**Chapter 6**

**Evaluation: User Study Design**

**6.1 Introduction**Uncertainty visualisation is one of the complex tasks in the visualisation domain, so designing a valid user study is also important. The study design usually prepares a particular set of questions that depends on the nature of the research, goal of the research, and the availability of resources, etc. There are various types of user studies such as experimental/interventional studies, descriptive studies, observational studies, and so on. Since our research domain is in visualisation, we have done a within-subject comparative study with the following measures:

* Task time
* Error Rate
* Subjective assessment (NASA-TLX, SUS)

**6.2 Background and Goal**

We have implemented a novel approach of uncertainty visualization and uncertainty data is generated from some existing machine learning predictive models. We then visualise the data itself in web platform in terms of Chromatic Aberration in an interactive fashion. This simulated Chromatic Aberration (CA) artificially separates the Red, Green, and Blue components of colors spatially around visualisation elements such as squares and circles. The effect is a particular kind of blurriness of color perception. The idea is that the more uncertainty there is in a single predicted datapoint, the more its visual representation will be affected at its outer edge by this artificial chromatic aberration, with the intent of conveying that sense of uncertainty to the viewer through the visual channel.

The purpose of this study is the test whether in fact chromatic aberration can be used successfully to represent uncertainty and determine how accurately viewers can estimate the degree of uncertainty based on a given level of chromatic aberration applied to representative visual elements of predicted data values. This will be determined interactively with users through a web-based visualization system. We have found a publication Correll et al. [35] that also visualises uncertainty in an alternate fashion called Value-Suppressing Uncertainty Palettes (VSUP). So, the prime goal of this study is to compare of chromatic aberration with that existing approach VSUP.

**6.3 Research Questions**

For our research, we have several research questions:

1. How Visualising Uncertainty with Chromatic Aberration works in web platform compare to VSUP in terms of user perception and detection accuracy?
2. Which representation is more efficient in terms of user response times?
3. How do the two representation compare in terms of user preference?

**6.4 Study Material**

We have developed a dynamic webpage with the content of study materials to seamlessly conduct the study session entirely remotely online. It helped to save both participant’s and researcher’s travelling time to meet in a common place and eliminate the risk of health issues due to pandemic which was still guide restriction at the time the study was designed and submitted to ethics. That’s why it was mandatory for each participant to have a Computer/Laptop and a fast internet connection to share participant’s screen and allow for uninterrupted audio conversation.

**6.4.1 Technology and Browser**

The webpage is developed with HTML, CSS, JavaScript, and D3.js for frontend and PHP in backend, deployed in the webspace (web.cs.dal.ca) allocated to the student by the Department of Computer Science, Dalhousie University. Since the webspace has public access over internet, anyone could access the page from anywhere which helped to remain inside the COVID-19 safety guidelines defined by the Nova Scotia Health department and provincial authority.

We also note that, we used CSS color blending to represent Chromatic Aberration that does not work properly in Google Chrome/ Safari. It is a well-known issue that they can’t render the blended color properly and when there are large number of cells with color blending in a chart, Chrome crashes. We found Firefox and Microsoft Edge works without issue and served our purpose, and Firefox/Edge are easy to install on any computer having an internet connection. For this reason, we made either Firefox/Edge mandatory for the participation.

**6.4.2 Components**  
As already noted, VSUP is the closest publication that also presented a technique for uncertainty visualisation but that uses only a grid-chart representation. In other words, the smallest unit of their representation is a square shape. But in our study, we broadened the test cases somewhat using circles and squares and that’s why we created the following core components of our study:

* **CA + Bubble**: Chromatic Aberration is applied on circles in a bubble chart.
* **CA + Grid**: Chromatic Aberration is applied on squares in a grid chart.
* **VSUP + Bubble**: Uncertainties are presented with circular shapes.
* **VSUP + Grid**: Uncertainties are presented with square shapes.

So, first two components use CA and the last two use VSUP representation. In other words, two representations are implemented in four different components with slightly modified approaches.

