Abstract

CÁRDENAS, LINA MARIA. Evaluation of Variability in Visual Assessment of Small Color Differences. (Under the direction of Drs. Renzo Shamey and David Hinks).

Several psychophysical methods and the associated intra and inter-observer variability in assessing small color differences of textile samples were investigated in three phases:

- Development and validation of a novel perceptually linear gray scale and incorporation into a visual assessment protocol;
- Development and determination of a robust visual assessment protocol via testing different psychophysical methods and viewing conditions;
- International replication experiment employing a controlled methodology to assess the repeatability of results and ascertain the degree of variability among observer panels from four continents.

During the first phase, a perceptually linear gray scale, tested under simulated illuminant D65 and a neutral gray background, was developed and validated for the assessment of magnitude of perceived color differences. The perceptually linear gray scale comprised ten discrete pairs of gray samples mounted onto a suitable support structure in a perceptual order of linearly increasing contrast. This scale was used as a reference in the development of visual assessment protocols for small color differences.

During the second phase, five visual assessment methodologies: AATCC standard gray scale, Jumbo scale, Jumbo scale with gap, novel perceptually linear scale, as well as pair comparison were evaluated to determine the method generating the lowest intra- and inter-observer variation. Thirty one polyester knitted samples around 13 color centers were assessed by panels of 25 observers. The visual assessment based on the use of the Jumbo scale, on average, produced the largest visual color differences for all pairs, whereas the assessments based on the standard AATCC scale produced the lowest visual color difference. A comparison of visual data obtained from different methods based on the STRESS function showed that various methodologies produced comparable visual data. However, intra- and inter-observer variability significantly decreased when the perceptually linear gray scale was employed relative to when any of the geometric scales were used.

A highly controlled replication study was carried out to test, systematically and for the first time, the range and magnitude of variability among four panels of observers from different regions of the world: Colombia, Czech Republic, USA, and a group of Chinese born and raised observers. Sixty nine polyester knitted samples around 13 color centers were assessed by 25 observers in each location using the same method.

On average, all panels demonstrated good agreement considering the perceptual nature of the study. The agreement between visual data from the individual panels and the total observer population increased for all observer panels suggesting that observer panels based on a mixed ethnic makeup may be better suited for investigations that involve human perceptions. Although there were differences in variability for each pair amongst observer panels, the degree of variation (intra-group standard deviation) among responses from any panel compared to the mean response of that panel was not significantly different amongst panels.

The performance of color difference equations was tested against the average visual data for each of the observer panels as well as the combined dataset using PF/3 and STRESS functions. Results showed that the ΔE^*_{ab} equation performed significantly poorer than $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models. In addition the STRESS function showed no significant difference between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models for the visual data obtained. Moreover, the inclusion of upper and lower confidence levels in the combined visual data did not change the PF/3 value significantly (<1 unit) while the change in STRESS values was also insignificant (<0.1).

Evaluation of Variability in Visual Assessment of Small Color Differences

by
Lina María Cárdenas

A dissertation submitted to the Graduate Faculty of
North Carolina State University
In partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy

Fiber and Polymer Science

Raleigh, North Carolina

February 24th, 2009

APPROVED BY:

Dr. Renzo Shamey Committee Co-chair	Dr. David Hinks Committee Co-chair
Dr. Warren Jasper	Dr. Roger Woodard

Dedication

To my husband Ignacio, for being my major support and motivation during this journey and to my Mom and Claudia who are no longer with us, but will always live in my heart.

Biography

Lina María Cárdenas was born in Sogamoso, Colombia, in 1979. She received her B.S in Textile Design from Universidad de los Andes, Bogotá, Colombia in 2002, and conducted her undergraduate research on rabbit hair felt used in the production of hats.

Acknowledgements

The author would like to thank Drs. Renzo Shamey and David Hinks, chairman and co-chairman of her advisory committee, for their advice, support, patience, and guidance throughout this research; Dr. Roger Woodard, member of the advisory committee, for the great contribution to the author's understanding, development, and interpretation of the statistical analysis. Their continuous encouragement and enthusiasm made an incredible difference for the author. Also, I would like to extend my gratitude to Dr. Warren Jasper, member of the advisory committee, for his valuable suggestions and encouragement. The author would also like to thank Prof. Rolf Kuehni for his invaluable ideas and comments.

Special thanks go to the National Textile Center, USA, for financial support of this project, without which this project would not have been possible.

In addition, the author thanks Dr. Ing. Michal Vik and Ing. Martina Viková (The Technical University of Liberec, Czech Republic), Ms. Sandra Gómez Guerrero (LaSalle College of Design, Colombia), and Drs. Roy Berns and Mark Fairchild (Munsell Color Science Laboratory, Rochester Institute of Technology), for their time, hosting the replication experiment and for providing the author with an opportunity to conduct visual assessments in their respective institutions.

Also, I would like to thank Dr. Maria Oliver-Hoyo, Ms. Laleh Shamey, Miss Li Liu, and Mr. Gang Fang for helping the author recruit Chinese observers.

The author is very grateful for the special paint developments made by the staff of Porter Paints at the Raleigh's office for the development of the perceptual linear gray scale.

The author would also like to thank all observers who participated in the visual assessments and those who collaborated in one way or another with the success of this study

A special thanks to Mr. John Darsey and colleagues at DyStar for generation of the dyed fabric samples.

Finally, the author would like to thank her husband, all members of her family, Miss Rebecca Klossner, Miss Reid Clonts, Mr. Aaron Horton, and Sandra Lopez for their continued love, support, help, and encouragement throughout this study.

Table of Contents

List of Figures

List of Tables

I. Introduction

Color is an integral component of the textile industry and is often considered to be a primary criterion in the selection of a product, thus significantly affecting sales volume. The main objective in color technology is to control and reproduce a target color on a specified material under a set of specified conditions and in the shortest possible period. In the textile industry, effective color control and communication between designer, dyer and retailer are critical. Color communication throughout the supply chain is a dynamic process. Attempts at optimizing color control must focus on the variability that arises due to the complex interaction between supplier and consumer. However, communication of the 'right' color is often subject to a large array of variable factors. Significant opportunities, therefore, exist in reducing color variability throughout the supply chain. Ideally, the 'right' color is one perceived by an observer (e.g. a consumer) that is as near as possible the color specified in the original product design. In many textile operations today, the control of color is commonly achieved via visual assessment with verification by color measurement. The visual assessment of color, even when conditions are closely controlled, is subjective and highly variable.

A practical approach to understanding the complexity of the number of variables in color control involves the use of a flowchart such as a fishbone diagram or cause-and-effect diagram. The fishbone diagram is a tool to identify the potential or real causes that contribute to a single outcome [1, 2].

The causes are organized in order of significance creating a hierarchical structure in relation to the outcome.

In color communication, a significant array of variables exist that need to be clearly identified, understood and ultimately minimized. A fishbone diagram can be used to depict some of the most important factors in the communication and control of color in the textile supply chain. The sources of variability in the control of color through the supply chain can be broken down into the following broad categories:

Concept;

Manufacturing;

Human Factor;

Color Quality Control and;

Point of Sale

Figure 1, developed as a component of this study, illustrates some of the most important causes of variability in the control and communication of color within the textile supply chain. Some of the important components of variation according to various sources have been included in this diagram.[3-5].

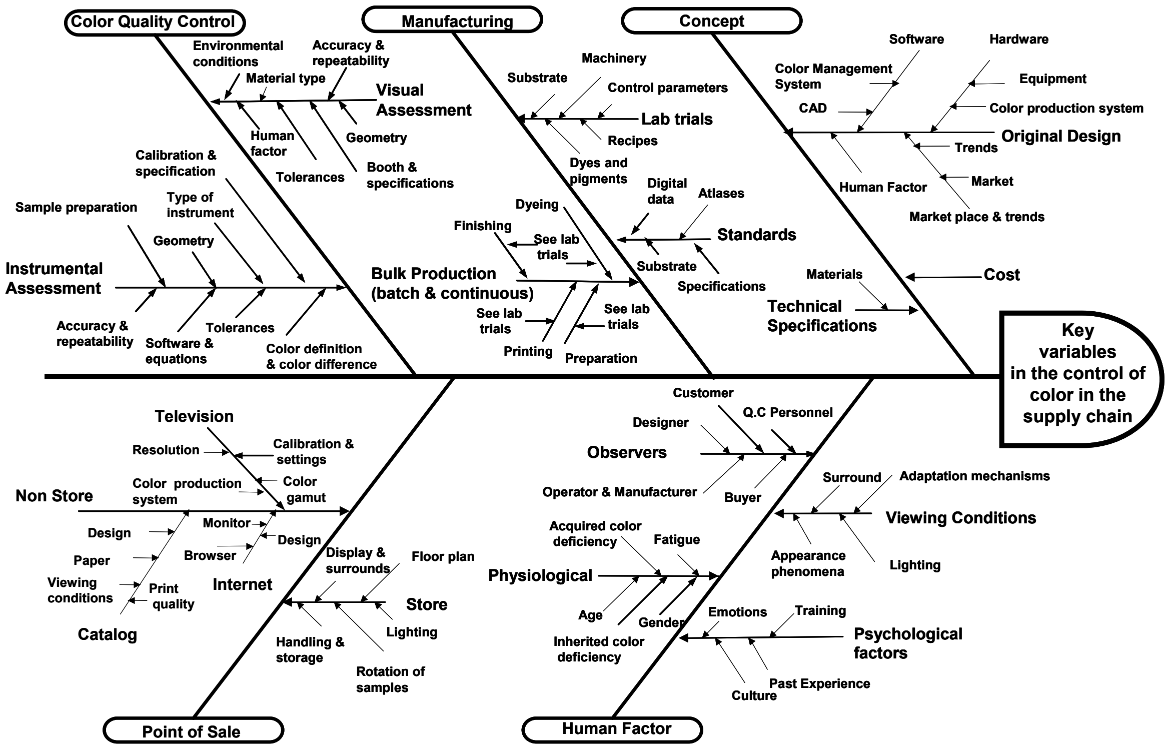


Figure 1. Key variables in the control of color in the supply chain.

The illustration of all or even the majority of variables would make such diagrams very complex, as exemplified by a comprehensive study of the variables of one batch dyeing process (which may form one small part of the supply chain fishbone diagram shown in Figure 1) [6]. However, a less complex design can be used as a framework to evaluate the significance of procedures and to indicate stages within the supply chain where critical problems may have been overlooked. Where necessary, a complete fishbone diagram can be developed for each sub-factor.

Arguably, the most important factor within the textile supply chain is the accurate assessment of color differences between two textile materials. Optimization of the correlation between the visual assessment of color and mathematical models that predict color differences is fundamental to any digital system of color management. However, this requires a clear understanding of the scope and limitations of the assessment methodology and optimization of numerical models.

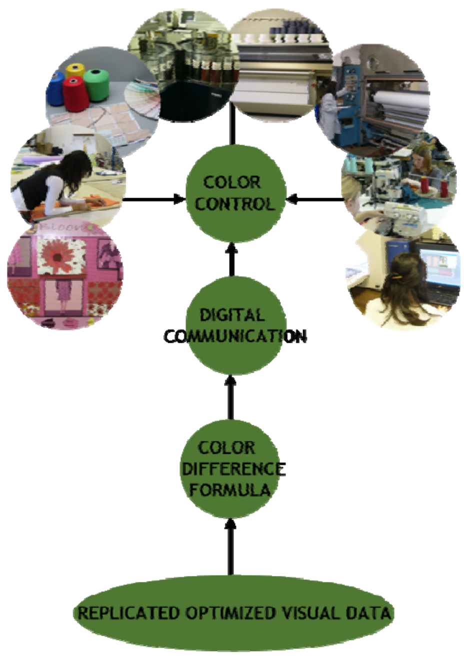


Figure 2. Flowchart of steps required for optimization of color control in the textile supply chain**.**

The existing models used to predict color differences and color variation are not optimized because they are based on different sets of perceptual data that have been established under various experimental conditions. Currently, the CMC (2:1) color difference formula is the most widely used standardized color difference model in the world, and is incorporated into standards by many international organizations [7]. In 2001, however, the International Commission on Illumination (CIE) recommended the CIEDE2000. A large dataset was used in the development of the CIEDE2000 formula. The dataset combined four separate experimental data. In addition, Luo et al. reported accuracy of prediction for several formulas using the PF/3 measure of performance. A value of 32.6 for the CIEDE2000 formula vs. 37.9 for CMC (1:1) was reported [8].

The visual assessment of color, even when conditions are closely controlled, is variable due to the subjective experiences of even well-trained expert assessors. Individuals tend to vary considerably when judging the perceived color difference between two colored samples [9]. A good example of such variability can be seen when individuals describe color without a reference sample as in the case of unique hues (UH) [9]. Controlled studies regarding the selection of UH have shown large variation among observers with normal color vision [10, 11]. Based on this observation, a degree of disagreement between colorists assessing a colored sample under the same viewing conditions could be expected [9]. Hence, it is critical to ascertain the level of variability in observers using highly controlled conditions of observation.

There is limited information regarding variation of individuals in perception of small color differences [9]. Information regarding observer variation may be limited because studies involving perception of small color differences employed a limited number of observers. Most of the studies reported have shown inter-observer variability of about 30% and intra-observer variation of about 50% [9].

In addition, variance can be added with different experimental conditions employed in each study. Studies in perception of small color differences have been carried out using a wide range of experimental conditions that not only vary in term of observers but also in terms of psychophysical methods [9], viewing parameters [12-14] and physical presentation of samples [15].

In order to ensure and determine if an improvement to the existing color difference models is possible, it is necessary to carry out new research to understand how differences in experimental methodology and some of the viewing parameters affect visual assessments. It is also necessary to establish a well designed and controlled experiment using identical surround conditions, sample and reference display and illuminations conditions [9]. Furthermore, to evaluate the reliability of the experimental conditions a replication study is needed to determine the degree of replicability with different panels of observers.

1. Research Proposal

1.1. Objectives

The primary goal of the proposed research is to optimize the experimental methodology for the assessment of small color differences and establish the minimum repeatable variability possible among a statistically significant set of observers under highly controlled viewing and illumination conditions. The specific objectives are:

- To carry out a complete analysis of published literature pertaining to color communication and control;
- To investigate the role of observer size and diversity in determining validity of visual assessments:
    - Examine the statistical significance of using expert vs. naïve observers.
- To investigate observer variability in small color difference assessments using the following psychophysical methods:
    - Pair Comparison Method.
    - Gray Scale Method with particular attention to:
        - Type of Scale (standard geometric scale and a new proposed scale).
        - Size of Scale.
        - Sample Separation.
        - Observer Population Size and Diversity.

- To replicate the highly controlled method in different parts of the world,
- To evaluate the reliability of the experimental conditions.
- To evaluate the performance of existing models and visual datasets based on the results of the replication study.
- To recommend procedures for visual assessment of colored objects that produce minimum variability among observers, and provide maximum correlation with mathematical models.
- To highlight areas that are need to be examined further in future to resolve outstanding issues pertaining to the assessment and communication of color.

II. Literature Review

1. Human Color Vision

It is perhaps surprising that there is not a complete explanation of how the human color vision system works. Some aspects of the process are now well-established, while others currently remain as ideas, particularly in the area of how we experience the sensation of color [16]. The visual mechanism involves four main areas: The anatomy of the eye, the relationship between the physics of light and the interaction with the eye, the biological process of transforming photons entering the eye into electrical signals relayed to the brain, and the psychological aspect of how we experience sensations [16]. In order to make significant progress in color-related research, it is necessary to understand what is known about the human vision mechanism.

1.1. The Eye

The eyes are the window to an outside world filled with information available to us. Thanks to a complex mechanism, humans experience vision. One of the principal components of human vision is the eye. Figure 3 shows a representation of the anatomy of the human eye. It is almost spherical and has a diameter of about 20 to 25 mm [17]. A camera functions in much the same way as the human eye [16-18] . A number of components inside the eye have an important role in visual perception.

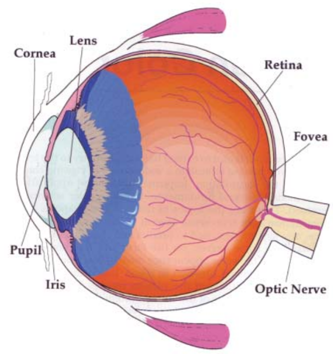


Figure 3. Human eye and its principal parts [17].

1.1.1. The Sclera

The sclera is a strong outer membrane that we perceive as the ‘white’ of the eye. Its shape is maintained by the pressure of the internal eye fluids [17].

1.1.2. The Cornea

The cornea is located in the front of the eye and is the only component externally exposed. It is a transparent surface layer allowing the passage of light into the internal area of the eye [16-18]. This membrane is approximately 0.5 mm thick and provides approximately two thirds of the optical power of the eye [19]. Variations in the shape of this membrane can lead to errors in vision [18].

1.1.3. The Iris

The iris is located behind the cornea. It is a colored sphincter membrane with a circular aperture in the center, and its appearance is what determines a person's eye color. The function of the iris is to control the amount of light that enters the eye. This control of light is done through the circular opening aperture called pupil [18, 19]. The opening of the pupil is usually determined by the level of illumination. The size of the pupil varies from approximately 1.5 mm when in bright light to approximately 8 mm when in darkness [19].The opening of the pupil can change due to non-visual experiences such as arousal [17, 18].

1.1.4. The Lens

The lens is a malleable transparent component of the eye located behind the iris. Its concave curvature can be changed by zonula muscles in order to focus images on the retina located at the back of the lens, when viewed at different distances [17].

This variable optical power is known as accommodation. The lens accounts for approximately the remaining third of the optical power of the eye [17]. The function of the lens tends to diminish with age, as the flexibility of the lens is lost due to dying cells in the interior of the lens that harden over time.

This explains why, when people reach the age range of 40-50, the ability to focus near objects often becomes difficult and reading glasses are needed [17, 19]. An important aspect of color vision related to the lens is its transparency across the visible spectrum. The lens tends to transmit less light in the blue end of the spectrum, which produces a measurable yellowish hue. The optical density of yellowness increases with age, which may result in different color experiences [17].

1.1.5. The Humors

The chamber between the lens and the cornea is filled with a clear fluid called aqueous humor. It is a clear liquid that serves as a source of nutrition of the lens and the cornea [18, 19]. The large chamber between the lens and the back of the eye is filled with another transparent fluid called the vitreous humor. This fluid has a higher viscosity than the aqueous humor and occupies almost 60% of the volume of the eye [17, 18].

1.1.6. The Retina

The components of the eye described so far are all involved in the process to form and focus an image on the retina, which is located at the back of the eye. It is a thin layer of cells approximately 0.1 mm thick [18, 19], in which the farthest layer of the retina contains photosensitive receptors known as rods and cones owing to their shape under magnification, as indicated in Figure 4. Section 1.2 describes the functions of the photoreceptors in detail.

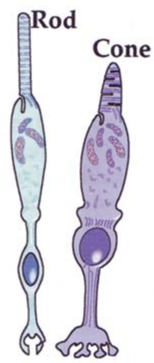


Figure 4. Shape of rods and cones [17].

Following the retina there is a layer called the pigmented epithelium. This dark layer inhibits light scattering by absorbing all the light that travels through the retina, i.e., all light that is not absorbed by the photoreceptor cells [18].

1.1.7. The Fovea and Macula

One of the most important sections of the retina is located around the optic axis, where images are focused. The optic axis is an imaginary line that passes through the center of the pupil and intersects the retina in a small area called the fovea [17, 19], as shown in Figure 5.

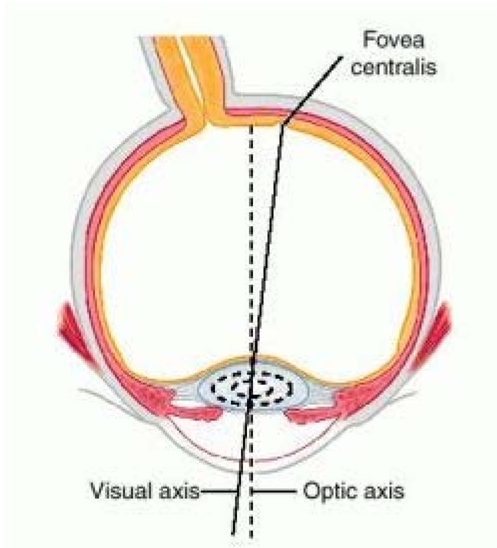


Figure 5. Optic axis.

There are approximately 7million cones and no rods located in the fovea of the human eye. The rods gradually begin to appear when moving away from the fovea [16], which is approximately at an angle subtended to the retina of five degrees. The macula is a larger region of the retina, subtending approximately 20 degrees away from the visual axis, and beyond this region is what may be considered to be peripheral vision [19].

