any higher or lower in the final satisfaction ranking activity than those who first saw the B graphs. The C group did not rank either of the graph 1 or graph 2 significantly differently by design. Because the C group saw only 4 graphs rather than 6, the total rank points possible for that group was lower and therefore cannot be compared to the other groups.

In the qualitative data, most codes concerning understanding and satisfaction, though emergent, described various features of the graphs, related both to the information itself and features of the presentation of the information. Different features were associated with ease or difficulty (high or low satisfaction) for different graphs (see Table 8). Because the interview format pivoted around the comparison of the graphs, features of each graph were often described as being similar or in opposition to the same graph of an alternate design.

Table 7

Satisfaction Ranking of Graphs by Group (rank points)

CONDITION		1A	1B	2A	2B	3A	3B	1C	2C
A	Mean	3.55	3.34	3.97	5.00*	2.17	2.97		
	Std. Dev.	1.90	1.72	1.18	1.04	1.49	1.48		
B	Mean	2.77	2.88	4.04	5.04*	2.42	3.85*		
	Std. Dev.	1.42	1.88	1.31	1.04	1.17	1.85		
A+B	Mean	3.18	3.13	4.00	5.02*	2.29	3.38*		
	Std. Dev.	1.72	1.80	1.23	1.03	1.34	1.71		
C	Mean	2.00		2.56				2.56	2.89
	Std. Dev.	1.12		0.73				1.42	1.17

Note. The C condition did not include graph 3 and only the C condition included the C graphs. Because of this, the C group maximum rank is 4 while A and B maximum rank is 6, so these values are not directly comparable.
* Indicates significant A/B graph difference, $p < 0.05$

The actual data representation was by far the most difficult feature for graph 1 and especially graph 1B. Participants were completely unfamiliar with the point estimate and error design of graph 1B, and many answered the data extraction questions using the length of the error line. According to their feedback, this was likely due to the greater visual salience of the error line, association with more common graphs like bar and line graphs, and lack of brackets on the ends of the lines to trigger recognition of error bars. Based on this feedback, the major adjustments made to create graph 1C included making the point estimates much larger and more salient and adding brackets to the end of the error bars. However, the error bars generally were unfamiliar, with several participants in the A condition also being confused by them or reading them as the primary data. Participants were also confused by the representation of time in the bar/point estimate format, as referenced in the example data for graph 1B difficulty in Table 8, likely also related to their familiarity with line graphs.

Graphs 2A and 2B were rated most highly for satisfaction at all stages of the study and very few participants had any difficulty understanding the contents of the graph. Instead, praise and criticism were primarily in opposition to the other of the two graphs, with different participants having different preferences for several features of the graph, including the separation of graphs (layout) and extra axis text (amount of information). Many participants had difficulties understanding the bottom decadal average bar, mostly due to lack of familiarity, but others preferred to use it rather than the high variation of values in the annual graph.

Table 8

Qualitative results: Satisfaction

Graph		Prominent Codes	Example
1A	Ease	Markers/data representation, organization, amount of information, units/axes, representation of uncertainty, complexity	"1A ends the bars at the middle point… It seems like more complete than this, because you have these little drop-offs on 1B, where it's like, 'what about all this data in between 0 degrees to positive .5?'" (p104)
	Difficulty	Markers/data representation, organization, representation of uncertainty, comparisons to line graphs	"Natural forcings, I didn't realize that, since there wasn't a color there, that there's actually substance to these things…It went over my head that [they] were just brackets, that could be possible." (p073)
1B	Ease	Use of color, organization, markers/data representation, observed warming, precision	"1B is a bit simpler in design, but also the color coding is super nice. It's like orange, me, green, not me." (p080)
	Difficulty	Markers/data representation, observed warming, organization, amount of information, familiarity, comparison to line graphs	"I didn't understand what these lines meant. Where the starting and ending of the changes occurred. I only read the temperature, I only know how to interpret it, I guess." (p103)
2A	Ease	Familiarity, trends/values of data, organization, amount of information, markers/data representation	"It may just be because I'm familiar with line graphs the most, but it's, I think, also just a very simple two-axis graph. It's easier to understand." (p081)
	Difficulty	Organization, units/axes, amount of information, use of color	"2A is a little bit more confusing because the graphs are like mashed together. It seems a little bit like the line dividing them could be an axis, so this could be positive and this could be negative." (p105)
2B	Ease	Amount of information, organization, use of color, units/axes, familiarity, trends/values of data	"I think 2B, the yearly average because this is so easy to visualize. You look at that and there's clearly a spike. Honestly, we could just do away with all these and have 2B. That gets across the point." (p090)

2B	Difficulty	Trends/values of data, markers/data representation, use of color, amount of information, precision	“Probably 2B because it was like scribble-scrabble. I was trying to figure out, what do I look at the most? Which point do I look at, or which color do I look at?” (p056)
3A	Ease	Organization, amount of information, use of color, markers/data representation, data salience, precision	“Seems more straightforward, I guess. Because it labels things more clearly. I like how it breaks apart the compounds by color… seems easier to get all your information instead of having to jump around B.” (p102)
	Difficulty	Amount of information, use of color, markers/data representation, language, organization, precision	“A lot of words in different colors and numbers. You had a lot of small print that you had to read. It has a lot of different sections… A lot of information, almost too much information in one graph to handle.” (p102)
3B	Ease	Amount of information, organization, use of color, precision, language, key	“The overall trend in 3B that’s graphed for halocarbons is a little bit more easy to understand because I don’t have all of the extra shown up there. If I do want that information, it’s over here on the right for me, which is nice.” (p083)
	Difficulty	Amount of information, organization, language, markers/data representation, units/axes	“You have to look more at the little numbers next to everything. If I was just looking at 3B I wouldn’t pay attention to any of this part because it doesn’t look like it has anything to do with the graph at all.” (p073)

Graphs 3A and 3B were described as difficult by many participants. Though the markers of 3B were highly similar to 1B, many participants used the numerical data instead of the visual representation to answer questions, and therefore did not report difficulty with the data representation. Instead, participants were generally overwhelmed and unsure how to handle the vast amount of information presented on both graphs and drew little meaning from the language used such as “radiative forcing” and various chemical compounds mentioned. The use of color in both, primarily graph 3B, helped make a connection to warming and cooling and participants were familiar with carbon dioxide and methane, but otherwise it was difficult for participants to draw meaning from the data.

There were strong relationships between understanding (satisfaction) and the topics of risk and credibility. In general, higher understanding was associated with higher risk and credibility, because participants felt able to consider these topics only after understanding and being able to evaluate the content of the graph. However, some participants also had strong associations between risk or credibility and a lack of understanding of the graphs. The emergent code for this phenomenon was, "I don't get it, therefore". A lack of understanding or complexity of the graphs was associated with scientists and credibility overall, described by participant p090:

> "…my gut tells me that this one looks more trustworthy, because I don't know what it's saying…B looks more like scientists would make or use it? You obviously expect them to know what they're doing, and I would expect not to know what they're doing because I don't have a background in science… If someone explained brain surgery to me, I would expect not to know at all what they're talking about, but I would expect it to be credible because they know what they're talking about if it's a doctor that is telling me..."

This participant describes a relationship between author expertise and the product they create, and another participant, p074, describes a potential cause when they talked about scientists as communicators:

> "If I look at it and don't understand it, just because the information in it is probably going to be more credible, or if scientists would just write an article or themselves, as scientists, they wouldn't think about making it simple for non-scientists to read it… They're thinking about, 'this is easy for me, because I know what this is, I know what it's saying.'"

Risk assessment of climate change, though more strongly related to high understanding of the graphs, was also important for participants' perceptions of difficult graphs, for various reasons. Participant p089 describes the difficulty itself as worrying, as well as specific features:

> "The amount of information and the difficulty to understand stresses me out. Also, the fact that they use radiation and emissions more than just warming… If it's more difficult to understand, I think I just assumed it'd be bad. Whatever they're talking about is so hard, I can't even read it, and it must be bad because these bars are long. Then with B I can read it and I can tell, ok, I know exactly how much CO2 radiation there is, or how much is emitted. It worries me less that I can understand it."

More often, however, greater understanding was related to higher risk, as the participants could draw meaning from the graphs and connect it to global warming, the likely reason that graphs 2A and 2B were rated most worrisome on average.

Attention

On average, the eye-tracker captured 92.0% of participants' weighted gaze samples (*SD* = 6.18%) and there was no significant difference between groups. Visual attention to the graphs is measured by the eye-tracker both in terms of fixation count within an AOI and total duration of those fixations. Typically, and for this study, these measures are highly correlated (see Figure 11). Because they are highly correlated, this study uses total fixation duration as the primary attention metric for subsequent analyses.

Figures 12, 13, and 14 show the differences between the A and B condition total fixation duration (in seconds) for each AOI with a one-to-one reference line. In each of these graphs, each data point represents an AOI from one of the data extraction questions posed to participants. If a point is above the one-to-one reference line, it means that AOI has a higher B group duration, and therefore that participants in the B condition extraction questions asked in the activity spent more time on average viewing that AOI. The Y-distance to the reference line, then, would show how many seconds longer the B group viewed the AOI. Points below the line were more attended to by the A group by the X-distance more seconds.

