MASTER

Optimal color strategy for complex data visualization

Chang, HanWen

Award date:
2019

Eindhoven, April 2019

Optimal color strategy for complex data visualization

HanWen Chang

Eindhoven University of Technology

Identity Number 1241786

in partial fulfilment of the requirements for the degree of

Master of Science

in Human-Technology Interaction

Supervisors (TU/e)
Prof. Dr. Wijnand IJsselsteijn
Prof. Dr. Ir. Raymond Cuijpers

Supervisor (ASML)
Jacqueline Galeazzi

Preface

Using the idea of psychology to find out some sparkles behind the design. This project is the perfect summary of my bachelor study as an industrial designer and my master study as a psychological researcher.

This thesis would not have been possible without the big support of my mentor at ASML, Jacqueline Galeazzi, my supervisors at TU Eindhoven, Wijnand IJsselsteijn and Raymond Cuijpers. Dank u wel!

Finally, I would like to thank my firends, especially to Daan and Bas, I couldn't enjoy my life in the Netherlands more without your company. And of course, my family who always be my strongest backup. Thank you, I have an incredible adventure.

Now it is time to be on the next step.

HanWen Chang

Eindhoven, April 2019

Abstract

Today, the high accessibility of data has shifted the business look into data-driven decision making. The colormap takes an important role in data visualization for humans to rapidly comprehend information. Over three decades, extensive research has recommended various perceptually-correct colormaps for gaining higher decision making quality, yet, we found out that there is a gap between expectation and reality. Which is, most of the people are still in favor of using the rainbow colormap as the main tool even though many published researches claim its deficiencies. Prior research about how human cognition may influence colormap's readability remained limited, it has also been inconclusive whether the mental model could explain why the rainbow is still served as default. Hence, the current research is the first study to assess the effect of mental model on colormap research. We expected that the difference of the expertise between experts and novice would affect performance while using the rainbow colormap. In order to verify this hypothesis, we conducted an experiment to compare the performance between the expert and novice groups and we varied the tasks and the used colormaps. Results suggested that mental model effect of the rainbow colormap make an effective interpretation, but it didn't enhance the accuracy of giving right answer. Based on the performance of the tasks, the diverging colormaps significantly outperformed other schemes in the Heatmap Pattern Recognition, the rainbow and diverging colormaps were recommended in the Vectormap Pattern Recognition, and the alternative sequential colormap which significantly outperformed other colormaps in the Abnormal Spot Finding. We analyzed these findings and composed color strategy guidelines which can be applied in scientific domain.

Contents

1 Introduction

Data, stored by the world's digital universe is growing by 40 percent each year and it is predicted to contain 44 zettabytes of information by 2020. This is nearly as many bits as the number of stars in the universe (Harvey., 2018). With vast volumes of data being generated every second, companies from almost all industries nowadays focus on exploiting data as their competitive benefit (Borkin et al., 2011; McAfee & Brynjolfsson, 2012). Data visualization, as a practical tool is being applied more extensively, using visualizations as a means to reduce and understand the massive amounts of information. It greatly increases prediction and decision making quality (Lohse, 1997). In this regard, how people perceive, and process visualizations of data is an important aspect in influencing the interpretation and judgement of the results (Tory, & Möller, 2004).

When inspecting a visualization of data, sensory input arrives at the viewer's eyes, guided by visual attention. In the visual areas, there are several distinct channels to process the visual features, such as form (orientation and shape), color, size or motion. The low-level visual features such as edge detection, perception of saturation and hue are more intrinsically attention-grabbing than others (Ware, 2013). This is known as bottom-up processing of the visual sense data. Experience or prior knowledge to a particular stimuli may also guide visual attention in a top-down fashion, forming a way of interpretation of what the relevant parts of a visualization may be, or what part of the data may contain a clue towards the goals one has in mind. Visual memory representations are dynamically updated in an interplay between bottom-up and top-down processing (Shipstead, Harrison, & Engle, 2012). What is important to note here, is that the interpretation of visual information is not merely determined by the sense data, but (inter)actively determined between the visualization and the knowledge of the viewer. In this thesis, we investigate the role of the use of different colormaps (bottom up) as well as viewer expertise (top down) in arriving at a correct and efficient interpretation of complex data sets.

To interpret data visualization, color is one of the most fundamental and important criteria in this process. Although extensive research has been done on the effectiveness of color strategies for data visualization, the consideration of exploring effects of human factors is still limited (Tory, & Möller, 2004). Many studies reveal significant results but offer no behavioral explanation as to how humans perceive and process data when it is represented with color.

There are many different colormap designs available today. Nonetheless, research shows that it is difficult to claim one optimal color design as a default colormap, because its effectiveness depends on the viewer's task and goal (Borland & Ii, 2007; Ware, 2004). Due to these inconsistencies, current color strategies are mainly based on designers' intuition, or simply use a default colormap – the rainbow color, which is prevailing in technical computing software (Moreland, 2009). Many studies consistently point out that the rainbow colormap is not an ideal tool (Moreland, 2009; Borkin, 2011; Stauffer, Mayr, Dabernig, and Zeileis, 2015; Reda, Nalawade, and Ansah-Koi, 2018). With an excessive numbers of unordered hues, it might mislead people to naturally perceive and interpret information. This additional information acts like noise and hinders performance, which may cause cognitive overload (Klapp, 1986). Also, it falsely segments the data with its non-uniform lightness increased, indicating data features that are not really there. Shortcomings of the rainbow colormap have been identified for over thirty years, nonetheless, it commonly appears in data visualizations of many different scientific and engineering communities. The reason may be that users with a professional background have obtained a rich mental model of certain stimuli through long-term usage (Toker and Moseley, 2013), and in this case, the rainbow colormap has been used for decades, this has resulted in a strong model. When the nature of a task changes, the mental model needs to be updated and this switch comes at a cost (Parker et al, 2005). It is a cognitive factor that might explain why previous researches cannot be widely applied in real situation.

Surprisingly, there is hardly any work on the effect of mental model on colormap strategy research yet. Also, to the best of our knowledge, despite the fact that general perceptual theories have been applied in various colormap research (Ware, 1988; Moreland, 2009; Reda, Nalawade & Ansah-Koi, 2018), there has been little research available on the performance of colormaps within a real-life scientific setting. The current thesis tried to bridge the gap between theory and practice by addressing the question how a mental model affects the interaction between humans and data visualization. Further investigation is thus needed to validate the issue, and to establish evidence-based guidelines for color strategy in the scientific field. Together, this leads to the following research question: “What are the effects of expertise on decision making with regard to effectiveness and accuracy for colormaps strategy on scientific data visualization?”

This research was carried out at ASML in Veldhoven. ASML, as the world leader in supply and maintenance of photolithography systems for the semiconductor industry. At ASML, data visualization is of great importance for engineers to interpret the data status and patterns. Therefore, the results in the thesis could be applied as a new guideline to enhance the data readability.

2 Literature Review

2.1 Data visualizations

Data visualizations can serve as practical tools to provide insight and overcome the limitations of traditional descriptive statistics cannot achieve. Today, we have more sophisticated ways to display data by using infographics, geographic maps, heat maps, vector maps etc. With these visual characteristics, data visualization is an accessible way to observe and understand trends, outliers, and patterns in data (Kandogan, 2001).

Data visualization is defined as data that can be estimated or represented by any form of graphic in any discipline (Few, 2009), whereas scientific visualization is a term that describes visualization of physical or scientific data (Card et al., 1999). In scientific research, visualization as a communication tool is essential to decode tremendous amounts of data into effective graphics (Ware, 2004), assisting researchers to interpret scientific information and make data-driven decisions. It is easier for the brain to comprehend an image versus words or numbers (Cukier,2010). The topic of visualization has been examined in a diverse range of books (Ware, 2004; Few, 2004; Strange, 2007; Steele, 2010; Rosenberg & Grafton, 2012; Knaflic, 2015) and journal articles, where the scientific visualization discussion can be either general (Kosslyn and Chabris, 1992), discipline-specific (e.g. Puhan et al., 2006), or written from a theoretical or psychological perspective (Spence and Lewandowky, 1991; Cleveland and McGill, 1984, 1987; Kosslyn, 1989). In this study, we investigated a potential gap between theory and reality of data visualization which has not been explained by previous research. We aim to give a further explanation and new insights of why those recommended colormaps in previous research have not gained broad acceptance.

In the current study, in order to evaluate the performances of different colormaps in actual usage scenario, we determined two types of data visualizations in the scientific field: the heatmap and the vectormap. The heatmap, which portraits varying gradients based on the strength of correlation and emphasizes potential attributes that have strong correlations amongst themselves. It has been widely used, not only for scientific purposes (e.g. object pressure, frequency estimation), but also to describe aspects of our daily life (e.g. weather information, population distribution). The characteristics of the heatmap provide an immediate visual summary of information, allowing the viewer to understand complex data sets. The vectormap uses glyphs such as arrows, streamlines and streaklines, particle tracing, line integral convolution (LIC) and topological methods to visualize the strength and direction from data point to point in vector fields. This visualization type is mostly applied to indicate direction or tendency over a period of time and assists viewers to rapidly interpret data which contains intensities and vector quantities at the same time.

2.2 Colormaps

One of the most fundamental and critical criteria in data visualization is the strategy of color usage. There are many different colormap designs available today for visualizing and differentiating data in graphical representations. The main principle of color mapping is that it enables viewers to evaluate quantities, distributions, and patterns as present in the original data (Zhou & Hansen, 2016). The color mapping can be applied as color-coded vector fields in wind direction visualization, color density gradients used for city population maps, or color categories to highlight different types of data charts. Considering its diversity, it is difficult to state one optimal correct color scale as a default colormap, because its effectiveness depends on the viewer's goal and the type of plots (Borland & Taylor, 2007; Ware, 2004). In this regard, the meaning of perceptual correct colormap in the current thesis will be defined by the suitable perceptual theory which can support the specific task and goal.

Extensive research is available with respect to colormap strategies for aiding data visualization. For instance, Moreland (2009) proposed a diverging color map as an ideal color scale which can be used for metric comprehension and form comprehension scientific visualization. It utilizes the theory from Ware (2004), and Mullen (1985) that human perceptive ability is sensitive to the changes on luminance, especially on high frequency data visualization. However, simply relying on luminance like a greyscale colormap may increase perceptual errors when having a large simultaneous contrast shifts (Ware, 1988). The diverging colormap instead, adds two major color components on each side of the map. The hue changes from both sides to an unsaturated middle (white or light yellow). Therefore, it enables people to gain quick identification whether values are close to the center or at the extremes of the scale. This colormap received high recommendation for scientific visualization by Moreland (2009), however, the study did not provide empirical evidence. The research from Borkin, et al. (2011) found further evidence that the diverging colormap performs better than the rainbow colormap in regard of identifying pattern region. One alternative colormap strategy involves varying hues instead of lightness. The study by Ware (1988), states that these kind of colormaps are good for metric comprehension due to people being good at differentiating multiple colors. This theory has been further supported by Reda, Nalawade, and Ansah-Koi (2018), indicating that hue-varying will lead to more accurate quantity estimation.

However, regardless of these studies findings, current scientific data plots are mostly applied the rainbow colormap which has a lack of perceptual ordering and a fully saturated color sequence. Instead of portraying regularly increasing lightness like other perceptually correct colormaps, the rainbow contains varying saturated-hues which approximates to the electromagnetic spectrum's visible wavelengths. This characteristic may be misleading since its unnatural hue ordering creates irregular color gradients. Viewers experience a troublesome effort when trying to interpret the data correctly (Rogowitz & Treinish, 1996; Ware, 1988). Also, the rainbow colormap obfuscates, rather than clarifies, the display of data in a variety of ways (Light & Bartlein, 2004; Moreland, 2009). Lastly, the rainbow colormap emphasizes a sharp visual boundary which hinders noticeability of small details in the data that fall within a single color range in the colormap (Borland & Ii, 2007).

Shortcomings of the rainbow colormap have been discussed for over thirty years, nonetheless, it is still the default in technical computing software (Moreland, 2009). In a study by Correl et al. (2011), despite the result recommended one banded saturation palette in the new development, users still preferred to use the original rainbow color map, as their familiarity with the colormap helped them to gain insight from the data. In another study, Borkin et al. (2011) showed that people thought they could perform better while using the default colormap- the rainbow color - whereas in reality they did not perform as well as the diverging colormap group. This phenomenon reveals that human cognition might influence users' performance in regard of colormap usage.

2.3 Mental model

The mental model is a psychological term which people naturally use to learn to gain faster accessibility of reading data visualization. When our eyes perceive objects then a mental representation or a mental model is formed of the perceptual objects in order to help to interpret information (Marr, 2010). When people are interacting with stimuli, they will continually construct a meaning of description, a set of fundamental representation—a mental model of stimuli (Johnson-Laird, 1995). These models reduce the time for processing data and provide predictive and explanatory power for understanding interactions with the world (Norman, 1983).

The concept of mental model has been widely discussed in various disciplines, since it is a theory to explain how knowledge or information can be organized into structured patterns (Johnson-Laird, 1983; Rouse & Morris, 1984). There are a variety of studies that discuss how expertise results from a strong mental model. These studies show that participants' mental models can be diverse on their levels of the experience toward the stimulus, and this difference influences how experts and novice interpret information (Winn, 2004; Ross, Shafer & Klein, 2006). Studies also revealed that mental models were one of the reasons for experts' superiority since it helps experts to understand, investigate, and foresee a problem situation holistically. From a data analysis perspective, it helps experts to hold a rich internal representation to find leverage points, identify data patterns, and to make better decisions than novices (Phillips, Klein & Sieck, 2008). However, studies also bring out a discussion of the time period of training in mental model. Novices who have a poorer mental model can, with practical training and substantial time spent on practicing, gain a richer mental model and acquire a higher performance (Borgman, 1986; Stefanidis et al, 2007).

However, when the nature of a task changes, the mental model also needs to be updated and this switch comes at a cost (Parker et al, 2005). The rebuilding mental model is required when the information to be acquired is inconsistent with existing experience or with the current structure of a theory (Parker et al, 2015). The mental rebuilding is difficult because the prediction and explanation of information based on everyday experience and tied to years of confirmation (Vosniadou, and Brewer, 1992). It may explain why the color strategy research field encounters a substantial difficulty in introducing a perceptually correct colormap while people have been using the rainbow colormap for decades.

The current study investigates the mental model effect by studying whether the performance of interpreting data plots while participants having different levels of familiarity to the rainbow colormap. We expect to find different patterns while using the rainbow colormap between expertise and novice groups.

2.4 Expertise

Colormaps have been studied for decades, aiming to find optimal color sequences to achieve better readability and usability in the field of data visualization. However, to our best knowledge, the colormap research has yet investigated how the expertise may influence the readability with colormaps. Most of the studies conducted experiments without putting subject's factors into consideration. One of a few exceptions is the study by Borkin et al, (2011), which investigated the professional performance from doctors while using various colormaps to identify patterns from a medical graph. Furthermore, a study by Rheingans (2000), summarized the rules for colormap coding and took the factor of culture connotations into account. Nonetheless, the effect of expertise on the colormaps remains unknown.

Expertise, is a general term that involves a number of areas related to problem solving, knowledge learning, and decision making, etc. For the past four decades, the terms - expert and expertise have focus more on comparison between highly and less skilled individuals. Thus, experts and novices show obvious differences in regard of organizing the quantity of meaningful patterns for information processing (Shanteau, 1987).

A more detailed explanation to distinguish experts and novices' behavior is that novices have no experience with the situations in which they are expected to perform tasks (Benner, 1982). The core difficulty is that they don't know how to use discretionary judgment. Since novices have no experience with the situation they encounter, they can only depend on the ability to conduct a rational analysis of the decision situation (Cater, 2014). Therefore, it lessens efficiency since no strategy can guide novice to solve the problem (Meleis, 2010).

Experts, in contrast to novices, not only have a richer store of related knowledge accessible in memory but also structure knowledge in more complex ways. At the expert level, the performer no longer relies on an analytical principle (rule, guideline, maxim) to connect her/his understanding of the situation to an appropriate action (Benner, 1982). In particular, experts construct comprehensive mental models that combine various information and schemas used by novices into integrated meaningful patterns, while at the meanwhile effectively getting rid of irrelevant information that novices attend to keep (Meleis, 2010). Thus, experts develop mental models, representing new and usable patterns of perceiving, thinking, and organize, process the new information (Glaser and Chi, 1988). For example, in the study by Mohammadi et al, (2015), the nurse has an intuitive grasp on the situation and focusses on the accurate region of the problem without wasteful consideration of a large range of unfruitful possible problem situations. In another example, firefighters' quickly notice details and detect differences which are overlooked by the novices (Klein, 1990)

In the current study we define experts as having a high level of experience with the default colormap in data visualization software (i.e. the rainbow color). By contrast, the novices are defined as having a low or no experience with the colormap. We expect to observe differences by expertise while interpreting colormaps, and whether the performance would be consistent after practice and learning.

2.5 Learning

Learning is a ubiquitous human behavior that helps individuals to recognize certain situations having common features and then developing an appropriate behavior for each class of situations. To monitor and quantify the progress of learning over time, psychologist Hermann Ebbinghaus first introduced the concept of the learning curve in the psychological field. It is a measurement that describes the performance improvement when output or time is doubled (Argote, 2012). The resulting function relates the number of learning trials to performance on one or multiple dependent measure (see Figure 1).