1.2. The Eye and the Brain: How Color Vision Experience is Produced

The processing of visual information starts at the retina and it is known to be highly complex. Once the light enters the eye, it passes through a net of retinal neurons before it reaches the photoreceptors where the visual image is translated into electrical signals.

While the precise pathway through the neural cells is not completely understood, it is likely to be a combination of interconnections between the cells. Figure 6 shows a diagram of the cross section of the retina.

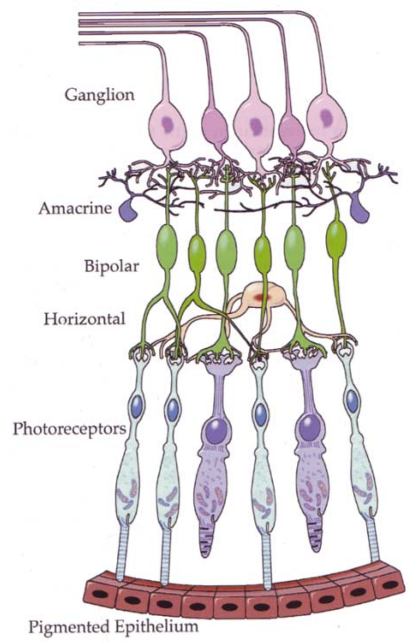


Figure 6. Cross-section of the retina [17].

Following the photoreceptors, the middle layer of the retina consists of bipolar cells. These cells are neurons that have two ends. One of the ends connects to the photoreceptors and the other one connects to the ganglion cells [17, 18]. Together, these cells form a vertical pathway. However, there are other retinal cells that also interconnect across them.

Near to the photoreceptors there are the horizontal cells. Other types of connectors are found between ganglion and bipolar cells. They are called amacrine cells [17].

Once the light reaches the photoreceptors a series of chemical transformations occur. The resulting electrical signals are processed by the layers of cells located in the retina (horizontal, bipolar, amacrine, and ganglion). The ganglion cells group together forming the optic nerve [18]. The optic nerve is a neural pathway that exits the eye through an orifice of the retina and the sclera. The electrical data carried through each axon is a combination of the data gathered by the responses from different photoreceptors located in a small region of the retina. The optic nerves from each eye come together to a point called the optic chiasm. At this point the optic nerves cross to the opposite side of the brain and lead to the lateral geniculate nucleus (LGN) which is located in the thalamus, as indicated in Figure 7 [17].

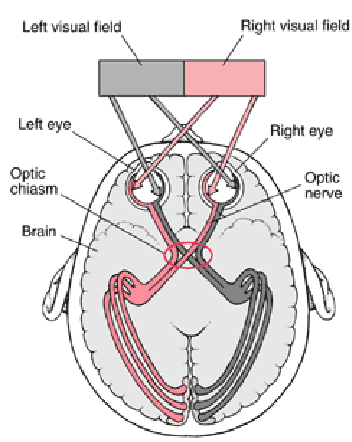


Figure 7. Diagram of the visual pathway to the LGN.

The cells in the LGN project the information to the occipital lobe of the cortex in an area called visual area V1. There are near to 30 visual areas in the cortex. The processing that happens inside the different visual areas includes a set of connections that finally allow the humans to have visual perception [18]. Even though it is not well understood, there is evidence that the main visual area related to color perception is V4. It was reported that a woman that suffered a lesion in the right side of the brain close to area V4 was able to see colors with her right eye, but unable with the left eye [17].

1.2.1. Receptive Fields

Receptive field is a term that refers to the concept of how the cells respond to a visual stimulus. The receptive filed is a representation of an area in the visual field for which a cell reacts. The type of response can be positive, negative or spectral bias and it is represented in a specific area of the receptive field. The response of a cell is determined from the input of different photoreceptors. Figure 8 shows a receptive field of a ganglion cell.

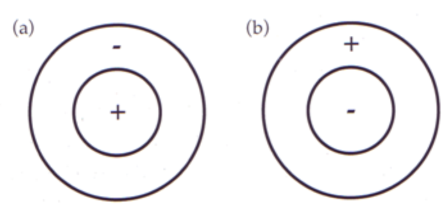


Figure 8. Receptive fields from a ganglion cell ((a)on-center, (b) off-center) [17].

The Figure illustrates the opposition between center and surround. The positive central response represents the positive input from a single cone surrounded by a negative response input from several cones. This type of response is known as on-center. The negative central response shown in Figure 8b demonstrates response of contradictory charge. This type of response is known as off-center. On-center and off-center responses form part of a dynamic system that allows perceiving the changes in a visual scene [18].

There are several types of ganglion cell responses that may occur during the visual process. However, the concept of receptive fields works similarly for all of them [18]. Also, the concept of receptive fields is used at the LGN. The cells at LGN work similarly to the ganglion cells in the retina. The LGN receptive fields can be thought as representing a massed retinal input since multiple neural fibers from the retina converge to each LGN [20].

1.2.2. Rods and Cones

The photoreceptors convert the light entering the eye into electrical pulses that are processed by the brain. There are approximately 120 million photoreceptors in each retina, although the quantity varies from individual to individual [16]. The quantity and distribution of rods and cones varies with the angle subtended to retina, with zero degree being the visual axis. One of the most important distinctions between rods and cones is their sensitivity to light.

The presence of two types of photoreceptors suggests two different visual functions. Around the 1860's Max Schultze, a retinal anatomist found that animals that function better during nighttime, such as owls, have only rods in the retina. However, Schultze found that animals entirely active during the day, such as pigeons, have only cones, while those animals that are active during both the day and night, such as monkeys, have retinas with rods and cones [17]. These discoveries led to the duplex retina theory of vision. In this theory the rods are used for vision under dim light (low luminance) and the cones are used for bright light. Vision where only the rods are active is called scotopic, whereas vision due to cone responses only is called photopic. Mesotopic vision is when both rods and cones are active.

The photo-active component of both rods and cones is a pigment-protein complex, which absorbs light within a particular wavelength range depending on the nature of the protein. The pigment-protein complex involved in the rods (scotopic vision) is called rhodopsin [21].

In normal photopic color vision, three pigment-protein complexes are present in the cones. Hence, there are three types of cones, called erythrolabe (red), chlorolabe (green) and cyanolabe (blue) [22]. The excitation of cells by light results in a similar change in both rods and cones, although in each case wavelength of maximum absorption is different. When photons are absorbed the pigment undergoes reversible cis-trans isomerization, which is the trigger for sending an electrical pulse in through the nerve cells in the retina, and ultimately onto the brain [23].

1.2.3. Monochrome Vision

The process resulting scotopic vision using only the rods can be generally described in three very rapid steps: isomerization via photon absorption, protein conformational change due to isomerization, and the translation of the physical change within the photoreceptor to an electrical signal that is received by the brain. When a photon is in the visible region of the spectrum the free rotation occurs resulting in the formation of all-trans retinal [23].

1.2.4. Color Vision

A similar three-step process for monochrome vision applies to photopic vision, although three cones with differing spectral sensitivity are involved. It is because of the differences in spectral sensitivity of the cone cells that humans with normal color vision are able to distinguish a vast range in colors.

Each cone has a different protein bound to the same pigment, 11-cis retinal, and the complex has a particular absorbance spectrum [24], with each complex having different absorbance maxima: 426 nm (blue), 530nm (green), and 560nm (yellow). The last pigment has absorbance sensitivity up to approximately 700 nm, thereby facilitating the perception of red hues [23]. Common abbreviations for these pigments are S, M and L: S corresponds to short wavelength absorption, M for medium wavelength absorption and L (long wavelength absorption).

1.3. Mechanisms of Color Vision

No unified theory of color vision exists today. However, there are three theories that have been shown to explain certain key aspects of visual perception pertaining to color.

1.3.1. Trichromatic Theory

The existence of three cone photopigments in the retina implies that humans are trichromatic. Thomas Young was the first to suggest that human color vision was thricromatic. However, this fundamental theory was developed more comprehensively by Clerk Maxwell and Hermann von Helmhotlz in the mid 19th century. This theory states that humans perceive color as the result of a mixture of three primary wavelengths corresponding approximately to the sensitivity of red, green and blue regions of the spectrum [18]. The theory has merit as evidenced by the later determination of the three types of cones by George Wald and Paul Brown, in 1959, at Harvard and William Marks, William Dobelle, and Edward MacNichol at Johns Hopkins [25]. In fact, this theory forms the basis of colorimetry today (see section 3 for a discussion of colorimetry).

1.3.2. Hering's Opponent – Color Theory

About the same time that the trichromatic theory was developed, the German scientist, Ewald Hering, hypothesized a different mechanism of color vision. His observations were not based on the physiological factors but in the experiences associated with color [17, 18].

According to Hering, when mixing yellow and red lights, a yellowish-red perception will result. However, when mixing blue and yellow lights a perception close to white may be experienced. Hering proposed that such colors ‘oppose’ each other. His theory was based on the existence of three opponent color channels located in the retina: red versus green, blue versus yellow and black versus white. These channels respond according to the stimulus wavelength with different polarities [19].

1.3.3. Recent Developments in Color Vision Mechanisms

In 1956 Svaetichin obtained results supporting Hering’s opponent theory. He studied the responses of the retina of goldfish, which have trichromatic vision [18]. In 1958, DeValois et al. also found an opponent system from responses of the LGN cells of macaque monkeys [18]. In 1955, Jameson and Hurvich developed a significant amount of data on trichromacy using human observers via hue-cancellation experiments. Based on these findings and additional research done since then, a modern opponent theory of color vision has been developed. This theory is also known as “Stage Theory”. The theory states that the color separation from the three receptors is not transmitted directly to the brain. The neurons located in the retina predetermine the color information into opponent signals. The sum of the three inputs produces the achromatic responses. The three opponent pathways have different temporal and spatial characteristics that allow the perception of color appearance [18].

1.4. Mechanisms of Adaptation

The human visual system is sufficiently complex that not only discrete visual processing should be considered, but how the color vision process responds to specific viewing conditions.

1.4.1. Dark Adaptation

Dark adaptation refers to changes in visual sensitivity when the level of illumination decreased. The visual mechanism responds to the absence of illumination by increasing the sensitivity in order to identify visual stimuli [18]. Perhaps one can notice this visual adaptation when walking into a dark room after being exposed to high levels of illumination. At first, the dark room will appear completely dark, but in a matter of minutes one is capable to distinguish forms, shapes, and any objects present in the room. As shown in Figure 9, during dark adaptation, the cones develop a high sensitivity until they level off within a couple of minutes and visual sensitivity becomes stable. After about 10 minutes, the rods start to surpass gradually the cones until they become the ones controlling the visual sensitivity [18].

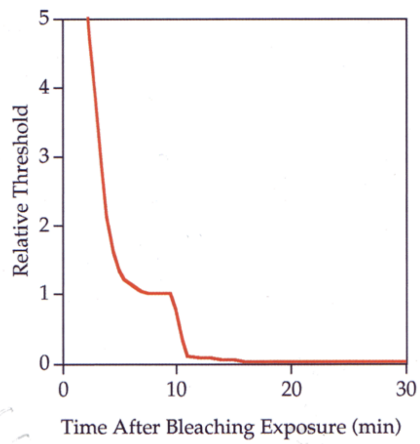


Figure 9. Dark adaptation recovery [17].

1.4.2. Light Adaptation

Light adaptation refers to the mechanism involved with visual sensitivity when the level of illumination increased, the opposite of dark adaptation, although light adaptation occurs much faster than dark adaptation. The visual mechanism responds to the presence of illumination by decreasing sensitivity in order to identify visual stimuli. This is achieved by shifting from the rod system to the cone system. Light adaptation takes place within 5 minutes. It is comparatively faster than dark adaptation [18].

1.4.3. Chromatic Adaptation

Chromatic adaptation refers to the independent sensitivity control of the three cone responses [18].

One can experience chromatic adaptation when looking at a white object under different types of light, with widely differing spectral power distributions (see Section 3). Although lights differ in terms of wavelength energy, a white object usually appears white despite remarkable changes in incident illumination, providing that the source has emission across the entire visible spectrum. This effect is due to a complex compensatory process that occurs in the cones [18].

1.4.4. Visual Mechanisms Affecting Color Perception

1.4.4.1. Color Memory

This phenomenon is related to the association of an object with an individually perceived 'ideal' color. For example, most humans with normal color vision will recall the color of red apples and they are able to reproduce this color without seeing a real object. However, it has been found that the color 'remembered' is usually different from the typical real object [18]. Color memory is used in some color matching techniques for the development of; for example, color inconstancy models (see Section 4.2.5).

1.5. Color Vision Deficiencies

Color perception varies widely from individual to individual because it is a personal physiologically-based experience [17], and therefore we cannot refer to any absolute value of color perception.

We can only refer to colors in relation to other colors. However, among the population there are several physiologically-based deviations from the average population that could lead to problems in color perception, some of which are related to color discrimination. Some of these problems can be genetically inherited or acquired.

1.5.1. Inherited Color Vision Deficiency

In relation to the trichromatic theory of color, a number of genetically inherited color deficiencies exist, although their relative abundance in the human population is dependent on geographic location. An extreme example of color vision deficiency is known as achromatopsia in which the cones that do not function, and therefore sight is dependent solely on the rods. People that only use the rod system and have no color discrimination ability and find bright light very uncomfortable. On the other hand, their night vision is normal [17].

A second type of deficiency is when only one type of cone is working in addition to the rods. Individuals who inherit this deficiency are known as monochromats and they have monochromatic vision, resulting in severe problems in color discrimination. However, vision is possible under both photopic and scotopic conditions [17].

The third type of deficiency is when one of the three types of cones lacks a particular photopigment.

These individuals are known as dichromats and they have some ability to discriminate color because they still have two types of cones that work. There are three types of dichromats: Protanopes, Deuteranopes and tritanopes. Protanopia is when the individual has an L-cone that is not functioning properly, which results in insensitivity to long wavelengths. A protanope has difficulty perceiving reddish and greenish hues [17, 18]. A well known case of Protanopia is the chemist John Dalton [17], whose published work on color vision deficiencies became known as Daltonism [26].

Deutaronopia is the most common form of dichromacy. A deuteranope is missing the M-cone photopigment; these individuals are still able to respond to green light. However, they have problems discriminating green from some combinations of red and blue [17, 18]. The last type of dichromacy is tritanopia. This type of deficiency is due to the lack or malfunctioning of the S-cone and it is the least common. In this case, the individuals see the long wavelengths as red and the shorter ones as bluish-green [17, 18].

Even though is impossible to describe accurately how dichromats experience color, Figure 10 provides an indication of hues that cannot be discriminated by dichromats.


(a)
(b)
(c)
(d)

Figure 10. (a) Normal vision (b) protanope (c) deuteranope (d) tritanope [17].

The fourth type of color vision deficiency is known as anomalous trichromatism. While individuals with this problem have trichromatic vision, they have a reduced ability to discriminate hues due an alteration in the spectral sensitivities of the photopigments. There are three types of anomalous trichromacy: protanomaly, deuteranomaly, and tritanomaly. The protanomalus has a weak L-cone photopigment or the L-cone absorption is shifted toward shorter wavelengths. The deuteranomalus has a weak M-cone photopigment or the M-cone absorption is shifted toward long wavelengths. The tritanomalus has a weak S-cone photopigment or the S-cone absorption is shifted toward long wavelengths [17, 18]. Table 1 shows the incidence of color vision deficiency in the UK between males and females. Figure 11 shows the incidence of color vision deficiencies in the world [27].

Table 1. Incidence of color vision deficiency in the UK[27].

Condition		Proportion (%)	
		Males	**Females**
Protanopia	Red-blind	1.0	0.01
Deuteranopia	Green-blind	1.0	0.01
Tritanopia	Blue- blind	very small	very small
Protanomaly	Red weak	1.0	0.03
Deuteranomaly	Green weak	5.0	0.35
Tritanomaly	Blue weak	very small	very small

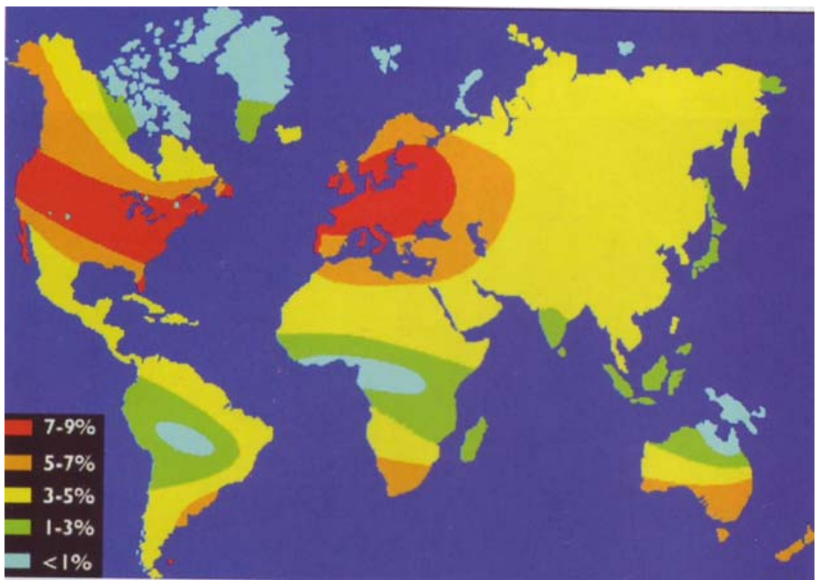


Figure 11. Percentages of Incidence of defective color vision around de world[27].

1.5.2. Acquired Color Vision Deficiency

Some color vision deficiency problems are not inherited. They are usually the result of a disease, toxicity or an accident.

Table 2 summarizes some of the main differences found between deficiencies acquired and inherited.

Table 2. Summary of common differences between hereditary and acquired color vision deficiencies.

Hereditary defects	**Acquired defects**
Mainly problems discriminating red-green	Mainly problems discriminating yellow and blue
More common in males	Common in males and females
Usually present in both eyes	Regularly presents a difference in severity between eyes
Problem not associated with color naming	It can be associated with color naming errors
Deficiency stable through time	Deficiency may present differences through time
Identifiable with standard color test	It requires more than clinical color test to identify them
Not related to illness or toxicity	Associated with ocular or systematic disease or toxicity

1.5.2.1. Chromatopsia

Another possible acquired color deficiency is chromatopsia. This problem is associated with toxicity, the effect of medication, hallucinogenic substances, or cataract extraction. Different types of chromatopsia are named for the colors seen. For example the use of Viagra may result in cyanopsia (blue) [22], which results in objects temporarily appearing to have a bluish tint.

1.5.2.2. Cerebral Achromatopsia

This deficiency can be acquired following a lesion (perhaps following surgery) on the area V4 of the brain, which is the area believed to control color processing [16].

1.6. Color Testing Methods

1.6.1. Ishihara test

The Ishihara test is one of the most widely used methods to determine red-green color vision deficiency, although it does not provide data on the degree of severity [28]. The test consists of printed screening and diagnostic plates with a Figure, usually a number, to be differentiated from the background [29]. The diagnostic plates enable the determination between protan and deutan, and mild and strong defects. Although the Ishihara method has shown good performance in detecting red-green color vision deficiencies, studies have shown high variability when diagnosing specific types of defects [28]. Nevertheless, it is commonly used as a rapid screening test for selecting observers with normal color vision in studies involving color perception. Figure 12 shows an example plate from the Ishihara pseudo-isochromatic confusion plates.

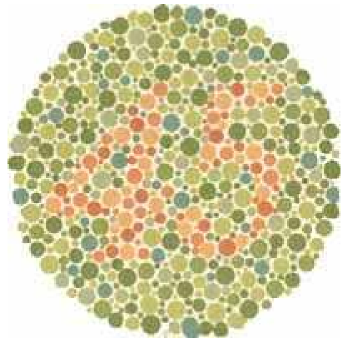


Figure 12. An example of pseudo-isochromatic Ishihara confusion plates.

1.6.2. Farnsworth –Munsell 100-Hue Test

This test requires observers to order a series of colored chips of constant lightness and chroma but varying hues of just noticeable difference. It was designed to determine the level of hue discrimination among observers, and enables the categorization of observers with normal color vision into different levels of hue discrimination capability, relative to the average population: superior, average and low. Also, the test allows the determination of hue regions of confusion for people that have color deficiency [29].

1.6.3. Neitz Test

The Neitz test is similar in to the Ishihara test. It is a simple test developed to detect the two main types of color deficiency: red-green and blue-yellow. It also classifies the subtype or red-green deficiencies and their severity. It usually takes less than 5 minutes to take it and it can be administered to groups of individuals. It is usually used as a screening test [30].

1.6.4. Cambridge Color Test (CCT)

The Cambridge Color Test was developed with the idea of measuring hue discrimination in a spatial and luminance noise situation [29]. CCT is a computerized test that allows the change of parameters in the stimulus and the threshold detection between pairs of target and background hues [29].