As shown in Figures 12, 13, and 14, there were several statistically significant differences between groups for view time in specific AOIs. During the data extraction questions for graph 1 (Figure 12), the B group paid significantly more attention to the possible answers to questions 1 and 2, $p = .019$, $p < .001$, and the data, $p < .001$, x-axis, $p < .001$, and correct answer, $p = .017$, AOI for question 2 (all Tukey HSD post hoc for significant ANOVA). The B group also spent longer in one of these same AOIs than the C group (not shown graphically), specifically the question 2 data, $p = .028$. Both the A, $p < .001$, and B group, $p = .003$, spent less time viewing the question 3 correct answer AOI than the C group. View times for graph 2 were more equal (Figure 13) with only a few significant A/B contrasts, specifically in time spent viewing the question 1 y-axis (A longer, $p = .010$) and the question 2 annual data (B longer, $p = .028$). However, the C condition paid significantly more attention to the decadal data during questions 1 and 3 than either the A, $p = .005$, $p < .001$, or B conditions, $p = .018$, $p = .013$. When answering the data extraction questions for graph 3 (Figure 14), the B group paid more attention to the correct answer and y-axis to question 1, $p = .002$, $p = .004$, and the title, y-axis, and question text for question 2, $p < .001$, $p = .011$, $p = .018$. The A group paid more attention to both the x-axis for question 2, $p = .004$, and the data for question 3, $p < .001$.

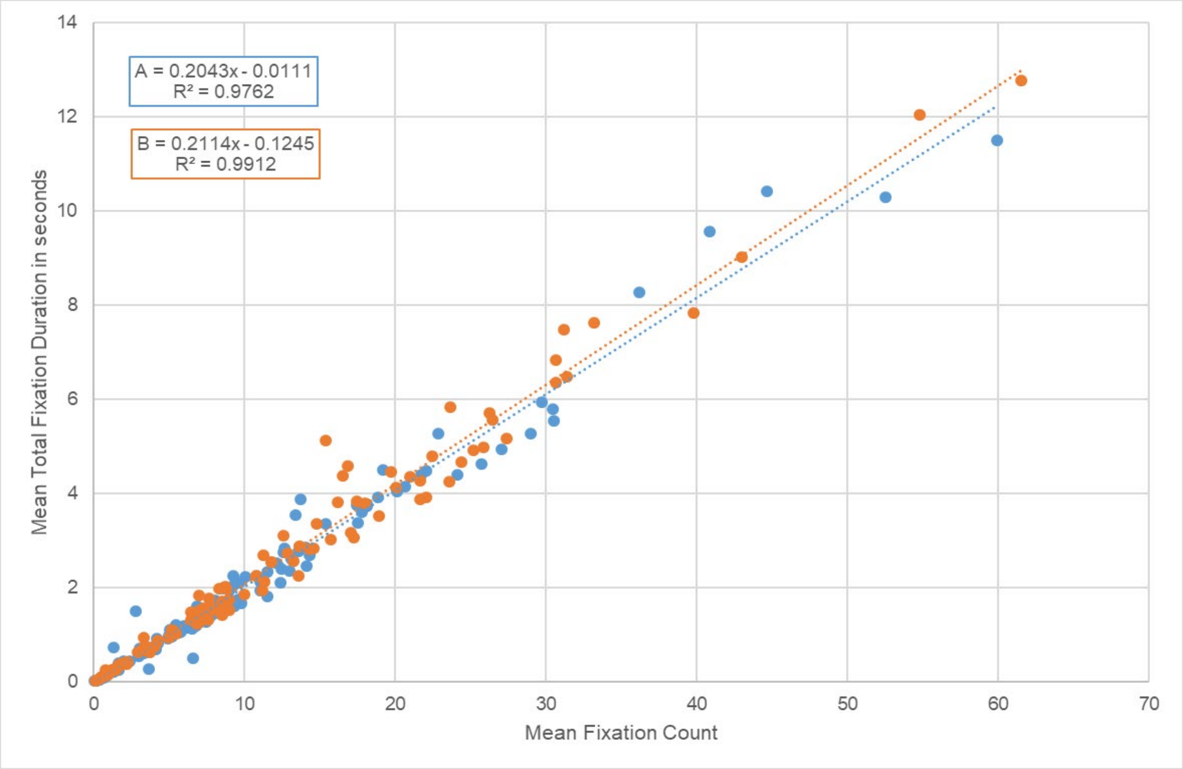


Figure 11. Mean fixation count and total fixation duration for each AOI by condition.

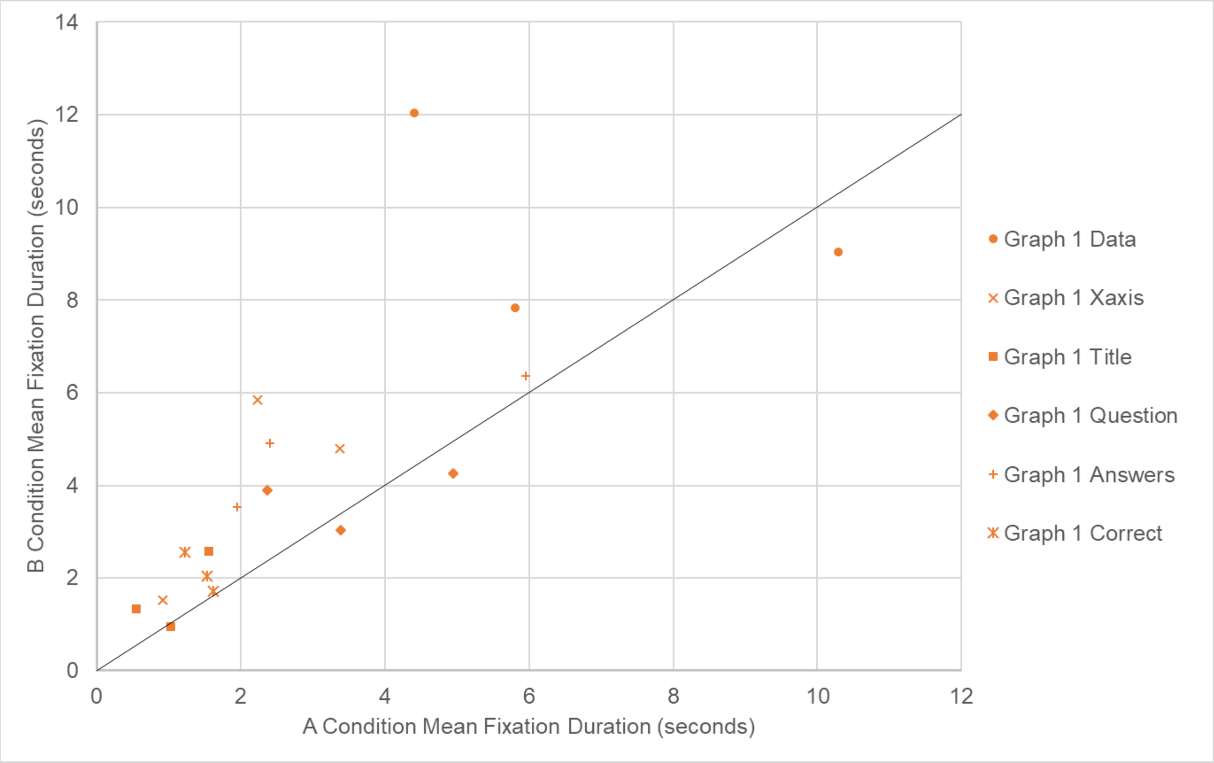


Figure 12. Comparison of A and B fixation duration by AOI type for graph 1 questions

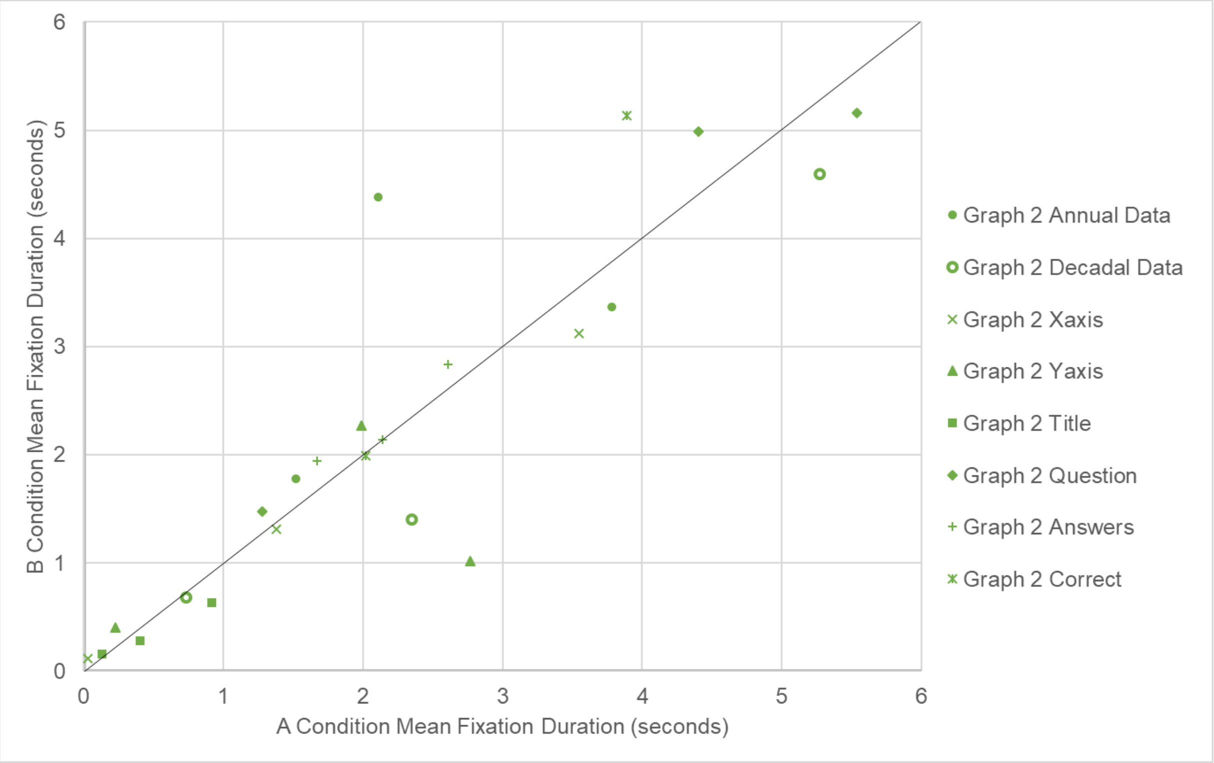


Figure 13. Comparison of A and B fixation duration by AOI type for graph 2 questions