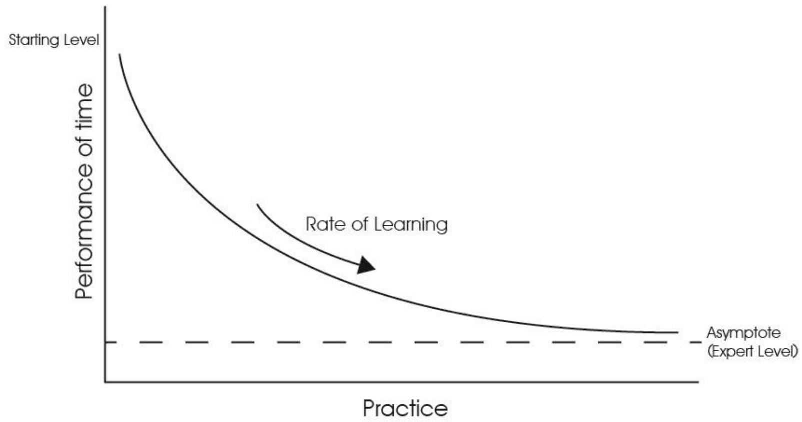


Figure 1. Learning Curve

Since its introduction, the learning curve has been applied in various fields, such as electronic, automotive, construction, software, and chemical companies (Anderson,1982; Lieberman, 1984; Adler and Clark, 1991; Vits and Gelders, 2002; Jarkas, 2010; Weber and Fayed, 2010). This popularity can be attributed to the fact that progress functions were a simple tool to describe complex phenomenon. The study of learning curve plays an important role in explaining the cost of production, by validating the relationship between worker performance and manufacturing rate (Ziemer and Kelly, 1993). It enables managers to scale the manufacturing cost per unit for any cumulative production quantity. In the economic perspective, it can determine competitive behavior in the markets. When there is a learning curve, the short-run output decision is a type of investment decision. It affects the accumulated output and through it, future costs and market position. It the psychology research, learning curve provides a rigorous conceptual lens that enables investigating longitudinal studies (Roediger and Smith, 2012).

As a previous section has mentioned, people can update and improve their mental models through repetition of the decision task. Learning causes a gradually improvement in strength (Keppel and Underwood, 1962). Therefore, the characteristics of the learning curve can be used as the ideal measurement to report the time spent in a sequence of trails to observe the improvement of performance and mental model (Roediger & Smith, 2012). The function can also serve as a comparison to validate the learnability between subjects or objects. As shown by the study of Yohannes et al, (2002), which compared the learning curve between two different surgeries and two types of groups based on their previous knowledge.

With subsequent little improvements while performing repeated trails, the learning curve will approach a plateau in the end. (Schimidt & Lee, 2005). It is a nature phenomenon when people stop learning as much as quickly. Study, for example, will have reached the plateau effect in language learning at some points (Keller, 1958; Richards, 2008). Plateau can be used as a parameter to indicate a level at which performance flattens.

Although learning curves have been used for numerous researches, they were rarely quantified for analyzing the certain period of time of learning progress. To gain an approximate numbers of trails we needed for subjects to learn colormaps, the current research used a method by fitting an inverse function to subjects' learning curves (Feldman, Cao & Fried, 2009).

2.6 Signal Detection Theory

Signal detection theory (SDT) is a method to measure individual's ability to distinguish stimuli. It is widely utilized in the field of human factors by psychologists. The book by Green and Swets, (1966) which describes SDT and its application to psychology, has been cited over 14,000 times . In the field of psychological research, a common issue is that the potential biases which are presented in the percentage calculation of giving right answer is hard to exclude, for example, participants may depend on guessing to give the response (Shapiro, 1994). The characteristics of SDT can provide a more accurate measurement to achieve a robust result.

SDT assumes that the decision maker is not a passive receiver of information, but rather an active decision-maker who makes difficult perceptual judgments under conditions of uncertainty (Ivancevic & Ivancevic, 2010). The theory is determined by two parameters: the Sensitivity (d') and the Criterion bias (β'). That is, a discriminability to give a correct response (Sensitivity); a tendency of individual with respect to giving response (Criterion bias). In the current study, we used sensitivity as one of our performance variables to analyze the effectiveness of colormaps.

SDT can be applied whenever two possible stimulus types must be discriminated. Normally, stimuli can be separated into signals and noises. Take visual signals for instance, when the intensity of the signal is low (e.g. hard to differentiate the pattern on the image), noise may cover the signals and lead subjects to make a wrong response. The d' represents the sensitivity toward stimuli by calculating the distance between the signal and the noise distributions corresponding to the effect of the signal. When the intensity of signal gets stronger, the overlap of signal and noise distributions will get smaller. The higher the d' value is, the more sensitive subjects are to the signal (Figure 2).

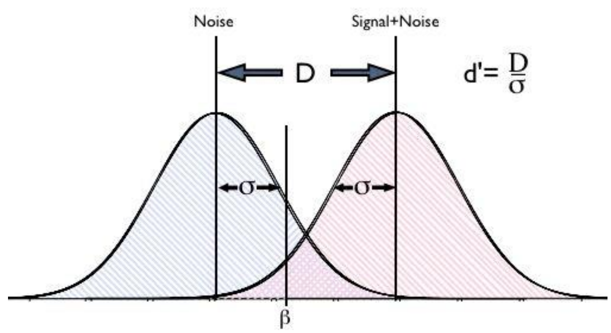


Figure 2. the diagram for the Signal Detection Theory model

2.7 Research question and hypotheses

The study expected that participants should have a better performance while using the perceptual correct colormaps than the rainbow in all of the tasks. However, one of the missing aspect from previous research is that whether a mental model of the rainbow colormap may cause a difference in performance between experts and novice. As participants learn over time this effect should be most pronounced at the beginning. We therefore come to the following research question, and corresponding hypotheses:

Which colormap(s) supports users to achieve better decision making with regard to effectiveness and accuracy, for quantitative data visualization in vector and surface plots?

H1. Is the performance of decision making in the expert condition influenced by previous experience with the rainbow colormap?

> H1a. We expect that performance with an alternative colormap is inferior to the performance with the rainbow colormap in the expert condition at the beginning of learning curve.
>
> H1b. We expect that performance with perceptual correct colormap(s) is superior to the performance with the rainbow colormap in the expert condition after practice.

H2. Is the performance of decision making in the novice condition remains consistently since there is no previous experience with the rainbow colormap?

> H2a. We expect the performance of using perceptual correct colormap(s) is consistently superior to the rainbow colormap in the novice condition.

3 Method

The current research aims to discover whether different colormaps for data visualization influence participants' decision making. We divided participants into two independent groups (experts and novices) as previous research has not investigated whether different expertise levels would affect performance. The experiment consisted of three independent tasks, with each five colormaps and multiple trails.

3.1 Design

A within-subject design was used, comprising 5 different colormaps as within-subject factor and the performance – time (time to submit an answer) and sensitivity (numbers of correct answer) of 2 trails (the beginning and the end of the trail) as the dependent variable.

3.1.2 Colormaps

We selected 5 colormaps (listed in Table 1 and illustrated in Figure 1) based on the design strategies below:

- Rainbow (RB): default colormap in most analytical software.
- Alternative diverging (AD): similar hues as rainbow with diverging uniformly-stepped luminance.
- Diverging (DG): dissociated two hues from two sides of AD .
- Sequential (SQ): monotonically increasing luminance over a one single hue
- Alternative sequential (AS): embed two additional hues into the Sequential colormap.

All colormaps were created in the Matlab R2016b. The RB colormap followed the Jet color which used as the default colormap in Matlab R2016B. The DG colormap followed the DG colormap study by Moreland (2009). The AD colormap followed the color research by Reda et al., (2018). The SQ colormap followed the Color Brewer (Brewer, 2006). The AS followed the Viridis color scale (Rudis, Ross, & Garnier, 2018).

Table 1. the five colormaps evaluated in this study

Colormap	Design Strategy	Luminance Control	Hues
 **RB:** Fully saturated hue colormap which is commonly used in analytical software.	Hue Rotation	None	Saturated RGB (6 color points)
 **AD:** Symmetrically increasing luminance with darker ends and a bright midpoint. Hue followed the settings of RB with a soft yellow middle.	Diverging	Uniform, mid peak, darker ends, Symmetric	Limited RGB (5 color points)

	**DG:** Symmetrically increasing luminance with darker ends and a bright midpoint. Diverging with blue and red at the end sides and soft yellow at the middle.	Diverging	Uniform, mid peak, darker ends, Symmetric	Blue, red, yellow (3 color points)
	**SQ:** Monotonically increasing luminance over a single blue hue.	Sequential	Monotonic increase	Blue (1 color point)
	**AS:** Monotonically increasing luminance. Hue from dark blue to green and yellow added at the top.	Sequential	Monotonic increase	Blue, green, yellow (3 color points)

3.1.3 Structure and ordering considerations

Each group received 5 x 3 = 15 blocks of 2 trails. In each block one of five colormaps is used for one of three tasks. The order of the trails is randomized. Then, we assigned half of the participants from each group took the first task as the starting point and the other half took the third task at the starting point. The second task was put in between. Furthermore, to eliminate the ordering effect of colormaps, we used Latin- Square to generate an ordering list to make sure there is no duplicate tasks. Finally, for the effect of trails ordering, we used semi-Latin-Square, which was, controlling the order of the first trail, and made sure the rest of trails were fully randomized.

3.2 Participants

A power analysis was used to estimate the sample size. We used an effect size of 0.45, based on a previous study by Borkin et al (2011) that compared the performance of various colormaps for data visualizations as used in the medical field. Since in our case we have 5 colormaps in a repeated design, to validate the performance within each group, we required a sample size of 12, with a power of 95% and an error probability rate of 0.05. Because the experiment was conducted with two independent groups, twice as many participants were required (total: 24 participants).

The actual number of participants in our sample was 30. The participants were split into two equal groups based on their previous answer in the intake questionnaire. The rubrics of grouping can see Table 2. The Expert group consisted of 3 women and 12 men. The Average years of working in the industry was 5.3 years. Average years of using analytical software was 6.7 years. Over 50% of the participants had to interpret plots more than 3 times a week. By contrast, the Novice group consisted of 5 women and 10 men. Average years at related industries was 3.6 years. Average years of using analytical software was 2.2 years, and only 27% of the participants had to interpret data plots more than 3 times a week.

We recruited subjects through the company. Some participants recruited each other through snowball sampling uncoordinated by the experimenter. All participants shall pass the Ishihara test for color blindness and Jaeger Chart for near visual acuity. The experiment lasted 40-70 minutes per participant.

Table 2.Rubrics of grouping

Novices	Experts
Related working experience < 2 years	Related working experience > 2 years
Software experience < 2 years	Software Experience ≥ 2 years
Frequency of seeing plots < 3 times a week	Frequency of seeing plots ≥ 3 times a week
Frequency of interpreting plots < 3 times a week	Frequency of interpreting plots ≥ 3 times a week

3.3 Apparatus and Materials

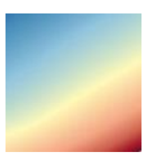
3.3.1 Platform

The experiment platform was programmed in HTML5, CSS6 and Javascript. All the three tasks can be performed through this platform. To maintain a precise record of the time when participants completed each trail, the program is able to record when one trail ended and the next one started. All study sessions were conducted in the similar meeting room within the company with identical lighting, and o n the same Lenovo laptop (T480) on the full screen Chrome browser. An example of the three tasks can be found in Figure 3.

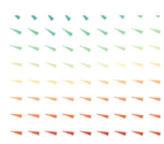
3.3.2 Plots

We used 6 datasets rendered with 5 colormaps for 3 different types of the tasks (total: 90 plots). All plots were generated with fake data, to avoid any confidentiality issues. Which were created separately through Matlab R2016B (see Table 3). For the Abnormal Spot Finding task, we had to do some additional editing with Photoshop to create outliers.

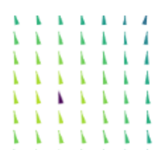
Table 3. Overview of 3 types of plots



Heatmap Pattern Recognition



Vector Pattern Recognition



Abnormal Spot Finding

3.3.2 Tasks

Heatmap/Vectormap Pattern Recognition Task

The goal of the task (see Figure 3) was to interpret the right graph (which correctly portraited the “neutral pattern” of the data) out of four options. The neutral pattern represented a contour map when the data was separated equally into three portions. Participants had to follow features from the map to interpret which option had the right pattern from the data.

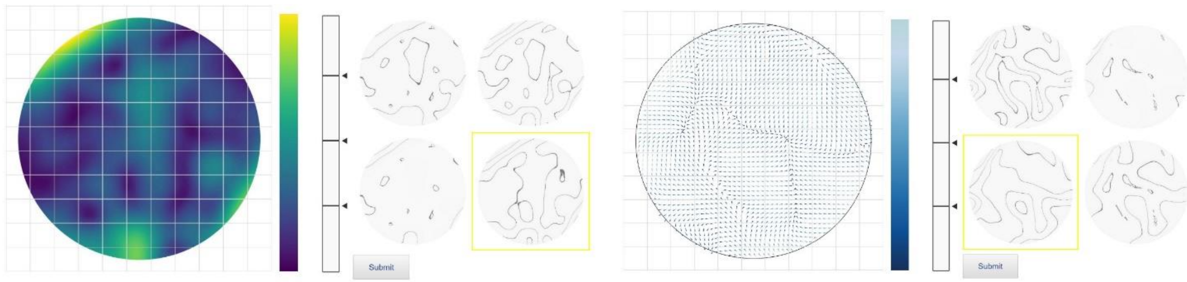


Figure 3. Pattern Recognition task: Heatmap (left); Vectormap (right)

Abnormal Spots Finding Task

The goal of the task (see Figure 4) was to select (with mouse-click) between one and seven data spot(s) which they were considered to be an abnormal outlier. We defined that the abnormal spot should contain a value that is beyond 6 σ of its surrounding data (Gregg, 2014). If participants wrongly selected a spot, they could always click on a "clear" button to restart the current trail.

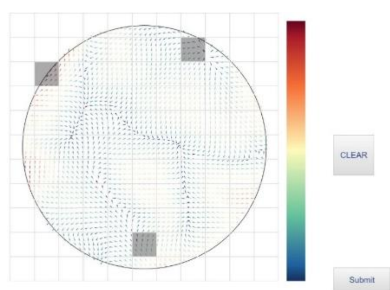


Figure 4. Abnormal Spot Finding task

3.3.4 Questionnaire

The intake survey contained 4 questions related to familiarity with the RB colormap, which was measured by: years in the industry; years of using of analytical software; frequency of seeing/ interpreting scientific plots.

The outtake survey is related to the tasks that the participants had just completed, they were asked to respond on a 7=point Likert scale (1-strongly disagree, 7 = strongly agree) to six statements that assess the participants' opinions of each colormap, such as confidence of performance, efficiency, and preference (see Borkin et al, 2011). This scale was used to assess participants' qualitative opinion of the five colormaps.

3.4 Measurement

3.4.1 Expertise

Before the experiment, subjects responded to four custom items to retrieve their familiarity levels of the Rainbow colormap (see Table 4).

Table 4. the survey before the experiment.

Question	Type of Question
year(s) of working experience in analytical industries	free input
year(s) of experience of analytical software	free input
frequency of seeing plots	Optional ☐ less than 3 times a week ☐ greater than or equal to 3 times a week
frequency of interpreting plots	Optional ☐ less than 3 times a week ☐ greater than or equal to 3 times a week

3.4.2 Performance

There were two variables we took as the measurement of participants' performance.

3.4.2.1 Time:

Per trial we recorded the time, from the moment the trial started, till the moment when the participant submitted the answer. The measurement indicates how efficiently participants can interpret a colormap.

3.4.2.2 Sensitivity:

It is the term represented a level of sensitivity to each stimulus (i.e. colormap) which based on the Signal Detection Measurement by Stanislaw and Todorov (1999). In Heatmap/ Vectormap Pattern Recognition tasks, we made use of Forced-Choice Tasks, which indicated that each trail presents one signal and one or more noise stimuli. The subjects had to identify which stimulus was the signal. Then, we recorded the number of correct choices participants given. Alternatively, the Abnormal Spots Finding Task, we used the structure of Yes/No Tasks, which is one type of task from the signal detection theory that presents one or multiple signals with one or multiple noises in a trial. Participants were asked to distinguish signals and noises. In this sense, the answer will be transfer into the value of sensitivity, the range of the score can be found in Table 5. We collected four parameters from the task to retrieve the subject's performance of sensitivity for further analysis. The parameters (see Table 6) were Hit (there is a signal, and been correctly selected), False Alarm (there is a noise, but been wrongly selected), Miss (there is a signal, but hasn't been selected) and Correct Rejection (there is a noise, and hasn't been wrongly selected). Then, using the analytical calculator which designed by Gaetano, Lancaster and Tindle (2015), we could transfer the numbers from the decision matrix into the value of sensitivity.

Table 5. The overview of each task in the experiment

Task	Type of the task	The range of the score (sensitivity), from worse to best
Heatmap/Vectormap Pattern Recognition	Force-Choice	-0.72 ~ 2.03

Abnormal Spot Finding	Yes / No	-4.86~5.13

Table 6. The decision matrix from the signal detection theory

		Signal	*Noise*
Response State	*"Yes"*	True Positive *'Hit'* (H)	False Positive *'False alarm'* (FA)
	"No"	False Negative *'Miss'* (M)	True Negative *'Correct rejection'* (CR)

After each task, the participants had to complete a questionnaire, which consisted of 4 items measured on a seven-point Likert scale. The aim was to collect qualitative opinions of each colormap right after each task. It would allow us to interpret differences between subject's mental perspective and actual performance.

3.5 Procedure

3.5.1 Pretest

We considered that repetitive activities may deplete an individual's resources, leading to fatigue (Yazdi & Sadeghniiat-Haghighi, 2015). To avoid this potential concern, we conducted a pretest to determine a maximum number of trails for the actual experiment. We selected the heatmap pattern recognizing study as our study for pretest.