**6.4.3 Counter Balancing**

## Each component consists of eight questions. The order of the questions is selected randomly that means no two participants would get the questions in same order and the components themselves were presented to the participant in “Balanced Latin Squares” method of counter balancing mechanism proposed [67] to give equal emphasis to each component throughout the study and balance the learning effect.We have explained in detail about the study procedure in Appendix-E but for reference Figure 6.1 shows an example of balanced-latin-square.

A picture containing text, crossword puzzle, dark, clipart

Description automatically generated Shape

Description automatically generated with medium confidence

Figure 6.1: Latin Square (left), Balanced Latin Square (right)

If we consider four components as A, B, C, and D then the first participant will have the order of the first row, second participant will have the order of second row etc. This approach ensures, no two consecutive participants will get the same order of components and 8 participants among 32 will get each component first.

Many empirical evaluations of input devices or interaction techniques are comparative. A new device or technique is compared against alternative devices or techniques. There are two common designs for such experiments *within-subjects design* and *between-subjects design*. We have used the former because we were able to test every component of the system by every participant.

However, counterbalancing conditions using a Latin Square does not fully eliminate the learning effect noted earlier.  Note in Figure 6.1 (left), the 4 × 4 latin square design, component A follows component B for three of the four groups of participants.  Thus, there is a tendency for better performance on component B simply because most participants benefited from practice on Component A prior to testing on Component B.  This phenomenon is mitigated using a *balanced Latin Square* Figure 6.1 (right).

Let us consider the following abbreviation of the modules to easier accommodation in tabular representation of task arrangement of user study:

CA+Bubble = CB, CA+Grid = CG, VSUP+Bubble = VB, VSUP+Grid = VG

|  |  |  |
| --- | --- | --- |
| **PID** | **Modules Order** | **Questions Order(random)** |
| P1 | CB => CG => VG => VB | Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8 |
| P5 | Q2, Q3, Q1, Q5, Q4, Q7, Q8, Q6 |
| P9 | Q4, Q2, Q3, Q8, Q5, Q6, Q7, Q1 |
| P13 | Q1, Q3, Q2, Q4, Q6, Q7, Q5, Q8 |
| P17 | Q6, Q4, Q7, Q3, Q8, Q2, Q1, Q5 |
| P21 | Q4, Q3, Q2, Q1, Q5, Q7, Q6, Q8 |
| P25 | Q3, Q1, Q5, Q4, Q2, Q7, Q8, Q6 |
| P29 | Q2, Q5, Q1, Q4, Q6, Q8, Q3, Q7 |
|  | | |
| P2 | CG => VB => CB => VG | Q5, Q4, Q8, Q2, Q1, Q7, Q6, Q3 |
| P6 | Q1, Q5, Q2, Q3, Q4, Q6, Q8, Q7 |
| P10 | Q4, Q2, Q3, Q8, Q5, Q6, Q7, Q1 |
| P14 | Q1, Q3, Q2, Q4, Q6, Q7, Q5, Q8 |
| P18 | Q6, Q4, Q7, Q3, Q8, Q2, Q1, Q5 |
| P22 | Q4, Q3, Q2, Q1, Q5, Q7, Q6, Q8 |
| P26 | Q3, Q1, Q5, Q4, Q2, Q7, Q8, Q6 |
| P30 | Q2, Q5, Q1, Q4, Q6, Q8, Q3, Q7 |
|  | | |
| P3 | VB => VG => CG => CB | Q2, Q1, Q5, Q4, Q3, Q7, Q8, Q6 |
| P7, | Q2, Q6, Q1, Q5, Q4, Q7, Q8, Q3 |
| P11 | Q4, Q2, Q3, Q8, Q5, Q6, Q1, Q7 |
| P15 | Q1, Q3, Q2, Q4, Q6, Q7, Q5, Q8 |
| P19 | Q6, Q4, Q7, Q3, Q8, Q2, Q1, Q5 |
| P23 | Q4, Q3, Q2, Q1, Q5, Q7, Q6, Q8 |
| P27 | Q3, Q1, Q5, Q4, Q2, Q7, Q8, Q6 |
| P31 | Q2, Q5, Q1, Q4, Q6, Q8, Q3, Q7 |
|  | | |
| P4 | VG => CB => VB => CG | Q7, Q8, Q3, Q4, Q5, Q6, Q1, Q2 |
| P8 | Q2, Q3, Q1, Q5, Q4, Q7, Q8, Q6 |
| P12 | Q8, Q2, Q3, Q4, Q5, Q6, Q7, Q1 |
| P16 | Q1, Q6, Q2, Q4, Q3, Q7, Q5, Q8 |
| P20 | Q6, Q4, Q7, Q3, Q8, Q2, Q1, Q5 |
| P24 | Q4, Q3, Q2, Q7, Q5, Q1, Q6, Q8 |
| P28 | Q3, Q1, Q5, Q4, Q6, Q7, Q8, Q2 |
| P32 | Q1, Q5, Q2, Q4, Q6, Q8, Q7, Q3 |