1.6.5. Color Vision Test Made Easy

The "Color Vision Test Made Easy" is a test designed for children. It is based on the same principles of the pseudo isochromatic Ishihara test, although the plates used in this test contain symbols and shapes that young children with normal color vision can easily identify [29].

2. Psychophysics

Color is a personal, physiologically-based experience. Researchers have been investigating links between stimulus and the human experience for centuries. At the heart of color science lies the critical issue investigating light interaction with objects and the visual experience. ‘Color’ can be expressed quantitatively as, for example, the amount of light reflected by an object, but the resultant human experience when the visual stimuli is perceived by the eye and brain is not objective, is complex, and varies widely from person to person. Different approaches have been developed in order to understand and ultimately quantify the relationship between experience and stimuli that in some way are defined the composition of matter.

2.1. Definition

Psychophysics has been defined as the “study of the relationship between the physical stimuli in the world and the sensations about them we experience” [31]. Hence, psychophysical techniques allow quantitative determination of a person’s experience to a given stimulus.

2.2. Historical Overview

Psychophysics is a relatively young discipline. In the 1850’s, Gustav Fechner, a German scientist and philosopher, studied the relationship between stimuli and perception, and is considered by some as the father of psychophysics.

He believed that there was a need for a scientific approach to understand the philosophical issues between mind and body [31]. For this purpose, he proposed three approaches: Detection, Discrimination and Scaling. Detection has as a goal the development of a method to measure the minimum amount of stimuli that can be perceived. Discrimination is the development of a method to determine how much a given stimuli must be varied in order to perceive a difference, and scaling is the development of an approach to quantify a particular dimension of our sensation (e.g. lightness) [31].

Numerous developments have been made since Fechner's first approaches to understanding perception. Around the 1960's Stanley Smith Stevens advanced Fechner's idea of scaling. Stevens studied the relationship between stimulus intensity and perceptual magnitude using a magnitude estimation technique. His findings allowed him to hypothesize that the relationship between stimulus intensity and perceptual magnitude resulted in a power law with various exponents according to perception studied. Stevens' precepts are important for the perceptual phenomena of color and his statements are found to be elementary in understanding color measurement [18].

A fourth approach was found to be necessary in the relationship stimuli and perception. Such approach is called Identification. Identification "involves being able to attach a learned label or category name to the stimulus we have encountered" [31]. This approach has been widely used in recent years because of the importance of the relationship between the things people are familiar with and what we can recognize [31].

The various methods developed, summarized below, enable collection of useful perception data, although different approaches will lead to different aspects of perception.

2.3. Detection Techniques

This group of psychophysical techniques is focused in how the sensory systems respond to changes of energy. Energy can be found in different forms of stimulation such as: electromagnetic (light), mechanical (sound, touch, movement), chemical (taste, smells) or thermal (hot, cold). The main idea of detection theory focuses on how much energy change, starting at zero, is needed in order for the change to be perceived by a person [31].

In color perception, these types of techniques are very useful for evaluating tolerances especially in color difference studies. A number of threshold experiments with different levels of complexity in experimental design have been developed with varying levels of significance of data obtained. Some of the most important threshold techniques are: method of constant stimuli, method of limits, and the method of adjustment [18].

2.3.1. Method of Adjustment

This is one of the easiest techniques. In this method, the observer controls the stimuli magnitude and changes it until the point where is just noticeable or different from a given reference or starting point. The threshold is calculated based on the average of many trials by one or more observers.

One of the disadvantages of this technique is that the observer has control over the stimuli and can bias the results due to observer variability. This method is sometimes used in color perception studies as a starting point to design more complex experiments [18].

2.3.2. Method of Limits

This method is more complex than a simple method of adjustment. In this method, the experimenter presents the stimuli at different levels of intensity, the levels of which have been determined prior to executing the experiment. The sequence of stimuli can decrease or increase as necessary. If using an ascending sequence, the researcher starts with a stimulus intensity that is not noticeable and asks the participants to determine and respond if they perceive the stimulus. The stimulus intensity is increased until the stimulus is obviously perceptible. The threshold is calculated by taking the average stimulus intensity in which the transition from not perceptible to perceptible took place. To minimize adaptation effects, it is suggested to do runs of ascending and descending sequences and then average the results [18].

2.3.3. Method of Constant Stimuli

In this method, about seven stimulus levels are pre-selected close to the threshold intensity. Then, the selected stimuli are presented to each observer in a random order several times. After many repetitions, the frequency in which each stimulus level is perceived can be determined.

The data acquired can be used to produce a psychometric function also known as a *frequency of seeing* curve. This curve allows determination of the threshold and its ambiguity. Usually a psychometric function can be obtained for an individual person (through multiple repetitions) or for a population of people (one or more repetitions for participant).

The method of constant stimuli can be classified into two methods according to the type of response acquired during the study: The yes-no method and the forced choice method [18]. The yes-no method is extensively used in color science studies where the measurement of visual tolerances is required. For this purpose, an anchor pair with a predetermined color difference is fixed and different pair samples with various color differences are presented to the observer. The observer is asked to determine if the presented stimulus is below the intensity of the anchor pair (pass) or above the anchor pair (fail) [18].

2.4. Matching Techniques

This type of psychophysical technique was developed to determine when two stimuli are not perceptibly different.

The measurement of the variability can be used to determine thresholds. These set of techniques also have been used in studies related to color appearance models and chromatic adaptation [18].

2.4.1. Asymmetric Matching

This method is mostly used for color appearance and chromatic adaptation studies. A color match is made between to objects or stimuli using different viewing conditions. The data generated can be used to develop and test the effect of changes in viewing conditions on color matching, and is important for the development of color appearance models [18].

2.4.2. Memory Matching

This type of technique can be used in studies related to color appearance. In this type of experiment, the participant is asked to produce a match base on a previously memorized color. Usually this type of study is carried out with the purpose of studying observer dependence on specific viewing conditions. This technique is often carried out using mental stimuli that can be an ideal achromatic color (gray) or a unique hue [18]. For a discussion of unique hue, see Section 4.1.1.

2.5. Scaling Methods

These types of psychophysical methods use the basic mathematical premise of a scale. Scaling techniques were developed to determine how much of a stimulus level humans can perceive. In other words, the goal of scale is to provide a measure of a sensation magnitude. However, psychological magnitudes have to be measured according to the need and the type of experience [31]. A change in a given stimulus may produce a change in appearance.

For example, changing the wavelength of light may cause the light to change from red to green. This change cannot be determined quantitatively in a perceptual way, only qualitatively [31].

2.5.1. Types of Scales

2.5.1.1. Nominal Scales

This type of scale is related to naming. This scale will classify the stimuli into a category and it is usually used to classify data into groups. For example, in color studies it can be used to classify an arrangement of colors into reds, yellows, greens etc. [18]

2.5.1.2. Ordinal Scales

This type of scales classifies the stimuli in ascending or descending order based on a specified characteristic. For example, a given set of colored objects can be organized from light to dark. The spacing between the stimuli is not necessarily specified in this scale [31].

2.5.1.3. Interval Scales

This type of scale is related to distance between stimuli. The scale is constructed using equal intervals between stimuli in such a way that the difference between them is perceptually equal. Some of the examples of these scales are the Celsius temperature scale and the Fahrenheit scale [18].

2.5.1.4. Ratio Scales

This type of scale has all the characteristics of the scales mentioned above with the difference that it has a defined zero point. This type of scale is not used in color research because it is hard to achieve a meaningful zero point. An example of such difficulty can be seen when trying to develop a scale of hue where there is zero point in hue [18].

2.6. One Dimensional Scaling

This special type of scaling method is based on the assumption that the characteristic to be scaled and the physical change of the stimulus are one dimensional. The participants are asked to make their judgment only in a particular perceptual characteristic. For example: lightness of a sample compared with others. A variety of one dimensional scaling tests have been developed [18].

2.6.1. Rank Order

In this method, observers organize a set of samples according to an increasing or decreasing level of a variable of interest. For example, observers may arrange a series of gray samples from dark to light. However, interval scales should not be derived from ranking experiments [18].

2.6.2. Graphical Rating

In this method, participants observe a stimuli and they then determine the magnitude of their perception on a one-dimensional scale with fixed end points. This method allows derivation of an interval scale [18].

2.6.3. Category Scaling

In this method, the participant is asked to categorize various stimuli into different groups. The number of times each stimulus is place in a group is recorded. This method allows producing ordinal or interval scales [18].

2.6.4. Paired Comparisons

In this method, observers are presented with a series of pairs of samples. The pairs comprise all possible combinations of samples from a given set. Each pair is individually presented to the observer usually using a third stimulus as reference. The observer is asked to judge if the test pair is greater or smaller in some attribute than each other sample. The proportion of times a sample is judged greater is written down. Data collected with this method can be applied to the law of comparative judgment which allows producing interval scales [18].

2.7. Techniques Used In Color Research

Color research may employ widely differing psychophysical techniques in order to understand different parameters affecting human perception of color and to develop methods and models that specify more accurately how color stimuli is perceived.

2.7.1. Studies Comparing Psychophysical Techniques

Psychophysical methodology is one the most important issues regarding color perception phenomena. Information obtained from different studies can vary widely from method to method. In addition, it is important to understand the scope and limitations of different methodologies. The two methods most commonly used are: Pair Comparison and Gray Scale Assessment.

The pair comparison method involves assessment of visual difference for a test pair of samples compared to a fixed reference pair of samples. The gray scale method involves assessment of the magnitude of visual difference for a test pair of samples compared a reference scale, which usually comprises a set of a pairs of gray samples of sequentially increasing lightness contrast. Figure 13 shows an example of typical sample arrangement for pair comparison method [12] Figure 14 shows an example of sample arrangement using the gray scale method [12].

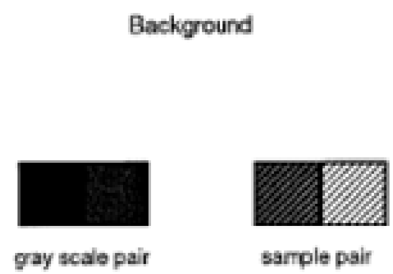


Figure 13. Sample arrangement used in a pair comparison experiment [12].

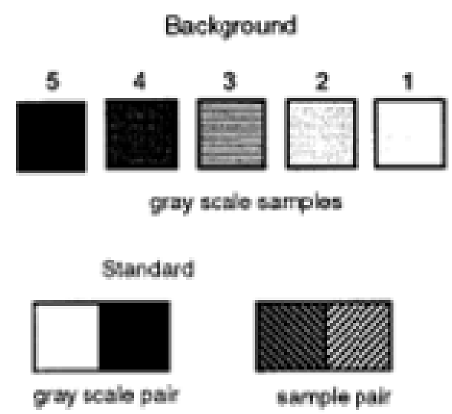


Figure 14. Sample arrangement used in a gray scale experiment [12].

Guan and Luo used a gray scale method and pair comparison method to assess large color differences in wool samples [13].

In a separate study, Guan and Luo used the same techniques to evaluate parametric effects in the assessment of small color differences [12]. Montag and Wilber used the gray scale method and method of constant stimuli to determine color tolerances in three color centers [32]. Kim studied the pair comparison method and the gray scale methods to determine the influence of parametric effects in the assessment of small color differences [33]. Witt and Döring produced perceptibility and acceptability ellipsoids around a recommended CIE green center. In this studied a group of 67 glossy painted samples were assessed using two different methods, absolute threshold and pair comparison [34]. Xu et al. used interleaved staircase and constant stimuli methods to study color discrimination threshold and suprathreshold color-difference comparison respectively [35].

2.7.2. Studies Using Gray Scale Method for Perceptibility Studies

In addition to the uses given in the previous section, this method has been the most widely used. While the actual gray scale used in experiments may differ in each study, the most commonly used is the gray scale for color change recommended by the International Standards Organization (ISO) and the American Association of Textile Chemists and Colorists (AATCC) [36].

For example, a scale similar to the ISO scale was used by Luo and Rigg to develop chromaticity-discrimination ellipsoids around 70 color centers [37]. It has been also used for the assessment of small color differences in various experiments [38-41].

Furthermore, the gray scale method has been used to determine the influence of parametric effects in the assessment of medium [14, 42] and large [13] color differences, for instance to evaluate crispening effect in CRT displayed color samples [43], daylight simulators [44, 45], to quantify metamerism [46] , and to determine the performance of different color difference formulas [47-49].

2.7.3. Studies Using Pair Comparison Method for Perceptibility Studies

While the pair comparison method has been used widely in color research, it has some limitations and the selection of anchor pair is a key consideration in order to obtain effective results [12]. Rich and Billmeyer studied small color difference for the development of color difference perceptibility ellipses using the pair comparison method with three neutral gray reference pairs [50].

Stroka et. al used this method to evaluate the influence of experimental parameters on the accuracy of color-difference ellipsoids [51]. Qiao et. al also used this method to determine hue suprathreshold tolerances in colored samples [52].

2.7.4. Other Psychophysical Methods

A number of studies in the perception of color difference have used methods different from those described above. Kuehni and Marcus used a ranking method, subjective estimation and acceptability judgments in their study of visual scaling of small color differences [53].

Alder et. al, Cheung and Rigg, and Coates et. al used a ratio estimation approach in their studies [54-56]. Witt used a simple detection technique in a study to determine the threshold color difference for four CIE color regions [57]. In his study, the observers were presented with sample pairs and they were asked to determine if there was a perceptible color difference.

Other studies have used different methodology to determine color difference acceptability judgments rather than perceptibility judgments [58-60]. These types of studies usually involve some form of a pass/fail assessment method.

2.7.5. Methodologies Recommended for Selection and Training in Color Assessment

The American Society for Testing and Materials (ASTM) Standard Guide for Selecting, Evaluating and Training Observers, E1499-97, explains parameters and psychophysical methods for individuals whose tasks involve perception and evaluation of color perception and appearance [61]. Regarding psychophysical methods, the ASTM recommends a length estimation test and a color estimation test.

The length estimation method is a simple magnitude scaling method in which the observer is asked to match a given anchor scale with a perceive difference in stimuli [61]. The color estimation test uses a set of 20 cards from the Japanese Color Aptitude Test [61]. The results of this test are comparable to those obtained using the Farnsworth Munsell 100 Hue Test.

2.7.6. Methodology Recommended for Visual Assessment of Color Difference

The ASTM Standard Test Method for Evaluation of Visual Color Difference with a Gray Scale, D2616-95, provides guidance for the visual evaluation of small-to-medium color differences in samples that are not self luminous. The visual assessment is carried out using a scale that consists of nine gray pairs with increasingly larger lightness differences [62]. The color difference from the pair samples in the grayscale increases geometrically (e.g., 0.5, 1, 2, 4, 8 arbitrary units) [62]. The grayscale used in this method is comparable to the gray scale used in Evaluation Procedure 1, Gray Scale for Color Change, recommended by the American Association of Colorists and Chemists (AATCC)[63].

3. Colorimetry

The measurement of color involves a wide range of disciplines and techniques, such as chemistry, physics, physiology, and psychology just to name few [18]. Colorimetry refers to the measurement of color. An important difference was made by Wyszecki when describing basic and advanced colorimetry. Basic colorimetry is defined as: “Colorimetry, in its strict sense, is a tool used to making a prediction on whether two lights (visual stimuli) of different spectral power distributions will match in colour for certain conditions of observation. The prediction is made by determining the tristimulus values of the two stimuli. If the tristimulus values are identical to those of the other stimulus, a color match will be observed by an average observer with normal colour vision” [64]. However, Wyszecki describes advanced colorimetry as: “Colorimetry in its broader sense includes methods of assessing the appearance of color stimuli presented to the observer in complicated surroundings as they may occur in everyday life. This is considered the ultimate goal of colorimetry, but because of its enormous complexity, this goal is far from being reached. On the other hand, certain more restricted aspects of the overall problem of predicting colour appearance of stimuli seem somewhat less elusive. The outstanding examples are the measurement of color differences, whiteness, and chromatic adaptation.”[64].

3.1. Color

Before any discussion can be made regarding colorimetry, it is necessary to comprehend how color is experienced.

The perception of color in every day experiences results from the interaction between a light source, the surface of an object, and the stimulus response of the eye-brain of an observer with normal color vision [19]. The energy emitted from a light source is modified by an object's chemical and physical properties. That energy is then detected in the eye by the photoreceptors and encoded by the brain to produce a sensation of color [18, 19]. Figure 15 illustrates the interaction among the three components.

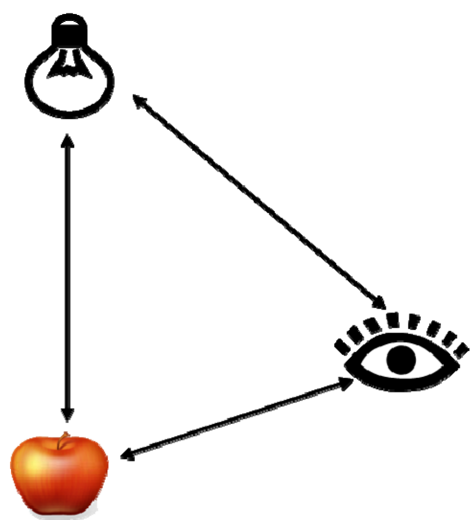


Figure 15. The three basic components needed for color perception.

In order to measure color, therefore, it is usually necessary to model or measure these three factors. Light emitted from a source can be quantified by measurement of their spectral power distribution (SPD) across the visible spectrum.

Objects can be specified according to their reflectance ratio in comparison with a standard white tile, for instance, across the visible spectrum. The reflectance spectrum of an object is constant regardless of the light source, provided the object is illuminated by a continuum of emission across the visible spectrum. The reflectance of an object is defined by its surface geometry, surface reflection (gloss), and the amount of energy absorbed, scattered or transmitted through the material.

While the human eye-brain system is highly complex, the visual system can be expressed in basic terms in regard to color matching properties representing the absorption of light by the three cones under normal luminance conditions.

3.1.1. Illuminants and Light Sources

3.1.1.1. Light

In 1672, Sir Isaac Newton demonstrated that white light or nearly white light can be decomposed into the visible spectrum using a prism [19, 65]. Visible light is a type of electromagnetic radiation detectable by the human eye, and can be defined in terms of wavelength using a nanometer as the unit of length [19, 65].

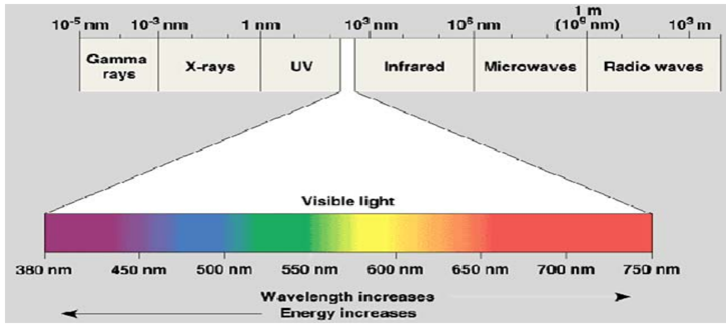


Figure 16. The visible spectrum in relationship to other types of electromagnetic radiation.

Due to the three cone-shaped photoreceptors in the human retina, wavelengths in the range 380 to 780 nm can be detected under normal luminance conditions. Light from any source can be quantified in radiometric terms and photometric terms. Radiometry may be defined as the science that studies the measurement of electromagnetic radiation including visible light [19]. Photometry may be defined as the science that studies the measurement of visible light weighted by the perception of brightness [19]. Table 3 shows the terms and units used in radiometry and photometry respectively.

Table 3. Terms and units used in describing light intensity [19].

Radiant energy term	**Unit**	**Luminous energy term**	**Unit**
Radiant flux	W	Luminous flux	Lm
Radiant intensity	W sr^{-1}	Luminous intensity	cd (lm sr^{-1})
Radiance	W sr^{-1} m^{-2}	Luminance	cd m^{-2}
Irradiance	W m^{-2}	Illuminance	lm m^{-2} (lux)
Radiant exposure	J m^{-2}	Light exposure	lux s

Illuminance is related to the perception of brightness of a source and is the amount of light falling on a surface weighted by a photometric brightness constant, and its unit is the lux (lm m^{-2}) [17].