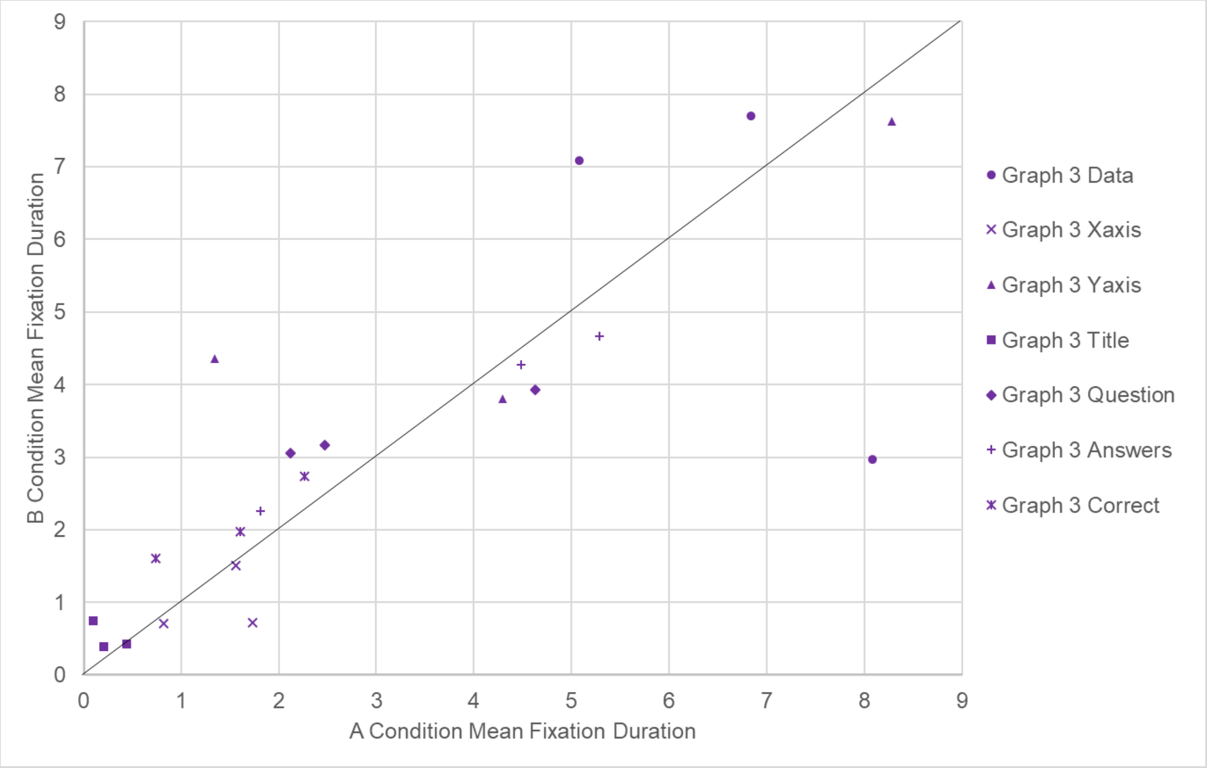
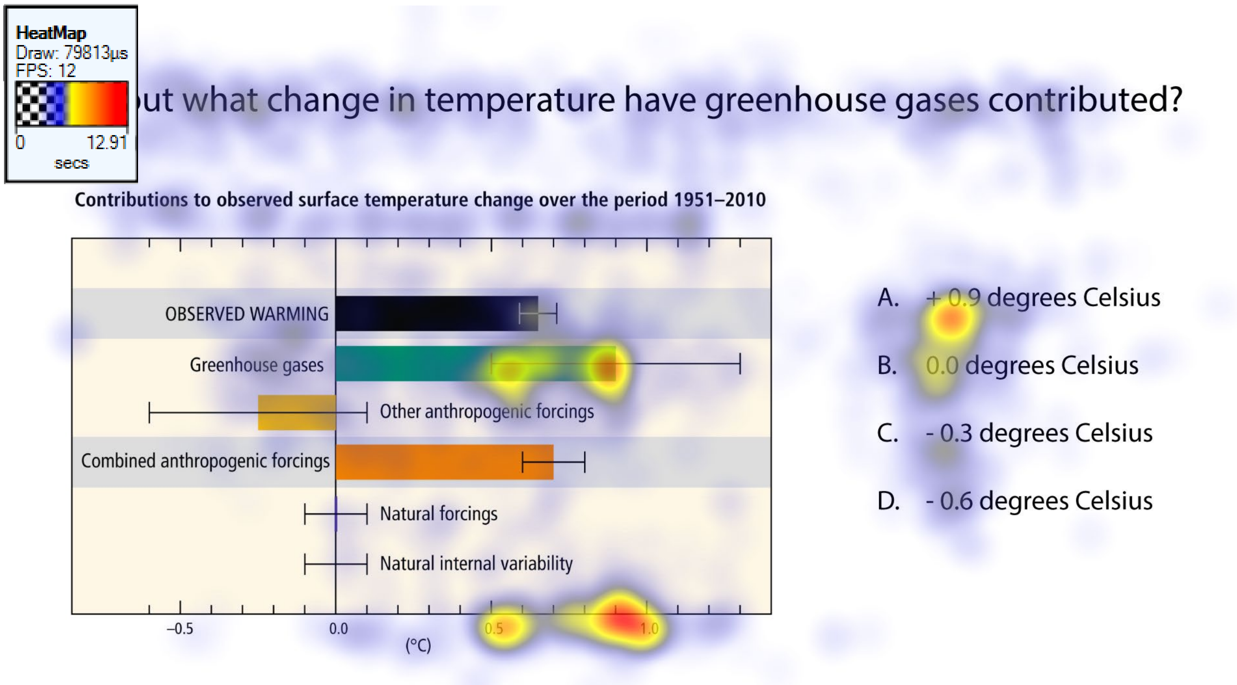


Figure 14. Comparison of A and B fixation duration by AOI type for graph 3 questions

Overall, when answering data extraction questions for graph 1, the group using the B graphs spent significantly more time than group A on the data, $p = .038$, the x-axis, $p = .001$, and the answers, $p = .007$. On graph 2, groups A and B were equivalent, but B spent longer looking at the potential answers than the group viewing the C graphs, $p = .028$. During the questions for graph 3, group B paid more attention to the y-axis on graph 3B than the A group did on graph 3A, $p = .016$. The B group also paid more attention to the AOI drawn around the data concerning the correct answers for the graph 3 data extraction questions, $p = .011$, however, the correct answer AOIs are extremely irregular by question and by graph so any differences should be interpreted with caution.

As reported above, graph 1, especially question 2 of graph 1, had the greatest group discrepancies in view time, both overall and for specific AOIs. Figures 15 and 16 show the A and B group heat maps for that question, respectively. In graph 1A, attention was focused mostly on the question-relevant point estimate and the lower end of the error bar, as well as the corresponding values on the lower x-axis. Those same features are the most-viewed areas of graph 1B as well, however, because there was a significantly longer total view time, the colors of the heatmap are weighted to represent longer durations (see figure keys, upper left corners). These observations align with the measured significant difference in attention to the data, x-axis, and possible answers noted above. In the Figure 16 of heatmap of Graph 1B, there is also greater apparent attention paid to irrelevant data (observed warming and combined anthropogenic), the y-axis labels, and multiple answer choices.


HeatMap
Draw: 79813µs
FPS: 12
0
12.91
secs
Contributions to observed surface temperature change over the period 1951–2010
OBSERVED WARMING
Greenhouse gases
Other anthropogenic forcings
Combined anthropogenic forcings
Natural forcings
Natural internal variability
−0.5
0.0
0.5
1.0
(°C)

ut what change in temperature have greenhouse gases contributed?