Table 6.1: The task arrangement of user study

**6.5 Recruitment**

Since the participants play a central role in any user study, it’s important to find the suitable participants for the study based on the attributed research domain.

As we have four independent components in our study and each component has eight random questions, we decided to hire (4 x 8 = 32) participants to give equal emphasis to every component and questions. The detail procedure and considerations about recruitment are described in the following sub-sections.

**6.5.1 Criteria**  
Given that our application is web-based and online, the population for our study is potentially all over the world including members of the Dalhousie University community. But we require participants to be fluent in English because there are questionnaires which needs to be understood correctly and answered accordingly. They all are at least post-secondary students or professionals who have some degree of computer experience as a user of common computer applications. In summary, here is the checklist of the criteria:

* **Age**: We wanted to eliminate participants of age lower than 17 years and higher than 60 years. Because younger participant might not have sufficient knowledge to understand the scope of the questions and elderly people might suffer from eyesight issues.
* **Education**: We required the minimum education level to be post-secondary level.
* **Experience**: We do not require any expertise in specific domains, but participants need to have minimum expertise in computer use, such as browsing websites.
* **Vision**: Participant’s eyesight must be reasonable to detect objects and pass our color blindness test to participate in the study, explained in section 6.4.
* **Head Mobility**: People who suffers from brain disorders are not considered for the study.
* **Physical Ability**: Participant must not be disabled which prevents them from using keyboard, mouse, browse the web or use computer.
* **Computer**: Participants must own a computer/laptop for the period of study session. Smartphones are not accepted to participate in the study due to the insufficient display size.
* **Internet**: Participants must have a good internet connection to continue the session without interruption with voice and video(screen) sharing.

**6.5.2 Hiring Procedure**

The internet is relatively cheap and easily accessible almost all over the world. So, with Covid-19 still an issue, we preferred to use online publicity as a recruitment strategy. We sent a recruitment notice to the Dalhousie University Computer Science undergraduate and graduate mailing list, Dalhousie Computer Science jobs email, physical bulletin boards on campus, in social media like LinkedIn. The recruitment notice outlined the study (process, eligibility criteria, data collection, compensation, and estimated time requirement) and instructions to contact the main researcher. Once a potential participant showed interest with a reply email, the main researcher emailed them a with more detail information and attached a consent form for their perusal and suggested to reply with three potential time slots if they agree with the detail requirements and a consent form content. Participation acceptance was done on a first-come first-serve basis until all places were booked. When participants either became sick or cancelled or did not continue interest up to the sessions, potential wait-list participants were called serially as per their time of participation confirmation.

**6.5.3 Making Schedule**

On confirmation of participation interest by the participant, the main researcher created an MS Teams/Skype event with the agreed time. Participants received a notification in their inbox with the detail of the event including a URL for the event. At the meeting time participants just needed to open the link in either a browser or installed desktop application of the relevant tool to start the session.

The main researcher used a MS Word document as a logbook to manage, track and keep the study process synchronized. The researcher always ensured that two participants participation time could not overlap each other and tried to keep a gap of 30 minutes between two schedules so that unexpected delays could be mitigated by that buffer time.

**6.6 Study Procedure**

The study session contains several stages such as a color blindness test, modular sessions for each of the core components, module introductions and clarifications before starting a module, and post-session questionnaires after completing core modules. In the following section we explain them briefly.