3.1.1.2. Black Body Radiators

A black body is a theoretical quantum mechanical concept that absorbs and emits light independent of composition. From this model, the emission of electromagnetic radiation as a function of temperature can be defined for a given wavelength, according to equation (1) [19].

$$M_e = \frac{c_1}{\lambda^5\left[\exp\left(c_2/\lambda T\right)-1\right]} \tag{1}$$

Where:

M_e= emittance per wavelength interval (W m^{-3})

T= absolute temperature of the source (K)

λ=wavelength of the radiation band considered (m)

c_1= $2\pi c^2 h$=3.741 83 x 10^{-16} W m^2

c_2=1.4388 X 10^{-2} m K

Hence, light and temperature are fundamentally related, and light sources can be defined in terms of a color temperature, although the emission of light from real sources is highly dependent on physical composition, and actual emission as a function of wavelength is not defined by equation (1). However, some materials are good approximations of black bodies. When a piece coal burns light is emitted and the color of the light changes as the temperature of the coal increases. At the beginning, it starts to glow with a reddish hue, and as temperature is raised the color changes from orange-red to yellow and it may eventually become ‘white hot’.

Since real light sources are not black bodies, the term correlated color temperature (CCT) can be used to define the spectral characteristics of the source. The correlated color temperature of the light source is the temperature of a black body radiator that has almost the same color of the source [18, 19, 65].

3.1.1.3. Standard Illuminants and Sources

The Commission Internationale de l’Eclairage (CIE) in 1931 quantifiably standardized three sources of illumination. A light source is defined as a physical emitter of radiation, whereas an illuminant is defined as a dataset that represents a light source. An illuminant may not be physically realizable.

Standard illuminant A was set to represent indoor artificial light with a CCT of approximately 2856 K. Standard illuminant B represents direct sunlight and had a CCT of 4900 K. Standard illuminant C represents average daylight with a CCT of 6800K [66].

However, research carried out during the 1950's and 1960's showed that illuminant B and C deviated from natural conditions [19].

In 1963 a new set of recommendations were made by the CIE that defined the standard D illuminants, which represent various types of daylight under different conditions [19]. In 2004, the CIE published a new report in colorimetry, entitled CIE Technical Report Colorimetry CIE 15:2004 [66] which defines the basic recommendations concerning modern colorimetry. Currently the recommended illuminant standards are:

CIE Standard Illuminant A: The relative SPD S_A (λ) can be described using equation (2):

$$S_A = 100\left(\frac{560}{\lambda}\right)^5 \times \frac{\exp\frac{1{,}435\times 10^7}{2848\times 560} - 1}{\frac{1{,}435\times 10^7}{2848\lambda} - 1} \tag{2}$$

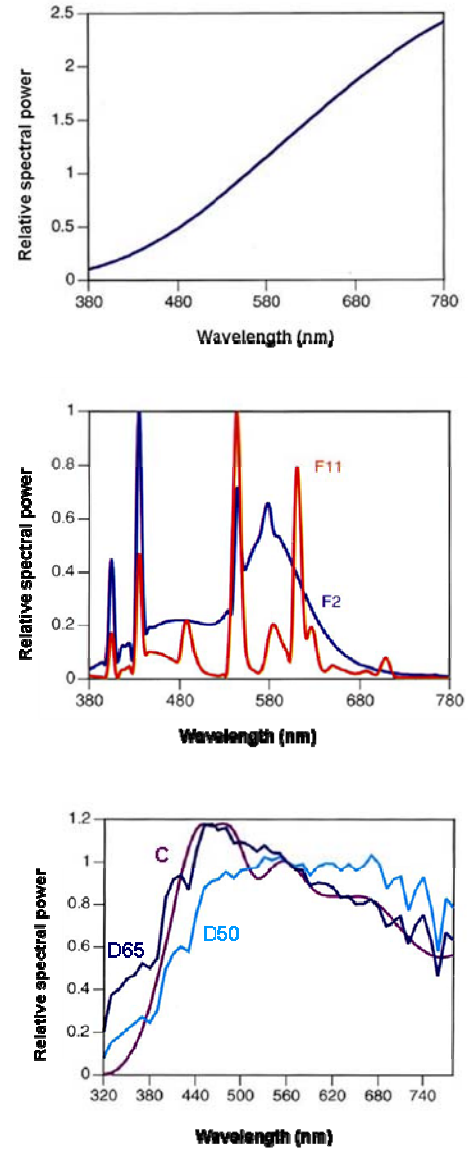
Where λ is the wavelength (nm) and the numerical values from the exponential terms are constants from the first definition of illuminant A in 1931 The remaining definitions regarding Illuminant A still in accordance with the original work published in 1931[66]. The CCT for illuminant A is 2856 and it is used for colorimetric calculations in which incandescent illumination is needed [18].

CIE Standard Illuminant D65: This illuminant corresponds to daylight with a correlated color temperature of 6500 K.

Other Illuminants D: While the most common illuminant to use for daylight is D65, other illuminants are specified and are used by some industries. For example, D50 is usually used in the graphic arts industry and D75 is used in some applications in the U.S. automotive industry.

Illuminant B and C: These illuminants were intended to represent daylight at CCT of 4900 K and 6800 K. However, their use is no longer recommended by the CIE [66].

Fluorescent light sources were introduced in the mid-20th century and these new sources require separate illuminants. A series of F illuminants have been standardized by the CIE, the most commonly used ones being, F2 (representing cool white fluorescent lighting) and F11 (representing energy efficient tri-band fluorescent lamps) [66]. Figure 17 shows the relative spectral power distribution for some of the standard illuminants.



(a)

(b)

(c)

Figure 17. Spectral power distribution of CIE standard illuminants (a) Illuminant A (b) Illuminants F2 and F11 (c) Illuminants C, D50, D65.

3.1.2. Colored Object

After the illuminant is defined it is usually necessary to determine the interactions between incident light and an object. Light can interact with an object in three ways keeping the principle of conservation of energy. Reflection, absorption and transmission are ways in which light interacts with an object. Reflectance, absorbance and transmittance are the measured quantities from the interactions. The incident energy is equal to the sum of the energy reflected, energy absorbed and transmitted [18]. Since the sum of the three is equal to 100%, is very common to measure either one of them according to the need. However, interactions between an object and light are not just explained by spectral data. Transmittance and absorbance are also dependent on the illuminant and the viewing geometry. Viewing geometry is important in cases in which the surface characteristics are determinant such as gloss properties [18]. Ten standard viewing geometries have been recommended by the CIE to be used in colorimetry trying to quantify the object's appearance effect based on reflectance, the most common of which are defined below [66].

3.1.2.1. Diffuse Eight-Degree Geometry Specular Component Included (di: 8º)

In this geometry, the sample should be uniformly illuminated from all angles using an integrating sphere coated with barium sulfate or polytetrafluoroethylene that has an opening (port) where the sample is placed [18, 66].

Through a second port in the sphere, all the light reflected from the surface of the object at 8 degrees from the normal angle is measured by a sensor [18]. This geometry is suited to non-textured and matt surfaces. The influence of any specular reflection or texture is included in the measurement, and may lead to significant error with materials that have these surface characteristics.

3.1.2.2. Diffuse Eight-Degree Geometry Specular Component Excluded (de: 8º)

This geometry has the same recommendations of the di: 8º. The difference is that specular reflection is excluded [65, 66], and this geometry is preferred for glossy or highly textured materials.

3.1.2.3. 45º Annular/Normal Geometry (45º a:0º)

This geometry specifies illuminating the sample uniformly from all directions [66] at a 45 degree angle and the sensor is placed normal to the plane of the material to be measured.

3.1.2.4. Normal/ 45° Annular Geometry (0º:45º a)

This geometry has the same viewing conditions that the 45 º a: 0º, but using a reversed path of light [66].

3.1.2.5. 45º Directional/Normal Geometry (45ºx:0º)

In this geometry is recommended to use the same angular and spatial conditions for the 45º a: 0º geometry. However, in this geometry the light is irradiated at one 45 degree azimuth angle, which excludes the specular component but accentuates texture and directionality [66]. This geometry is usually used for the measurement of appearance effects of textured materials that are visually assessed at a 45/0 geometry.

3.1.2.6. Normal/ 45º Directional Geometry (0º: 45ºx)

This geometry uses the recommendations given for the 45ºx: 0º geometry but with the light path reversed [66].

3.2. Tristimulus Values and Color Matching Functions

3.2.1. The 1931 CIE Standard Observer

In the 1920’s the CIE, while working to specify the red, green, and yellow colored lights used in railroads and later in road traffic lights, developed the concept of the standard observer and a standardized apparatus as a method for color specification [65]. Guild and Wright, in separate studies, measured the color matching functions of a set of 7 and 10 observers, respectively. Both of the experiments were conducted under similar viewing conditions using filtered red, green and blue lights, in which the stimulus intensity was recorded when modified to match a test stimulus [65].

The viewing field used in the experiments had a bipartite area limited to a 2° visual angle subtended on the retina, with the surround fixed with complete darkness. The experiment was based on Grassmann's laws of additive color matching. The observers were presented with a monochromatic spectral light on one side of the field and they were required to match the color by combining and modifying the intensity of the three additive primary lights of red, blue and green.

The amount of light selected for each of the lights was recorded for each wavelength of the visible spectrum [65]. Both of the experiments showed strong agreement and the CIE concluded that the seventeen color normal English males was a valid estimation of the human population [65]. The average visual data were based on a reliable set of primaries in order to produce good repeatability when used in colorimeters. The selected primaries were 435.8 nm, 546.1 nm, and 700 nm as illustrated in Figure 18. The set of color matching functions is known as $\bar{r}_\lambda, \bar{g}_\lambda$ and $\bar{b}_\lambda$, in which the bars represent an average of all the observers included in the study. These functions defined the tristimulus values of the spectral colors for this particular group of primaries. Tristimulus values define the quantity of a set of primaries used to describe a color match in a given stimuli [65].

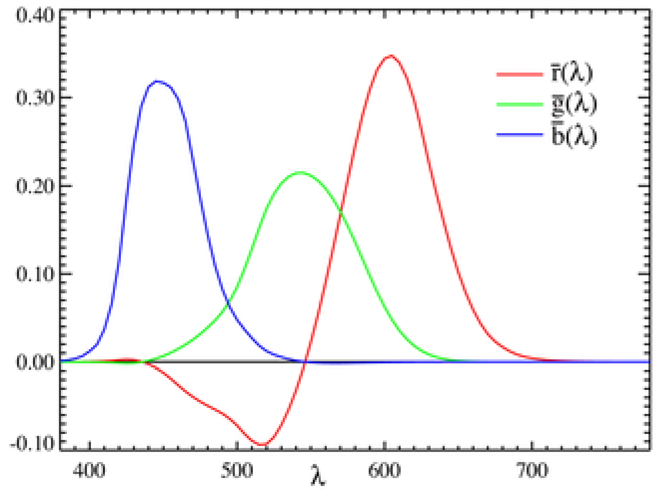


Figure 18. Color matching functions for the 1931 standard observer.

In 1924 the CIE developed a system of photometry. The spectral luminous efficiency function, V(λ), was developed for photopic vision. This function, if plotted as a function of wavelength, shows that the human visual system is more sensitive to the perception of brightness in spectral wavelengths located in the middle of the visual spectrum and less sensitive towards the extremes [18]. It is understood that the V(λ) corresponds to weighted sum of the three cone response functions.

Figure 18 shows that some of color matching functions have negative values. This presented problems to the CIE committee of the day, owing to the likely errors that would result from conducting many calculations by hand. To overcome this problem, the CIE transformed the RGB primaries to another set of primaries known as the XYZ.

This non-realizable (imaginary) set of primaries eliminated the negative values of the color matching functions. In addition, one of the color matching functions was forced to be equal to the photopic luminous efficiency function V(λ) [18], and thus comprised all luminance or lightness information. The corresponding color matching functions are known as $\bar{x}_{\lambda}, \bar{y}_{\lambda}$, and $\bar{z}_{\lambda}$. This system is known as the CIE 1931 2° standard observer. Figure 19 shows the plot of the 1931 color matching functions as a function of wavelength.

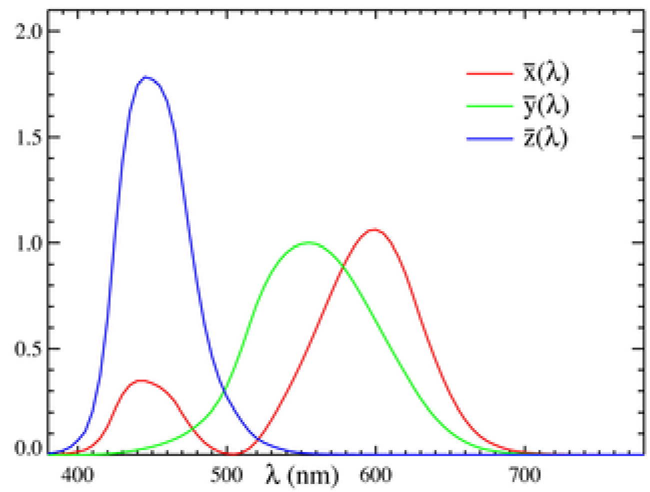


Figure 19. 1931 Standard observer.

3.2.2. The 1964 CIE Standard Observer

The experiments carried out in the 1920's that led to the development of the 1931 CIE standard observer were limited to the study of the fovea area, a 2^o subtended angle from the visual axis.

A number of applications require larger visual fields. While the CIE determined that the 1931 standard observer would predict accurately matches for a small visual field, this may not be the case for large-field color matching owing to the dramatically differing distribution of cones as the angle subtended increases [65]. Studies were carried out by Stiles, Burch, and also Speranskaya with the goal of establishing color matching functions for a 10^o field of view. The first two carried out the experiment using 49 observers and high level of illumination to avoid rod intrusion [67], while Speranskaya used 27 observers and lower levels of illumination [68]. The CIE removed the effect of rod intrusion in Speranskaya's data and combined the data from the two studies to develop what is known today as the 1964 CIE Supplementary Standard Observer, or more commonly known as the 10º standard observer. The color matching functions are denoted as $\bar{x}_{10\lambda}$, $\bar{y}_{10\lambda}$, and $\bar{z}_{10\lambda}$ [18]. These color matching functions are mostly used when the visual field angle exceeds 4º, which is the case for many applications. Another important issue with these groups of colorimetric functions is that $\bar{y}_{10\lambda}$ was not made to be equivalent to the luminance or lightness of a specified object or light source [65].

Color perception is a personal experience and it is expected to be different from the standard observers. However, this concept has been helpful for the development of colorimetry. Research has been carried out to find out if the established standard observers are truly the average and if a revision should be done. North and Fairchild [69, 70] studied the color matching functions of 18 observers that correlated similarly to the CIE 1931 standard observer. Using a different experimental methodology, Rich and Jalijali found similar results to those of North and Fairchild [71]. However, despite the emerging data, the CIE published in 2004 recommendations concerning the existing standard observer data [66]. Hence, the CIE 1931 standard colorimetric is recommended for visual color matching fields subtending a visual angle of about 1º to 4º at the eye of the observer. A 2º corresponds to a diameter of about 17mm at a viewing distance of 50 cm [66]. For a visual color matching field greater than 4º it is recommended to use the 10º observer. A 10º corresponds to diameter of about 9 cm viewed at a distance of 50 cm [66].

3.2.3. Calculating Tristimulus Values

A further step in the foundations of colorimetry was to broaden the tristimulus values so that they can be obtained for any stimulus characterized by a specific illuminant. In other words the tristimulus values described a colored object [18]. This was accomplished by multiplying the value of the color stimulus function $\phi_\lambda(\lambda)$ at each wavelength by the CIE color-matching functions, and integrating each set of results over the wavelength range corresponding to the visual spectrum (360 nm to 830 nm) [66]. For reflecting or transmitting colored objects, the CIE recommends replacing the color stimulus function $\phi_\lambda(\lambda)$ with the relative color stimulus function described as:

$$\phi_\lambda(\lambda) = R(\lambda) \times S(\lambda) \quad or \quad \phi_\lambda(\lambda) = \tau(\lambda) \times S(\lambda) \tag{3}$$

Hence, the equations needed to specify the tristimulus values of a reflecting colored object are given in equations (4-7) [65].

$$X = \sum_\lambda R(\lambda) S(\lambda) \bar{x} \Delta\lambda \tag{4}$$

$$Y = \sum_\lambda R(\lambda) S(\lambda) \bar{y} \Delta\lambda \tag{5}$$

$$Z = \sum_\lambda R(\lambda) S(\lambda) \bar{z} \Delta\lambda \tag{6}$$

$$k = \frac{100}{\sum_\lambda S_\lambda \bar{y}_\lambda \Delta\lambda} \tag{7}$$

In these equations, $R(\lambda)$ is the spectral reflectance factor of the stimuli evaluated with one of the CIE suggested geometries. $\tau(\lambda)$ is the spectral transmittance factor of the stimuli evaluated with one of the CIE suggested geometries.

$S(\lambda)$ is the relative spectral power distribution of the illuminant and is recommended to be one of the CIE standards [66].

The value, k, is a normalizing constant, and $\Delta\lambda$ is the wavelength interval which is usually 5, 10 or 20 nm [65]. For reflecting objects, Y has been assigned with a value of 100 that represents ideal white reflecting 100% at all wavelengths. For transmitting objects, Y also equal to a 100, but represents a colorless sample transmitting 100% at each wavelength. The k normalization constant accounts for this arrangement [65].

3.3. Chromaticity Diagrams

A colored object can be specified using the tristimulus values. These three values can be considered three variables that if mapped in a three dimensional space will define the position of such a colored object. Any objects with the same XYZ values for a given illuminant and observer combination will have the same color for the standard observer. In order to represent these values in a two dimensional way, for an easier visualization, chromaticity diagrams were developed. A chromaticity diagram is a two dimensional map in which the tristimulus values of an object are mathematically transformed into two variables. The transformation is done by a normalization that eliminates luminance information [18, 65]. The new variables are known as chromaticity coordinates defined by equations (8-10):

$$x = \frac{X}{X + Y + Z} \tag{8}$$

$$y = \frac{Y}{X + Y + Z} \tag{9}$$

$$z = \frac{Z}{X + Y + Z} \tag{10}$$

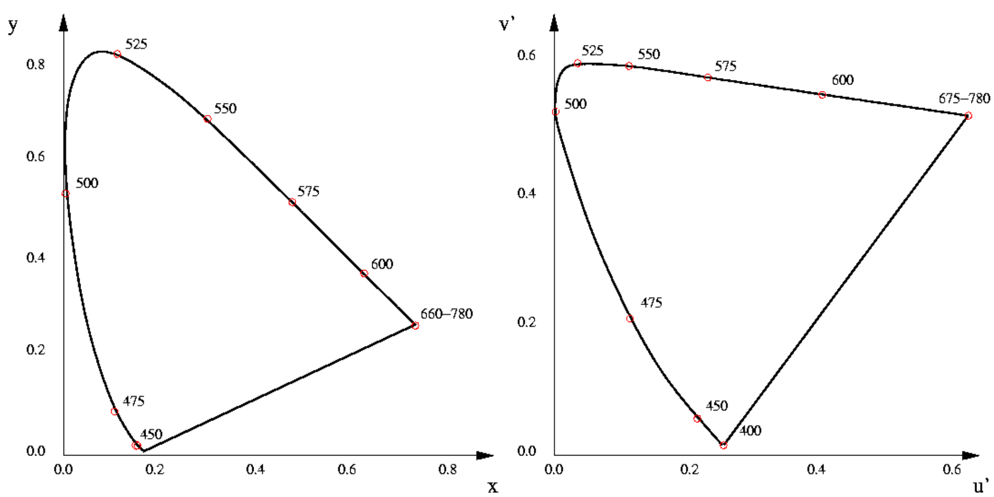
The sum of x, y, and z is equal to 1. Since there are only two chromaticity coordinates in a chromaticity diagram, the third coordinate can always be calculated by defining the other two.

Usually x and y are the coordinates plotted and this plot is known as the CIE chromaticity diagram [66]. The chromaticity coordinates of the visual spectral colors determine the boundaries of this diagram that looks like a horse-shoe shape [65]. The points making up the boundary of the diagram are known as the spectral locus [65]. Chromaticity diagrams for the 1931 and the 1964 CIE standard observers differ to some extent, but the basic feature remains the same [65]. While the space maintains additivity, it is not a perceptually uniform space. Over the last several decades, substantial effort has been made to transform chromaticity diagrams into perceptually more uniform spaces.

An important improvement was made and adopted by the CIE and it is known as the CIE 1976 Uniform Chromaticity Scales diagram [18]. This chromaticity diagram is recommended in cases where a projective transformation of the x,y diagram intending to have a color spacing perceptually more uniform is desired. The diagram is obtained plotting equations (11-12): Figure 20 shows the x,y and u',v' diagrams.

$$u' = \frac{4X}{X + 15Y + 32} \tag{11}$$

$$v' = \frac{9Y}{X + 15Y + 32} \tag{12}$$



(a) **(b)**

Figure 20. (a) 1931 x,y chromaticity diagram (b) u'v' uniform-chromaticity scale diagram.