Six individuals from the company were requested to complete 50 trials. To get a precise measurement of the learning curve, we accumulated all participants' data and applied the curve fitting function by Feldman, Cao, & Fried., (2009). Results (see Figure 5) showed that the AS colormap had a moderate fit (R^2=0.55) which indicated that subjects did learn from the colormap. AD (R^2=0.35), DG (R^2=0.26) and RB (R^2=0.21) had a fit between fair to poor which indicated that subjects could slightly learn from these colormaps. SQ had nearly no fitting (R^2=0.03) which meant subjects could hardly learn from the colormap. Furthermore, based on the patterns retrieved from the results, we discovered participants all got an ascending pattern in each colormap's task from fifth trail onward. This gave rise to a concern of subjects becoming fatigued during the experiment. Based on the result of the pretest, we decided to limit the number of trials to 6.

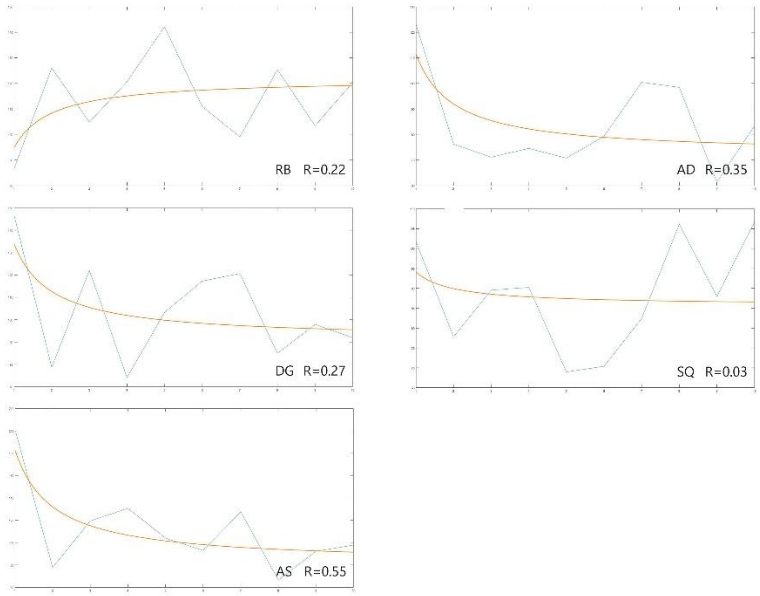


Figure 5. Overview of curve fitting in the colormaps

3.5.2 Experiment

When participants came into the office, they were asked to read and sign an informed consent form. After they consented to join the study, participants conducted a color blindness and a near-sight acuity test, to make sure that they have no vision problems. After that, they were required to fill out an intake survey, which was used to divide participants into two groups for later experiment.

At the start of the experiment, participants were asked to complete three tasks at a laptop separately. They received a tutorial before each task started. It described the goal of the specific task, how many trails were included, and the approximated time of the task may take. Additionally, experimenter demonstrated an example of the trail, instructing how to proceed the task.

Each task started when participants clicked on the button. Once participants finished the task, they were asked to fill out an outtake survey. After the whole tasks had been finished, participants were required to fill out one final survey. Finally, all participants were thanked for their participation.

3.6 Data Analysis Plan

The primary analysis will be a comparison of 5 colormaps in 3 different tasks, using a Repeated measure ANOVA (within factors) with 5 measurements (5 colormaps). To compensate the multiple independent tasks are tested, we use Bonferroni correction for the alpha level: alpha=0.05/3=0.0167. If the main effect is significant, we will follow this effect up with post hoc tests to discover which specific means differed.

4 Results

In our analyses, we tested the performance of different colormaps in three different tasks at the beginning and the final trail. We divided participants into two groups – the experts and the novices to further investigate whether the mental model of rainbow colormap would affect the performance differently.

4.1 Outlier Removal

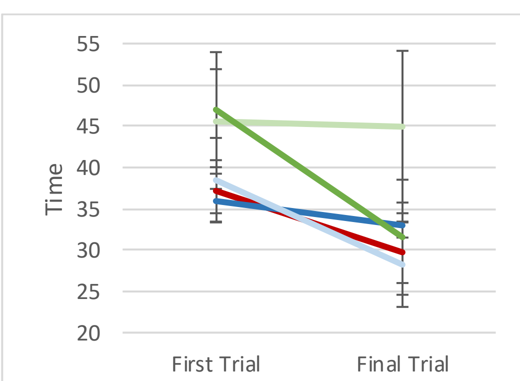
The total number of participants in our sample was 30 in each task. For one participant in the Vectormap Pattern Recognition task, the data could not be retrieved, due to the program overwrote the data by later result. In the Abnormal Spot Finding task, one participant showed an exceptionally slow performance (the modified z-score of time greater than 3.5). However, we further investigated the performance of sensitivity, finding that the participant's z-score was not beyond the boundaries (3.5 to -3.5). Therefore, he/she should not consider as an outlier. To sum up, the total number of participants in the data set used for analysis was 30 (for Heatmap Pattern Recognition), 29 (for Vectormap Pattern Recognition), and 30 (for Abnormal Spot Finding).

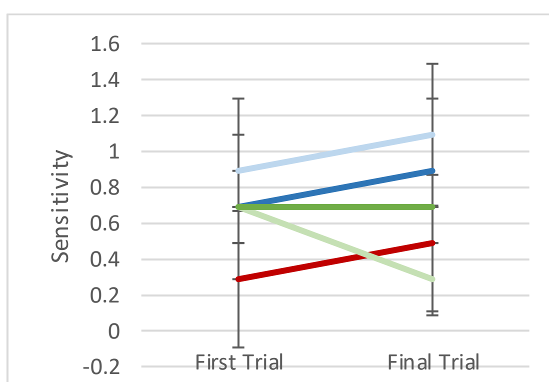
4.2 Hypothesis Testing

In order to test whether the effects of expertise of the rainbow colormap influence the decision making, we conducted a repeated measure ANOVA comparing the average performance in each colormap in the first and the final trail by two groups (the experts and novices).

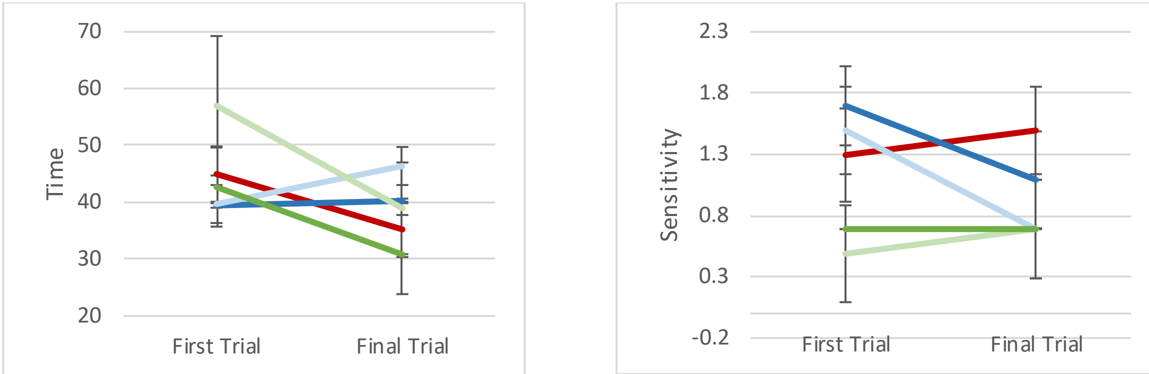
4.2.1 Task performance for experts

In the first hypothesis, we assumed that because the experts have a high expertise of the rainbow colormap, the usage of rainbow colormap should outperform the other colormaps at the beginning of trail. After practice, the perceptual correct colormap(s) should perform better at the final trail.

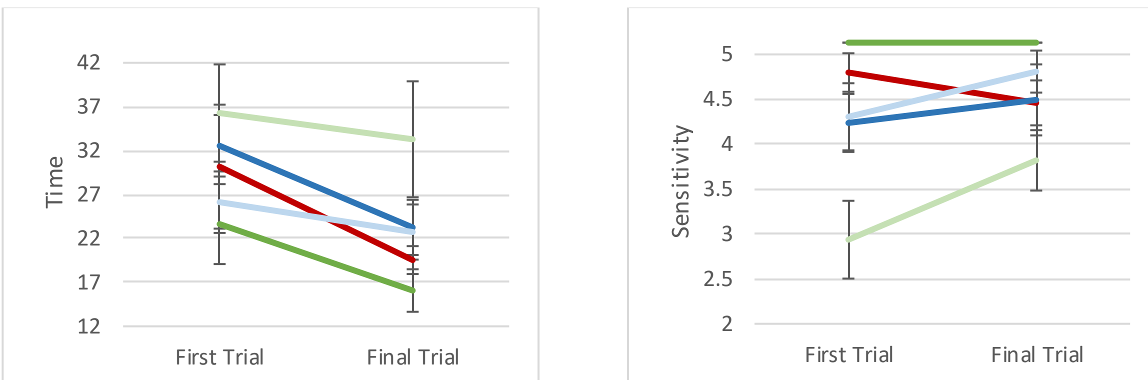




Experts' performance in the first and final trail in HPR task



Experts’ performance in the first and final trail in VPR task



Experts’ performance in the first and final trail in ASF task

Figure 6. the overview of the task performance in expert group

Heatmap Pattern Recognition

First Trial

Time

Estimated Colormaps cell means were RB: 39.05±22.62, AD: 38.10±21.79, DG: 40.05±23.12, SQ: 47.52±24.20, and AS: 49.63±25.92.

Muchly’s test indicated that the assumption of sphericity is not violated with p =0.122. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ = 1576.56, $F(4,52)=1.51$, $p=.22 > 0.0167$, and $\eta^2_{partial}$ =0.10.

Sensitivity

Estimated Colormaps cell means were RB: 0.36±1.50, AD:0.57±1.55, DG: 1.00±1.55, SQ: 0.57±1.55, and AS: 0.79±1.57.

Muchly’s test indicated that the assumption of sphericity is not violated with p =0.42. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ = 3.39, $F(4,52)=0.46$, $p=.76 > 0.0167$, and $\eta^2_{partial}$ =0.034.

Final Trail

Time

Estimated Colormaps cell means were RB: 30.03±23.84, AD: 35.11±24.10, DG: 29.92±25.33, SQ: 47.74 ±35.16, and AS: 33.63±21.08.

Muchly's test indicated that the assumption of sphericity is violated with p =0.014, therefore degrees of freedom were corrected using Huyhn-Feldt correction instead. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ = 3000.31, $F(4,52)=1.42$, $p=.25 > 0.0167$, and $\eta^2_{partial}$ =0.10.

Sensitivity

Estimated Colormaps cell means were RB: 0.57±1.55, AD:0.79±1.57, DG: 1.22±1.50, SQ: 0.36±1.50, and AS: 0.79±1.57.

Muchly's test indicated that the assumption of sphericity is not violated with p =0. 66. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ = 5.73, $F(4,52)=0.61$, $p=.61 > 0.0167$, and $\eta^2_{partial}$ =0.045.

Vectormap Pattern Recognition

First Trial

Time

Estimated Colormaps cell means were RB: 44.96±29.12, AD: 39.46±30.17, DG: 39.70±30.66, SQ: 56.99±47.52, and AS: 42.68±29.92.

Muchly's test indicated that the assumption of sphericity is not violated with $p = 0.104$. The main effect of Colormaps was not statistically significant with $SS_{colormaps} = 3116.09$, $F(4,56)=1.11$, $p=.36 > 0.0167$, and $\eta^2_{partial} = 0.07$.

Sensitivity

Estimated Colormaps cell means were RB: 1.29±1.47, AD: 1.70±1.25, DG: 1.49±1.38, SQ: 0.49±1.53, and AS: 0.69±1.56.

Muchly's test indicated that the assumption of sphericity is not violated with $p = 0.99$. The main effect of Colormaps was not statistically significant with $SS_{colormaps} = 16.30$, $F(4,56)=1.87$, $p=.13 > 0.0167$, and $\eta^2_{partial} = 0.12$.

Final Trail

Time

Estimated Colormaps cell means were RB: 35.23±28.00, AD: 40.25±38.63, DG: 46.39±55.80, SQ: 38.95±31.24, AS: 30.81±25.12.

Muchly's test indicated that the assumption of sphericity is not violated with $p = 0.19$. The main effect of Colormaps was not statistically significant with $SS_{colormaps} = 2027.03$, $F(4,56)=0.70$, $p=.57 > 0.0167$, and $\eta^2_{partial} = 0.05$.

Sensitivity

Estimated Colormaps cell means were RB: 1.29±1.47, AD: 1.70±1.25, DG: 1.49±1.38, SQ: 0.49±1.53, and AS: 0.69±1.56.

Muchly's test indicated that the assumption of sphericity is not violated with $p = 0.15$. The main effect of Colormaps was not statistically significant with $SS_{colormaps} = 7.78$, $F(4,56)=1.07$, $p=.38 > 0.0167$, and $\eta^2_{partial} = 0.07$.

Abnormal Spot Finding

First Trial

Time

Estimated Colormaps cell means were RB: 30.18±27.33, AD: 32.55±13.56, DG: 26.13±19.19, SQ: 36.28±21.45, and AS: 23.64±17.62.

Muchly's test indicated that the assumption of sphericity is not violated with $p = 0.315$. The main effect of Colormaps was not statistically significant with $SS_{colormaps} = 1516.87$, $F(4,56)=1.75$, $p=.16 > 0.0167$, and $\eta^2_{partial} = 0.11$.

Sensitivity

Estimated Colormaps cell means were RB: 4.80±0.83, AD: 4.24±1.52, DG: 4.31±1.45, SQ: 2.94±1.68, and AS: 5.13±0.00.

Muchly's test indicated that the assumption of sphericity is violated with p =0.04, therefore degrees of freedom were corrected using Huyhn-Feldt correction instead.

The main effect of Colormaps was statistically significant with $SS_{colormaps}$ = 42.00, F(4,56)=7.09, p < 0.001, and $\eta^2_{partial}$ =0.34. Then, we run post hoc tests, highlighting exactly where these differences occurred. We found out that in the average performance of sensitivity in the first trail,

- the RB with the SQ indicated a significant negative effect (p<0.001).
- the AS with the AD (p=0.028), the DG (p=0.04), and the SQ(p<0.001) had a significant negative difference.
- the SQ with the AD (p=0.028), and the DG (p=0.031) had a significant positive difference.

Final Trail

Time

Estimated Colormaps cell means were RB: 19.54±6.16, AD: 23.26±12.16, DG: 22.73±8.58, SQ: 33.29±25.56, and AS: 16.07±9.37.

Muchly's test indicated that the assumption of sphericity is violated with p <0.001, therefore degrees of freedom were corrected using Huyhn-Feldt correction instead.

The main effect of Colormaps was statistically significant with $SS_{colormaps}$ = 2490.47, F(4,56)=4.48, p=.023 < 0.05, and $\eta 2_{partial}$ =0.24. Then, we run post hoc tests that can highlight exactly where these differences occur. We found out that in the average performance of time in the final trail,

- the RB with the SQ indicated a significant negative effect (p=0.040).
- the AS with the AD (p=0.028), the DG (p=0.006), and the SQ(p=0.003) had a significant negative difference.

Sensitivity

Estimated Colormaps cell means were RB: 4.46±0.97, AD: 4.49±1.11, DG: 4.81±0.91, SQ: 3.82±1.30, and AS: 5.13±0.00.

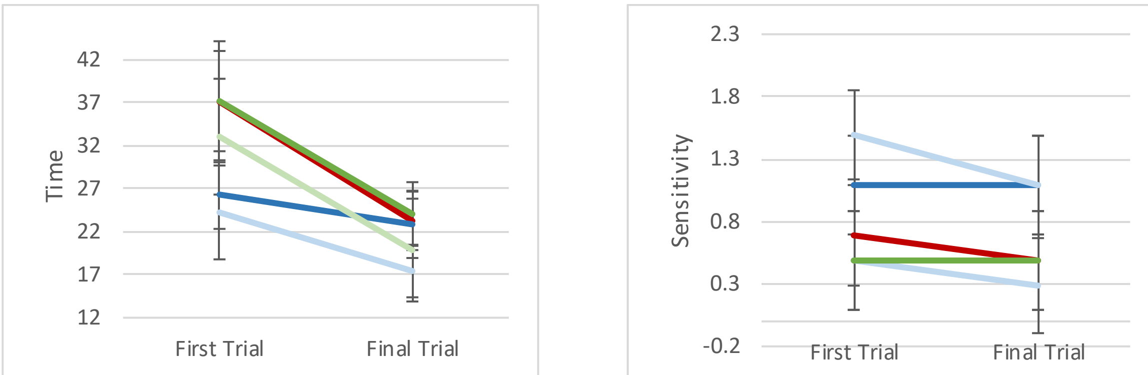
Muchly's test indicated that the assumption of sphericity is violated with p =0.001, therefore degrees of freedom were corrected using Huyhn-Feldt correction instead.