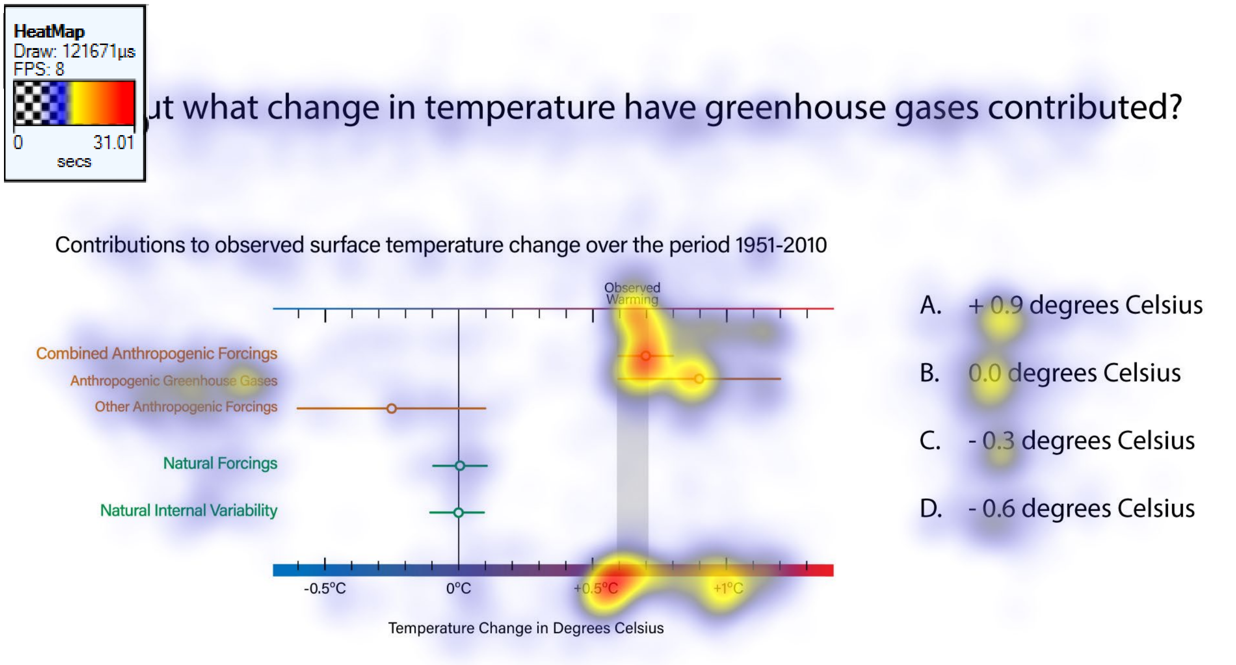
A. + 0.9 degrees Celsius

B. 0.0 degrees Celsius

C. - 0.3 degrees Celsius

D. - 0.6 degrees Celsius

Figure 15. Heatmap (absolute duration, seconds) of data extraction question 2 for Graph 1A


HeatMap
Draw: 121671µs
FPS: 8
0
31.01
secs
Contributions to observed surface temperature change over the period 1951-2010
Observed Warming
Combined Anthropogenic Forcings
Anthropogenic Greenhouse Gases
Other Anthropogenic Forcings
Natural Forcings
Natural Internal Variability
-0.5°C
0°C
+0.5°C
+1°C
Temperature Change in Degrees Celsius

ut what change in temperature have greenhouse gases contributed?

A. + 0.9 degrees Celsius

B. 0.0 degrees Celsius

C. - 0.3 degrees Celsius

D. - 0.6 degrees Celsius

Figure 16. Heatmap (absolute duration, seconds) of data extraction question 2 for Graph 1B

As shown in Table 9, there were also several significant differences between groups' first fixations in AOIs. On average, the A group fixated more quickly on the data of graph 1, p = .010, and the data, p < .001, and title, p = .014, of graph 3. The B group fixated on the x-axes, p < .001, and title, p = .019, of graph 2. However, the mean TtFF is not calculated for those who did not fixate in the AOI at all, so it is important to note how many users fixated in the AOI during the questions at all. Specifically, for each AOI with significant TtFF differences, an equal number or more B group participants fixated in the AOI for more questions (higher counts).

Table 9

Mean Time to First Fixation (seconds) and Use of AOIs (count), Data Extraction Questions

	Graph 1		Graph 2		Graph 3	
	Mean	Count	Mean	Count	Mean	Count
Graph A Data	1.21*	89	.482	90	.519*	90
Graph B Data	2.19	90	.576	90	1.32	90
Graph A X-axis	9.09	82	6.51	61	12.6	72
Graph B X-axis	8.13	87	4.20*	90	10.3	66
Graph A Y-axis	N/A		8.61	62	5.95	81
Graph B Y-axis			7.28	67	5.01	88
Graph A Title	2.08	77	6.65	33	7.06*	39
Graph B Title	2.98	78	2.63*	38	12.1	58

Note. Maximum count of 90 (30 participants across 3 data extraction questions). Count of less than 90 indicates at least one user did not fixate within the AOI during at least one question.
* Indicates significant A/B difference, $p < .05$

Time to first fixation is an especially relevant metric for participants' first exposures to each graph. As a reminder, each graph was first shown to each participant on its own, with no task or questions, for an unlimited time until the participant advanced the activity. There were no TtFF differences for graph 1, but when shown graph 2, the A group fixated more quickly on both the annual, p = .029, and decadal data, p = .016, and the B group fixated more quickly on the y-axes, p = .003. When viewing graph 3, the A group viewed the atmospheric driver data more quickly than the B group (*p* < .001). Participants are most likely to attend to more salient features first, including larger objects (Harold et al., 2016), and the graph 2 differences all corresponded to notable differences in AOI size (33-43%). However, the graph 3 atmospheric driver AOIs were of approximately equal sizes and represent the only major data location difference among the graphs.

Credibility

Participant perceptions of credibility of climate scientists were measured before and immediately after the computer activity with an 18-item 7-point Likert-style instrument (McCroskey & Teven, 1999). This study affirmed the reliability of the instrument overall (pre-survey α = .934, post-survey α = .928), as well as the three sub-factors (competence, pre α =

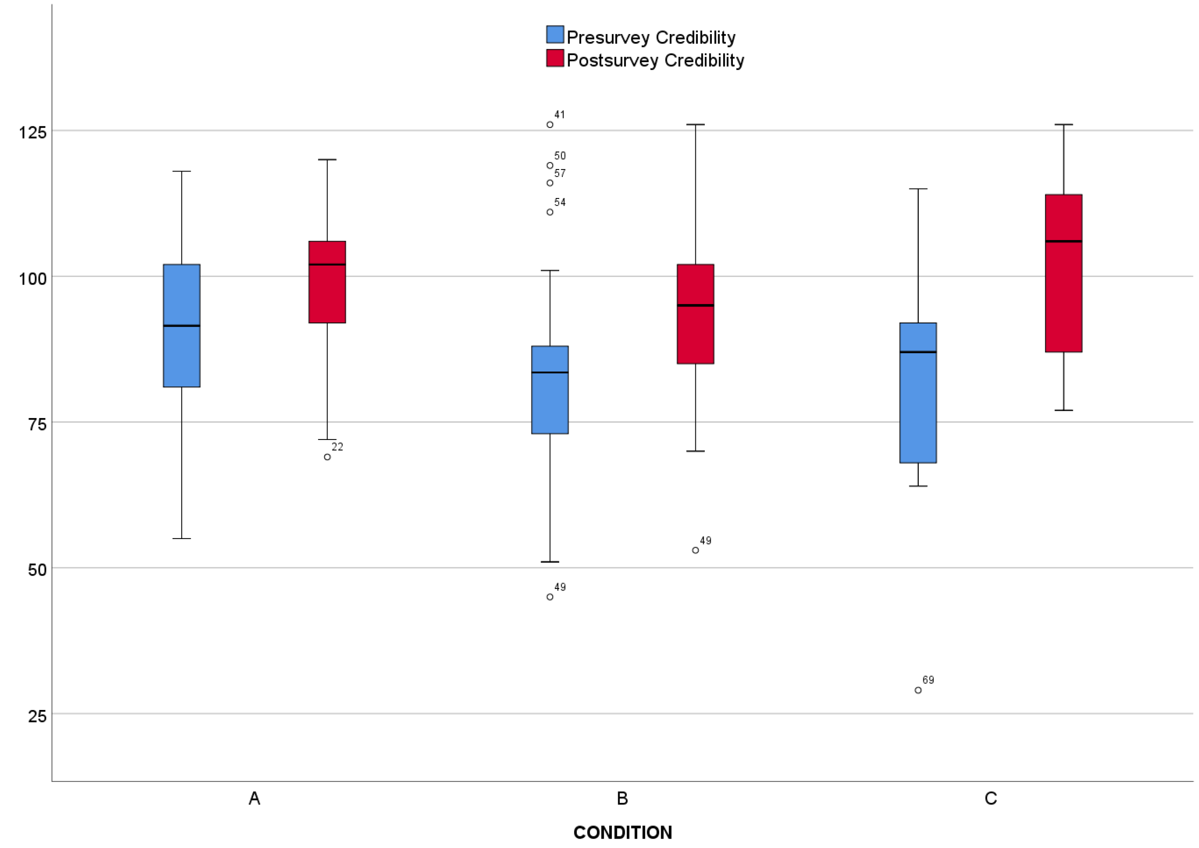


Figure 17. Boxplot of participant credibility ratings of climate scientists as measured by an instrument by McCroskey and Teven (1999) on the pre- and post-survey by condition. Maximum possible rating is 126.

.913, post α = .813; goodwill, pre α = .881, post α = .877; and trust, pre α = .916, post α = .922). In general, the participants had a significantly higher rating of the credibility of climate scientists after completing the activity than before, *F*(1, 66)= 39.31, *p* < .001. Effect size of this change varied by condition (A group Cohen's *d* = .58, B group *d* = .93, C group *d* = .87), however there was no significant interaction from condition, *F*(2, 66)= 1.95, *p* = .150 (see Figure 17). The trust sub-factor of the instrument had a significant interaction from group membership, *F*(2,66)= 3.75, *p* = .029, however, no Tukey HSD posthoc comparisons were significant.

Participant perceptions of graph credibility were measured at several points throughout the experiment. During the initial computer activity, there were no significant differences between trust ratings of A/B/C equivalent graphs (see table 10). Participants' total trust rating of the graphs during the activity were significantly correlated with both their perception of satisfaction with the graphs, *ρ* = .575, *p* < .001, and their performance on the tasks, *ρ* = .321, *p* = .007. During the interview portion of the study, differences in judgements arose, with results showing that between 60% and 94% of participants in each group rated the redesigned graphs as more credible (see Table 11).