3.4. Measuring Color

3.4.1. Visual Assessment

In the last several decades, color technology has evolved to a point where instrumental measurement can be reliable and visual measurements are either not needed or are required for verification only. However, visual assessments are still carried out in most industrial practices. Visual color assessment requires the preparation and standardized presentation and viewing of physical samples, which inevitably involves subjective judgments by trained observers [36]. AATCC Evaluation Procedure 9 provides a detailed account when evaluating and describing perceived color or color difference between two materials. The method defines the following: illumination conditions and the acceptable tolerances for each light source, viewing environment (color of the background), viewing geometries, presentation and preparation of the samples, and it also suggests a method to report the perceived color differences in terms related to hue, chroma and lightness [36].

In addition to standard practices for visual assessment, it is recommended when doing visual assessment to record all key viewing conditions used, including spectral power distribution of light source used, the correlated color temperature, illuminance, and lightness [65].

3.4.2. Instrumental Measurement

Instrumental measurement is now common practice in many situations requiring color quality control owing to the high variability in visual assessment and the need for accurate color assessment and communication. The interaction between observer, light and object are the basis for all instrumental measurement of color. Since direct measurement of color perception is not possible, the measurement of factors related to color perception can be used to numerically to describe color [65].

3.4.2.1. History Overview

Devices for color measurement date as early as 1900's and they were originally known as absorptiometers. They were used to establish, by a visual inspection, if two colored solution were equal in shade [72]. The first device designed for color measurement of opaque materials was developed in 1915-1920 and it was called reflectometer [72]. This device consisted of three colored filters and the apparatus was used to directly measure reflectance factors relative to the amount of theses primaries contained in the measured sample [72]. Following the development of the CIE 1931 standard observer, the tristimulus colorimeter was introduced [72]. This device replaced the existing reflectometer. In 1935, the General Electric Recording Spectrophotometer was introduced and this apparatus sustained the basis for modern spectrophotometry [72].

3.4.2.2. Tristimulus Colorimeters

Colorimeters are relatively simple instruments that provide numerical data representing the measured sample and a color difference in comparison to a standard sample [73]. A colorimeter contains three photodetectors similar to the eye: red, green, and blue that measure tristimulus values directly [19]. In a colorimeter, the light source typically illuminates the sample at an incidence angle of 45º. The reflected light from the sample is collected and passed to a detector, which consists of the three colored filters.

The combined response of the detector is then mathematically transformed to match the response of the eye [19, 72]. Figure 21 shows schematically the main features of a classical colorimeter.

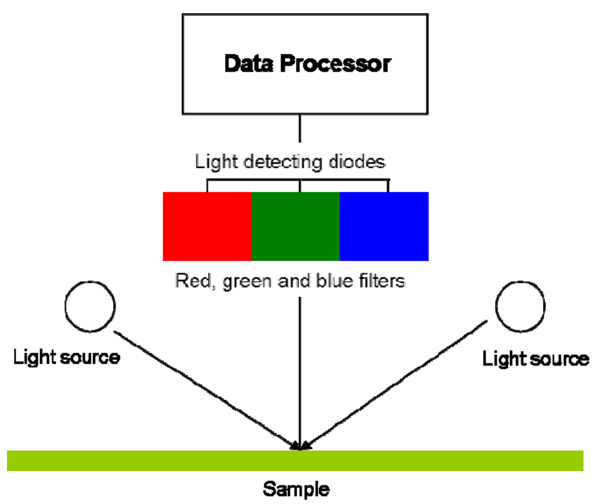


Figure 21. Scheme of a colorimeter.

Colorimeters are usually easy to operate, inexpensive, and fast [19, 73]. However, they are limited to a single illuminant/observer combination. Hence, it is impossible to determine metamerism and other important color attributes [19, 72].

3.4.2.3. Spectrocolorimeters

This device is to some extent an improvement of a colorimeter. A spectrocolorimeter can give colorimetric values such as XYZ or CIE L*a*b* for different CIE standard illuminants, but by far the most important class of instrument are spectrophotometers [72].

3.4.2.4. Spectrophotometers

Spectrophotometers can measure reflectance, transmittance or absorbance of a sample at different wavelengths [72]. When measuring reflectance, the spectrophotometer measures the reflectance of light, typically from 360 nm to 700 nm and provides as an output a reflectance ratio in comparison to a perfect diffuse white calibration tile. The reflectance factor is usually expressed in terms of percent reflectance [72]. All spectrophotometers have been designed with the same basic parts: a light source, a method to separate the light spectrum, a detection system, and a microprocessor for analysis of data and calculations. The layout of the parts and how it is operated determines the optical geometry of the instrument [19, 72]. Figure 22 shows the typical parts of a dual beam spectrophotometer.

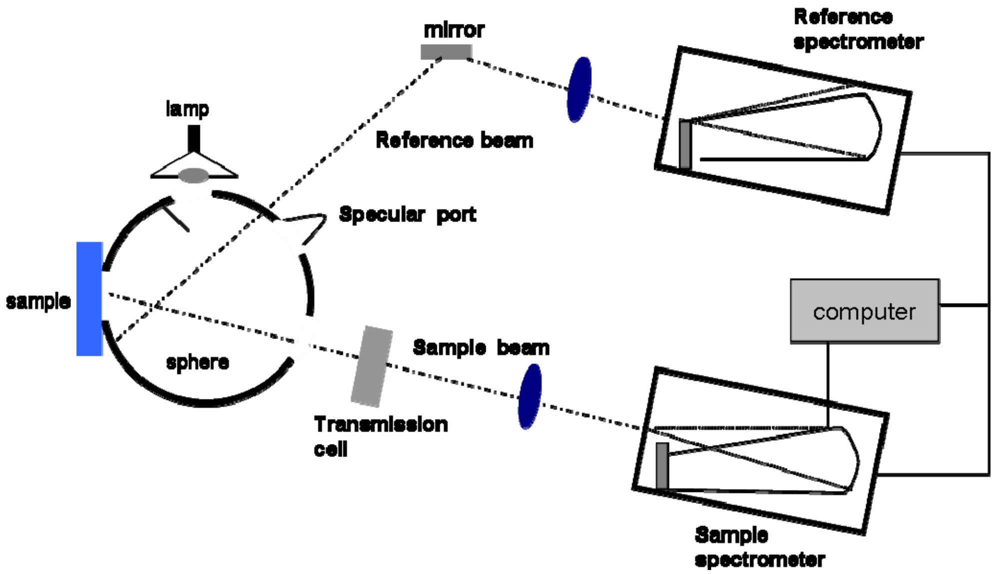


Figure 22. Diagram of a dual beam spectrophotometer.

This type of spectrophotometer contains a xenon lamp filtered to approximate illuminant D65, the light source also includes a filter wheel that control the UV content, which may be calibrated. In addition, the spectrophotometer includes a sphere approximately 15 cm in diameter which is covered in barium sulphate [19].

The reflectance curves from the visible region of the spectrum are in general smooth. This particular characteristic enables measurement of increments larger than 1 nm, with most modern spectrophotometers usually measuring the response in intervals of 5-20 nm [72], and are therefore known as abridged spectrophotometers. A range of 16 to 40 points are usually all that are needed in order to obtain useful data [19].

Some of the features found in modern spectrophotometers are the use of concave holographic diffraction gratings to diffract light into spectral bands. The spectral decomposition is then focused on to a 256-element diode array that is not linearly spaced. These features provide the basis for excellent inter-instrumental agreement, depending on the material to be measured [19].

3.4.2.5. Instrument Geometry

The instrument geometry used for measuring color was discussed in Section 3.1.2. The viewing geometry of the instrument determines the angles of illumination and the angles of observation. Figure 23 shows some of the most common viewing geometries used in color measurement.

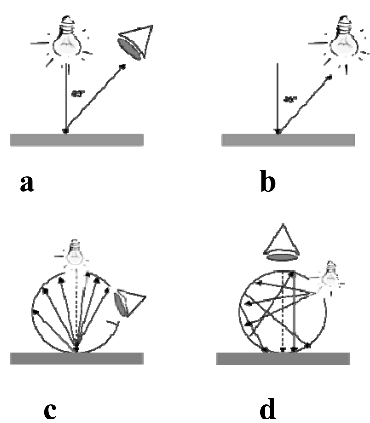


Figure 23. Most commonly used viewing geometries: a) 0/45 b) 45/0 c) 0/ diffused d) Diffuse/0.

3.4.2.6. Sample Presentation and Procedure for Instrumental Measurement

The colorimetric measurement of a sample should be representative of the appearance of an object [73]. A complete and detailed set of standard ways to present any kind of sample for color measurement and appearance can be found in the ASTM standards on color and appearance measurement. Standard method E 805-94, Standard practice for identification of instrumental methods of color or color difference measurement materials, describes the standard practice to identify the proper instrumental method for color and color difference measurement of samples. This method contains specific information regarding identification of samples and specimens and their main characteristics (such as if the specimen is opaque, fluorescent etc) the instrument geometry and the calibration for standards. The method does not contain specific information about how to carry out the measurement but instead establishes the critical criteria to be used when conducting a measurement [74].

For textiles color measurement, the AATCC has developed a set of recommended methods to present samples for reflectance measurement, Evaluation Procedure 6 (EP6), Instrumental color measurement [75], and Evaluation Procedure 7 (EP7), Instrumental assessment of the change in color of a test specimen [76]. EP6 recommends standard practices for instrumental measurement of reflectance or transmittance and includes related calculations. When measuring reflectance, the procedure suggests standard practices for sample preparation and instrument standardization according to the characteristics of the material to be measured [75].

EP7 was developed as an instrumental alternative to the visual assessment method of EP1, Gray scale for color change [63, 76].

Christment recommends specifying and controlling the following conditions: temperature, humidity, pressure, thickness, tension, and the presence of liquid, vapor, dust or other elements on the surface of the material. In addition, he includes a recommended procedure for colorimetric measurement [73]. Connelly also suggested methods for a good sample presentation in textile materials [77] and provides recommendations for measurement methods for fibers, yarns, wovens, knits, pile, textured and printed fabrics [77].

Hence, instrumental reflectance measurement of samples requires specification of all variables including measurement conditions using calibrated instruments. Standardization of methodologies used by various industries is necessary to ensure that all parties use the same 'color language' when communicating within the supply chain [3, 77-79].

3.4.2.7. Other Measurement Instruments

There are some instruments used in color measurement that do not necessarily measure color, but one attribute related to color perception, including the following.

3.4.2.7.1. Spectroradiometers

A spectroradiometer is an instrument that measures the spectral properties of light sources, i.e. light booths, projectors, monitors etc.

They are essentially spectrophotometers for quantifying light sources since they operate on the same basic principles. Spectroradiometers can take measurements at shorter wavelength intervals than conventional spectrophotometers [65]. There are two recommended geometries for the measurement of light: spectral radiance and spectral irradiance. The spectral radiance is the flux falling per unit area. Spectral irradiance is the flux "emanating from a surface or falling surface in a given direction, per unit of projected area of the surface as viewed from that direction, per unit of solid angle" [65]. Typically, both radiometric and photometric data are provided as outputs.

3.4.2.7.2. Photometers

A photometer is an instrument that measures the illuminance or the luminance of sources. Photometers have the same geometries established for spectroradiometers. The instrument is similar to a colorimeter, but with only one channel [65].

3.4.2.7.3. Imaging Systems

Imaging devices such as photographic film, television, cameras, digital cameras, scanners etc, can detain color as a function of spatial location. These devices offer new opportunities for reliable color measurement, although they typically exhibit specific limiting factors [65].

3.4.2.7.4. Goniophotometers

A goniophotometer measures goniometric properties of an object. Goniometric properties of an object are characteristics related to the particular regular and diffuse reflection of the surface such as gloss, specular gloss, haze, luster, sheen, and image distinction. Data are obtained via measurement of the reflectance ratio as a function of viewing angle [65].

3.4.3. Summary

When measuring color, it is important to consider which factors are important to achieving accurate, repeatable and reproducible measurements in order to establish which instrument and which conditions are needed. First, it is necessary to establish the type of spectra data required. Also, it is important to establish the nature of stimuli components. In addition, it is important to establish the instrument geometry. Furthermore is necessary to evaluate accuracy and repeatability of the measurement [65]. Finally, it is important to verify and calibrate all spectrophotometric and related optical devices in order to maintain reliability of output data.

4. Color Perception Phenomena

As already stated perception of color is a personal experience and cannot be described with the fundamental concepts of basic colorimetry alone. The CIE color description system is limited since it is based on color matching using colored lights and involves the definition of a standard observer [18]. Psychologists believe that personal experience of color is influenced by a large number of variables. For the past two centuries, these variables have been studied and it has been demonstrated that color experience goes beyond the interaction of light, visual system and object. Color perception is thus an experience influenced by the combination of complex variables [16].

Mathematical models constructed based on visual response of color normal observers can be useful under specific conditions but need to involve the role of various perceptual phenomena. Perceptual studies have shown that there is not a linear relationship between stimuli and perception [16]. Some of the factors that have shown to have an effect in human perception of color are related to: the retinal area of stimulation, viewing angle, luminance level, viewing surroundings, background, size, shape, surface properties, and viewing geometry among many others. If one of these factors is altered, two stimulus having exact tristimulus values may no longer be perceived as a match. Some of the most important factors influencing perception have been studied in depth and implemented in color appearance models [16].

4.1. Important Definitions

Study of the factors influencing color perception phenomena is an active area of research in color science. For a systematic study, it is necessary to use certain terms in order to describe precisely the phenomena studied. Color as a term is comprehended by most people; however, the description of color cannot be achieved without the use of an example [16, 18].

4.1.1. Hue

Hue is the attribute of a visual sensation that gives a particular color the name such as red, blue, yellow, etc. [18, 73]. The hue component is probably the most notorious of color attributes since all the hues are defined by the wavelengths of the visible spectrum. Some hues, however, do not exist in the spectrum, specifically purples, but they can be obtained by mixing lights from both ends of the spectrum. Hues have a circular nature thanks to reddish nature at the ends of the spectrum [16].

Hering observed an interesting hue phenomenon in the 19th century [80]. He defined four fundamental Unique Hues (UH) red, blue, yellow, and green that are observer dependent. Since that time, a number of studies have attempted to identify the location of UHs in relation to other colors as well as the perceptual ability of individual observers to identify them [16, 18]. In a recent study, examining the selection of unique hues, a large inter-observer variability was observed.

It was also shown that intra-observer variability accounted for 15% of the total variation, highlighting that unique hue selections are highly observer dependant [11].

Two phenomena related to the perception of hue are the Helson-Judd effect and the Bezold-Brücke hue shift effect. The Helson-Judd effect cannot be perceived under normal viewing conditions and it does not have much importance in practical situations. This effect is perceived when viewing neutral Munsell patches on a uniform background under nearly monochromatic illumination conditions. Under these conditions, observers have reported the perception of a glowing light of the hue complementary to the light source overlaying the samples that are darker than the background [18]. One assumption about hue is that it can be specified by the wavelength of monochromatic light. However, the Bezold-Brücke Hue shift effect shows a change in luminance results in a shift in hue of a monochromatic stimulus. This effect, however, does not occur in related colors, a concept that will be explained later [18].

4.1.2. Lightness

Lightness is the visual impression of brightness of a non white object compared to a series if grays ranging from black to white [16, 65]. This attribute of color is independent of hue and chroma and varies linearly [73]. Some of the phenomena related to the perception of lightness include the Helmontz-Kohlrausch effect and lightness crispening.

In the Helmontz-Kohlraush effect the perceived brightness is not directly proportional to stimulus luminance (or relative luminance).

A stimulus is perceived to be brighter as it gets more chromatic while the luminance is kept constant [16, 18]. Figure 24 shows two typical examples of the Helmontz-Kohlrausch effect. The colored squares and the surrounding gray squares have the same lightness values, however, the colored square is perceived to be brighter.

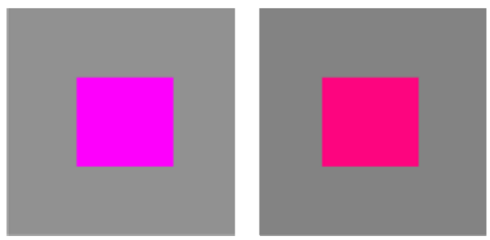


Figure 24. Colored squares at high saturation surrounded by gray square with the same lightness values (Helmontz-Kohlrausch effect).

Lightness crispening is a phenomenon that results in increasing the degree of perceived color difference between two samples when the background on which the samples are viewed is similar in lightness to the stimuli [18]. Semelmelroth was the first to discover this effect and he developed a model for its prediction [18, 43]. Xin et al. published a study in which they evaluated 38 neutral samples under 5 different backgrounds. The results confirmed the presence of a crispening effect when viewing samples similar in lightness to the background.

They also indicated that, among recent color difference formulas, CIEDE00 performed best since it accounts for the crispening effect [43]. Figure 25 shows an example of the lightness crispening effect.

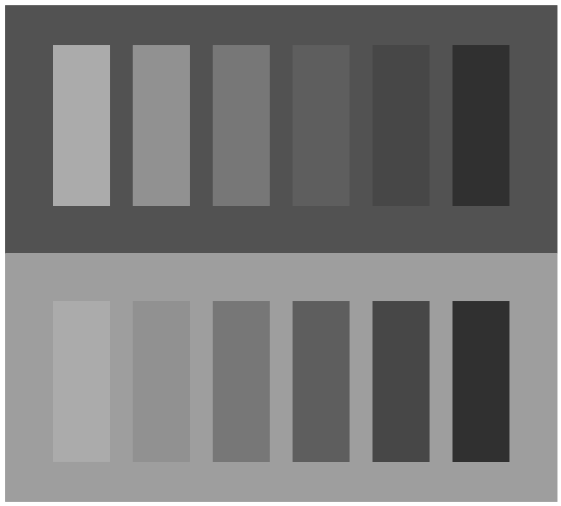


Figure 25. Lightness crispening effect.

4.1.3. Chroma

Chroma is also known as saturation or purity of a color [73]. Chroma is the color attribute that gives the visual sensation of saturation or relative colorfulness in a stimulus [16, 18]. Chroma is a quantitative attribute of color and has been found to be the most difficult to evaluate [18].

An important phenomena related to chroma is chromatic crispening. In this phenomenon an increased degree of perceived color difference between two chromatic samples is observed when the background on which the samples are viewed is similar in chroma [16].

4.1.4. Unrelated and Related Colors

Unrelated colors are not commonly seen. Unrelated colors can only be viewed in isolation from other colors [16, 18]. Their perception is accomplished under special viewing conditions [16]. On the other hand, related colors refer to perceiving colored stimuli when viewed in relationship to other colored stimuli [18]. It is not possible to see unrelated colors that seem to be brown or gray [18]. An object in a nature scene that appears to be brown in relationship to other colors, when viewed in isolation will appear orange or yellow [16].

4.2. Additional Perceptual Phenomena

4.2.1. Simultaneous Contrast or Chromatic Induction

In this phenomenon, an achromatic color shifts its appearance when the background is changed [16]. Simultaneous contrast owed its name to the fact that cause and effect happen at the same time. Figure 26 shows an example of simultaneous contrast. The gray squares placed on the top half of the picture are identical to the ones shown in the bottom on the yellow and blue background.

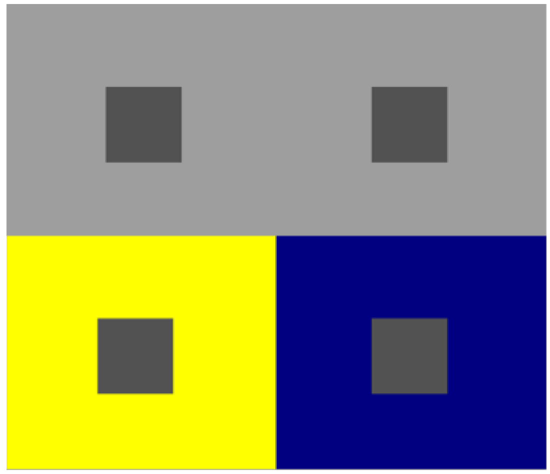


Figure 26. Simultaneous contrast.

Contrast phenomena can also affect perception of color differences [65]. Figure 27 shows a pair of samples that approximately match when viewed on a gray background; if on the other hand the background is changed to white the samples seem to match perfectly. Once the background is changed to black, the color difference of the pair is perceived to be larger.

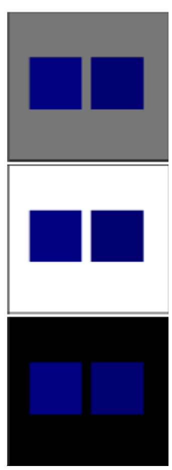


Figure 27. Simultaneous contrast in pairs.

Understanding the effect of simultaneous contrast is of significant importance for artists and designers and it has been a subject of research for nearly 150 years [16].