**6.6.1 Start Event**

Since the study was conducted online and schedules were made between two parties (researcher and participant) and an event is created through the online meeting platform or conferencing tool such as MS Teams, the participant just needed to be in front of a computer at the scheduled time of the day and click on the link he or she received in his or her email to meet in the meeting platform. When the participant logged-in to the system, it notifies the researcher that participant is waiting at the lobby, and he needs to be admitted. On approval, the event is instantly started at the online meeting room and both parties will be able to hear each other. Researcher greeted and welcomed the participant and exchanged their formal greetings. If participant faced any technical difficulty to join then the researcher tried to help by possible means.

**6.6.2 Briefing**

The researcher needed to brief the participant about the steps he or she had to go through and explained how he was going to conduct the session. Participants were also asked if his/her system has a Firefox/Edge browser installed which is mandatory for the study. If not he/she would be requested to install it and the researcher might instruct further if they needed any help regarding the installation of the browser. After confirming the browser is ready to go, the participant was requested to open it and informed him/her that researcher would give two URLs for the session i. for Color Blindness Test and ii. for the Questionnaire about the study.

**6.6.3 Color Blindness Test**

One of the prime requirements for the selection process is to test for color-blindness of the participants. The participants had to be capable to decern color to provide meaningful input through their participation in the study. To maintain similarity with Correll et al. [35], we presented a set of Ishihara plates [60] attached in Appendix E in a webpage. The URL of the page (Figure 6.1 shows a screenshot of the webform with plate and input fields) is given to the participant through chat(conversation) box of the conference tool and they were requested to fill the input field with what they saw in the image and click next to get next question. This would continue until it ended with all samples identified. We excluded those that misidentified values or who self-reported as having a color vision deficiency.

**Chart

Description automatically generated**

Figure 6.2: Example Color Plate in our portal

**6.6.4 Pre-Session Discussion**

After the color blindness test was passed successfully, the researcher asked participants about some basic questions which we thought to be relevant with their performance. Because it needs full concentration on the task to understand the question, find relevant values, and uncertainties. For instance, the following information were noted by the main researcher:

* Educational (science, arts, etc.) background
* Professional background (IT, Accountant, etc.)
* Computer skills (Basic, Intermediate, Expert)
* Mathematical and Geometric knowledge
* Visualization and Computer graphics knowledge
* Computer gaming skill
* Measurement knowledge (inch, feet, pixel, etc.)
* Physical Condition (Tired, Sleepy, Hungry, Fresh, etc.)

**6.6.5 Questionnaire**

There were two types of questions in our user study design, and they are as follows:

1. Component Questions
2. Post Session Questionnaire (PSQ)

As noted previously in 6.2.2, we have four components in our study. So, by Component Questions we refer to the questions relevant to those four core components. On the other hand, PSQ refers to the questions to obtain user feedback from the experience of using the four core components of the system. PSQ includes System Usability Scale (SUS) test questions and NASA-TLX Work-Load Scale test questions.

D

C

PSQ

B

A

Figure 6.3: Flow of Questionnaires for a participant

If we consider A, B, C, and D as four components of the study then Figure 6.3 shows the flow of the components that come one after another randomly during the session of the participants as discussed in counter-balancing section. It also shows that PSQ appears at the completion of four modules.

We intentionally placed PSQ at the end of core modules so that participants could give their fair ratings based on their immediate experiences gathered from the components. Since we have used our self-developed online web page to conduct the whole session, the system automatically and quantitatively captures the participant’s answers and saves in memory from time to time and stores in the server at the end of the study. Figure 6.4 shows System Usability Scale test related question sample for both CA and VSUP.

Graphical user interface, application

Description automatically generated

Figure 6.4: System Usability Scale (First Question Example)

Graphical user interface, text, application, email

Description automatically generated

Figure 6.5: NASA-TLX Work-Load Scale (First Question Example)

Since underlying mechanism is same for both CA+Bubble and CA+Grid, they are grouped together and placed at the top of the UI in CA section. Similarly, VSUP+Bubble and VSUP+Grid are grouped together for the same reason and placed at the bottom in VSUP section of the UI. For both cases, we have shown the same question and the scale of answer is 1 to 5 where 1 means ‘Strongly Disagree’ and 5 means ‘Strongly Agree’ and the rest of the scales 2, 3, 4 carries in between weights based on their values.