The main effect of Colormaps was statistically significant with $SS_{colormaps}$ = 14.22, F(4,56)=4.28, p = 0.018 < 0.0167, and $\eta^2_{partial}$ =0.234. Although we didn't find a significant difference with a strict Bonferroni Correction, but we did discover a significant trend. Then, we run post hoc tests, highlighting exactly where these differences occur. We found out that in the average performance of sensitivity in the final trail,

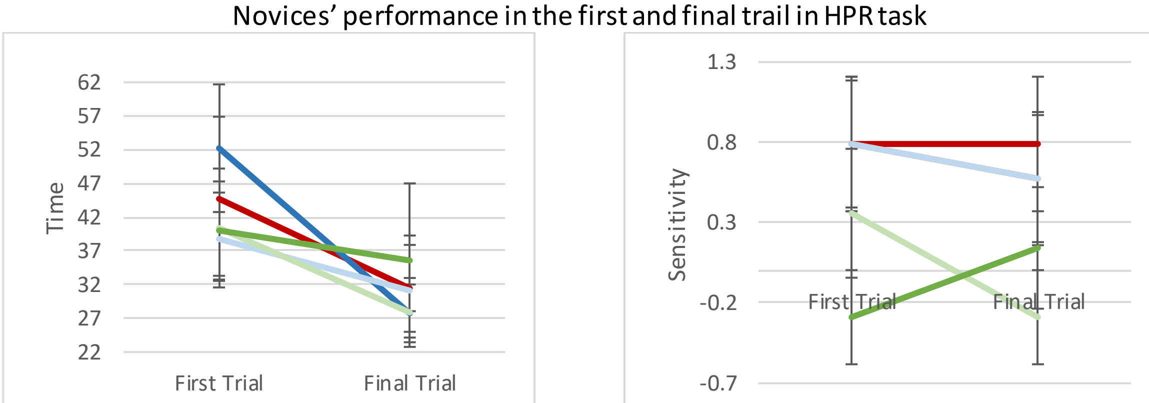
- the RB with the AS indicated a significant negative effect (p=0.018).
- the AS with the AD (p=0.043), and the SQ(p=0.002) had a significant positive difference.
- the DG with the SQ(p=0.039) had a significant positive different.

4.2.2 Task performance for novices

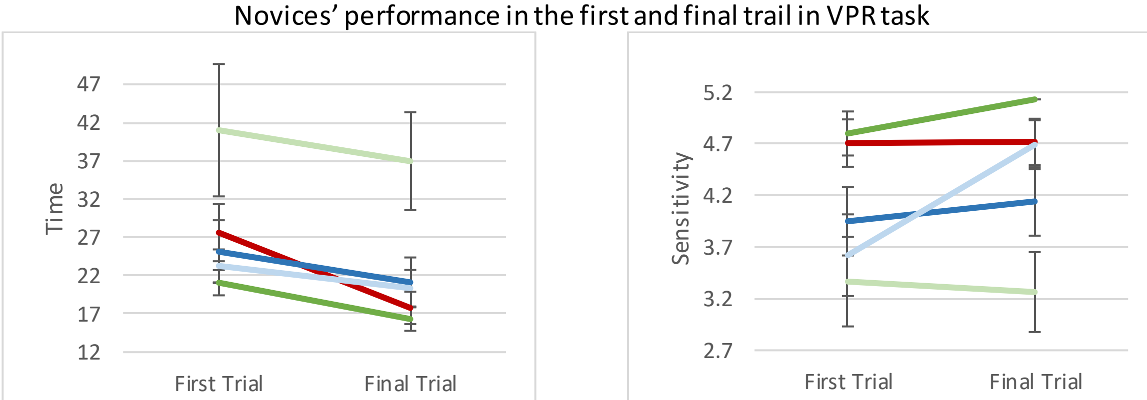
In the second hypothesis, we assumed that the novices with less or zero expertise of the rainbow colormap, the usage of rainbow colormap should perform less efficient and accurate than the perceptual correct colormap(s) at the first and final trail.



Novices' performance in the first and final trail in HPR task



Novices' performance in the first and final trail in VPR task



Novices' performance in the first and final trail in ASF task

Figure 7. the overview of the task performance in novice group

Heatmap Pattern Recognition

First Trial

Time

Estimated Colormaps cell means were RB: 37.14±27.30, AD: 26.34±15.47, DG: 24.26±21.16, SQ: 33.08 ±26.11, and AS: 37.22±22.60.

Muchly's test indicated that the assumption of sphericity is not violated with p =0.576. The main effect of Colormaps was not statistically significant with SScolormaps = 2191.264, F(4,56)=1.673, p=.169 > 0.0167, and η2partial =0.107.

Sensitivity

Estimated Colormaps cell means were RB: 0.69±1.56, AD: 1.09±1.53, DG: 1.49±1.38, SQ: 0.49±1.53, and AS: 0.49±1.53.

Muchly's test indicated that the assumption of sphericity is not violated with p =0.431. The main effect of Colormaps was not statistically significant with SScolormaps = 11.431, F(4,56)=1.633, p=.179 > 0.0167, and η2partial =0.104. Although we didn't find significant difference in general, in the post hoc test did highlight the average performance of sensitivity in the first trail that:

- the AS with DG (p=0.019) showed a significant negative effect.

Final Trail

Estimated Colormaps cell means were RB: 23.28±13.18, AD: 22.87±15.17, DG: 17.44±11.84, SQ: 19.88±23.17, and AS: 24.07±14.36.

Muchly's test indicated that the assumption of sphericity is not violated with p =0.257. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ = 462.206, F(4,56)=0.823, p=.516 > 0.0167, and $\eta^2_{partial}$ =0.056.

Sensitivity

Estimated Colormaps cell means were RB: 0.49±1.53, AD: 1.09±1.53, DG: 1.09±1.53, SQ: 0.29±1.47, and AS: 0.49±1.53.

Muchly's test indicated that the assumption of sphericity is not violated with p =0.462. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ = 8.512, F(4,56)=1.065, p=.382 > 0.0167, and $\eta^2_{partial}$ =0.071. Although we didn't find significant difference in general, in the post hoc test did highlight the average performance of sensitivity in the final trail that:

- the DG with the SQ(p=0.041) showed a significant positive effect.

Vectormap Pattern Recognition

First Trial

Time

Estimated Colormaps cell means were RB: 44.72±45.53, AD: 52.23±35.42, DG: 38.77±25.72, SQ: 40.40±32.97, and AS: 40.03±27.23.

Muchly's test indicated that the assumption of sphericity is violated with p =0.108. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ = 1698.741, $F(4,52)=0.62$, $p=.65 > 0.0167$, and η2partial =0.046.

Sensitivity

Estimated Colormaps cell means were RB: 0.79±1.57, AD: 0.79±1.57, DG: 0.79±1.57, SQ: 0.36±1.50, and AS: -0.29±1.10.

Muchly's test indicated that the assumption of sphericity is not violated with p =0.953. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ =12.508, $F(4,52)=1.56$, $p=0.199 > 0.0167$, and $\eta^2_{partial}$ =0.107. Although we didn't find significant difference in general, in the post hoc test did highlight the average performance of sensitivity in the first trail that:

- the AD with AS(p=0.019) showed a significant positive effect.

Final Trail

Time

Estimated Colormaps cell means were RB: 31.45±24.11, AD: 27.74±16.01, DG: 31.12±30.56, SQ: 27.88±19.12, and AS: 35.58±42.78.

Muchly's test indicated that the assumption of sphericity is not violated with p =0.089. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ =578.067, $F(4,52)=0.37$, $p=.829 > 0.0167$, and $\eta^2_{partial}$ =0.028.

Sensitivity

Estimated Colormaps cell means were RB: 0.79±1.57, AD: 0.57±1.55, DG: 0.57±1.55, SQ: -0.29±1.10, and AS: 0.14±1.42.

Muchly's test indicated that the assumption of sphericity is violated with p =0.126. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ = 10.423, $F(4,52)=1.156$, $p=.341 > 0.0167$, and $\eta^2_{partial}$ =0.082.

Abnormal Spot Finding

First Trial

Time

Estimated Colormaps cell means were RB: 27.61±14.45, AD: 25.13±15.90, DG: 23.27±8.33, SQ: 41.0 ±33.49, and AS: 21.07±6.38.

Muchly's test indicated that the assumption of sphericity is violated with $p < 0.001$, therefore degrees of freedom were corrected using Huyhn-Feldt correction instead. The main effect of Colormaps was statistically significant with $SS_{colormaps} = 3705.022$, $F(4,56)=3.865$, $p=.008$, and $\eta^2_{partial} = 0.216$. Then, we run post hoc tests, highlighting exactly where these differences occur. We found out that in the average performance of sensitivity in the first trail:

- the AS with AD (p=0.01), the DG (p=0.03), and the SQ(p=0.028) indicated a significant negative effect (p=0.018).

Sensitivity

Estimated Colormaps cell means were RB: 4.71±0.89, AD: 3.95±1.28, DG: 3.62±1.53, SQ: 3.37±1.68, and AS: 4.80±0.83.

Muchly's test indicated that the assumption of sphericity is not violated with $p = 0.431$. The main effect of Colormaps was statistically significant with $SS_{colormaps} = 24.761$, $F(4,56)=5.09$, $p=.001 < 0.0167$, and $\eta 2_{partial} = 0.267$. Then, we run post hoc tests, highlighting exactly where these differences occur. We found out that in the average performance of sensitivity in the first trail:

- the RB with AD (p=0.042), the DG (p=0.08), and the SQ(p=0.02) indicated a significant positive effect.
- the AS with the SQ(p=0.006) indicated a significant positive effect.

Final Trail

Time

Estimated Colormaps cell means were RB: 17.78±8.21, AD: 21.12±12.58, DG: 20.36±9.25, SQ: 36.95±24.80, and AS: 16.32±5.96.

Muchly's test indicated that the assumption of sphericity is violated with $p = 0.001$, therefore degrees of freedom were corrected using Huyhn-Feldt correction instead. The main effect of Colormaps was statistically significant with $SS_{colormaps} = 4134.085$, $F(4,56)=7.937$, $p<0.001$, and $\eta 2_{partial} = 0.362$. Then, we run post hoc tests that can highlight exactly where these differences occur. We found out that in the average performance of time in the finaltrail:

- the AS with RB (p=0.08), AD (p=0.018), the DG (p=0.04), and the SQ(p=0.02) indicated a significant negative effect.

Sensitivity

Time

Estimated Colormaps cell means were RB: 4.72±0.87, AD: 4.14±1.28, DG: 4.69±0.92, SQ: 3.26±1.50, and AS: 5.13±0.00011.

Muchly's test indicated that the assumption of sphericity is not violated with p =0.147. The main effect of Colormaps was not statistically significant with $SS_{colormaps}$ = 31.166, $F(4,56)=8.421$, $p<0.001$, and $\eta_{2partial}$ =0.375. We found out that in the average performance of sensitivity in the final trail:

- the RB with the SQ(p=0.003) indicated a significant positive effect.
- the AS with the AD (p=0.01 <0.05), SQ(p<0.001) indicated a significant positive effect.

4.3 Exploratory

4.3.1 Learning Effect

To validate whether the colormaps can be learned by practice, we conducted a repeated measures ANOVA to investigate the main effect of learning by analyzing the difference between the performance of time in the first trail and the final trail (Brengelmann, 1959).

Heatmap Pattern Recognition

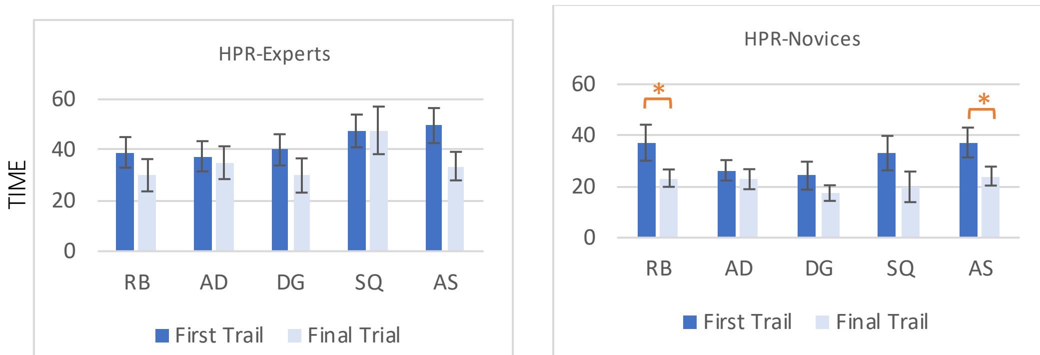


Figure 8.the improvement of the time performance in the Heatmap Pattern Recognition task

In the Heatmap Pattern Recognition task, the result showed that the main effect of the time in the RB colormap is not significant in the expert group: $F(1,13)=3.141, p=0.1$, $\eta_p^2=0.195$, observed power=0.375. However, in the novice group, the effect is significant: $F(1,14)=4.665$, $p=0.049$, $\eta_p^2=0.25$, observed power= 0.52.

Vectormap Pattern Recognition

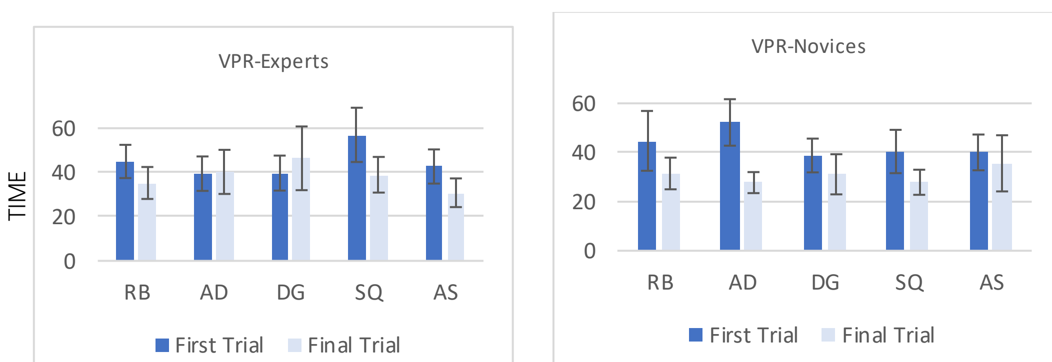


Figure 9.the improvement of the time performance in the Vectormap Pattern Recognition task

In the Vectormap Pattern Recognition, the main effect of time in the novice group, the AD is significant with $F(1,13)=8.309$, $p=0.013$, $\eta_p^2=0.39$, observed power=0.76.

In the Abnormal Spot Finding, the main effect of time in the expert group, the AD is significant with $F(1,14)=4.806$, $p=0.046$, $\eta_p^2=0.256$, observed power=0.532; the ASwith $F(1,14)=6.145$, $p=0.027$, $\eta_p^2=0.305$, observed power=0.636. By contrast in the Novice group, the DG with $F(1,14)=9.88$, $p=0.007$, $\eta_p^2=0.414$, observed power=0.832; the AS with $F(1,14)=9.5$, $p=0.008$, $\eta_p^2=0.404$, observed power=0.817.

Abnormal Spot Finding

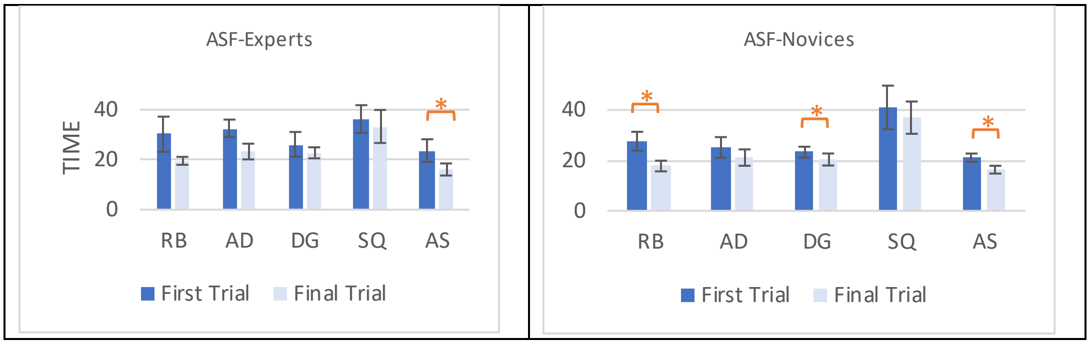


Figure 10. the improvement of the time performance in the Abnormal Spot Finding task

In the Abnormal Spot Finding task, the main effect of the time of the RB colormap is not significant in the expert group: $F(1,14)=2.669$, $p=0.125$, $\eta_p^2=0.160$, observed power=0.331. However, once again, in the Novice group, the effect is significant: $F(1,14)=6.01$, $p=0.028$, $\eta_p^2=0.3$, observed power=0.626.

4.3.2 Overall Performance

To investigate the performance among the 5 colormaps in average, analysis of repeated measure, and paired-samples t-test were performed.

Heatmap Pattern Recognition

Time

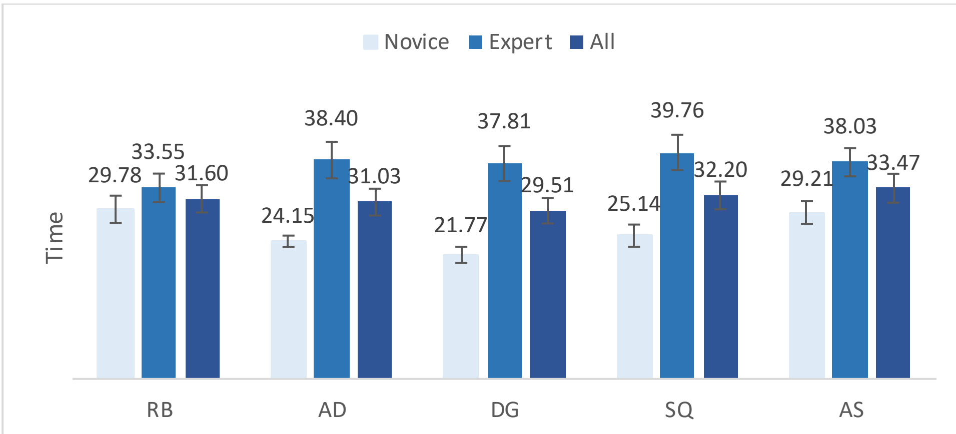


Figure 11.the comparison of the time performance among colormaps in the HPR task

All Participants

Results showed that there is no significant difference among the five colormaps in regard of time in average ($F(4,112)=0.558, p=0.694$, $\eta_p^2=0.024$, observed power=0.126).

Expert

In the Expert group, the average of time within five colormaps is 37.56±18.41 seconds. It showed that there is no significant difference among the five colormaps in regard of time in average ($F(4,52)=0.59, p=0.671$, $\eta_p^2=0.043$, observed power=0.183).