Table 10

Computer Activity Credibility Ratings of Graphs by Group (4-point Likert)

CONDITION		Graph 1	Graph 2	Graphs 1-2	Graph 3	Graphs 1-3
A	Mean	2.73	3.10	5.83	3.03	8.87
	Std. Dev.	.583	.481	.913	.615	1.38
B	Mean	2.53	3.13	5.67	3.03	8.70
	Std. Dev.	.629	.434	.844	.556	1.18
C	Mean	2.78	3.22	6.00		
	Std. Dev.	.441	.441	.707		
Total	Mean	2.65	3.13	5.78	3.03	8.42
	Std. Dev.	.590	.451	.855	.581	1.54

Note. The C condition did not include graph 3. No significant A/B/C differences.

Table 11

Pairwise Credibility Ranking of Graphs by Group (Proportion Redesigned Higher)

CONDITION		Graph 1	Graph 2	Graph 3
A	Proportion B	.63	.86	.93
B	Proportion B	.61	.87	.79
C	Proportion C	.94	.78	
Total	Proportion B + C	.66	.85	.86

Note. The C condition did not include graph 3.

Table 12

Credibility Ranking Results of Graphs by Group (rank points)

CONDITION		1A	1B	2A	2B	3A	3B	1C	2C
A	Mean	1.72	2.48	2.97	4.52*	3.72	5.59*		
	Std. Dev.	.922	1.38	1.30	.949	1.28	.946		
B	Mean	1.65	2.15	3.33	4.83*	3.81	5.23*		
	Std. Dev.	.977	1.05	1.43	.761	1.20	1.24		
A+B	Mean	1.69	2.33*	3.14	4.66*	3.76	5.42*		
	Std. Dev.	.940	1.23	1.36	.872	1.23	1.10		
C	Mean	1.44		2.33				3.00*	3.00
	Std. Dev.	1.01		1.12				.866	.667

Note. The C condition did not include graph 3 and only the C condition included the C graphs. Because of this, the C group maximum rank is 4 while A and B maximum rank is 6, so these values are not directly comparable.
* Indicates significantly A/B/C difference, *p* < .05

In the whole-group ranking activity, summarized in Table 12, the A group participants did not rank any of the graphs significantly differently than the B group. Graphs 1B, 2B, and 3B were all ranked significantly higher than the corresponding A versions, $Z = 2.38$, $p = .017$; $Z = 4.99$, $p < .001$; $Z = 4.811$, $p < .001$. Graphs 2B and 3B were ranked higher in both the A and B groups separately as well, but within only the A and B groups both designs of graph 1 were ranked statistically equally. In group C, graph 1C was ranked higher than 1A, $Z = 2.11$, $p = .034$. Within the A and B groups, the ranks of 2A were also significantly higher than 1B, $Z = 2.60$, $p = .009$, 3A higher than 2A, $Z = 2.28$, $p = .022$, and 3B higher than 2B, $Z = 3.22$, $p = .001$.

Therefore, there is a significant difference between every graph pair in order of mean and median ranks, so these participants seemed to agree on the trustworthiness order of these graphs, from 1A, 1B, 2A, 3A, 2B, finally to 3B.

Qualitative results indicate a variety of relationships between graph design and perceptions of credibility of graphs and their creators. Table 13 describes the most notable codes from participant descriptions of credibility as it pertains to design of particular graphs and communication of information more generally. Codes and codes commonly associated with them by participants are listed roughly in order of prominence in the data.

During the retrospective eye-tracking interview, when discussion of credibility first came up, many to most of the participants expressed frustration at the lack of a source or citation by which to judge credibility. However, for subsequent questions and sections, the vast majority of participants instead expressed their judgements of credibility based on aesthetics and the communication style and success of the graphs, though with occasional discomfort or lack of confidence in those judgements. Generally, aesthetic judgements of the graphs concerned associations with intellectual authorities such as scientists, teachers, professors, or other academic materials such as textbooks. These judgements didn’t usually correspond to any particular traits of those authorities that would deem them trustworthy, but some participants referenced potential authors as being well-informed (“know what they’re talking about”) or the graph being informed by more quality or plentiful data.

Many participants also cited the implied effort in creating the graphs as a sign of trustworthiness. In graphs, this was evidenced to participants by the presence of more thorough axis labeling, color-coding, and other features that were typically present in the B but not A graphs and were therefore salient in comparisons. Some participants relayed that increased effort in graph design was indicative of credibility because of the author's likely passion for the subject, and therefore likely high education in that field. More often, though, effort to create the graph was related to the author's desire and intent to communicate the subject effectively and was therefore tied to goodwill.

Table 13

Qualitative results: Credibility

Code (*emergent)	Example	Associated codes
*Official, professional, fancy, etc.: participants use descriptors that imply an association with authority or advanced status, usually based on aesthetics	"…but this one does look more trustworthy because it's more official… It's what we're all taught, to trust your, I don't know, the people in charge of you, the official people. You trust them." (p094) "Just some more complex graph would make it more credible. It doesn't make sense, but like, the more complicated it looks, the more that I think it has a fancy purpose, where I guess it's made by someone high up." (p070)	Aesthetics, organization, understanding, amount of information, science
Competence of author: referring to competence factor of McCroskey and Teven (1999); referring to author's education, experience, intelligence, etc.	"It's got the most, both of the information of course, it's the most thorough, it's the most well-presented, and it's the most clear. I think the mark of expertise is to get across what you are talking about to someone who knows nothing about what you are talking about, and I think this accomplishes it very nicely." (p076)	Amount of information, understanding, science, research, goodwill, took time/effort
*Took time/effort: References to an author putting more time, effort, or thought into creating a graph, not necessarily the research behind it	"Like I said, this looks like they took more time to prepare, have more information on it…If they took more time to do it and they obviously researched into it. They didn't just find a graph, put it on. They took the time to look into it and make sure that it told everything they wanted to tell." (p103)	Amount of information, understanding, organization, goodwill, use of color
Goodwill/intentionality: inspired by the goodwill factor of McCroskey and Teven (1999); references to the caring or sensitivity of the author, intentionality in communication	"It looks like they just put more effort into it because they wanted to actually do what they created the graph to do. In order for you to understand it better… if I knew for a fact that everything I'm putting on this graph is right, I would want it to be…very clear and make sure that everyone looking at it could understand exactly what it is." (p073)	Understanding, amount of information, took time/effort, organization, competence

*Academia/tests/textbooks: references to graphs or similarities to graphs in academic settings, including standardized tests	“Obviously, anything presented within a classroom by a professor, I’m going to believe. If I do trust something initially, there’s no need to make it seem more trustworthy as long as the information is easy to read.” (p099)	Aesthetics, understanding, familiarity, science, prior knowledge, organization
*Anyone/child/I could make this: verbatim or other references to a graph potentially being made by a non-professional, references to school or children	“The reason 1A is last is because the colors make it look more childish. Even though the bigger font might make it easier to read, it makes it also more childish, which I would find less credible.” (p086)	Use of color, complexity, aesthetics, organization, amount of information
*Research amount or quality: explicit comments that a graph must have been founded on more or better research or data	“I guess if you have more data, you’re going to think it’s more trustworthy because it has things to back it up. It might not be actual data but it’s more information than the other one had.” (p095)	Amount of information, competence, took time/effort, precision
*Science: related to perceptions of science as an institution, scientists as a type of person, scientific traits or appearances	“Trustworthy? Probably 1A because it looks like a scientist made it… This sounds bad, but…It’s not as appealing and sometimes I see scientific papers or graphs and it’s like, ‘Oh, that’s not very appealing, I really don’t want to look at that’.” (p080)	Aesthetics, academia, official, competence, amount of information
Honesty: drawn from the trustworthiness factor of McCroskey and Teven (1999); having to do with overall honesty and bias in graphs	“I feel like if you are not able to present your argument or your evidence in a way that is easy for someone else to understand, it makes you wonder, like, ‘are you trying to dupe me?’” (p090)	Amount of information, understanding, organization, precision, use of color

In comparing individual graphs, there were not major differences in which factors were important for assessing credibility (see Table 14). Generally, the amount of information, both data and context, was recurrent as an important criterion for participants to judge credibility. Amount of information was applied as a single code across data and contextual information because participants generally did not or could not distinguish between the two. For example, both graphs 1A and 1B were often judged untrustworthy because of their simplicity and scarce data, but graph 2B was often ranked above 2A because of the additional contextual data such as the additional axis labels and zero Y-axis intercept. For graphs 3A and 3B, participants were generally overwhelmed by the plentiful data, though amount of contextual information such as the full names of chemical compounds on 3B were sometimes important. There was the greatest divergence of judgement of credibility between and concerning graphs 3A and 3B, often manifesting as dissonance between the amount of information presented (associated with high credibility) and the difficulty to understand the information in either or both (associated with low credibility).