In 2004, Cunthasaksiri et al used the “Recognized Visual Space of Illumination” (RVSI). They used this theory to explain the simultaneous contrast on a simple center-surround sample configuration. They pointed out that that the visual system constructs an RVSI on the surround which then determines the perception of the center of the sample. The theory stipulates that the contrast between the center and the surround should get bigger as the size of the surround increases, which was proved later [81].

In 2005, Wu et al. investigated the simultaneous contrast phenomenon using 174 CRT and 154 surface colors. They found a strong lightness contrast effect for both sets of colors. They also showed that the lightness of a color decreases under a lighter surround. However, the effect was significantly larger on the CRT display, possibly due to the fabric texture effect and its interaction with light [82]. Wu and Wardman, in a separate work recommended a modification to the CIECAM02 color appearance model in order to predict the effect of the simultaneous contrast phenomenon [83].

4.2.2. Successive Contrast (Afterimage Effect)

An afterimage effect is the phenomenon that occurs in the viewing field after an observer is exposed for a prolonged period (about 20 seconds) to a relatively strong stimulus followed by exposure to a white background.

If the afterimage formed is composed of the complementary colors of the original stimuli, the afterimage is denoted as negative. On the other hand, if the afterimage is comprised of the same colors as the stimuli the effect is called positive, however, positive afterimages can only be obtained under very special viewing conditions.

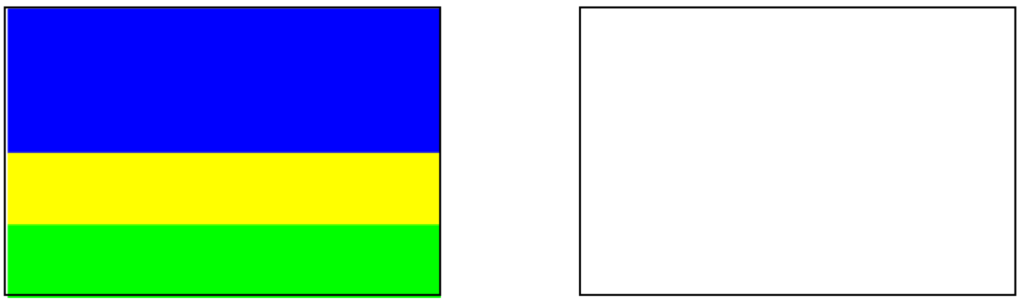


Figure 28. Afterimage effect. Please stare at the colored Figure for about 40 or more seconds then move your eyes away to the blank portion of the page. You should be able to see the Colombian flag.

4.2.3. Spreading

If a stimulus presenting simultaneous contrast effect increases its spatial frequency or becomes smaller, the simultaneous effect disappears and is swapped by another phenomenon known as spreading [18]. Spreading is the phenomenon in which spatial fusion is perceived, in other words, stimuli seems to blend with the background, although still distinctive in comparison to the background [18]. Figure 29 shows how achromatic rectangles with low spatial frequency (large rectangle) appear to be redder thanks to simultaneous contrast.

On the other hand, the rectangles with high spatial frequencies (rectangle made with lines) show the spreading effect. The rectangles are perceived to be greener and blend with the surround.

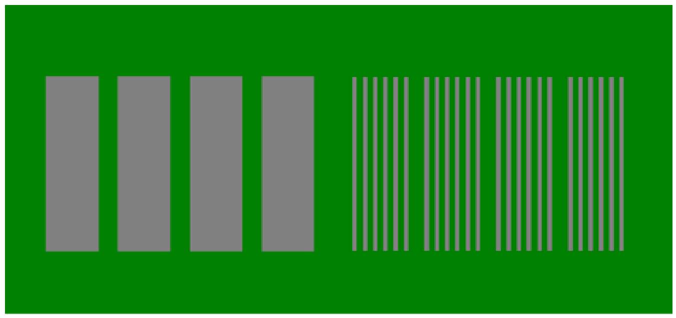


Figure 29. Spreading effect.

4.2.4. Mach Bands

This phenomenon can be better explained with a concept known as lateral inhibition [17]. Lateral inhibition occurs when light falls both into dark and light receptors creating competition between the two regions where one part of the receptor field wants to be active and another does not. Figure 30 shows the effect created by Mach bands. Each band seems to be a gradient bending inwards.

Figure 30. Mach bands.

4.2.5. Color Constancy

This phenomenon is one of the most important effects related to perception and assessment of color. Color constancy is not present in humans [18].

Color constancy is defined as the predisposition of a colored object to remain constant when the level and color of the illumination conditions are changed [18, 19, 65]. Fairchild quotes Evans' interesting description of this phenomenon: "…in everyday life we are accustomed to thinking of most colors as not changing at all. This is due to the tendency to remember colors rather than to look at them closely" [18]. It is common to hear descriptions such as white paper, red apple and so on. Most natural objects seem to have a constant appearance when viewed under different light sources regardless of the amount of light reaching the eye. This experience can be explained by the chromatic adaptation phenomenon. Compensation for the change in illumination conditions takes place through a complex communication and interaction between brain and the eye [18, 19].

Color constancy is experienced in everyday life and is more complex than is imagined [19]. Most objects tend to remain relatively color constant under normal viewing conditions, however; this is not always the case. The color of an item in a store may be different to that when viewed under a different set of viewing conditions. This phenomenon is known as flare or color inconstancy.

Assessment and prediction of color constancy is an important issue in mathematical models that describe color especially in color appearance models [16]. In 1998, Luo and Hunt proposed a chromatic adaptation transform (CAT) that was implemented in the CIECAM97. The transformation was used to derive the CMCCON97, a Color Inconstancy Index that accounts for illuminants involving different chromaticities [84, 85].

Wardman and Hallam evaluated color constancy using dichroic ladders against color inconstant samples using different light sources. The ladders consisted of a series of colored samples with small hue differences from one sample to the next. The samples used in the study were near the neutral area in the color space. The found that the proposed CMCCON97 agreed reasonably well with the visual assessments made under the experimental conditions. However, they questioned the performance of the model under different experimental settings [84].

4.2.6. Metamerism

This phenomenon is related to color inconstancy but involves two samples. A metameric pair of samples has identical color appearance under specific lighting and viewing conditions, but not when illumination or viewing conditions are changed [73]. Metamerism applies to both colored lights and objects [16].

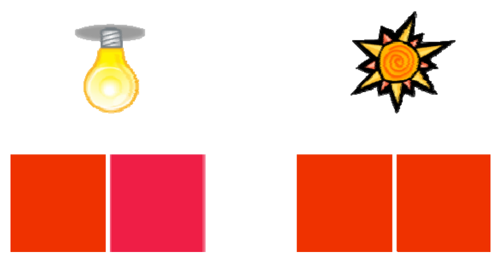


Figure 31. Metamerism Effect.

Metamerism phenomenon significantly affects color technology [65]. Colored objects with different reflectance data can still produce the same visual appearance. In fact, color matching is achieved using this particularity, allowing for matching colors without using the same materials [65]. Color reproduction can be achieved even if colorants and pigments have to be replaced because they are found to be toxic or to reduce costs [16, 65]. Metamerism is one of the reasons why some technologies have evolved, for example, photography, television, and movies just to name few.

These new technologies reproduce a whole chromatic world only with three or four primaries [16, 65].

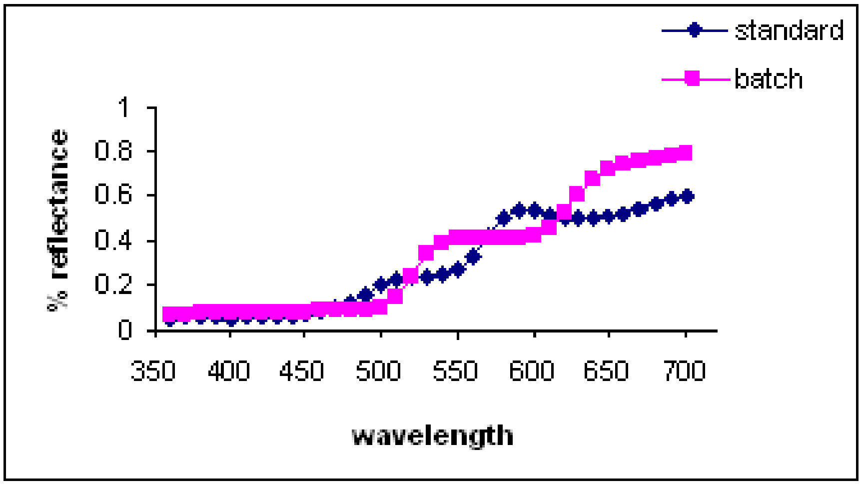


Figure 32. Reflectance curves for a metameric pair.

As stated earlier the perception of color is due to the interaction of light source, object and observer. A change in any of these factors will result in color disparity. If the pair no longer matches when the illuminant is changed, the effect is known as illuminant metamerism. If the effect is due to a change of observers, an observer metamerism is obtained [65]. Kuo and Luo studied different methods for quantifying metamerism in visual and instrumental assessment of color [46, 86]. Alfvin and Fairchild [87] showed intra and inter observer variability in metameric color matching due to a change of observers. They found that inter-observer variability was about twice the intra-observer variation.

Also, they found that the metameric matches for the recommended CIE standard observers were inside the 95% confidence ellipsoids of the distribution of the inter-observer matches, concluding that the CIE recommended standard observer is a reasonable representation of an average observer [87, 88]. Diaz et al. in 1998 recommended a method to assess observer metamerism [88].

Two other cases of metamerism have been reported by Günter Wyszecki [89]. One is a special kind of observer metamerism called "field metamerism" and is explained by the Maxwell spot [89]. This special type of metamerism occurs when viewing samples that are greater in size than a 4º field of view. Mathematically this can be treated as observer metamerism [89]. Another type of metamerism suggested by Wyszecki is related to viewing geometry and viewing conditions [89].

The perceived low/off quality of metameric products can often lead to consumer dissatisfaction. Metamerism can be desirable because it facilitates color reproduction, however, it should be noted that reproduced colors would only match the standard color under specific viewing/illumination conditions.

Indices of metamerism have been developed as metrics to assess the degree of metamerism during color matching. These indices are based on colorimetric concepts and coordinates recommended by the CIE [89].

The condition in which the match was made is known as the reference condition and the condition in which the match is evaluated is known as the test condition [65]. For example, a reference condition could employ D65 illuminant and the 1964 standard observer, and the test condition could use the illuminant A and the 1964 standard observer. In this case D65 is known as the primary illuminant and illuminant A is the secondary illuminant. Special cases may require more than two illuminants for color matching. The definition of the order of the illuminant in such cases is essential because it provides the hierarchy of importance of the illuminants [65].

Around 1980's new terminology related to metamerism was developed. The terms paramerism, parameric, and paramers were introduced to clarify some differences in color matching related to the reflectance data. Metameric samples have spectrally different information but the same tristimulus values. On the other hand, parameric samples present similar spectrophotometric curves and produce approximately the same color appearance under the same viewing conditions [65]. Residual color difference is defined as the difference found in metameric pairs that do not achieve colorimetric equality for a set of reference conditions [90]. Many methods to correct residual color difference have been suggested in the past. These include the CIE recommended multiplicative correction method, parameric decomposition [65], and the additive correction in L*a*b* [66, 90].

Li and Berns recently studied different methods to improve residual color difference in metameric matches. They found that the recommended CIE method performed poorly in comparison to the parameric decomposition [90].

In addition to the color appearance phenomena mentioned in this chapter, there are several other reported color perception phenomena including the Abney effect, Hunt effect, Stevens effect, Barteslon-Breneman contrast effect, effect of discounting illuminant among others which can be found elsewhere [15, 18].

5. Color Specification

5.1. Color Order Systems

As already discussed the description of color is a complex task, involves many dimensions, and requires a classification system. A color system is defined as a set of principles to arrange, organize, designate, and locate colors according to their attributes [19]. About 400 color order systems have been developed over the years [19], some dating back thousands of years [65]. One of the first color order systems was developed by Aristotle around 350 BC. The system was loosely three dimensional. Black and white were placed opposite to each other, and red was placed in between them to represent the sky between day and night [19].

Most of the existing systems are tridimensional with slight variations. This tridimensional arrangement is attributed to Forcius in 1611 with a later rediscovery made by Munsell. Most of the color systems are based on three methods: color mixing (color physics), color perception and color matching (color experimentation) [19]. Color systems based on color perception and color matching are discussed in this section.

5.1.1. Systems Based on Color Perception

These systems are based on perceptual principles of equal visual spacing [65]. Some of the most important color systems based on color perception are the Munsell Color System, the

Natural Color System (NCS), and the Optical Society of America Uniform Color Scales System (OSA-UCS).

5.1.1.1. Munsell Color System

This system is one the most widely used color ordering models. It was developed by a well known American artist, Albert Munsell, in the early 1900's as a teaching tool for art students [15, 65]. He developed a three dimensional color space using three attributes: hue, lightness and chroma. These attributes are known correspondingly as the Munsell Hue, Munsell Value, and Munsell Chroma. Each attribute is scaled in such a way that equal changes in any of the attributes results in the same degree of perceived color difference [18, 19].

Munsell's desire to numerically classify these attributes led to designating 10 main steps for each of the hue, value, and chroma attributes. The colors are arranged on a vertical cylinder where the vertical axis represents the Munsell Value (lightness). The lower end of the value axis (V=0) represents perfect black and the upper end (V=10) represents white or the perfect reflecting diffuser; the intermediate values represent a range of achromatic colors [19]. There are 10 principal hues: red (R), yellow-red (YR), yellow (Y), green-yellow (GY), green (G), blue-green (BY), blue (B), purple-blue (PB), purple (P) and red-purple (RP). The hues are arranged in a circular fashion and each individual hue can be further divided into 10 sub hues: i.e. 1R, 2R, etc.

In a defined hue and value setting, colors are separated with steps of nearly equal perceptual difference in order of increasing chroma when moving from the center to the outside of the cylinder as shown in Figure 33.

It is important to note that all hues are not extended to the same maximum chroma at each value. The original arrangement created by Munsell was later modified which led to the current specification of the Munsell system. Using the Munsell notation, 5R 4/6, denotes a red sample, with a value of 4, and a chroma of 6 [65].

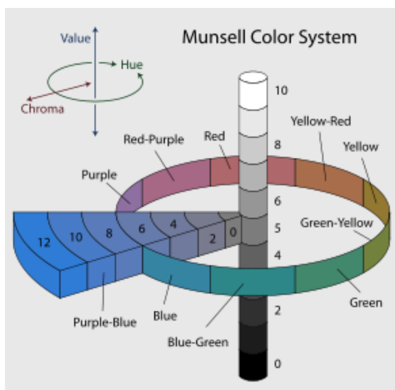


Figure 33. Munsell system showing the hue circle corresponding to a value of 5 and a chroma of 6.

5.1.1.2. Natural Color System (NCS)

The NCS color system, developed by Johanson and Hesselgren, is based on Hering's theory of opponent color channels [15, 65]. Johanson defined hue, saturation, and value (lightness) as the three attributes needed to characterize any perceived color [15]. This system is arranged into a tridimensional hexagon allowing for various combinations of attributes according to Hering's opponent color theory as shown in Figure 34. The concept of unique hues, defined in section 4.1.1, is used for the organization of colors. For instance, unique blue connects with green, red, black and white, but not with yellow [65]. Any given hue can be defined as a percentage of adjacent unique hues. For example, Y10R represents a color that contains 90% yellow and 10 % Red [65]. The color space can be sliced into planes resulting in color triangles. Using the position of a color in the triangle, the chromaticity (c), the blackness (s), and whiteness can be determined as a percentage. A color is defined by its blackness, chromaticness and hue attributes respectively. Using the NCS notation, 10, 80, G denotes a chromatic sample with10% blackness, 80% chromaticness, and a green hue [65].

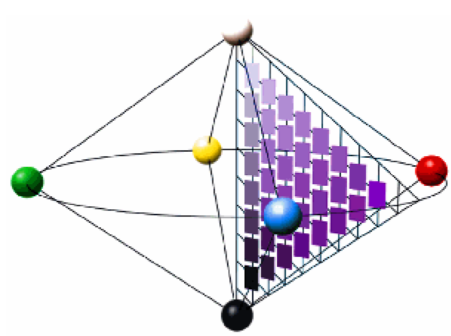


Figure 34. Natural Color System.

5.1.1.3. The OSA Uniform Color Scales System

The OSA UCS system is the result of the work of the Uniform Color Scale Committee which was established by the Optical Society of America in 1947. It is designed in a three dimensional Euclidean space with L, j, and g axes. L represents lightness; j denotes yellowness-blueness, and g signifies redness-greenness. This system is designed in such a way that any given color is separated from its neighbors in all three dimensions in equal perceptual color difference steps [18, 65]. The tridimensionality of equal spacing is achieved through a cubo-octahedron in which a sample is equally spaced from 12 nearest neighbors. This arrangement allows for rectangularly sampled planes that can be viewed from different angles [18, 65]. A schematic representation of the OSA UCS system is shown in Figure 35.

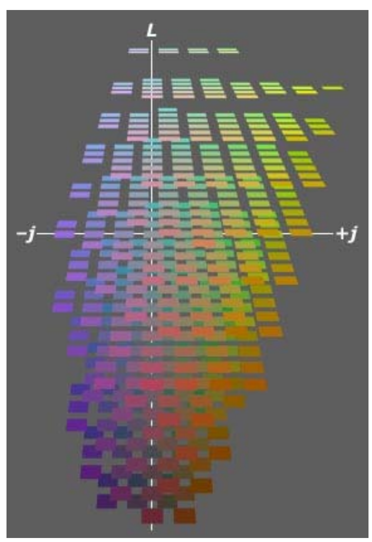


Figure 35. Structure of the OSA Uniform Color Scales System.

In addition to the systems described above there are several other systems based on color perception including: the Colorcurve System, the DIN System (Deutsches Institut für Normung), the Ostwald System, the Coloroid System, the Universal Color Language, and the ISCC-NBS Method for Designating Colors and a Dictionary of Color Names, amongst others [18, 65].

5.2. Color Spaces

Even though color order systems can be useful, it is necessary to have a collection systematically organized in a color space. Several color spaces have been developed some of which will be discussed in the following sections.

5.2.1. CIE xyY

The development to the CIE xyY color space, shown in Figure 36, was described in detail in section 3.2. As already discussed the CIE xyY is not perceptually uniform and therefore since its introduction many researchers have attempted to transform its coordinates or the tristimulus values into a more perceptually uniform space [19]. Some of the spaces based on such transformations are described in the following sections.

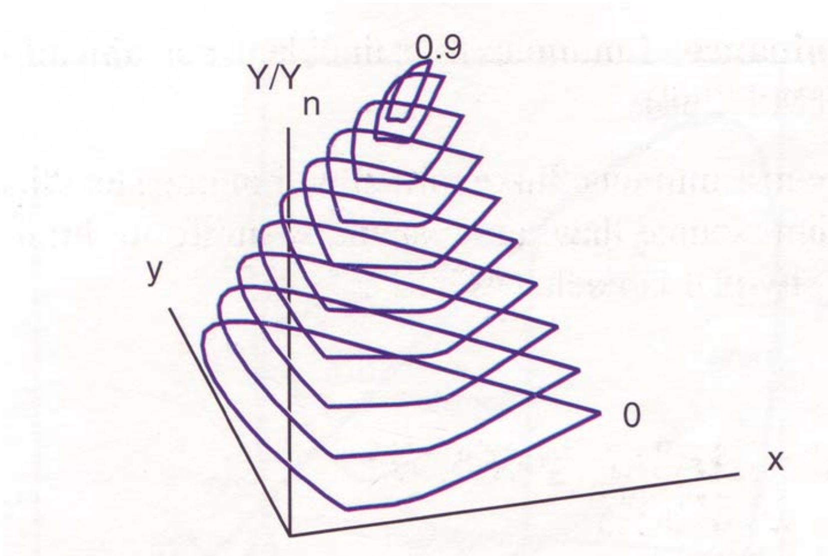


Figure 36. CIE xyY color space.

5.2.2. Judd and MacAdam UCS (Uniform-Chromaticity Scale) Diagrams

The Judd-MacAdam UCS model is an example of transformations used to improve the perceptual uniformity of the CIE x,yY system which was described in detail in 3.3 [19]. Two types of transformations have been generally employed: linear and nonlinear. The linear transformations maintain additive color mixing features and were used by Judd (1935) and MacAdam (1937). In 1960, the CIE adopted the UCS proposed by MacAdam which is known as the u,v chromaticity diagram. One of the most important shortcomings of this UCS diagram is the lack of perceptual uniformity as a function of luminance factor.