Novice

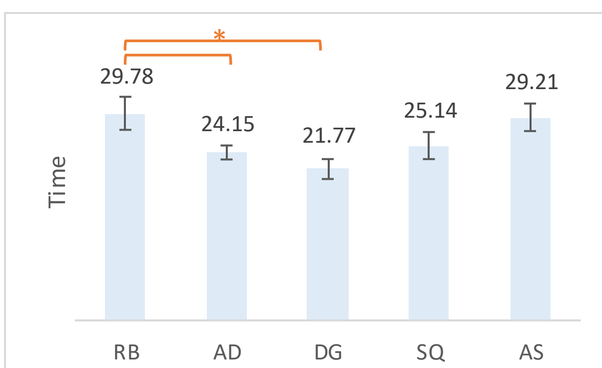


Figure 12.the comparison of the time performance in the novice group in the HPR task

In the Novice group (see Figure 12), the average of time within five colormaps is 26.01±16.00 seconds. It showed that there is no significant difference among the five colormaps in regard of time in average ($F(4,56)=2.153, p=0.086$, $\eta_p^2=0.133$, observed power=0.600). In the further investigation by conducting paired-samples t-test we found:

- The Rainbow colormap performed significantly slower than Alternative Diverging colormap; $t(14)=2.254$, $p=0.041$.
- The Rainbow colormap performed significantly slower than Diverging colormap; $t(14)=2.661$, $p=0.019$.

Sensitivity

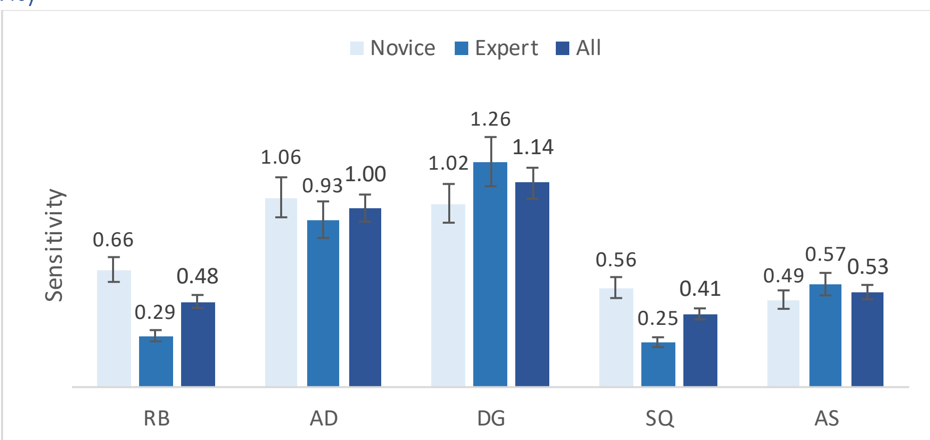


Figure 13.the comparison of the sensitivity performance among colormaps in the HPR task

All Participants

Figure 14 showed that there is a significant difference among the five colormaps in regard of the Sensitivity in average ($F(4,112)=8.086, p<0.001$, $\eta_p^2=0.22$, observed power=0.998). We found:

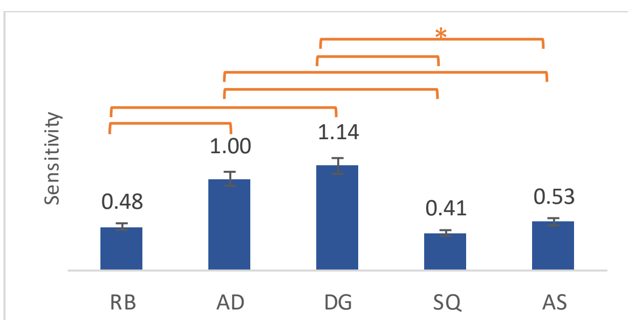


Figure 14.the comparison of the sensitivity performance from all participants in the HPR task

- The Rainbow colormap performed significantly lower than Alternative Diverging colormap; $t(28)=-3.20$, $p=0.003$.
- The Rainbow colormap performed significantly lower than Diverging colormap; $t(28)=-5.168$, $p<0.001$.

- The Alternative Diverging colormap performed significantly higher than Sequential colormap; $t(28)=3.12$, $p=0.004$.
- The Alternative Diverging colormap performed significantly higher than Alternative Sequential colormap; $t(28)=2.42$, $p=0.022$.
- The Diverging colormap performed significantly higher than Sequential colormap; $t(28)=4.11$, $p<0.001$.
- The Diverging colormap performed significantly higher than Alternative Sequential colormap; $t(28)=3.57$, $p=0.001$.

Expert

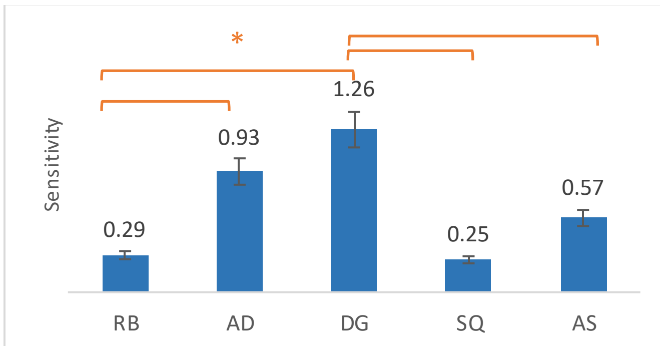


Figure 15. the comparison of the sensitivity performance in the expert group in the HPR task

In the expert group (see Figure 15), the average of sensitivity within five colormaps is 0.66 (N=70, SD=0.88). It showed that there is a significant difference among the five colormaps in regard of the Sensitivity in average ($F(4,52)=5.777, p=0.001$, $\eta_p^2=0.308$, observed power=0.972). We found:

- The Rainbow colormap performed significantly lower than Alternative Diverging colormap; $t(14)=-2.43$, $p=0.029$.
- The Rainbow colormap performed significantly lower than Diverging colormap; $t(14)=-5.21$, $p<0.001$.
- The Diverging colormap performed significantly higher than Sequential colormap; $t(14)=4.18$, $p=0.001$.
- The Diverging colormap performed significantly higher than Alternative Sequential colormap; $t(14)=2.75$, $p=0.016$.

Novice

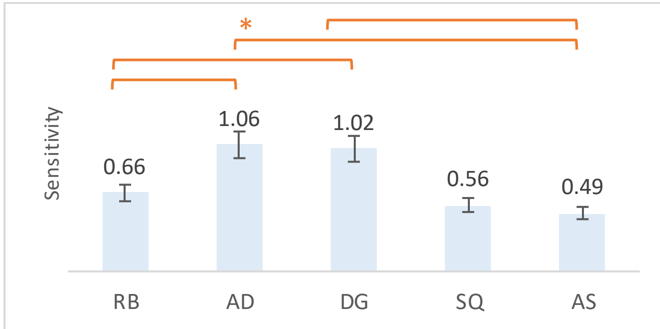


Figure 16. the comparison of the sensitivity performance in the novice group in the HPR task

In the novice group (see Figure 16), the average of sensitivity within five colormaps is 0.76 (N=75, SD=0.74). showed that there is a significant difference among the five colormaps in regard of the Sensitivity in average ($F(4,56)=3.141, p=0.021$, $\eta p2=0.183$, observed power=0.785). We found:

- The Rainbow colormap performed significantly lower than Alternative Sequential colormap; $t(14)=-2.26$, $p=0.041$.
- The Rainbow colormap performed significantly lower than Diverging colormap; $t(14)=-2.96$, $p=0.01$.
- The Alternative Diverging colormap performed significantly higher than Alternative Sequential colormap; $t(14)=2.61$, $p=0.021$.
- The Diverging colormap performed significantly higher than Alternative Sequential Colormap; $t(14)=2.36$, $p=0.033$.

Vectormap Pattern Recognition

Time

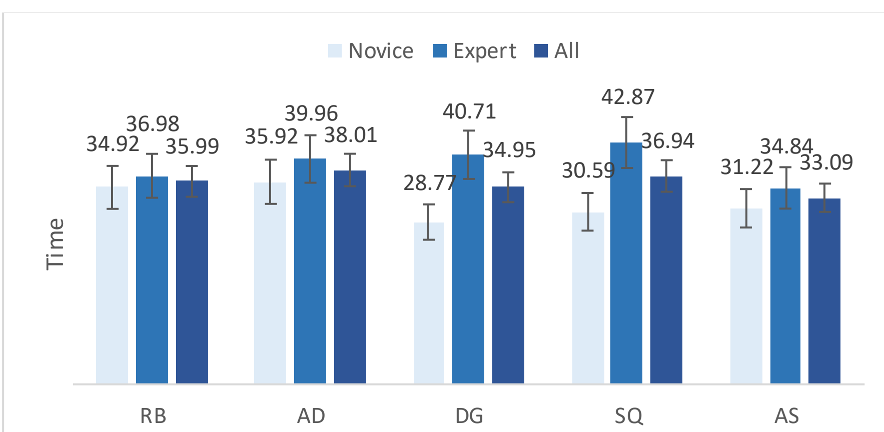


Figure 17. the comparison of the time performance among colormaps in the VPR task

All Participants

Results showed that there is no significant difference among the five colormaps in regard of time in average ($F(4,112)=0.557, p=0.694$, $\eta_p^2=0.02$, observed power=0.181).

Expert

In the Expert group showed that there is no significant difference among the five colormaps in regard of time in average ($F(4,56)=0.632$, $p=0.641$, $\eta_p^2=0.043$, observed power=0.195).

Novice

In the Novice group showed that there is no significant difference among the five colormaps in regard of time in average ($F(4,52)=0.982$, $p=0.426$, $\eta_p^2=0.07$, observed power=0.289).

Sensitivity

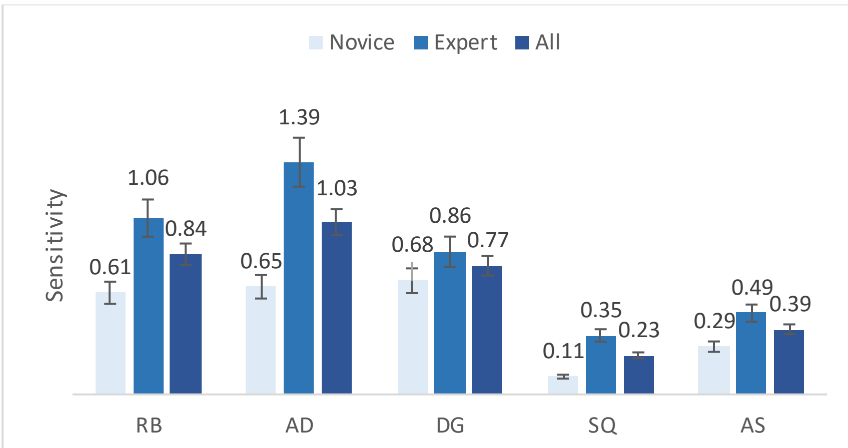


Figure 18. the comparison of the sensitivity performance among colormaps in the VPR task

All Participants

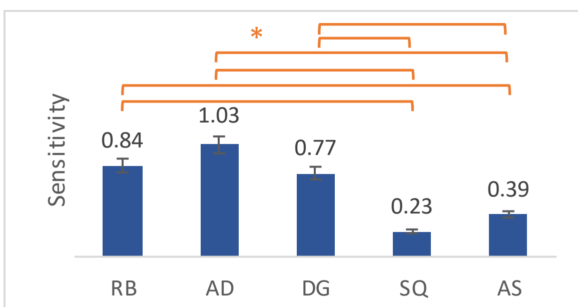


Figure 19. the comparison of the sensitivity performance from all participants in the VPR task

Figure 19 showed that there is a significant difference among the five colormaps in regard of the Sensitivity in average ($F(4,112)=10.02, p<0.001$, $\eta_p^2=0.263$, observed power=1.0). We found:

- The Rainbow colormap performed significantly higher than the Sequential colormap; $t(28)=5.141$, $p<0.001$.

- The Rainbow colormap performed significantly higher than the Alternative Sequential colormap; $t(28)=3.227$, $p=0.003$.
- The Alternative Diverging colormap performed significantly higher than the Sequential colormap; $t(28)=4.79$, $p<0.001$.
- The Alternative Diverging performed significantly higher than the Alternative Sequential colormap; $t(28)=3.972$, $p<0.001$.
- The Diverging performed significantly higher than the Sequential colormap; $t(28)=3.287$, $p=0.003$.
- The Diverging performed significantly higher than the Alternative Sequential colormap; $t(28)=2.856$, $p=0.009$.

Expert

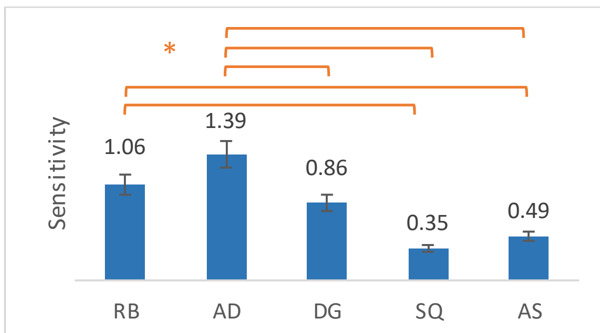


Figure 20. the comparison of the sensitivity performance in the expert group in the VPR task

The Expert group (see Figure 20) showed that there is a significant difference among the five colormaps in regard of the Sensitivity ($F(4,56)=7.747$, $p<0.001$, $\eta_p^2=0.356$, observed power=0.996). We found:

- The Rainbow colormap performed significantly higher than the Sequential colormap; $t(14)=5.501$, $p<0.001$.
- The Rainbow colormap performed significantly higher than the Alternative Sequential colormap; $t(14)=3.012$, $p=0.009$.
- The Alternative Diverging colormap performed significantly higher than the Diverging colormap; $t(14)=4.281$, $p=0.001$.
- The Alternative Diverging colormap performed significantly higher than the Sequential colormap; $t(14)=3.829$, $p=0.002$.
- The Alternative Diverging colormap performed significantly higher than the Alternative Sequential colormap; $t(14)=2.256$, $p=0.041$.

Novice

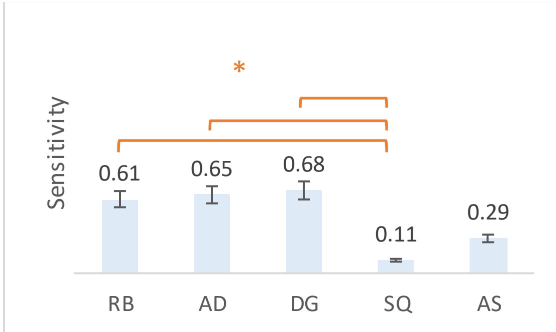


Figure 21. the comparison of the sensitivity performance in the novice group in the VPR task

In the Novice group (see Figure 21) showed that there is a significant difference among the five colormaps in regard of the Sensitivity in average ($F(4,52)=3.281, p=0.018$, $\eta_p^2=0.202$, observed power=0.802). We found:

- The Rainbow colormap performed significantly higher than the Sequential colormap; $t(13)=2.463$, $p=0.029$.
- The Alternative Diverging colormap performed significantly higher than the Sequential colormap; $t(13)=2.519$, $p=0.026$.
- The Diverging colormap performed significantly higher than the Sequential colormap; $t(13)=2.447$, $p=0.029$.

Abnormal Spot Finding

Time

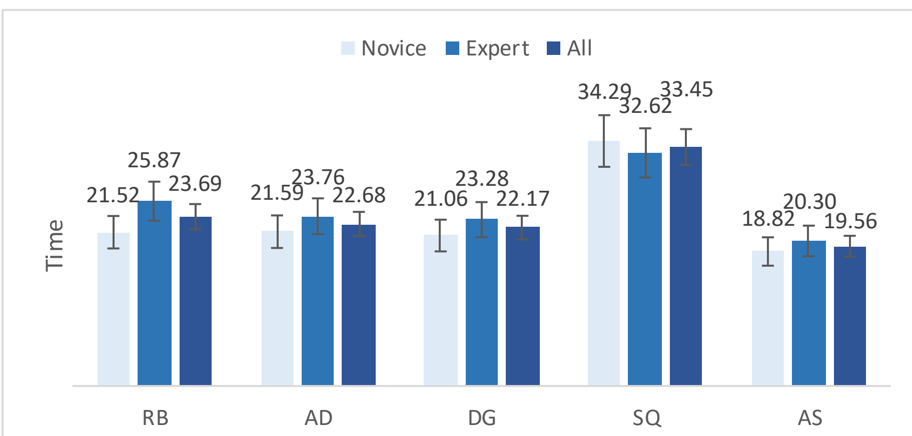


Figure 22. the comparison of the time performance among colormaps in the ASF task

All Participants

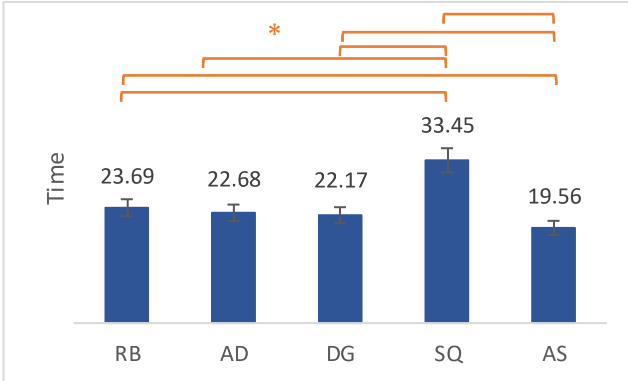


Figure 23. the comparison of the time performance from all participants in the ASF task

Figure 23 showed that there is a significant difference among the five colormaps in regard of the Time in average ($F(4,116)=13.47, p<0.001$, $\eta_p^2=0.317$, observed power=1.0). We found:

- The Rainbow colormap performed significantly faster than Sequential colormap; $t(29)=-3.756$, $p=0.001$.
- The Rainbow colormap performed significantly slower than the Alternative Sequential colormap; $t(29)=2.565$, $p=0.016$.
- The Alternative Diverging performed significantly faster than the Sequential colormap; $t(29)=-3.840$, $p=0.001$.
- The Diverging performed significantly faster than the Sequential colormap; $t(29)=-5.105$, $p<0.001$.
- The Diverging performed significantly slower than the Alternative Sequential colormap; $t(29)=2.064$, $p=0.048$.
- The Sequential colormap performed significantly slower than the Alternative Sequential colormap; $t(29)=5.891$, $p<0.001$.