Table 14

Qualitative results by graph: Credibility

Graph	Prominent Codes	Example
1A	Aesthetics, understanding, anyone/child/I could make this, use of color, complexity	"1A is, I guess, childish. I don't know. It's very simple. Although I do trust it, maybe not as much as if it was more professional-looking." (p099)
1B	Understanding, aesthetics, organization, official/professional, amount of information, complexity	"Then 1B, I think is actually more credible... Again, since it's more complicated to me it seems like it's a higher level of knowledge than the one I have, which seems more credible." (p084)
2A	Amount of information, understanding, organization, honesty, anyone/child/I could make this, aesthetics	"Because this one seems like something I would draw if I were trying to draw a graph really quickly and get out of the lab. It's not bad. It's got all the information there in a more-or-less understandable way. It's just, there's extra information that's not really talked about." (p083)
2B	Amount of information, understanding, took time/effort, research, honesty, organization	"2B because it's easy to understand. The deal with trustworthy is, if it's easy to understand, it's not trying to hide anything." (p067)
3A	Amount of information, use of color, understanding, aesthetics, took time/effort, competence, goodwill	"It looks more accurate-looking because it's numbers. It's not trying, not to help you out but it's raw numbers, the raw facts, and raw information without even that description, per se." (p056)
3B	Amount of information, understanding, organization, official/professional, goodwill, aesthetics, academia	"Whoever made 3B really took into consideration who would be looking at the graphs. There would probably be people who really don't read graphs and who can easily find the information they need just looking at a table." (p061)

Risk Perception

Participants' overall risk perception of climate change was measured in the pre- and post-surveys using several items from Leiserowitz et al. (2019) and frequently used by those authors. In our study, these items had scale reliabilities of α = .785 (pre-survey) and α = .844 (post-survey). Participation in the activity significantly increased participant risk perception, within-subjects *F* = 36.0, *p* < .001. There was no significant effect from group membership, *F* = 1.206, *p* = .306, and the effect sizes for the A, B, and C groups were *d* = .66, .72, and 1.22, respectively (see Figure 18). Participation in the activity also led to significantly higher perceptions of climate scientist consensus around climate change on both continuous and ordinal items, *F*(1,66)= 9.20,

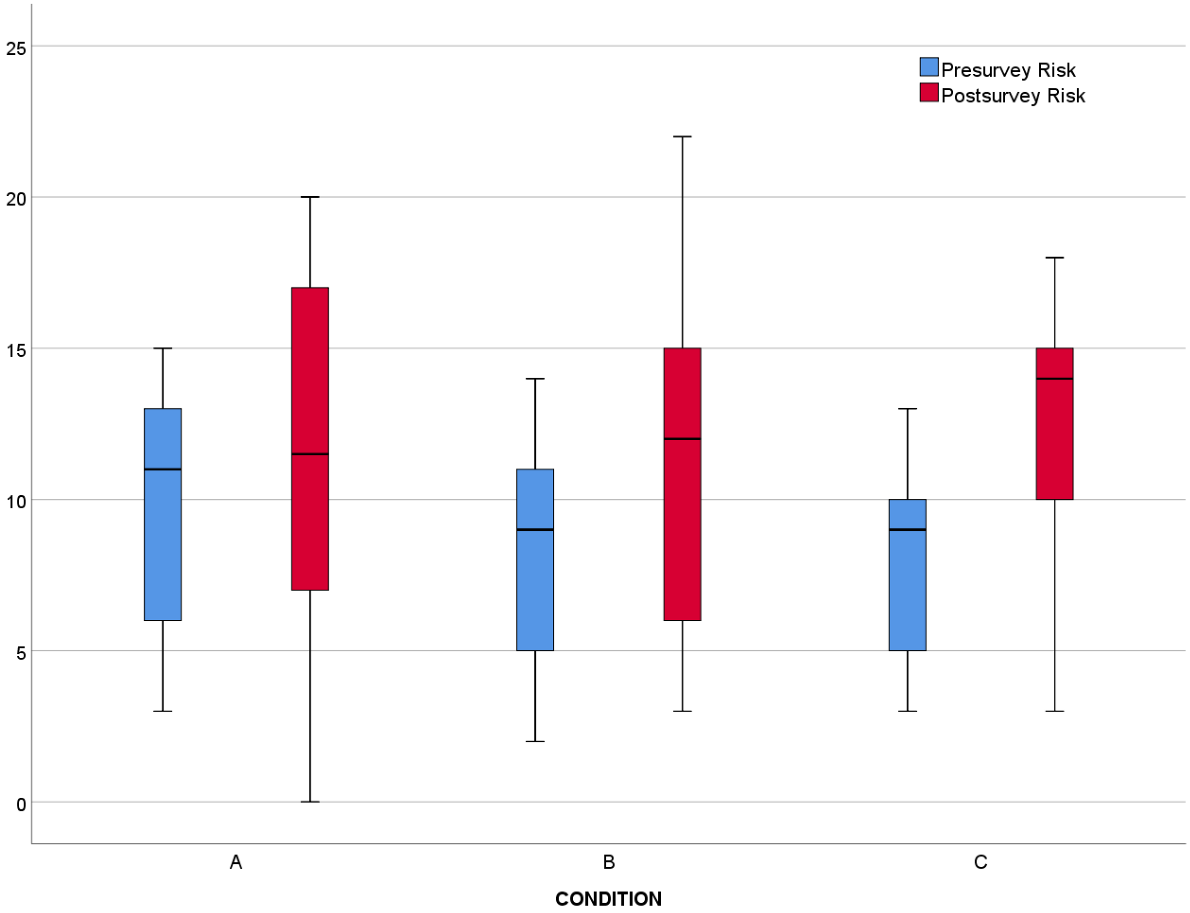


Figure 18. Boxplot of participant perceptions of risk from climate change as measured by items from Leiserowitz et al. (2019) on pre-survey and post-survey.

p = .003, and *Z* = 2.66, *p* = .008. For the continuous measure, group effect sizes were *d* = .27 (A group), *d* = .41 (B group), and *d* = 1.2 (C group).

Participants only compared graphs directly for risk in the final ranking activity (phrased as, 'which graphs make you the most to least worried about climate change?') and those results are shown in Table 15. Unlike satisfaction and credibility, participants did rank some graphs differently depending on their group membership. Specifically, group A ranked graph 1A higher than group B did, *Z* = 2.02, *p* = .043, and group B ranked graph 3B higher than group A, *Z* = 2.39, *p* = .017. In the A group, the B group, and both combined, graph 2B was ranked higher than 2A, *Z* = 3.98, *p* < .001; *Z* = 3.97, *p* < .001; *Z* = 5.60, *p* < .001. However, there were no other significant differences by design.

Table 15

Risk Ranking Results of Graphs by Group (rank points)

CONDITION		1A	1B	2A	2B	3A	3B	1C	2C
A	Mean	3.10	2.66	4.21	5.17*	3.19	2.67		
	Std. Dev.	1.42	1.10	1.34	1.28	1.75	1.78		
B	Mean	2.35	2.46	4.23	5.23*	3.02	3.67		
	Std. Dev.	1.38	1.41	1.23	1.27	1.60	1.38		
A+B	Mean	2.75	2.56	4.22	5.20*	3.11	3.15		
	Std. Dev.	1.44	1.25	1.28	1.26	1.67	1.67		
C	Mean	2.11		2.28				2.67	2.94
	Std. Dev.	1.27		1.03				1.12	1.07

Note. The C condition did not include graph 3 and only the C condition included the C graphs. Because of this, the C group maximum rank is 4 while A and B maximum rank is 6, so these values are not directly comparable.
* Indicates significant A/B difference, *p* < .05

Participant discussions of risk assessment did not inspire separate emergent codes. Participants sometimes pinpointed graph features in reference to risk assessment, but generally, risk was associated only with rising temperatures and occasionally the human causes of it.

Therefore, risk assessment was primarily related to the level of understanding of the content of the graphs, i.e. in what way each graph communicated climate change to the participant, if at all. Though there were some important exceptions as discussed earlier, greater understanding was generally associated with greater risk perception. Specifically, graphs 2A and 2B were ranked the highest, and 2B higher than 2A, which is related to high understanding of these graphs. Participants expressed this in terms of the content as well, often relaying that graph 2 most clearly expressed global warming. This is likely in part also because graph 2A and 2B showed warming itself, rather than causes and forcings, which the participants did not relate to worry. Participants also often attributed worry about graph 2 to the relatively long time span shown, which gave them more confidence in the reality of the trend. However, this may be at least a partial effect of their relative discomfort with non-line-graphs, since graph 3B was explicit that it technically covered a longer time period.

Low understanding and perceived simplicity of graph 1 was often referred to while explaining risk perceptions. Even for those participants who could understand the values of the graph, less worrisome meaning was drawn from them about the phenomena of global warming, again primarily because participants were not concerned as much by the causes of the warming. For graph 3, understanding and misunderstanding were more polarizing in the topic of worry. Some participants expressed that the complicated content and design of both 3A and 3B could be worrisome (see Table 16).

Table 16

Qualitative Results: Risk

Graph	Prominent Codes	Example
1A	Understanding, data salience/focus, trends/values of data, units/axes, time, use of color	"1A is the one that I've learned the most because you can see the natural forcings have almost no effect… Humans have the greatest effect. It just puts it in perspective, I guess." (p054)
1B	Understanding, trends/values of data, amount of information, units/axes, aesthetics	"It doesn't make me worry. If it gave me lines that went up, increasing or something, I would think, 'Ok, maybe this isn't good,' but they're straight lines, so it'd be staying the same." (p061)
2A	Trends/values of data, time, understanding, prior knowledge	"These two right here. These probably paint the clearest picture to me and it's not the prettiest of pictures. It tells us flat out this world is getting warmer. Things are melting. It's getting more tropical, which is not exactly a good thing." (p076)
2B	Trends/values of data, time, understanding, prior knowledge	"I put 2B as the most because it starts at 1850 so it has a lot of history. It goes into showing to almost modern times. It shows the trend. Trends can typically be pretty accurate." (p081)
3A	Understanding, amount of information, data salience, language, trends/values of data, use of color	"The reason why I chose 3A to be more worried about is because I do not know anything on that graph. I feel like if I did not know it, then it has to be something serious that only climate scientists knew. They were trying to tell us something." (p079)
3B	Understanding, amount of information, trends/values of data, language, data salience	"Probably 3B. Maybe that's just because there's so many different gases that they have included… it's hard to be able to see how that changes the temperature, but it still makes it look scary, just because they use fancy chemicals." (p082)

DISCUSSION

This study was explicitly targeting non-expert young adults who were previously unconcerned with climate change. Participating in the computer activity and interacting with climate change information significantly increased participant risk perception, scientist consensus estimate, and perception of credibility of climate scientists. The A group achieved higher accuracy overall, but there was variation in performance on individual questions. Because performance on each question did not co-vary, and because they were designed to test different features, considering the questions separately, and especially those two with a disparity in performance, may be valuable.