5.2.3. Hunter Lαβ and Scofield Lab Color Spaces

Hering's theory of opponent color vision led to the development of two independent approaches to develop uniform color spaces. Hering's theory includes three opposing pairs: blackness-whiteness, redness-greenness, and yellowness-blueness. Each color can be specified by quantifying these three attributes. Hunter proposed a formula in which lightness ($Y^{1/2}$) was quantified by the Munsell V function and using the CIE x, and y both of the other two opposing pairs were nonlinearly transformed into two variables called α and β respectively. Using XYZ tristimulus values he modeled redness minus greenness as a function of X-Y, and yellowness minus blueness with a scaling factor of 0.4 as a function of Y-Z to achieve uniformity in his color space [19]. Scofield modified Hunter's formula a year later. He adopted the variables used today: L for lightness, *a* for redness-greenness, and *b* for yellowness-blueness where $L = 10Y^{1/2}$ and *a* and *b* are functions of α and β [19].

5.2.4. Adams Chromatic Value Color Space

Adams in a separate development in 1942 suggested linking Hering's opponent color theory and the Young-Helmhotlz trichromatic theory. He proposed that the red-green opponent response was a function of the difference in red-green sensitivity of the eye receptors, and the yellow-blue opponent response was a function of the difference in green-blue sensitivity of the eye receptors [19].

These differences were modeled as X-Y and Z-Y. Adams's approach included two important features: the tristimulus values of a given color under a specific illuminant, and a specific observer normalized by dividing it by the perfect reflecting diffuser. In addition, Adams applied the Munsell function of the time ($Y^{1/2}$) to the X and Z tristimulus functions to define the three functions shown in equation (13):

$$V_x = (X/X_n)^{1/2}$$
$$V_y = V = Y^{1/2} \qquad (13)$$
$$V_z = (Z/Z_n)^{1/2}$$

(by definition $Y_n = 100$ for $0 \leq Y \leq 100$)

5.2.5. Hunter Lab Color Space

Around 1958, Hunter developed a new formula which was originally designed for use with illuminant C and the 1931 2º standard observer. A general form of the equation suitable for use under other conditions is shown in equation (14) [19]:

$$L_H = 100\, Y_R^{1/2}$$
$$a_H = 175\,(X_R - Y_R)\left(\frac{0.0102 X_n}{Y_R}\right)^{1/2} \qquad (14)$$
$$b_H = 0.4 \times 175\,(Y_R - Z_R)\left(\frac{0.00847 Z_n}{Y_R}\right)^{1/2}$$

L_H designates lightness, a_H redness-greenness, and b_H yellowness-blueness. The subscript H denotes that these are Hunter's coordinates. X_R, Y_R, Z_R, are the tristimulus ratios (X/X_n, Y/Y_n, and Z/Z_n) and specific factors were included for illuminant C and 2º observer [19].

5.2.6. Adams-Nickerson (ANLAB) Color Space

Around 1944 Nickerson and Stultz replaced the coordinates V_x, V_y, and V_z of the Adams chromatic space with a Judd polynomial obtained after substitution of Y for X_R, Y_R, Z_R, and V for V_x, V_y, and V_z. (this polynomial[1] was used to improve the perceptual spacing in the Munsell system) [19]. In addition, they suggested scaling the lightness coordinate by a factor of 0.23.

The three opponent scales were denoted as L, A, and B. The resulting formula is given in equation (15):

$$\begin{aligned} L_{AN} &= 0.23\, SV_y \\ A_{AN} &= S\,(V_x - V_z) \\ B_{AN} &= 0.4S\,(V_y - V_Z) \end{aligned} \tag{15}$$

The factor S is used to scale the entire color space according to various applications. Initially this factor was suggested to be 42. However, this was later changed to 40 [19].

[1] $Y = 1.2219V - 0.23111V^2 + 0.23951V^3 - 0.021009V^4 + 0.0008404V^5$

5.2.7. CIE 1960 UCS Diagram and CIE 1964 (U*V*W*) Color Space

In an attempt to standardize a single formula to represent color space coordinates, the CIE adopted the UCS model developed by MacAdam in 1937. In 1964 the UCS diagram incorporated a cube root transformation of the lightness denoted with W*. This coordinate represented the third axis located perpendicular to U* and V*. U* and V* are nonlinear transformations of the UCS variables u and v. The resulting space is known as the CIE 1964 (U*V*W*) uniform color space [19]. Equation (16) shows the transformations used:

$$\begin{aligned} W^* &= 25Y^{1/3} - 17 \\ U^* &= 13W^*(u - u_n) \\ V^* &= 13W^*(v - v_n) \end{aligned} \tag{16}$$

5.2.8. CIE 1976 UCS Diagram

The attempts to standardize color space models resulted in the development of several new formulas and extensive investigation of the existing models. In 1973, the CIE agreed to adopt a cube root version of the ANLAB. The resulting color space is known as the CIELAB [65]. The CIELAB does not employ a uniform chromaticity scale due to the use of the nonlinear transformations of the tristimulus values. However, the imaging industry representatives recommended that a chromaticity diagram be developed [65]. As a first attempt, Eastwood suggested a modified version of the 1960 UCS model incorporating a cube root of the lightness function similar to the ANLAB model as shown in equation (17).

Other modifications were also attempted, which are briefly described in the following sections:

$$
\begin{aligned}
&u' = u = \frac{4x}{P} = \frac{4X}{Q} \\
&v' = 1.5v = \frac{9y}{P} = \frac{9Y}{Q} \\
&where: \\
&P = 3 - 2x + 12y \\
&Q = X + 15Y + 3Z
\end{aligned}
\tag{17}
$$

5.2.9. CIELUV Color Space

The second recommendation to improve the uniformity of the color space was the addition of an axis for lightness perpendicular to u* and v* (derived from u' and v').

The resulting coordinates are known as the CIE 1976 (L*u*v*) color space which is abbreviated to CIELUV. Equations (18) show the component axes of this space:

$$
\begin{aligned}
&L^* = 116(Y_R)^{1/3} - 16 \qquad if\ Y_R > 0.008856 \\
&L^* = 903.3(Y_R) \qquad if\ Y_R \le 0.008856 \\
&u^* = 13L^*(u' - u'_n) \\
&v^* = 13L^*(v' - v'_n)
\end{aligned}
\tag{18}
$$

5.2.10. CIELAB Color Space

The third attempt to approximate a uniform color space resulted in the CIE 1976 (L*a*b*) color space which is officially abbreviated as CIELAB. This color space consists of three mutually perpendicular opponent color axes denominated as L*, a*, and b* as shown in Figure 37. Equation (19) shows the component axes for the CIELAB color space:

$$
\begin{aligned}
&L^* = 116\left(Y/Y_n \right)^{1/3} - 16 \qquad if \quad Y/Y_n > 0.008856 \\
&L^* 903.3\left(Y/Y_n \right) \qquad if \quad Y/Y_n \leq 0.008856 \\
&a^* = 500\left[\left(\frac{X}{X_n} \right)^{1/3} - \left(\frac{Y}{Y_n} \right)^{1/3} \right] \\
&b^* = 200\left[\left(\frac{Y}{Y_n} \right)^{1/3} - \left(\frac{Z}{Z_n} \right)^{1/3} \right]
\end{aligned}
\tag{19}
$$

Where Y_n is the tristimulus value of a specified white object (usually the perfect reflecting diffuser so that Y_n= 100).

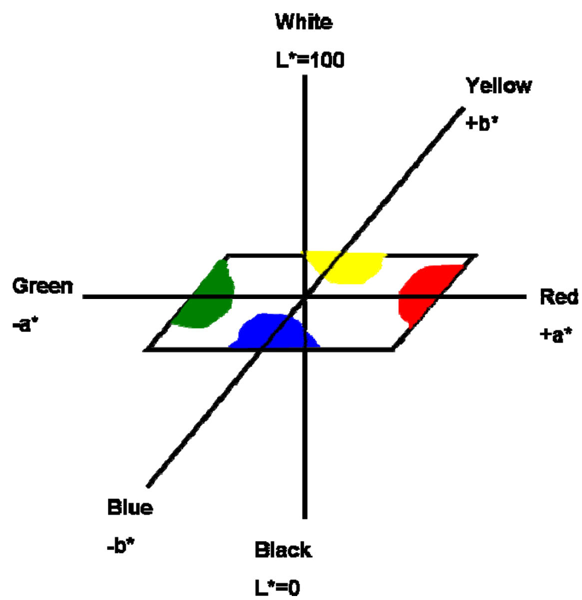


Figure 37. Representation of the CIELAB Coordinates.

It is much easier to visualize the structure of a color space if cylindrical coordinates (polar coordinates) are used. In 1976 the CIE defined the metric correlates of chroma and hue for the CIELUV color space [19]. These metrics are defined in equation (20):

$$C^*_{uv} = \left(u^{*2} + v^{*2}\right)^{1/2}$$
$$h_{uv} = \arctan\left(\frac{v^*}{u^*}\right) \tag{20}$$

In the CIELAB color space the metric correlates of chroma and hue are defined in equation (21):

$$C^*_{ab} = \left(a^*{}_2 + b^*{}_2\right)^{1/2}$$
$$h_{ab} = \arctan\left(a^*/b^*\right) \qquad (21)$$

The correlates of hue in both color spaces are expressed in degrees measured counter clockwise from the positive a*. The correlates of chroma in both color spaces are such that C*=0 represents an achromatic color and C*>0 denotes chromatic colors [19].

Despite all the efforts to develop a perceptually uniform color space, CIELAB is still not a fully uniform system [15]. In a uniform model all equi-chroma contours would be represented by full circles and all equi-hue lines would be straight. Figure 38, however, shows that the CIELAB system is still non-uniform. Kuehni suggests that for practical purposes, a uniform color space should be developed employing the most up to date knowledge of human color vision physiology to calculate threshold level color differences against a standard surround [15].

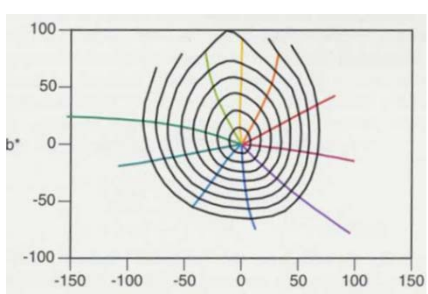


Figure 38. Schematic non-uniformity of the CIELAB color space.

5.3. Color Difference Formulas

Color is an integral component of many industries and is one of the main decisive factors in the selection of a product.

Manufactures need to control the production and delivery of colored products. This requires following protocols that ensure production remains within tolerance threshold. Quality control of color is commonly achieved via visual assessment often with verification through color measurement.

As a first approach to ensure color quality, it is important to recognize that there are two types of visual assessments: perceptibility and acceptability [65]. Industrial determination of quality is often based on acceptability judgments to determine commercial acceptability of the products [91]. On the other hand, perceptibility assessments are based on psychophysical techniques in which the intention is to scale the magnitude of the perceived color difference [91]. Ideally, the perceptibility data after conversion to visual color differences should approximate color differences derived from various mathematical models. Data used to develop color difference models usually includes perceptibility experimental data because acceptability data could introduce industrial bias [91]. More than 40 formulas have been developed since 1936. These formulas aim to represent the color difference between two samples with a single number, denoted with ΔE. Researchers have continuously attempted to improve the correlation of visual experimental data with calculated color differences.

In that regard, the development of color difference formulas can be divided into three time periods: pre 1976, during 1976, and post 1976 [91]. Some of the most important developments will be described in the following sections.

5.3.1. Tolerance

Before any further discussion related to color difference formulas, a clear definition of tolerance is warranted. Color tolerance is the permissible color difference between a standard and a batch [92]. There are two types of color difference tolerances: acceptability and perceptibility. Usually acceptability tolerances are obtained using pass/fail visual comparison of the pair to an anchor gray pair. In a perfect Euclidean space, the tolerance volume thus obtained would be spherical since the magnitude of color difference would not be directionally dependant. However, experimental studies have shown that tolerance volumes are ellipsoids centered about the standard. Ellipsoids also change in shape and size dependent on their position on the color space [19]. Experimental data obtained from visual assessment of tolerance volumes have been used to improve the performance of color difference formulas. Extensive research has been conducted in recent years to study the tolerance ellipsoids, the nature of the datasets (acceptability, perceptibility, combined) and the experimental conditions that have been shown to affect results [37, 38, 51, 53, 55, 57, 59, 93-97].

5.3.2. CIELAB and CIELUV Color Difference Formula

The color difference between two colored samples can be calculated using a number of methods. The CIELAB color difference is based on the Pythagorean theorem for calculating the distance between two points in a Euclidean space [73] as shown in equation (22).

These coordinates can be converted into polar coordinates as shown in equation (23).

$$\Delta E^*_{ab} = \left(\Delta L^{*2} + \Delta a^{*2} + \Delta b^{*2}\right)^{1/2} \tag{22}$$

Where:

$\Delta L^* = L_2^* - L_1^*$
$\Delta a^* = a_2^* - a_1^*$
$\Delta b^* = b_2^* - b_1^*$
2 represents batch and 1 represents the standard sample.
This arrangment is often used in all color difference equations.

$$\Delta E^*_{ab} = \left(\Delta L^{*2} + \Delta C^{*2}_{ab} + \Delta H^{*2}_{ab}\right)^{1/2} \tag{23}$$

$$\Delta H^*_{ab} = 2\left(C^*_{ab,1}\, C^*_{ab,2}\right)^{1/2} \sin\left[\left(h_{ab,2} - h_{ab,1}\right)/2\right]$$

$$\Delta H^*_{ab} = \left(\Delta E^{*2} - \Delta L^{*2}_{ab} - \Delta C^{*2}_{ab}\right)^{1/2}$$

As can be seen ΔH^*_{ab} requires an indirect calculation method [91]. The second method shown above does not produce a negative sign and therefore is not recommended. CIE recommends a number of methods for the calculation of ΔH^*_{ab}. For small color differences away from the achromatic axis ΔH^*_{ab} may be calculated using equation (24) [66].

$$\Delta H^*_{ab} = \left(C^*_{ab,2}\, C^*_{ab,1}\right)^{1/2} \Delta h_{ab} \tag{24}$$

Despite the fact that the CIELAB formula was recommended by the CIE, the need for additional modification to account for various parametric factors was acknowledged [98].

The CIELUV color difference formula is derived from the CIELUV color space. The color difference can be calculated using the Euclidean distance between the points represented in the space as shown in equation (25):

$$\begin{aligned} &\Delta E^*_{uv} = \left[(\Delta L^*)^2 + (\Delta u^*)^2 + (\Delta v^*)^2\right]^{1/2} \\ &or \\ &\Delta E^*_{uv} = \left[(\Delta L^*)^2 + (\Delta C^*_{uv})^2 + (\Delta H^*_{uv})^2\right]^{1/2} \\ &where: \\ &\Delta H^*_{uv} = 2\left(C^*_{uv,1}\, C^*_{uv,2}\right)^{1/2} \times \sin\left[\left(\frac{h_{uv,2} - h_{uv,1}}{2}\right)\right] \end{aligned} \tag{25}$$

Alternative methods to calculate $\Delta H^*_{uv,}$ may be found elsewhere[66].

5.3.3. JPC79 Color Difference Formula

Many sets of experimental data on color discrimination have been obtained since 1976. Most of the studies have been conducted using large surface color samples viewed under typical industrial conditions for visual assessment of color [91].

One of the most important datasets was developed by McDonald [91, 99]. He produced over 600 polyester thread samples surrounding 55 color centers. The assessment was carried by a group of 8 expert observers using a pass/fail judgment [99]. In addition, a larger set of data using 8000 pairs around 600 color centers was obtained by one single observer [91]. These visual results were used to derive the JPC79 formula shown in equation (26) [19]:

$$\Delta E_{JPC79} = \left[\left(\frac{\Delta L}{S_L} \right)^2 + \left(\frac{\Delta C}{S_C} \right)^2 + \left(\frac{\Delta H}{S_H} \right)^2 \right]^{\frac{1}{2}} \tag{26}$$

Where:

$$S_L = \frac{0.08195 L*_{ab,1}}{1 + 0.01765 L*_{ab,1}}$$
$$S_C = 0.638 + \frac{0.0638 C*_{ab,1}}{1 + 0.0131 C*_{ab,1}}$$
$$S_H = S_C + T$$
$$T = 1 \; if \; C_S < 0.638$$
$$T = k_1 + \left| k_2 \cos(H_S + k_3) \right| \quad if \; C_S \geq 0.638$$
$$k_1 = 0.36, k_2 = 0.4, k_3 = 35 \quad if \quad H_S \leq 164 \; or \; H_S \geq 345$$
$$k_1 = 0.56, k_2 = 0.2, k_3 = 168 \quad if \quad 164 < H_S < 345$$

The JPC79 formula has two major problems. The first problem is the breakdown in the evaluation of lightness differences of dark samples. The second problem involves a chromatic discontinuity of fairly neutral samples which was introduced by the definition of *T* [19, 91].

5.3.4. CMC (l:c) Color Difference Formula

The CMC (l:c) formula is based on the CIELAB color space and was recommended by the Color Measurement Committee of the Society of Dyers and Colourists [65]. This formula is a modified version of the JPC79 formula that includes correction for the problems related to the evaluation of dark as well as near neutral colors [65]. In addition, it was noticed that the lightness weighting was not very strict for non textile materials. These small modifications to the JPC79 resulted in a new formula known as the $CMC_{(l:c)}$ which is defined in equation (27):

$$\Delta E_{CMC} = \left[\left(\frac{\Delta L}{l S_L} \right)^2 + \left(\frac{\Delta C}{c S_C} \right)^2 + \left(\frac{\Delta H}{S_H} \right)^2 \right]^{1/2} \qquad (27)$$

Where:

$$S_L = \frac{0.040975 L_{ab,1}}{1 + 0.01765 L_{ab,1}} \quad unless\ L^*_{ab,1} < 16\ when\ S_L = 0.511$$

$$S_C = 0.638 + \frac{0.0638 C_{ab,1}}{1 + 0.0131 C_{ab.1}}$$

$$S_H = S_C (Tf + 1 - f)$$

$$f = \left[\frac{{C^*_{ab,1}}^4}{\left(C^*_{ab,1} + 1900\right)} \right]^{1/2}$$

$$T = k_1 + \left| k_2 \cos\left(h_{ab,1} + k_3\right) \right|$$

$$k_1 = 0.36, k_2 = 0.4, k_3 = 35 \quad if \quad H_S \le 164\ or\ H_S \ge 345$$

$$k_1 = 0.56, k_2 = 0.2, k_3 = 168 \quad if \quad 164 < H_S < 345$$

The parametric factors (k_L, k_C, and k_H) are the weighting functions used to adjust the importance of lightness, chroma and hue attributes in the equation. For perceptibility assessments the recommended ratio of k_L:k_C weighting factors (in this case l:c) is 1:1. On the other hand, for acceptability assessments, a 2:1 (l:c) ratio is used. The CMC does not provide a variable weighting factor for k_H which is always set to one.

The CMC formula is used for the calculation of small color differences and is recommended by the ISO and the AATCC for assessment of color difference of textile samples [7, 100]. One of the problems related to this formula is that when the standard and batch in a pair are interchanged the resulting color difference values may not be identical [91].

5.3.5. BFD (l:c) Color Difference Formula

The BFD (l:c) formula is based on several datasets collected by Luo and Rigg at the University of Bradford[91, 101, 102] The datasets included small to medium color differences from different surfaces (paint, textile, etc). Luo and Rigg developed a perceptibility dataset based on 4000 samples as well as an acceptability dataset using 2000 samples [91]. Results of additional experiments involving color assessment of 500 wool textile sample pairs were also included. Results were combined to generate a common visual scale [101]which was used to formulate the BFD (l:c) color difference formula shown in equation (28).

$$\Delta E_{BFD} = \left[\left(\frac{\Delta L_{BFD}}{l} \right)^2 + \left(\frac{\Delta C_{ab}}{cD_C} \right)^2 + \left(\frac{\Delta H *_{ab}}{D_H} \right)^2 + R_T \left(\frac{\Delta C *_{ab} \, \Delta H *_{ab}}{D_C D_H} \right) \right]^{1/2} \tag{28}$$

Where:

$$L_{BFD} = 54.6\log(Y + 1.5) - 9.6$$
$$D_C = 0.521 + \frac{0.035\overline{C}*_{ab}}{1 + 0.00365\overline{C}*_{ab}}$$
$$D_H = D_C(GT' + 1 - G)$$
$$G = \left[\frac{C*_{ab}{}^4}{(\overline{C}*_{ab} + 14000)}\right]^{1/2}$$

$$T' = 0.627 + 0.055\cos(\overline{h}_{ab} - 254^\circ) - 0.040\cos(2\overline{h}_{ab} - 136^\circ) + 0.070\cos(3\overline{h}_{ab} - 32^\circ)$$
$$+ 0.049\cos(4\overline{h}_{ab} + 114^\circ) - 0.015\cos(5\overline{h}_{ab} - 103^\circ)$$
$$RT = R_C R_H$$
$$R_H = -0.260\cos(\overline{h}_{ab} - 308^\circ) - 0.379\cos(2\overline{h}_{ab} - 160^\circ) - 0.636\cos(3\overline{h}_{ab} + 254^\circ)$$
$$+ 0.226\cos(4\overline{h}_{ab} + 140^\circ) - 0.194\cos(5\overline{h}_{ab} + 280^\circ)$$
$$R_C = \left[\frac{\overline{C}*_{ab}{}^6}{(\overline{C}*_{ab}{}^6 + 7\times 10^7)}\right]^{1/2}$$

$\overline{h}_{ab}$ and $\overline{C}*_{ab}$ denote the mean hue and chroma between a standard and batch. The use of a mean value improved the shortcoming found in CMC. In order to predict acceptability using the BFD equation, the l and c weighting functions are set to 1.5 and 1 respectively; for perceptibility they are both set to 1[101].