Expert

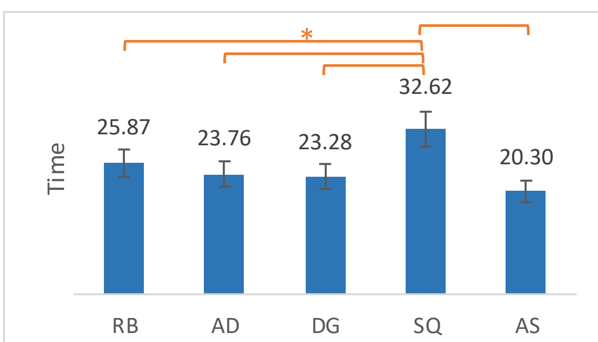


Figure 24. the comparison of the time performance in the expert group in the ASF task

In the Expert group (see Figure 24) showed that there is a significant difference among the five colormaps in regard of the Time in average ($F(4,56)=4.361, p=0.004$, $\eta_p^2=0.238$, observed power=0.912). We found:

- The Rainbow colormap performed significantly faster than Sequential colormap; $t(14)=-2.457$, $p=0.028$.
- The Alternative Diverging performed significantly faster than the Sequential colormap; $t(14)=-1.946$, $p=0.072$.
- The Diverging performed significantly faster than the Sequential colormap; $t(14)=-3.162$, $p=0.007$.
- The Sequential colormap performed significantly slower than the Alternative Sequential; $t(14)=3.940$, $p<0.001$.

Novice

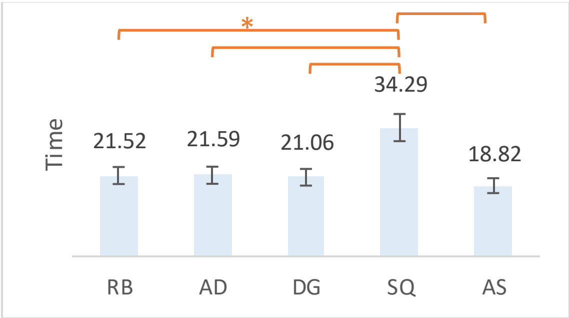


Figure 25.the comparison of the time performance in the novice group in the ASF task

In the Novice group (see Figure 25) showed that there is a significant difference among the five colormaps in regard of the Time in average ($F(4,56)=10.276, p<0.001$, $\eta_p^2=0.423$, observed power=1.0). We found:

- The Rainbow colormap performed significantly faster than the Sequential colormap; $t(14)=-2.918$, $p=0.011$.
- The Alternative Diverging colormap performed significantly faster than the Sequential colormap; $t(14)=-3.766$, $p=0.002$.
- The Diverging colormap performed significantly faster than the Sequential colormap; $t(14)=-3.993$, $p=0.001$.
- The Sequential colormap performed significantly slower than the Alternative Sequential colormap ; $t(14)=4.305$, $p=0.001$.

Sensitivity

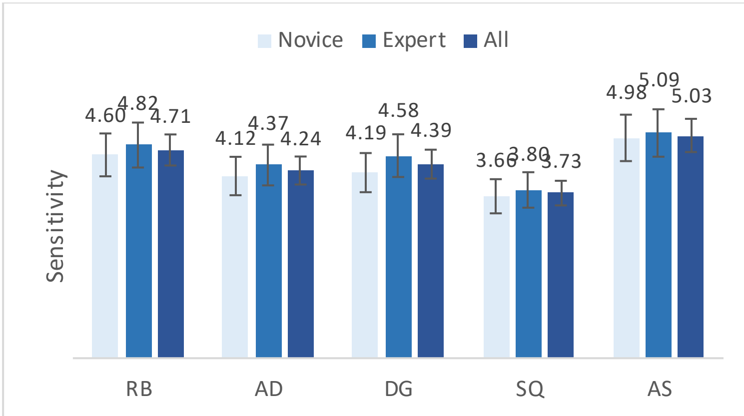


Figure 26. the comparison of the sensitivity performance among colormaps in the ASF task

All Participants

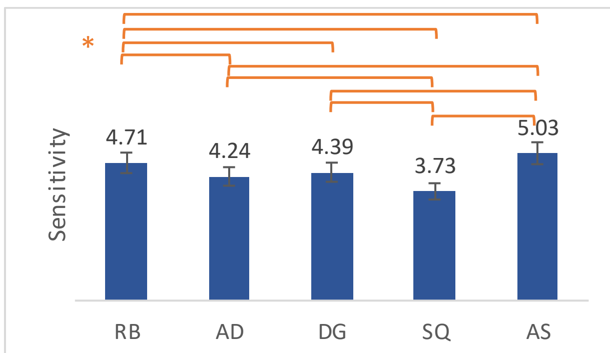


Figure 27. the comparison of the sensitivity performance from all participant in the ASF task

Figure 27 showed that there is a significant difference among the five colormaps in regard of the Sensitivity in average ($F(4,116)=31.57$, $p<0.001$, $\eta_p^2=0.521$, observed power=1.0). We found:

- The Rainbow colormap performed significantly higher than Alternative Diverging colormap; $t(29)=4.716$, $p<0.001$.
- The Rainbow colormap performed significantly higher than Diverging colormap; $t(29)=3.319$, $p=0.002$.
- The Rainbow colormap performed significantly higher than the Sequential colormap; $t(29)=6.697$, $p<0.001$.
- The Rainbow colormap performed significantly lower than the Alternative Sequential colormap; $t(29)=-3.757$, $p=0.001$.
- The Alternative Diverging colormap performed significantly higher than Sequential colormap; $t(29)=4.066$, $p<0.001$.
- The Alternative Diverging colormap performed significantly lower than Alternative Sequential colormap; $t(29)=-5.803$, $p<0.001$.
- The Diverging colormap performed significantly higher than the Sequential colormap; $t(29)=4.542$, $p<0.001$.

- The Diverging colormap performed significantly higher than the Alternative Sequential colormap; t(28)=4.11, p<0.001.
- The Sequential colormap performed significantly lower than Alternative Sequential colormap; t(29)=-7.762, p=<.001.

Expert

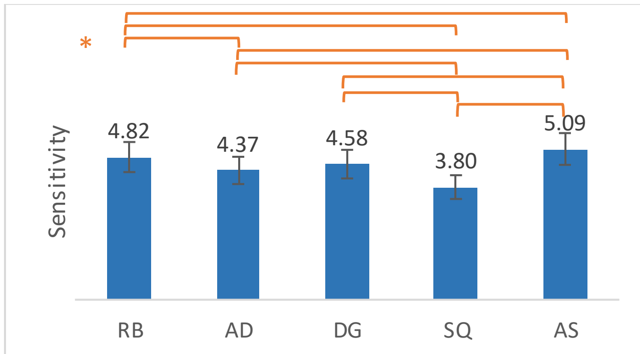


Figure 28. the comparison of the sensitivity performance in the expert group in the ASF task

In the Expert group (see Figure 29) showed that there is a significant difference among the five colormaps in regard of the Sensitivity in average ($F(4,56)=17.650, p<0.001$, $\eta_p^2=0.558$, observed power=1.0). We found:

- The Rainbow colormap performed significantly higher than Alternative Diverging colormap; t(14)=3.105, p=0.008.
- The Rainbow colormap performed significantly higher than the Sequential colormap; t(14)=5.322, p<0.001.
- The Rainbow colormap performed significantly lower than the Alternative Sequential colormap; t(14)=-2.871, p=0.012.
- The Alternative Diverging colormap performed significantly higher than Sequential colormap; t(14)=3.028, p=0.001.
- The Alternative Diverging colormap performed significantly lower than Alternative Sequential colormap; t(14)=-3.685, p=0.002.
- The Diverging colormap performed significantly higher than the Sequential colormap; t(14)=3.963, p=0.001.
- The Diverging colormap performed significantly lower than the Alternative Sequential colormap; t(14)=-4.493, p=0.001.
- The Sequential colormap performed significantly lower than Alternative Sequential colormap; t(14)=-5.783, p=<0.001.

Novice

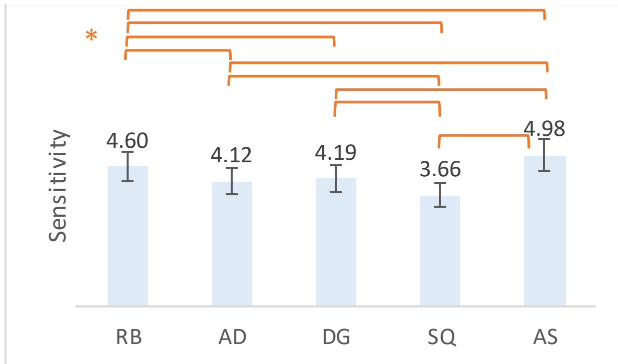


Figure 29. the comparison of the sensitivity performance in the novice group in the ASF task

In the Novice group (see Figure 29) showed that there is a significant difference among the five colormaps in regard of the Sensitivity in average ($F(4,56)=14.08, p<0.001$, $\eta_p^2=0.501$, observed power=1.0). We found:

- The Rainbow colormap performed significantly higher than Alternative Diverging colormap; $t(14)=3.462$, $p=0.004$.
- The Rainbow colormap performed significantly higher than Diverging colormap; $t(14)=2.56$, $p=0.023$.
- The Rainbow colormap performed significantly higher than the Sequential colormap; $t(14)=4.131$, $p=0.001$.
- The Rainbow colormap performed significantly lower than the Alternative Sequential colormap; $t(14)=-2.58$, $p=0.012$.
- The Alternative Diverging colormap performed significantly higher than Sequential colormap; $t(14)=2.63$, $p=0.02$.
- The Alternative Diverging colormap performed significantly lower than Alternative Sequential colormap; $t(14)=-4.417$, $p=0.001$.
- The Diverging colormap performed significantly higher than the Sequential colormap; $t(14)=2.498$, $p=0.026$.
- The Diverging colormap performed significantly lower than the Alternative Sequential colormap; $t(14)=-4.417$, $p=0.001$.
- The Sequential colormap performed significantly lower than Alternative Sequential colormap; $t(14)=-5.089$, $p<0.001$.

4.3.3 Prediction

To further investigate whether a mental model of the rainbow colormap affected how participants predict their performance, a simple linear regression was calculated to predict the performance (time & sensitivity) based on the Confident (from one of the items in the outtake survey- How confident you are of giving the answer by using the rainbow colormap in this task).

Heatmap Pattern Recognition

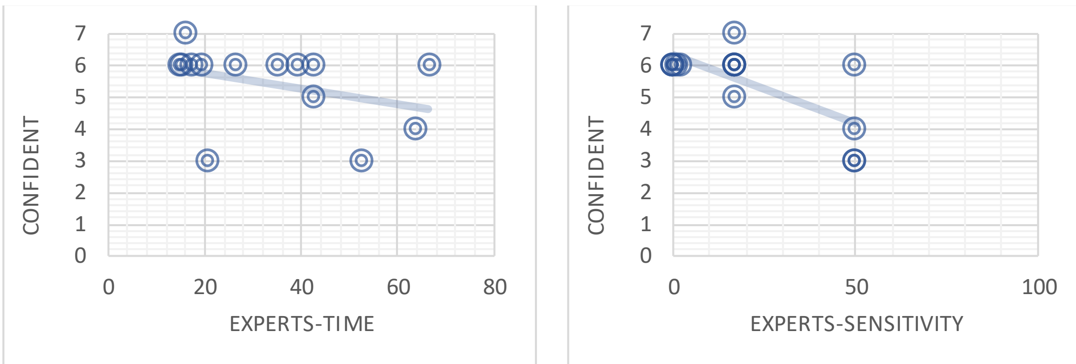


Figure 30.prediction of the experts' performance in HPR task

Experts- Time

The residuals had a normal distribution as determined with a Shapiro-Wilk test, with w=0.885 and p=0.069. We found a non-significant regression equation ($F(1,12)=0.685$, $p=0.426$), with and $R^2=0.059$. Expert group predicted their performance of time is equal to 55.47 – 4.16 (confident) seconds when confident is measured in scores. Expert group's performance of time decreased 4.16 for each score of confident.

Experts- Sensitivity

The residual normal distribution as determined with a Shapiro-Wilk rejected the null-hypothesis, with w=0.862 and p=0.033. Also, We found a non-significant regression equation ($F(1,12)=0.70$, $p=0.420$), with and $R^2=0.060$. Expert group predicted their performance of Sensitivity is equal to 1.41 – 0.21 (confident) scores when confident is measured in scores. Expert group's performance of Sensitivity decreased 0.21 for each score of confident.

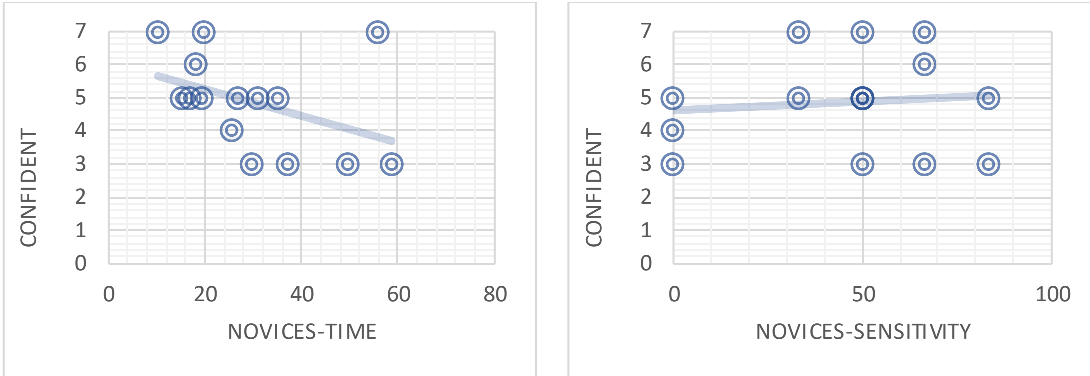


Figure 31.prediction of the novices' performance in HPR task

Novices- Time

The residual normal distribution as determined with a Shapiro-Wilk rejected the null-hypothesis, with w=0.854 and p=0.020. We found a non-significant regression equation ($F(1,13)=2.666$, $p=0.126$), with and $R^2=0.170$. Novice group predicted their performance of time is equal to 50.328 – 4.222 (confident) seconds when confident is measured in scores. Novice group's performance of time decreased 4.222 for each score of confident.

Novices- Sensitivity

The residuals had a normal distribution as determined with a Shapiro-Wilk test, with w=0.930 and p=0.270. We found a non-significant regression equation ($F(1,13)=0.136$, $p=0.718$), with and $R^2=0.010$. Novice group predicted their performance of time is equal to 0.37 + 0.59 (confident) scores when confident is measured in scores. Novice group's performance of Sensitivity increased 0.59 for each score of confident.

Vectormap Pattern Recognition

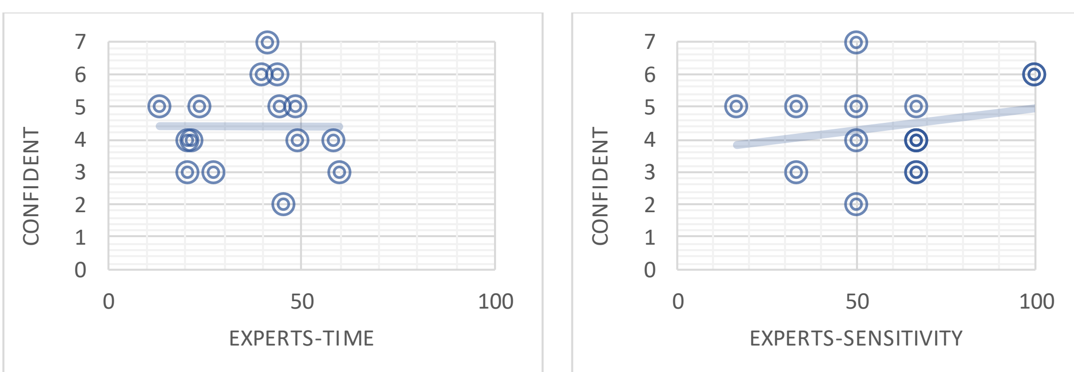


Figure 32.prediction of the experts' performance in VPR task

Experts- Time

The residuals had a normal distribution as determined with a Shapiro-Wilk test, with $w=0.923$ and $p=0.218$. We found a non-significant regression equation ($F(1,12)<0.001$, $p=0.989$), with and $R^2<0.001$ Expert group predicted their performance of time is equal to 37.17 – 0.04 (confident) seconds when confident is measured in scores. Expert group's performance of time decreased 0.04 for each score of confident.

Experts- Sensitivity

The residuals had a normal distribution as determined with a Shapiro-Wilk test, with $w=0.960$ and $p=0.688$. We found a non-significant regression equation ($F(1,12)=0.70$, $p=0.418$), with and $R^2=0.051$. Expert group predicted their performance of Sensitivity is equal to 0.56 – 0.11 (confident) scores when confident is measured in scores. Expert group's performance of Sensitivity decreased 0.11 for each score of confident.