The computer activity questions indicated no satisfaction differences between graphs or conditions, however graphs 2B and 3B were consistently ranked higher than their A counterparts by both participant groups in the interview stage of the study. This rating discrepancy could be related to any of a myriad of factors, however it is reasonable that participant perceptions of each graph would have shifted over the course of spending 20-40 minutes engaging with the graphs more thoughtfully and thoroughly during the interview portion as opposed to the quick and task-based computer activity. For example, one of the participants' primary complaints about graph 1B was unfamiliarity, which would likely shift after engaging with the graph further. In interviews, higher satisfaction with 2B and 3B were attributed to similar features on both, namely the temperature color-coding, the separation of graph sections, and additional text defining abbreviations, adding context to axes, etc. It is encouraging that these differences are those explicitly encouraged by the guidelines set forth in Harold et al. (2016), suggesting that the guidelines may lead to improved user satisfaction.

Graph 2 and especially 2B were rated highest in satisfaction overall which coincided with the greatest mean accuracy and efficiency in all groups. The participants were extremely familiar and comfortable with line graphs, and the additional contextual information on graph 2B assisted with difficulties in comprehending temperature anomalies. Many participants had trouble with graph 1 and graph 3. In particular, participants were overwhelmed by graph 3 and confused by unfamiliar chemical compounds and units, but comfortable with the representation of the data, especially because the graphs provide numbers in addition to visual representation. In contrast, the unfamiliar data representation, primarily of uncertainty, of graph 1 (especially 1B) was very challenging for many participants.

Considering each of these metrics in combination lends insight into the experience of these participants in engaging with a graph. For example, one of the two questions on which the A group performed significantly better than the B group, the second question for graph 1, was also one of the two questions which the B group took significantly more time to answer. Additionally, graph-by-graph, B took more time to answer the questions for graph 1 overall, the same graph for which there was a significant performance difference. Further, for the graph 1 questions overall and question 2 specifically, the B group paid significantly more attention to the data and x-axis. The qualitative data suggests that the greatest challenge of graph 1B was determining the values of the data, a task which would require information from both the data and x-axis, and particularly relevant for the second question of this graph. Qualitative data provides crucial context to eye-tracking data, especially because longer view times can correspond with greater difficulty understanding material or greater interest in the information (Bergstrom & Schall, 2014).

However, the insight generated by an evaluation such as this one is of no use without application. After the primary data collection stage of this study, a new graph C design was created for graphs 1 and 2 based on the feedback of the first 60 participants. The 9 participants using the newest graphs had the highest mean data extraction score, most data extraction questions with perfect scores, lowest mean time spent on questions, highest computer activity credibility ratings, and the highest pre/post change in both risk index and climate scientist credibility. The ranking results were not directly comparable to the A and B conditions, but within the C condition, the C graphs had higher mean rankings for satisfaction, credibility, and perceived risk. Unfortunately, obtaining statistical significance is difficult due to the small sample size.

Satisfaction was related to participant perceptions of credibility and climate change risk as well, so considering participants' varying experience with each graph is important. For example, like satisfaction, no differences in credibility were detected between graphs during the computer activity but strong differences arose during the ranking activity. This could be relevant to participant experiences using the graphs or could be an artifact arising from limitations of this study discussed below. Nonetheless, as indicated by the ranking activities and associated qualitative data, the graphs redesigned with consideration for cognitive science research as synthesized by Harold et al. (2016) were perceived as more credible and satisfactory by the participants of this study. Previous research suggested that complexity and difficulty may have strong ties to perception of higher credibility (McMahon et al., 2016); however, this was not the case with the participants of this study, and instead, greater usability enhanced perceptions of credibility. The C condition findings suggest that designing graphs with regard for both cognitive science and user testing may lead to even greater gains in both user perceptions and performance.

CONCLUSION

This study was conducted to examine the role of design in influencing non-expert undergraduates use and perceptions of climate change graphs. The framework used to make this comparison was the application of guidelines published by Harold et al. (2016) which were created through the synthesis of decades of research from the fields of cognitive science, computer science, and climate change communication research. This study is unique in applying usability evaluation to climate change graphs as well as in adding measurements of affect such as perception of credibility and risk. The strong influence of culture and worldview on climate change beliefs (Kahan et al. 2011) necessitates this integration.

While this study did not investigate any potential influences of culture, the study was conducted in a state with lower-than-national-average belief and worry concerning anthropogenic climate change (Howe et al., 2015) and the selection of participants with lower risk perception was intentional to include those whom climate change communicators may target. Additionally, this low initial measurement left ample room for change and variation in the experiment. Not every individual became more worried about climate change as a result of completing this study, but mean risk assessment did increase. This is impressive because the subject matter of the particular graphs (temperature change and forcings) are not in-and-of-themselves risky; a study involving graphs showing future projections or costs of climate change may be far more impactful on participants. It is encouraging that simply exposing individuals to information about climate change, which the participants said they very rarely hear about, impacted their perceptions and hopefully likelihood to take action on the topic.

Graphs are used to communicate complex information because visuals are thought to help viewers by allowing the "offloading of cognitive processes onto perceptual processes" (Hegarty, 2011, p. 451). Previous studies have shown that differences in graph design can significantly impact comprehension (Carpenter & Shah, 1998; Hegarty, 2011; Renshaw et al., 2003; Shah & Carpenter, 1995). These studies vary from simple, controlled changes, to entirely different designs such as the present study, and have shown that creating visualizations for the best outcomes does not always align with our intuitions about design. Further, user-testing can be implemented to inform and improve graph design, leading to higher performance, as suggested by the C group results in this study and shown in other research (Grant & Spivey, 2003).

Several of the guidelines for graph accessibility, as outlined by Harold et al., seemed to have a positive impact on participant understanding. The redesigned graphs 2B and 3B were praised by users for many of the features encouraged by the guidelines, such as color-coding axes and temperature values. Though participants using these graphs spent more time in many AOIs, on some questions they spent significantly less time viewing axis AOIs, potentially implying that the color-coding helped participants encode and remember the meaning. 2B and 3B also had significantly more text, which can improve novice task accuracy (Gegenfurtner, Lehtinen, & Saljo, 2011). On these same questions, the B-group participants paid more attention to the data. This may have been caused by unfamiliarity with the data representation, leading to increased cognitive load and more time required to make sense of the data. However, it should not be assumed that longer view times always indicate difficulty with a task – comparatively higher attention, especially to important features, has been observed in high and expert performers in a number of graph studies (Atkins & McNeal, 2018; Gegenfurtner et al., 2011; Ho et al., 2014; Okan, 2016).

Additionally, the re-designed graphs were rated higher than the originals for participant perceptions of credibility and worry concerning climate change. Participants were in high agreement surrounding credibility perception and most participants associated credibility with high usability. Since the SPMs are made for public consumption, this is a positive finding, suggesting that credibility does not have to be sacrificed for usability. Participant perceptions of risk concerning climate change, and how the graphs affect perceptions of risk, diverged. Judgements of risk were the only ranking results that significantly varied by participant condition, i.e. which graphs they saw first during the computer activity, specifically for graphs 1A and 3B. Qualitative data did not provide answers to why this discrepancy may have occurred but future work may explore such a question.

Eye-tracking was crucial in informing the evaluation and redesign of the graphs. The most distinct example of the value of this tool was in the use of graph 2. While participants performed statistically equally on graph 2 overall, there was a discrepancy in performance for question 2 of that graph. Eye-tracking heat maps revealed that those who answered correctly in the B condition paid most of their visual attention to the more-relevant decadal anomaly graph whereas those who answered incorrectly looked only at the annual graph. This finding informed the layout change for graph 2C, namely, moving the two graphs back closer together to encourage use of both. In the C condition, only one of nine answered incorrectly, the lowest proportion of any condition.

The overall results of the C condition are a testament to the power of testing communication products with the intended audience. In this study, redesigning the graphs based on generalized cognitive science research alone shifted participant affect while redesigning graphs based on both research and user testing (as suggested by the Harold et al. guidelines) may have both shifted affect and improved performance. Future work will include more C condition data collection to reach appropriate sample sizes for statistical comparison.

From these results, the authors recommend several practices for more effective design and evaluation of climate change graphs:

- Among the various alterations made to the graphs, color-coding axes was the most user-praised change. Contextual scaffolding such as "warmer than average" text was usually highly praised, but this addition also risks adding visual clutter.
- Use of color was very important to participants both for understanding and judgements of credibility. Minimal and meaningful use of color (i.e., color-coding variables to show values or relationships) is perceived as highly credible.
- Use graphics formats that are familiar to your audience. Participants were confused by even minor changes to common graph formats, e.g., error bars on bar graph 1A. If a less familiar format is used, include a key, such as on graph 3B. The key was perceived as helpful and also more credible.
- Use eye-tracking with retrospective interview, which adds an excellent qualitative explanation to eye-tracking data without added distraction during the activity. Additionally, participants may be more prepared to share descriptions after seeing other graphs and gaining context for them.
- Pilot test early and thoroughly. Participant comprehension errors that appeared in the first several participants typically persisted, especially because the population being sampled was relatively homogenous.
- Design iteratively based on input from the target population of the communication tool. Different audiences may have very different content knowledge and graph experience, as well as very different associations with culturally-bound perceptions such as risk and credibility. The changes made to the graphs in this

study based on user feedback were seemingly more impactful than those based on research alone.