Again, Figure 6.5 showsNASA-TLX Work-Load test related questions where CA and VSUP sections are placed in the top and bottom respectively since they are two groups of four components just like SUS test explained in previous section. But in this case the number of questions is six and the scale range is 1(Very Low) to 22 (Very High).

We have not shown all questions here for none of SUS or NASA-TLX as we mentioned the complete list can be found in APPENDIX E.

We can summarise the questionnaire as follows:

* Each component would appear to the participant conforming counter balancing rule as stated in 6.4.3.
* Every component had eight random ordered questions but the content of the question for every participant were same.
* The Post Session Questionnaire had two types of questions i. SUS and ii. NASA-TLX
* SUS had 10 questions and NASA-TLX has 6 questions.
* Total number of questions = 4 x 8 + 2 (10 + 6) = 64

**6.7 Data Collection and Storing**We have developed the webpage by ourselves, so we have implemented it in such a way that the system can automatically track the status of every answer whether correct or wrong. That means it keeps a record of every question from starting to end in a JSON object. The structure of the sample JSON data is given below:

*Data = {  
 participant-index: 1,*

*ca-bubble: {*

Q1**: { // single variable and single target selection**

status: true, ca: 52, mode: "single-var-one",

selected: ['Vietnam'], options: ['Vietnam', 'Canada', 'Philippines', 'Kazakhstan', 'Palestine', 'Colombia']

},

…

Q3**: { // single variable and all target selection**

status: true, ca: 90, mode: "single-var-all ",

selected: ['Mongolia', 'Argentina', 'Russia', 'Peru'], options: ['Argentina', 'Mongolia', 'Peru', 'Russia']

},

…

Q8**: { // double variable with single target selection**

status: true, ca: 71, mode: "double-var ",

selected: ['Peru'], options: ['Mongolia', 'Peru']

},

single-var-one-time: 4.9, // time required for questions of single variable-single answer

single-var-all-time: 5.7, // time required for questions of single variable-all answers

double-var-time: 6.5 // time required for questions of double variable-single answer

},

*ca-grid: {… same structure of ca-bubble ...},*

*vsup-bubble: {… same structure of ca-bubble …},*

*vsup-grid: {… same structure of ca-bubble ….},*

nasa-ca: {1: 14, 2: 13, 3: 14, 4: 15, 5: 15, 6: 15}, // answers of NASA-TLX for CA components

nasa-vsup: {1: 13, 2: 12, 3: 13, 4: 12, 5: 13, 6: 12}, // answers of NASA-TLX for VSUP components

sus-ca: {1: 2, 2: 3, 3: 4, 4: 3, 5: 4, 6: 3, 7: 2, 8: 3, 9: 2, 10: 3}, // answers of SUS for CA components

sus-vsup: {1: 2, 2: 1, 3: 2, 4: 1, 5: 2, 6: 3, 7: 2, 8: 3, 9: 2, 10: 3} // answers of SUS for VSUP components

*}*

*}*

## After completion of the entire questionnaire, the generated JSON data is stored on the server with the email address provided by the participant. In the above sample structure, we see for every component it has its own block with the common set of properties for each. The above structure is designed with some self-descriptive properties that it would be helpful later in the results and numerical analysis phase.

**6.8 Session Ending**

Once the participant completed the post-session questionnaire, the page immediately informs the participant with following message:

Graphical user interface, text, application

Description automatically generated

Finally, participants were given whole-hearted thanks for their dedication and participation in the study and immediately sent the promised $10 e-gift card (Amazon) to their email address. A sample of such gift card is attached in APPENDIX K.

Amazon gift-card, will be added in APPENDIX K in main document.

Graphical user interface, text

Description automatically generated

Figure K.1: Amazon gift-card (Claim Code redacted since it is sent to the participant)