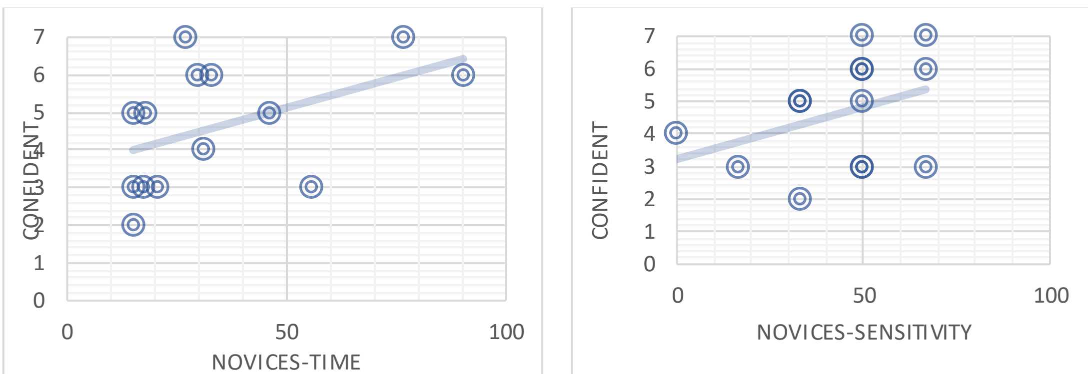


Figure 33.prediction of the novices' performance in VPR task

Novices- Time

The residuals had a normal distribution as determined with a Shapiro-Wilk test, with $w=0.899$ and $p=0.107$. We found a non-significant regression equation ($F(1,12)=3.368$, $p=0.091$), with and $R^2=0.219$. Novice group predicted their performance of time is equal to 3.327 + 6.81 (confident) seconds when confident is measured in scores. Novice group's performance of time increase d 6.81 for each score of confident.

Novices- Sensitivity

The residuals had violated the normality with a Shapiro-Wilk test, with $w=0.958$ and $p=0.691$. We found a non-significant regression equation ($F(1,12)=1.954$, $p=0.187$), with and $R^2=0.14$. Novice group predicted their performance of time is equal to -0.001 + 0.13 (confident) seconds when confident is measured in scores. Novice group's performance of time increased 0.13 for each score of confident.

Abnormal Spot Finding

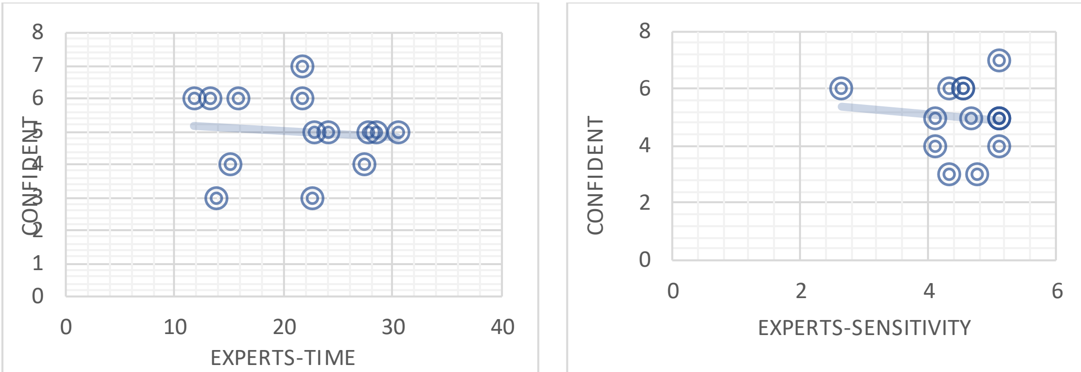


Figure 34.prediction of the experts' performance in ASF task

Experts- Time

The residuals had violated the normality with a Shapiro-Wilk test, with w=0.641 and p<0.001. We found a non-significant regression equation (F(1,12)=0.165, p=0.692), with and R^2=0.013. Expert group predicted their performance of time is equal to 19.40 + 1.37 (confident) seconds when confident is measured in scores. Expert group's performance of time increased 1.37 for each score of confident.

Experts- Sensitivity

The residuals had a normal distribution as determined with a Shapiro-Wilk test, with w=0.96 and p=0.688. Also, We found a non-significant regression equation (F(1,12)=1.184, p=0.296), with and R^2=0.084. Expert group predicted their performance of Sensitivity is equal to 5.18 – 0.08 (confident) scores when confident is measured in scores. Expert group's performance of Sensitivity decreased 0.08 for each score of confident.

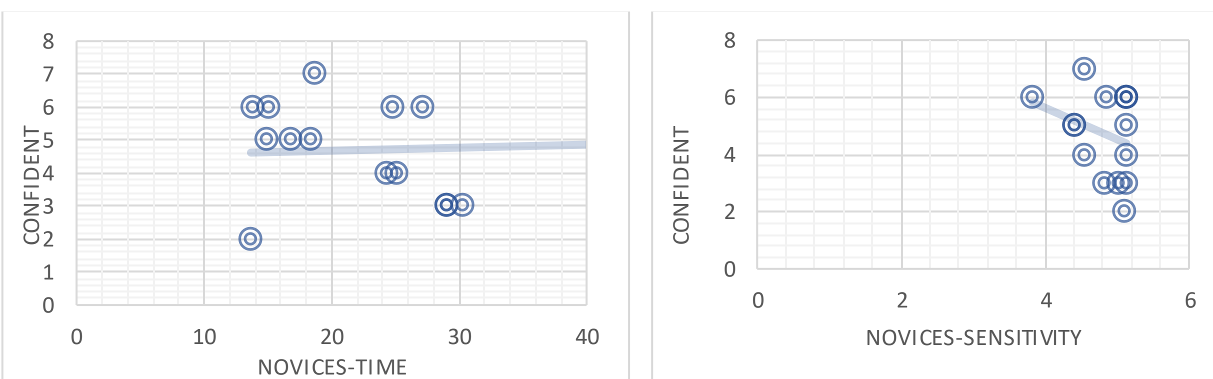


Figure 35.prediction of the novices' performance in ASF task

Novices- Time

The residuals had a normal distribution as determined with a Shapiro-Wilk test, with w=0.918 and p=0.206. We found a non-significant regression equation (F(1,12)=0.119, p=0.737), with and R^2=0.10. Novice group predicted their performance of time is equal to 23.78 – 0.53 (confident) seconds when confident is measured in scores. Novice group's performance of time decreased 0.53 for each score of confident.

Novices- Sensitivity

The residuals had violated the normality with a Shapiro-Wilk test, with w=0.835 and p=0.014. We found a non-significant regression equation ($F(1,12)=0.152$, $p=0.704$), with and $R^2=0.012$. Novice group predicted their performance of time is equal to 4.88 – 0.06 (confident) scores when confident is measured in scores. Novice group's performance of Sensitivity decreased 0.06 for each score of confident.

5 Discussion

This paper aims to give a further explanation and new insights on how the mental model impacts the interpretation of data visualization. Additionally, these findings can be used to establish the colormap strategy guidelines for practical usage in scientific field.

5.1 Hypothesis Testing

In the hypotheses, we assumed that the colormaps which are in accordance with the perceptual theories should outperform the rainbow colormap in the current experiment (Borkin et al, 2011; Moreland, 2009; Rogowitz and Treinish, 1996). However, since the expert group had a rich mental model of the rainbow colormap, which may influence the performance at the beginning.

Our results partially supported the hypotheses, demonstrating that from the Heatmap Pattern Recognition and Abnormal Spot Finding task, the experts performing in faster rate with the Rainbow colormap in the first trail. Which indicated that a higher familiarity on the default colormap can have an intuitive grasp of interpretation (Klein, 1990). But when reaching the final trail, the perceptually correct colormap(s) outperform the Rainbow colormap. Especially, in the Abnormal Spot Finding task, the difference of time between the rainbow and the perceptually correct colormaps (Alternative Sequential colormap in this task) was close to statistically significant from the first to the final trial ($p=0.218$; $p=0.058$). Moreover, the performance of sensitivity between the Rainbow and the Alternative Sequential colormap revealed a significant trend in the last trail ($p=0.018$). Therefore, we suggest that a rich mental model assisted experts to interpret data in a more effective way with the Rainbow colormap in the beginning, but after practice, the perceptually correct colormap could reach a higher performance.

In the second hypothesis, since the novices had no prior knowledge of the rainbow colormap, a weak mental model may not influence their interpretation (Glaser and Chi, 1988). As such, we assumed that the perceptually correct colormap(s) always outperform the Rainbow colormap. Our results although were not significant but supported the trend, we found that the perceptually correct colormap(s) performed better than the Rainbow colormap in respect of the time and sensitivity at the beginning and the end of the trails. For the Heatmap Pattern Recognition (HPR) and Abnormal Spot Finding (ASF) task, we found a nearly significant difference of the time between the rainbow colormap and the perceptually correct colormap from the first to the end of the trail (HPR: from $p=0.069$ to $p=0.056$, ASF: from $p=0.079$ to $p=0.06$). Which indicated that the differences between the colormaps became nearly significant overtime.

The study from Heatmap Pattern Recognition and Abnormal Spot Finding task showed the difference in results between the two groups as the hypotheses expected. However, an inverse result had been found in the Vector Pattern Recognition (VPR) task, revealed that regardless the groups, the Rainbow colormap outperformed other colormaps in the final trail. In theory, the irregular increase in lightness from the rainbow colormap may cause difficulty (Borland and Li, 2007), but then, the hue-varying and saturated colors can assist people to distinguish partial data more effectively (Ware, 1988). As such, the Rainbow colormap could still work well for this specific task.

To sum up, the performances of the expert and novice group were different at the first trail but showed a similar pattern at the end. Even though, part of the results was not significantly different, the direction of the effect is in parallel with what could be expected from earlier research. As such, the perceptually correct colormap(s) could do a better job than the rainbow colormap for interpreting scientific data. In the meanwhile, the presence of a mental model of the rainbow colormap might influence how people perform at the beginning.

Table 7. Overview of the Hypothesis Testing

	H1-1(the first trial)	H1-2(the final trial)
	Heatmap Pattern Recognition (HPR)	
H1(Experts)	Experts performed faster but less sensitivity with the Rainbow colormap than the Diverging colormap.	Experts performed faster and more sensitive with the Diverging colormap than the Rainbow colormap.
	Vectormap Pattern Recognition (VPR)	
	Experts performed slower and less sensitive with the Rainbow colormap than the Alternative Diverging colormap.	Experts performed faster and more sensitive with the Rainbow colormap than the Alternative Diverging colormap.
	Abnormal Spot Finding (ASF)	
	Experts performed slower and less sensitivity with the Rainbow colormap than the alternative Sequential colormap.	Experts performed faster with the Alternative Sequential colormap than the Rainbow colormap. Additionally, experts performed high sensitivity with the Alternative Sequential colormap than the Rainbow colormap.
	Heatmap Pattern Recognition (HPR)	
H2(Novices)	Novices performed faster and higher sensitivity with the Diverging colormap than the Rainbow colormap.	
	Vectormap Pattern Recognition (VPR)	
	Novices performed faster and higher sensitivity with the Diverging colormap than the Rainbow colormap at the first trial. After practice, the usage of Rainbow colormap showed a faster and higher sensitivity at the final trial.	
	Abnormal Spot Finding (ASF)	
	AS showed a significantly faster and higher sensitivity than the Rainbow colormap. The difference of the sensitivity was significant different at the final trail.	

5.2 Exploratory findings and future research

5.2.1 Guidelines for a Colormap Strategy

The research casts new light on how performance can be varying while applying colormaps in various scientific data visualization. We established the colormap strategy guidelines by analyzing the average performance of colormaps. The results also led to some surprising findings while comparing the HPR and VPR task.

Heatmap Pattern Recognition (HPR)

The HPR task made participants to interpret the correct neutral pattern based on the data plot rendered by one of the five colormaps. We found the Diverging and Alternative Diverging colormap had the highest score of sensitivity as well as the shortest of time to interpret. The result can support the perceptual theory that symmetrically increasing luminance can increase accuracy and efficiency at recognizing the pattern. By contrast, the Sequential and Alternative Sequential colormap, two monotonically increasing luminance on colormaps performed relatively weaker, and so did the Rainbow colormap in the HPR task.

Guideline #1:

To interpret patterns of scientific data with the heatmap setting, we suggest symmetrical lightness increasing as the critical strategy for generating the colormap. The ideal colormap for example, the Alternative Diverging and the Diverging colormap from this research. By contrast, the Rainbow, Sequential, and the Alternative Sequential colormaps shall be avoided to harness.

Vectormap Pattern Recognition (VPR)

The VPR task had a similar goal as the HPR task (i.e. the pattern recognition). The major difference between these two maps is that instead of having consistent gradient changing by shape in the heatmap, the vectormap had scattered gradient changing by arrow spots. In the results, we found there was no different in term of the time among the five colormaps. However, the Rainbow, the Diverging, and the Alternative Diverging shared the same highest score of sensitivity. Although the results were not entirely consistent with expectation, it is still clear enough that colormaps with symmetrically increasing luminance can be more accurate and efficient than monotonic lightness changing. Also, since the density of graphic was sparser than the heatmap (from shapes to arrows), we assume the contrast sensitivity theory (Ware, 1988) played an important role to assist the process of distinguishing pattern boundaries that created by arrows. Therefore, the Rainbow with varying saturated hues can stand out in the VPR task.

Guideline #2:

To interpret patterns of scientific data with the vectormap setting, we suggest the symmetrical lightness increasing should be the crucial consideration in the colormap design strategy. Moreover, a colormap having the hue variation can also improve the readability. The ideal colormap for instance: the Alternative Diverging, Rainbow, and Diverging colormap.

Abnormal Spot Finding (ASF)

From the task- ASF which required participants to find out as many as possible of spots with suspicious values on the data plot rendered by one of the five colormaps. We obtained the Alternative Sequential had the highest score of sensitivity as well as with the shortest time among colormaps. The result was supported by color theory that hue- and saturation-varying will lead to more accurate quantalities estimation at specific locations (Rogowitz et al, 1996).

Guideline #3:

To differentiate or quantify the data spots like outlier finding for instance, the colormaps with varying of hue and saturated colors can achieve a higher accuracy and effectiveness. The ideal colormap for instance: the Alternative Sequential.

5.2.2 Mental Model

We further investigated several aspects of data between experts and novices which may provide additional support of how the mental model of the rainbow colormap influence subjects perceiving data physically and psychologically.

Learning Effect

Firstly, to validate whether participants can learn while using the colormaps during the experiment, the results showed that the novices had larger improvement in regard of the time than the experts in the three different tasks. Especially in the Heatmap Pattern Recognition and Abnormal Spot Finding task, the experts had no significantly difference while the novices showing a statistically difference of the improvement. It provided a support to the effectiveness of the experiment grouping approach, explaining that the novice group with weaker mental model of the rainbow colormap can be improved by practice. Nonetheless, for the expert group with a richer mental model of the rainbow, there was a weaker learning effect can be observed.

Overall Performance

Secondly, we further analyzed a general performance of colormaps between experts and novices by comparing the average time and sensitivity.

Heatmap Pattern Recognition (HPR)

We found in average the novices performed faster and more sensitive than the experts. Statistically, the overall sensitivity between novices and experts is not significantly different with $p=0.474$, while the time is significantly different with $p<0.001$. This is an interesting but counter-intuitive finding since many researches stated that experts make decision more effectively than novices. However, a potential explanation suggested that experts monitor the strategies utilized to solve the problem more carefully than novices (Voss, Greene, Post, & Penner, 1983; Newell& Simon, 1972). Thus, experts devote proportionately more time to create a representation of problem than the novices (Andreas Sofroniou, 2017). The HPR task is featured with multiple randomized, irregular geographic-like contours as the data plots. Experts may tend to spend more time to establish a right solving strategy while novices creating a weaker structure but faster decision making strategy instead. Also, the design of the task was referred to experts' daily work which may add extra pressure to them for trying to perform equally well. We don't have a clear interpretation of a reason why the experts had slightly lower sensitivity than the novices. One explanation is that the way of interpretation which experts use all the time in their daily w ork may have some flaws which could also explain why the Rainbow colormap was significantly lower than Alternative Diverging and Diverging with sensitivity. Experts tended to give a decision without rational analysis process (Franklin, 2014). Hence, with an unsound mental model they couldn't make an accurate decision but performed slightly worse than the novices.

Vectormap Pattern Recognition (VPR)

In general, the novices performed faster but less sensitive than the experts. Statistically, the difference of time between the experts and novices is not significant with p=0.082, but the sensitivity is significantly different with p=0.004. Comparing to the similar task – Heatmap Pattern Recognition, we found an opposite result here. One possible explanation to the difference between these two pattern recognition tasks is that the strategy of interpretation is fundamentally different. The vector plot contains scattered arrows to portrait patterns which may require two different perceptual skills: the pattern re cognition and quantity estimation. Since experts construct a meaningful strategy of interpretation to cope with complex situation more effectively than novices (Meleis, 2010), they can notice patterns without wasteful consideration (Klein, 1990). Moreover, from the perceptual theory by Ware (1988), suggested that hue-varying and saturated-colors like the Rainbow color is an ideal colormap for quantity estimation. In this regard, the experts with a rich mental model of the Rainbow colormap could be expected to outperform the novices in this specific task.

Abnormal Spot Finding (ASF)

In general, the novices performed faster but less sensitive than the experts in this task. Statistically, the difference of sensitivity and time between the groups were not significant with p=0.093, p=0.425. The results showed that experts performed slightly better than novices in regard of the time and the sensitivity. The similar patterns in the VPR and ASF, indicated that experts had a better interpretation strategy for the vectormap task.

Prediction of Performance

To further investigate whether a mental model of the rainbow colormap effected the performance, we analyzed subjects' prediction of their personal results of tasks by interpreting a regression by one criterion from the survey – the Confidence (How confident you are of giving the answer by using the rainbow colormap in this task) and the two variables of performance. We uncovered some patterns that showed an interesting difference between the participants.