Reliability and Validity

The methodology of this study provides several lines of evidence for the reliability and validity of the results. As mentioned above, the measurement scales used (credibility, risk) all have acceptable Cronbach's α values ranging from .78-.93 at various implementations and the credibility scales have had equal or higher values in use by other authors (McCroskey & Teven, 1999). Internal validity was supported by the triangulation of multiple measurements throughout the experiment and assuring approximate equivalence of participant condition groups. Specifically, the surveyed experience and demographics of the groups were compared and found to be statistically equal. All participants were also provided a vocabulary primer and a practice question to mitigate any differences arising from those usage factors. The order of the graphs was also alternated to prevent effects from relevant content. The specific eye-tracking system used may improve external validity over other systems because it allows greater natural physical movement. The independent variable, graph design based on the Harold et al. (2016) guidelines, was reviewed by several authors at both the rubric and graph creation stages.

The validity of pre-survey and post-survey differences was threatened by variation in the amount of time between surveys, i.e., some students participated in the study one week after completing the pre-survey while some others had up to a 12-week gap. The results of this study best represent large four-year university traditional undergraduate students and may not be generalizable to the general public. Considering this study also examined the role of credibility, and the participants found scientists and universities highly credible, the on-campus setting of this study may also limit external validity. The culturally-charged nature of climate change may also lead to higher self-selection in research participation than other topics.

Limitations and Future Work

Though the satisfaction and credibility constructs were measured in several ways, this study is limited by the inadequate measurement at some of these stages. For example, during the computer activity, both satisfaction and credibility were measured with only one item per graph. Besides the risk of misinterpretation of individual questions (potentially low validity), there is also little statistical power offered by single ordinal items (low reliability). Further, the quick pace of the questions in the computer activity may more accurately reflect the average person's experience interacting with graphs in the real world. This fact serves as a reminder that, as with any communication tool, user testing must be highly specialized to the goals and context of the tool to achieve maximum real-world applicability.

Guided by these lessons, future work may involve more robust quantitative measures and a more realistic approximation of the specific format in which a communication tool might be presented to the intended audience. Additionally, though many of the examined variables are related, this study did not investigate participant motivation or likelihood to act to mitigate or adapt to climate change, the ultimate goal of climate change communication. The content matter of these graphs was not best suited to that topic either, so future work may explore graphs and behavior change, potentially with very different graphs. Lastly, this study was limited by the small sample size of the C condition, which will be rectified with additional data collection in the near future.

Experts are known to sometimes prefer visual representations that may "actually impair comprehension" (Harold et al., 2016, p. 1082) which proved true in the initial redesign phase of this study. The original graphs used in this study, published in the IPCC Summaries for Policymakers, underwent thorough expert review for content and clarity, but improvements can still be made through both testing with target audiences and the application of relevant cognitive science research. Many of the differences between graphs are aesthetic, which is encouraging, as aesthetic adjustments are relatively easy to make for climate change communicators around the world. While public understanding of the content of data visualizations is the primary goal, affective judgements such as those surrounding credibility and risk are also important to consider for any climate change-related communications. Participant perceptions lend insight into the public's relationship with intellectual and scientific authority and may play an important role in achieving any progress mitigating the effects of climate change.

REFERENCES CITED

Aksit, O., McNeal, K.S., Libarkin, J.L., Gold, A.U., and Harris, S.E. (2017). The influence of instruction, prior knowledge, and values on climate change risk perception among undergraduates. *Journal of Research in Science Teaching*, *55*(4), 550-572.

Ancker, J. A., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design Features of Graphs in Health Risk Communication: A Systematic Review. *Journal of the American Medical Informatics Association*, *13*(6), 608–618.

Atkins, R.M. & McNeal, K.S. (2018). Exploring differences among student populations during climate graph reading tasks: An eye tracking study. *Journal of Astronomy & Earth Sciences Education, 5*(2), 85-114.

Bergstrom, J.R. & Schall, A.J. (2014). *Eye Tracking in User Experience Design.* Waltham, MA: Morgan Kaufman, an imprint of Elsevier.

Bojko, A. (2013). *Eye Tracking the User Experience: A Practical Guide to Research.* New York, NY: Rosenfeld Media.

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*, *20*, 150-166. doi:10.1016/j.learninstruc.2009.02.014

Chin, J., Diehl, V., & Norman, K.L. (1988). Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. In *Proceedings of the SIGHI conference on Human factors in computing sysems* (pp. 213-218). ACM.

Connolly, R. & Bannister, F. (2007). Consumer trust in Internet shopping in Ireland: towards the development of a more effective trust measurement instrument. *Journal of Information Technology, 22*(2), pp 102-118.

Cook, J., Oreskes, N., Doran, P.T., Anderegg, W.R.L., Vergheggen, B., Maibach, E.W., ... & Nuccitelli, D.. (2016). Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters, 11*(4), 1-7.

Creswell, J.W., & Clark, V.L.P (2017). *Designing and conducting mixed methods research* (3rd ed.). Los Angeles, CA: SAGE Publications, Inc.

Duchowski, A.T. (2007). *Eye Tracking Methodology: Theory and Practice* (2nd ed.). London, England: Springer-Verlag.

Freedman, E.G., & Shah, P. (2002). Toward a model of knowledge-based graph comprehension.

In *Diagrammatic Representation and Inference: Second International Conference, Diagrams 2002 Callaway Gardens, GA, USA, April 18-20, 2002 Proceedings* (18-30). https://doi.org/10.1007/3-540-46037-3_3

Gegenfurtner, A., Lehtinen, E., Saljo, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Education Psychology Review, 23*, 523-552.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*, 631–645.

Goldberg, J. H., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In Hyönä, J., Radach, R. and Deubel, H. (Eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movements* (pp. 493-516). Amsterdam, The Netherlands: Elsevier Science BV.

Grant, E.R., & Spivey, M.J. (2003). Eye Movements and Problem Solving: Guiding Attention Guides Thought. *Psychological Science, 14*(5), 462-466.

Harold, J., Lorenzoni, I., Shipley, T.T., & Coventry, K.R. (2016). Cognitive and psychological science insights to improve climate change data visualization. *Nature Climate Change, 6*(12), 1080-1089.

Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. *Topics in Cognitive Science, 3*(3), 446-474.

Howe, P., Mildenberger, M., Marlon, J., & Leiserowitz, A. (2015) Geographic variation in opinions on climate change at state and local scales in the USA. *Nature Climate Change*, *5*, 596-603.

IPCC. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Geneva, Switzerland: IPCC.

Kahan, D.M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. J*ournal of Risk Research, 14*(2), 147-174.

Leiserowitz, A., Maibach, E., Rosenthal, S., Kotcher, J., Goldberg, M., Ballew, M., … Bergquist, P. (2019). *Politics & Global Warming, December 2018*. New Haven, CT.

Leiserowitz, A., & Smith, N. (2010). *Knowledge of Climate Change Across Global Warming's Six Americas*. New Haven, CT: Yale Project on Climate Change Communication.

Libarkin, J.C., Gold, A.U., Harris, S., McNeal, K.M. & Bowles, R. (2018). *Utilizing Rasch analysis to develop a measure of climate change understanding.* Manuscript submitted for publication.

McCright, A.M., Dunlap, R.E., Xiao, C. (2013). Perceived scientific agreement and support for government action on climate change in the USA. *Climatic Change, 119*(2), 511-518.

McCroskey, J., & Teven, J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communication Monographs, 66*(1), 90-103.

McMahon, R., Stauffacher, M., & Knutti, R. (2015). The unseen uncertainties in climate change: reviewing comprehension of an IPCC scenario graph. *Climatic change*, *133*(2), 141-154.

McMahon, R., Stauffacher, M., & Knutti, R. (2016). The scientific veneer of IPCC visuals. *Climatic change*, *138*(3-4), 369-381.

Okan, Y., Stone, E. R., & Bruine de Bruin, W. (2018). Designing Graphs that Promote Both Risk Understanding and Behavior Change. *Risk Analysis*, *38*(5), 929–946.

Olsen, A., & Strandvall, T. (2010). Comparing different eye tracking cues when using the retrospective think aloud method in usability testing. *Proceedings of the 24th BCS Interaction Specialist Group Conference* (pp 45-53). British Computer Society.

Poushter, J., & Huang, C. (2019). Climate Change Still Seen as the Top Global Threat, but Cyberattacks a Rising Concern. Retrieved December 3, 2019, from http://www.pewglobal.org/2019/02/10/climate-change-still-seen-as-the-top-global-threat-but-cyberattacks-a-rising-concern/

Renshaw, J.A., Finlay, J.E., Tyfa, D., & Ward, R.D. (2003). Designing for visual influence: An eye tracking study of the usability of graphical management information. In M. Rauterberg et al. (Eds.), *Human-Computer Interaction – INTERACT'03* (pp. 144-151). Amsterdam, Netherlands: IOS Press.

Shah, P., & Carpenter, P. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, *124*(1), 43-61.

Stofer, K., & Che, X. (2014). Comparing experts and novices on scaffolded data visualizations using eye-tracking. *Journal of Eye Movement Research, 7*(5), 1-15.

van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2014). How to communicate the scientific consensus on climate change: plain facts, pie charts or metaphors?. *Climatic Change*, *126*(1-2), 255-262.

van der Linden, S.L., Leiserowitz, A.A., Feinberg, G.D., & Maibach, E.W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PLoS ONE 10*(2): e0118489.

van der Linden, S., Leiserowitz, A., & Maibach, E. (2017). Scientific agreement can neutralize politicization of facts. *Nature Human Behavior, 2*(1), 2-3.