In the HPR task, results revealed that the experts couldn't precisely predict their actual performance with the rainbow colormap. The linear regression showed that with more confident the experts were, they performed the task faster but less accuracy.

By contrary, the novice group had a rational pattern of colormaps in the sensitivity and the time. The linear regression indicated that as more confidence the novices had, the faster interpretation and higher sensitivity they performed.

To sum up, the results suggested that the difference between participants with different levels of mental model is that the experts thought they did well by using the Rainbow colormap even when in reality they did not perform as well as using the perceptual correct colormap(s).

In the VPR and ASF task, we found a weak difference between groups to the Rainbow colormap prediction. It seemed like the experts could have a rather rational prediction than the HPR task. One plausible explanation is that the visual characteristics of the vectormap have a perfect fit with hue-varying and saturated colors in colormaps (Ware, 1988). Therefore, using the Rainbow colormap in the two tasks did not mislead the experts' prediction.

5.3 Limitation & Future Works

There are some considerations and limitations that need to be taken care in future research.

Firstly, during the experiment, an experimenter stayed in the meeting room with the participants, explaining the tutorial before each task started. Although the purpose was to check whether the study procedure was on track, it may also bring out a concern on the anxiety of participants that being monitored by experimenters. Future study should set up a better experiment flow to control this factor.

Secondly, the research lessened the amount of trails to prevent an overlong study that lead to tiredness. Nonetheless, it left a question that whether the colormaps had reached its learning plateau at the end of trail or it may still have a space for improving. Future research could either focus on a single task or split all the tasks into several independent studies, so that the results are able to be determined the ultimate performance of the colormap.

Thirdly, the experiment was tried to replicate the actual task setting to gain the genuine performance to each colormaps. However, to minimize and control all the variables in plots, we had to simplify the information in the plots and the task setting itself. Therefore, the experts may encounter a difficulty to fill in the gap between the experiment and the actual task. The results may not be able to apply into a real case.

The last but not least, it is surprising that the pattern recognition task applied by two different types of data visualizations indicated the opposite performance ranking from the same colormaps. Which evokes an interest that the original pattern estimation theory needed to have a further investigating in different types of data visualization. The perceptual density or types of data map might require different perceptual theories to gain the optimal performance.

6 Conclusion

We conducted three experiments to examine the performance of the colormaps and the effects of mental model. The results revealed that the subjects with a rich mental model of the rainbow colormap made an effective interpretation but it didn't enhance the accuracy of giving right answer. The effect also biased experts' prediction of their actual performance. Furthermore, we found that three experiments required different perceptual theory for establishing the optimal colormap strategy. In the Heatmap Pattern Recognition, the design of colormap should follow symmetrical and rigorous lightness increased; In the

Abnormal Spot Finding, the design of colormap should consider the hue variation; In the Vectormap Pattern Recognition, the design of colormap should take both the lightness and hue variation into account. Which led to a new update and recommendations for current data visualization design guidelines.

7 References

1. Adler, P & Clark, K. (1991). Behind the Learning Curve: A Sketch of the Learning Process. *Management Science. 37*. 267-281.
2. Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological Review, 89, 369-406.
3. Benner, P. (1982) From novice to expert. *American Journal of Nursing* 82(3), 402–407.
4. Borgman, C. L. (1986). The users mental model of an information retrieval system: An experiment on a prototype online catalog. *International Journal of Man-Machine Studies,24*(1), 47-64. doi:10.1016/s0020-7373(86)80039-6
5. Borkin, M., Gajos, K., Peters, A., Mitsouras, D., Melchionna, S., Rybicki, F., Feldman, C. and Pfister, H. (2011). Evaluation of Artery Visualizations for Heart Disease Diagnosis. *IEEE Transactions on Visualization and Computer Graphics,* 17(12), pp.2479-2488.
6. Borland, D., & Ii, R. T. (2007). Rainbow Color Map (Still) Considered Harmful. *IEEE Computer Graphics and Applications, 27*(2), 14-17. doi:10.1109/mcg.2007.323435
7. Brengelmann, J. C. (1959). *The effect of repeated electroshock on learning in depressives*. Berlin (etc.): Springer.
8. Brettel,H., Viénot, F., Mollon, J. D. (1997). Computerized simulation of color appearance for dichromats. *J. Opt. Soc. Am.*, vol. 14, no. 10, pp. 2647-2655.
9. Brewer, C. (2006). COLORBREWER 2.0. Retrieved November 15, 2018, from http://www.ColorBrewer.org/
10. Card, S. K., Mackinlay, J. D. & Shneiderman, B. (1999) *Readings in information visualization: using vision to think.* San Francisco, California: Morgan Kaufmann Publishers. 686p.
11. Correll, M., Witmore, M., & Gleicher, M. (2011). Exploring Collections of Tagged Text for Literary Scholarship. *Computer Graphics Forum,30*(3), 731-740. doi:10.1111/j.1467-8659.2011.01922.x
12. Cukier, K. (2010). *A special report on managing information*. The Economist 394 (8671), 3-18.
13. Feldman, L., Cao, J., & Fried, G. (2009). 27. A Method to Characterize the Learning Curve for Performance of a Fundamental Laparoscopic Simulator Task: Defining "Learning Potential" and "Learning Rate". *Journal of Surgical Research,151*(2), 183-184. doi:10.1016/j.jss.2008.11.043
14. Few, S. (2004). *Show Me the Numbers: Designing Tables and Graphs to Enlighten.* Analytics Press, Oakland, CA
15. Few, S. (2009). Introduction to Geographical Data Visualization. *Visual Business Intelligence Newsletter* [online]. Dostupné z: http://www.perceptualedge.com/articles/visual_business_intelligence/geographical data visualization.pdf
16. Franklin, C. L, II. (2013). Developing expertise in management decision-making Acad. *Strategic Manage.* J., 12 (1) , pp. 21-36
17. Gaetano, J. M., Lancaster, S., Tindle, R. (2015). *Signal detection theory calculator 1.0* [Excel workbook downloaded from https://www.researchgate.net/profile/Justin_Gaetano2/].

18. Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), The nature of expertise (pp. xv-xxviii). Mahwah, NJ: Lawrence Erlbaum Associates.
19. Gregg, B. (2014). Frequency Trails: Detecting Outliers. Retrieved October 21, 2018, from http://www.brendangregg.com/FrequencyTrails/outliers.html
20. Gupta, A., & Thirani, V. (2017). The value of data. Retrieved February 10, 2019, from https://www.weforum.org/agenda/2017/09/the-value-of-data/
21. Harvey, C. (2018, May 23). Big Data Market 2018: Growing and Moving to the Cloud. Retrieved February 15, 2019, from https://www.datamation.com/big-data/big-data-market-2018-growing-and-moving-to-the-cloud.html
22. Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement. Routledge. p. 141. ISBN 0-415-47617-8.*
23. Ivancevic, V. G., & Ivancevic, T. T. (2010). *Computational mind: A complex dynamics perspective*. Berlin: Springer.
24. Jarkas, A. M. (2010). Critical Investigation into the Applicability of the Learning Curve Theory to Rebar Fixing Labor Productivity. *Journal of Construction Engineering and Management,136*(12), 1279-1288. doi:10.1061/(asce)co.1943-7862.0000236
25. Johnson-Laird, P. N. (1995). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge: Harvard University Press.
26. Kandogan, E. (2001). Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 01*. doi:10.1145/502512.502530
27. Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning & Verbal Behavior, 1*(3), 153-161.
28. Klapp, O. E. (1986). Overload and boredom: Essays on the quality of life in the
29. Klein, G. (1990). Knowledge engineering. Beyond expert systems. 16. 27-41.
30. information society: Greenwood Publishing Group Inc.
31. Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. Hoboken, NJ: John Wiley & Sons.
32. Lant, T. K., & Argote, L. (2000). Organizational Learning: Creating, Retaining, and Transferring Knowledge. *Administrative Science Quarterly, 45*(3), 622. doi:10.2307/2667112
33. Lieberman, M. B. (1984). *The learning curve and pricing in the chemical processing industries*. Stanford: Graduate School of Business, Stanford University.
34. Light, A., & Bartlein, P. J. (2004). The end of the rainbow? Color schemes for improved data graphics. *Eos, Transactions American Geophysical Union, 85*(40), 385. doi:10.1029/2004eo400002
35. Lohse, G. L. (1997). The role of working memory on graphical information processing. *Behaviour & Information Technology, 16*(6), 297-308.
36. Lyster, R.: *Differential effects of prompts and recasts in form-focused instruction.* In: *Studies in Second Language Acquisition* #26 (2004), pp. 399-432.
37. Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.
38. McAfee, A. & Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review, 90, 60--68.

39. Meleis, A. I. (2010). Transitions Theory: Middle-range and situation-specific theories in nursing research and practice. New York, NY: Springer Publishing Co.
40. Mohammadi E, Abbasi K, Saadati M. (2015) Barriers to implementation of nursing theories in clinical practice from the perspective of nurses. *Iranian journal of nursing research*, 10, 13-24
41. Moreland, K. (2009). Diverging Color Maps for Scientific Visualization. *Advances in Visual Computing Lecture Notes in Computer Science,*92-103. doi:10.1007/978-3-642-10520-3_9
42. Mullen, K.T. (1982) The contrast sensitivity of human colour vision to red–green and blue–yellow chromatic gratings. *The Journal of Physiology 359*, 381–400
43. Newell, A., & Simon, H. A. Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery*, 1976,19, 113-126.
44. Norman, D. A. (1983). "Some observations on mental models". In *Mental models*, Edited by: Gentner, D. Ö. and Stevens, A. L. 7–14. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers Inc..
45. Parker, L. E., Schneider, F. E., & Schultz, A. C. (2005). Multi-robot systems: From swarms to intelligent automata. Dordrecht: Springer.
46. Phillips, J. K., Klein, G., & Sieck, W. R. (2008). Expertise in Judgment and Decision Making: A Case for Training Intuitive Decision Skills. *Blackwell Handbook of Judgment and Decision Making,*297-315. doi:10.1002/9780470752937.ch15
47. Reda, K., Nalawade, P., & Ansah-Koi, K. (2018). Graphical Perception of Continuous Quantitative Maps. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18*. doi:10.1145/3173574.3173846
48. Rheingans, P. L. (2000). Task-based color scale design. *28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*. doi:10.1117/12.384882
49. Roediger, H. L., & Smith, M. A. (2012). The "pure-study" learning curve: The learning curve without cumulative testing. *Memory & Cognition,40*(7), 989-1002. doi:10.3758/s13421-012-0213-5
50. Rogowitz, B. E., Treinish, L. A., & Bryson, S. (1996). How Not to Lie with Visualization. *Computers in Physics,10*(3), 268. doi:10.1063/1.4822401
51. Rogowitz, B., & Treinish, L. (1998). Data visualization: The end of the rainbow. *IEEE Spectrum,35*(12), 52-59. doi:10.1109/6.736450
52. Rosenberg, D., & Grafton, A. (2012). *Cartographies of time: A history of the timeline*. New York: Princeton Architectural.
53. Ross, K. G., Shafer, J. L., & Klein, G. (2006). Professional judgments and "naturalistic decision making". In R. R. Hoffmann (Ed.), *The Cambridge handbook of expertise and expert performance* (pp. 403-420). New York: Cambridge University.
54. Rouse, W. B., & Morris, N. M. (1984). On looking into the black box: Prospects and limits in the search for mental models. *PsycEXTRA Dataset*. doi:10.1037/e506172009-012
55. Rudis, B., Ross, N., & Garnier, S. (2018). The viridis color palettes. Retrieved November 10, 2018, from https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html
56. Schmidt, R. A., & Lee, T. D. (2005). *Motor control and learning: A behavioral emphasis*. Champaign, IL: Human Kinetics.

57. Shanteau, J. (1987). Psychological Characteristics of Expert Decision Makers. *Expert Judgment and Expert Systems,*289-304. doi:10.1007/978-3-642-86679-1_16
58. Shipstead, Z., Harrison, T. L., & Engle, R. W. (2012). Working memory capacity and visual attention: Top-down and Bottom-up guidance. *Quarterly Journal of Experimental Psychology, 65,*401–407.
59. Stauffer, R., Mayr, G. J., Dabernig, M., & Zeileis, A. (2015). Somewhere Over the Rainbow: How to Make Effective Use of Colors in Meteorological Visualizations. *Bulletin of the American Meteorological Society,96*(2), 203-216. doi:10.1175/bams-d-13-00155.1
60. Steele, J. (2010). *Beautiful visualization: Looking at Data Through the Eyes of Experts*. Sebastopol, CA: OReilly.
61. Stefanidis, D., Korndorffer, J. R., Heniford, B. T., Markley, S., Sierra, R., & Scott, D. J. (2006). Closing the gap in operative performance between novices and experts: Does harder mean better for laparoscopic simulator training? *Journal of the American College of Surgeons,203*(3). doi:10.1016/j.jamcollsurg.2006.05.201
62. Toker, S. and Moseley, J. (2013). The Mental Model Comparison of Expert and Novice Performance Improvement Practitioners. *Performance Improvement Quarterly*, 26(3), pp.7-32.
63. Tory, M., &Möller, T. (2004). Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics,* 10(1), 72–84. https://doi.org/10.1109/TVCG.2004.1260759
64. Vits, J., & Gelders, L. (2002). Performance improvement theory. *International Journal of Production Economics* 77(3): 285-298.
65. Vosniadou, S., & Brewer , W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24, 535-585.
66. Voss, J. F., Greene, T. R., Post, T. A., & Penner, B. C. (1983). Problem-Solving Skill in the Social Sciences. *Psychology of Learning and Motivation,*165-213. doi:10.1016/s0079-7421(08)60099-7
67. Ware, C. (1988). Color sequences for univariate maps: Theory, experiments and principles. *IEEE Computer Graphics and Applications,8*(5), 41-49. doi:10.1109/38.7760
68. Ware, C. (2004) *Information Visualization: Perception for Design. 2nd edn.* Morgan Kaufmann ISBN 1-55860-819-2.
69. Ware, C. (2013). *Information visualization: Perception for design*. Amsterdam: Elsevier.
70. Weber, C., & Fayed, A. (2010). Optimizing Your Position on the Operating Curve: How Can a Fab Truly Maximize Its Performance? *IEEE Transactions on Semiconductor Manufacturing,23*(1), 21-29. doi:10.1109/tsm.2009.2039248
71. Winn, W. (2002). Current trends in educational technology research: The study of learning environments. *Educational Psychology Review, 14*(3), 331–351.
72. Yohannes, P., Rotariu, P., Pinto, P., Smith, A. D., & Lee, B. R. (2002). Comparison of robotic versus laparoscopic skills: Is there a difference in the learning curve? *Urology,60*(1), 39-45. doi:10.1016/s0090-4295(02)01717-x
73. Zhou, L., & Hansen, C. D. (2016). A Survey of Colormaps in Visualization. *IEEE Transactions on Visualization and Computer Graphics, 22*(8), 2051-2069. doi:10.1109/tvcg.2015.2489649
74. Ziemer, R., and Kelly, P.(1993). The deflation of military aircraft, in Price Measurements

and Their Uses, Foss M., Manser M. and Young A. (Eds.). Chicago: University of Chicago Press.

Appendix A: Intake Questionnaire& Grouping Rubric

All data obtained during this study will be reported anonymously and will be in no way traceable back to you.

1) What is your job title at Company?

   ____

2) Year(s) at Company (including similar industries)?

   ______yr(s)

3) How long you have been using any analytical software (e.g. Matlab)?

   ______yr(s)

4) Frequency of encountering/ seeing data plots

   ☐ Never/ Less than 3 times a week
   ☐ More than 3 times a week

5) Frequency of reading, interpreting, analyzing data plots

   ☐ Never/ Less than 3 times a week
   ☐ More than 3 times a week

Appendix B: Outtake Questionnaire

Heatmap/ Vectormap Pattern Recognition

The following statements are rated on a 7-point Likert Scale (1= strongly disagree, 7= strongly agree)

	Colormap #01 	Colormap #02 	Colormap #03 	Colormap #04 	Colormap #05 
I found it easy to identify patterns					
I was able to perform the task efficiently.					
I am confident that all the patterns I identified are accurate.					
I prefer using this colormap.					

Abnormal Spot Finding

The following statements are rated on a 7-point Likert Scale (1= strongly disagree, 7= strongly agree)

	Colormap #01 	Colormap #02 	Colormap #03 	Colormap #04 	Colormap #05 
I found it easy to identify abnormal spots.					
I was able to perform the task efficiently.					
I am confident all the places I marked are really abnormal.					
I prefer using this colormap.					

Appendix C: Informed Consent Form

CONSENT FORM FOR PARTICIPATION IN RESEARCH

The purpose of this form is to inform you of the nature of the research conducted at [company] and to ask for your informed consent to participate.

Your participation in this experiment is strictly voluntary. You may choose not to participate, and you may withdraw at any time during the study. More information about this study has been explained to you through the tutorial. If you have any hesitations about participating or you have any questions about the research, please feel free to ask the researcher at [email] or approach the supervisors [name] [email],[name] [email] and [name] [email].

Note that the study will be expected to take 50 minutes which includes intake and outtake questionnaires, 3 independent experiments (each will take around 10-15 minutes) in total. You are able to take a rest in between.

Regardless, all data obtained during this study will be reported anonymously and will be in no way traceable back to you.

All research conducted at the [company] adheres to the Code of Ethics of the NIP (Nederlands Instituut voor Psychologen – Dutch Institute for Psychologists).

CERTIFICATE OF CONSENT

I, (NAME)………………………………….. have read and understood this consent form and have been given the opportunity to ask questions. I agree to voluntarily participate in this study carried out by the research group at [company] and Human Technology Interaction of the Eindhoven University of Technology.

____________________________ ______________________

Participant's Signature Date