5.3.6. CIE94 (K_L:K_C:K_H) Color Difference Formula

The CIE94 formula was developed by Berns *et. al* [93] and included visual assessments using glossy paint samples.

The dataset included 156 individual-color tolerances around 19 color centers. The formula is similar in structure to the CMC (l:c), but uses simpler functions. This formula, shown in equation (29), was recommended by the CIE in 1995[66, 103].

$$\Delta E_{94} = \left[\left(\frac{\Delta L^*}{k_L S_L} \right)^2 + \left(\frac{\Delta C^*_{ab}}{k_c S_C} \right)^2 + \left(\frac{\Delta H^*_{ab}}{K_H S_H} \right)^2 \right]^{1/2} \tag{29}$$

Where:

$$S_L = 1$$
$$S_C = 1 + 0.045 C^*_{ab,1}$$
$$S_H = 1 + 0.015 C^*_{ab,1}$$

The parametric factors were included to account for variation in experimental conditions. This formula is currently obsolete and has been replaced by the CIEDE00 [66].

5.3.7. CIEDE2000 (K_L: K_C: K_H) Color Difference Formula

After the development of the CIE94, there were two color difference formulas that were approved by different organizations: the CMC (l:c) approved by the ISO and the CIE94 approved by the CIE [91]. However, neither one of these formulas matched consistently the experimental results in the blue region and the near achromatic region. The inconsistencies in the existing color difference formulas warranted the need for a new formula. In 1998 a CIE technical committee was created to correct these deficiencies.

The formula created as a result of this study is known as the CIEDE2000 [8, 91] which is shown in equation (30).

$$\Delta E_{00} = \left[\left(\frac{\Delta L'}{k_L S_L} \right)^2 + \left(\frac{\Delta C'}{k_C S_C} \right)^2 + \left(\frac{\Delta H'}{k_H S_H} \right)^2 + R_T \left(\frac{\Delta C'}{k_C S_C} \frac{\Delta H'}{k_H S_H} \right) \right]^{1/2} \tag{30}$$

Where:

$$S_L = 1 + \frac{0.015\left(\overline{L}'-50\right)^2}{\sqrt{20 + \left(\overline{L}'-50\right)^2}}$$

$$S_C = 1 + 0.045\overline{C}'$$

$$S_H = 1 + 0.015\overline{C}'T$$

$$T = 1 - 0.17\cos\left(\overline{h}'-30^\circ\right) + 0.24\cos\left(2\overline{h}'\right) + 0.32\cos\left(3\overline{h}'+6^\circ\right) - 20\cos\left(4\overline{h}' -63^\circ\right)$$

$$R_T = -\sin\left(2\Delta\theta\right)R_C$$

$$\Delta\theta = 30\exp\left\{-\left[\frac{\left(\overline{h}'-275^\circ\right)}{25}\right]^2\right\}$$

$$R_C = 2\sqrt{\frac{\overline{C}'^7}{\overline{C}'^7 + 25^7}}$$

$$L' = L*$$

$$a' = (1+G)*a*$$

$$b' = b*$$

$$C' = \sqrt{a'^2 + b'^2}$$

$$h' = \tan^{-1}\left(\frac{b'}{a'}\right)$$

$$G = 0.5\left(1 - \sqrt{\frac{\overline{C}*^7}{\overline{C}*^7 + 25^7}}\right)$$

The CIEDE2000 included five corrections to the CIELAB: a lightness weighting function represented by S_L, a chroma weighting function denoted by S_C, a hue weighting function S_H, an interactive term, RT, between hue and chroma that improves the performance for blue colors, and a factor, 1+G, that improves the performance for colors close to the achromatic point. This formula was developed to match four color discrimination datasets [8, 91]. Results of PF/3 for different datasets including the combined dataset used for the development, show that CIEDE2000 performs either best or second best with an average performance of 67.4% [8].

5.3.8. Performance of Color Difference Equations

The introduction of the latest formula by the CIE highlighted important issues: such as whether the CIEDE2000 will be the last recommendation and does it provide a significant improvement over existing models to justify a change [9]. A number of independent studies have been carried out to determine the performance of the existing CIELAB-based color difference formulas and the relative importance of the terms included [47, 48, 59, 104]. It has been shown that, compared to CIEDE2000, the CMC $_{(2:1)}$ formula provides an improved performance for the prediction of color differences of textiles [48, 59]. Some of the conclusions also point that models based on CIELAB color space may not be suitable for the development of future color difference equations [9, 91, 104].

It should be pointed out that a completely accurate prediction of color differences may never be modeled; however, further modification of the existing models to provide consistently more reliable results correlating with 80-85% of visual assessments would be highly desirable. This would in turn facilitate color control and communication throughout the color supply chain [9].

6. Statistical Methods for Evaluating Color Difference Formulas and Color Difference Assessments

In an ideal situation, a reliable color difference equation should predict values, ΔE, that correlate to perceived color difference, ΔV (visual difference), producing a straight line between the sets of values. However, in practice this is not the case. Correlation of color difference formulas and visual differences produces scatter plots [91]. The evaluation of color difference models does not employ regular statistics because, in practice, the required assumptions do not apply [91].

In this section, some of the most used methods in evaluation of color difference models will be discussed. In addition, regular statistical methods for the analysis of variance of observer variability are reviewed.

6.1. Statistical Methods Used in Color Difference Evaluation

6.1.1. Correlation Coefficient

The correlation coefficient (r), shown in equation (31), was used to test the performance of color difference formulas based on acceptability or perceptibility data [91]. A perfect agreement between the two datasets results in $r = 1$.

$$r = \frac{N\sum(X_iY_i) - \sum X_i \sum Y_i}{\sqrt{\left[N\sum X_i^2 - \left(\sum X_i\right)^2\right]\left[N\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}} \tag{31}$$

Where X_i and Y_i are values measured by the instrument and those obtained from the visual assessments (% A of ΔV) respectively for sample *i*. N is the total number of sample pairs.

6.1.2. Wrong Decision Measurement

This method was developed by McLaren and has been used to test the performance of the color difference models based on acceptability data [91]. If the assessment of a pair of samples results in more than 50% acceptance, it is considered to be a visual pass. Samples with less than 50% are considered visual fail. Pass/fail visual decisions corresponding with above/under a given DE are counted as *wrong decisions* (WD). A small number of WD indicates a good performance for a formula. On the other hand, a large number of WD indicates poor color difference performance [91].

6.1.3. Performance Factor (PF/4 and PF/3)

The performance factor (PF) was developed by Luo and Rigg [101] to compare two sets of data. PF shown in Equation (32) was designed to minimize problems associated with using different statistical methods to report summaries. PF combined four weighted measures into a single value.

$$PF = 100\,(\gamma + V_{AB} + CV/100 - r) \tag{32}$$

Where r represents the correlation coefficient, CV and γ were proposed by Coates et. al [56] and V_{AB} is a statistical method developed by Schultze [105], as shown in equations (33-38), respectively.

$$r = \frac{N\sum(X_iY_i) - \sum X_i \sum Y_i}{\sqrt{\left[N\sum X_i^2 - \left(\sum X_i\right)^2\right]\left[N\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}} \tag{33}$$

$$CV = \frac{\sqrt{\frac{1}{N}\sum\left(X_i - fY_i\right)^2}}{\overline{X}} \times 100 \tag{34}$$

and

$$f = \frac{\sum X_iY_i}{\sum Y_i^2} \tag{35}$$

$$\log_{10}(\gamma) = \sqrt{\frac{1}{n}\sum\left(\log_{10}\left(\frac{X_i}{Y_i}\right) - \log_{10}\left(\frac{X_i}{Y_i}\right)\right)^2} \tag{36}$$

$$V_{AB} = \sqrt{\frac{1}{N}\sum\frac{\left[X_i - (FY_i)\right]^2}{X_iFY_i}} \tag{37}$$

and

$$F = \sqrt{\sum \frac{X_i}{Y_i} \Big/ \sum \frac{Y_i}{Xi}} \tag{38}$$

Where N is the number of sample pairs, and X_i and Y_i are values for sample pair *i*.

However, Guan and Luo [12] stated that in some cases the correlation coefficient (r) was inconsistent with other statistical methods, and removed this component leaving only the remaining three measures, as shown in Equation (39).

$$PF/3 = 100[(\gamma - 1) + V_{AB} + CV/100)]/3 \tag{39}$$

This modified formula was denoted as PF/3. For a perfect agreement between datasets, X and Y will have the same size; CV =0, V_{AB} =0, γ =1 and PF/3 = 0.

PF/4 and PF/3 have been widely used to calculate observer accuracy and repeatability [12, 13, 38, 43-45, 47, 106, 107], performance of different formulas [12, 13, 38, 43-45, 47-49, 101, 104, 107, 108], and correlation between two datasets [12, 13, 44-46].

6.1.4. VM Index

PF/3 has been widely used to test the performance of color difference formulas; however, one of its main problems is that it cannot indicate the significance of the difference between two datasets. Alman, in a personal communiqué, proposed a new method based upon statistical F test [109]. This method was developed to examine the significance of CIEDE2000. Luo et. al modified the original calculation proposed by Alman and removed the intercept, assuming that ΔE and ΔV both pass through the origin [109].

The calculation involves the following steps and equations (40, 41) are used:

1. Definition of null and alternative hypotheses:

   $H_0 = V_A = V_B$ (There is no difference between the two datasets)

   $H_A = V_A \neq V_B$ (There is significant difference)

2. Calculation of F using $F = V_A/V_B$

3. Rejection of null hypothesis (H_0) if $F < F_C$ or $F > 1/F_C$

$$V_M = \sum \frac{\left(\Delta V_i - a_M \Delta E_{M,i}\right)^2}{\left(N-1\right)} \qquad M \in \{A, B\} \tag{40}$$

and:

$$a_M = \frac{\sum(\Delta E_{M,i} \Delta V_i)}{\sum(\Delta E_{M,i})^2} \tag{41}$$

Where F_C = F(df$_A$, df$_B$, 0.025) is the lower critical value of a two-tailed F distribution with 95% confidence level and df_A and df_B are the degrees of freedom. F_C can be calculated using different statistical software packages or it may be found in statistical textbooks. V_A and V_B represent the residual error variances after scaling correction for models A and B, respectively.

The a_M is the slope between the visual results ΔV and the measured values ΔE for models A and B, respectively. N is the number of samples in the dataset, and $df_A = df_B = (N - 1)$ [109, 110]. The results can be divided into five categories as shown below:

● Model A is significantly better than model B when $F < F_C$

● Model A is significantly poorer than model B when $F > 1/F_C$

● Model A is insignificantly better than model B when $F_C \leq F < 1$

● Model A is insignificantly poorer than model B when $1 < F \leq 1/FC$

● Model A is equal to model B when F=1

6.1.5. Standardized Residual Sum of Squares (*STRESS)*

A new index named Standardized Residual Sum of Squares (*STRESS*) was recently proposed by Garcia et. al as a statistical tool to test the performance of different color-difference formulas. It was developed with the aim to analyze the performance of various formulas and to determine the statistical significance of improvements in newly developed formulas [110]. The new proposed tool is based on Multidimensional Scaling techniques (MDS). STRESS is a loss function commonly used in MDS. In MDS, differences between pairs of objects are given by an appropriate coefficient and are then approximated by distances between corresponding pairs of entities in a visual representation.

The value of such approximation is expressed by the loss function (*STRESS*), which gives the optimal arrangement at its minimum. STRESS is also known as normalized *STRESS* or Kruskal's *STRESS.* The STRESS function is given in equation (42).

$$STRESS = \left(\frac{\sum (\Delta E_i - F_1 \Delta V_i)^2}{\sum F_1^2 \Delta V_i^2} \right)^{1/2}$$
$$F_1 = \frac{\sum \Delta E_1^2}{\sum \Delta E_i \Delta V_i} \qquad (42)$$

Equations (43) and (44) are two alternative forms of *STRESS,* which Garcia *et. al* proved [8] lead to the same results as those obtained from equation (42).

$$STRESS = \left(\frac{\sum (F_2 \Delta E_i - \Delta V_i)^2}{\sum \Delta V_i^2} \right)^{1/2}$$
$$F_2 = \frac{\sum \Delta E_i \Delta V_i}{\sum \Delta E_1^2} = \frac{1}{F_1} \qquad (43)$$

$$STRESS = \left(\frac{\sum (\Delta E_i - F_3 \Delta V_i)^2}{\sum \Delta E_1^2} \right)^{1/2}$$
$$F_3 = \frac{\sum \Delta E_i \Delta V_i}{\sum \Delta V_i^2} \qquad (44)$$

Where F_i (1,3) are factors minimizing STRESS. Notice that $F_2 = a_M$ in the V_M index.

A loss function such as *STRESS* can also be generalized to weigh the contribution of individual pairs of objects leading to weighted normalized *STRESS*. This feature is very useful since it can be used to account for the importance of color difference of individual pairs. The weighted normalized function is shown in equation (45):

$$\textit{Weighted normalized STRESS} = \left(\frac{\sum w_i (\Delta E_i - f \Delta V_i)^2}{\sum w_i \left(f \Delta V_i^2 \right)} \right)^{1/2} \qquad (45)$$

Where w_i indicates the individual weights for each colored pair and the *f* function indicates the use of a scaling factor between ΔV and ΔE.

Some of the key properties of this new statistical tool are that it is symmetric and therefore ΔV and ΔE can be interchanged. In addition the results obtained from STRESS are bound between 0 and 1, and it can be used to perform F-tests with two color difference formulas. Since $F_2=a_M$ it can be deduced that: [110]

$$F = \frac{STRESS_A^2}{STRESS_B^2} \tag{46}$$

6.2. Regular Statistical Methods used for Analysis of Variance

6.2.1. Paired t-test

A paired t-test is used to determine whether the difference between two paired datasets X_i and Y_i of n measured values, differ from each other in a significant way under the assumptions that the paired differences are independent and identically normally distributed. The paired t-test is generally used when measurements are taken from the same population when one or more variable is changed. The t-value is calculated using the equation (47) [111].

$$t = \frac{\bar{d}}{\sqrt{s^2/n}} \tag{47}$$

Where $\bar{d}$ is the mean difference, s^2 is the sample variance and n is the sample size. The t-value can be used to determine the p-value associated with the test. The p-value is the probability of obtaining results at least as extreme as the data observed under a null hypothesis. Generally, the null hypothesis assumes no difference between the datasets. If the p-value is below the threshold selected for statistical significance (usually 0.05 for a 95% confidence level), the null hypothesis is rejected in favor of an alternative hypothesis, which usually states the two datasets are different [111, 112].

6.2.2. Measurement Errors

There are two categories for the sources of measurement errors: systematic errors and random errors. Systematic errors, also known as offsets or bias, are constant and on average represent the offset of an instrument when a true value is being measured. Random errors are those that can be caused by differences among instruments and operators, as well as instability over time, different set ups, and environmental changes. Random errors are difficult to predict. However, their level can be estimated [113].

6.2.2.1. Systematic Errors

Systematic errors can be used to represent accuracy. Accuracy refers to the degree to which measurements X tend to agree with a true value, $\bar{X}_m$, [105, 113] as shown in Equation (48).

$$\text{Accuracy} = \overline{X}_m - X \tag{48}$$

This measure could be used to analyze intra-group as well as observer variation.

6.2.2.2. Random Errors

Random errors are due to parameters that cannot be predicted, however, their level can be estimated [113]. The analysis of error is based on a basic model such as that shown in equation (49) which shows a measured response is composed of a true response and an error.

$$X_m = X + \varepsilon \tag{49}$$

However, it is possible to include the role of various components that contribute to the error. The magnitude of random errors can be determined by their variation but not their average and is as likely to be positive as negative, so that their expected value is zero. It is usually assumed that different sources of error are independent of one another.

6.2.3. Estimating Measurement Error

6.2.3.1. Error Variance

The complexity of a measurement system usually increases the magnitude of the error. The measurement error of a simple system (one operator, one part, one instrument) will be smaller than the ones encountered in complex systems.

However, this complexity can be described by incorporating suitable components within general models and when combined with experimental data the resulting error can be quantified. Equation (50) shows the model for a measurement that takes into account various components such as the effect of part (αi), operator (βj), and the interaction (γij) between operators and parts. Models can be written to account for greater or fewer sources of error [113, 114]. However, to provide a general description of the analysis, the model shown in equation (50) is discussed in detail in the following sections.

$$X_{mij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij} \tag{50}$$

Where:

X_{mij} = measured value of the ith part by the j^{th} operator;

μ = overall part measurement mean;

α_i = random effect due to i^{th} part;

β_j = random effect due to j^{th} operator;

γ_{ij} = random effect due to the interaction of i^{th} part and the j^{th} operator;

ε_{ij} = error not explained by the model.

6.2.3.1.1. Variance Component Analysis

As already mentioned, the magnitude of random errors is shown by the variation rather than the average of errors. Using the measurement model stated in equation (50), it can be assumed that the variance of a sum of different sources of random variables is equal to the sum of the individual variances:

$$X_{mij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij} \rightarrow V(X_{mij}) = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2 \quad (51)$$

Where V(X_{mij}) is the variance of the measured values and σ_x^2 represent the variance of each of the components of the model.

This type of analysis is known as variance component analysis or analysis of variance with random effects [113, 114]. The main goal of variance components analysis is to estimate the amount of variability introduced into the measured values from various factors. This approach also determines the factors that truly affect the magnitude of error. The experimental design used to examine the variance component analysis model shown in equation (51) is called a two-factor crossed design with replication [113].

The total variation among the measured values $V(X_{mij})$ is obtained using Equation (52):

$$\text{SST} = \sum (X_{mij} - \bar{\bar{X}})^2 \quad (52)$$

The abbreviation SST represents "sum of squares" of total variation and $\overline{\overline{X}}$ is the grand average of the results. The summation extends over all the measurements. The sum of squares of total variation can be decomposed into components to estimate the effects of different factors. Table 4 includes the equations required to obtain variance component analysis. Each entry in the column "mean square" is a numeric measure of variability. The variance components model states that this mean square is created by summing an appropriate combination of the variances: $\sigma_{\epsilon}^2, \sigma_{\alpha}^2, \sigma_{\beta}^2, \sigma_{\gamma}^2$. As a result, an algebraic combination of mean squares can be used to estimate the variance of components themselves. A more detailed explanation of the variance components methodology is given elsewhere [113, 114].

Table 4. ANOVA calculations for the two factor cross design with repeat measurements.

Source	**DF**	**Sum of Squares**	**Mean Square**
Parts	k-1	[1] $SSP = nc\sum_{i=1}^{K}\left(\overline{P}_i - \overline{\overline{X}}\right)^2$	[2] $MSP = \frac{SSP}{k-1}$
Operators	c-1	[3] $SSO = nK\sum_{j=1}^{c}\left(\overline{O}_j - \overline{\overline{X}}\right)^2$	[4] $MSO = \frac{SSO}{c-1}$
Interaction	(c-1)(k-1)	[5] SSI = [6] SSM- SSO-SSP	[7] $MSI = \frac{SSI}{(c-1)(k-1)}$
Error	ck(n-1)	SSE= [8] SST- SSM	[9] $MSE = \frac{^{10}SSE}{ck(n-1)}$

Where :

k is the number of parts and *i* represents the part number;

c is the number of operators and j represents the operator number;

n is the number of replications;

[1]SSP is the sum of all squared deviations of the average of the n parts from the grand mean;

[2]MSP is the sum of squares for the parts divided by its degrees of freedom;

[3]SSO is the sum of all squared deviations of the average of the *j* operators from the grand mean;

[4]MSO is the sum of squares for the operators divided by its degrees of freedom;

[5]SSI is the sum of all squared deviations for the interaction;

[6]SSM is the sum of all squared deviations of the average of the n measurements made by the j^{th} operator on the part i^{th};

[7]MSI is the sum of squares for the interaction divided by its degrees of freedom

[8]SST is the sum of squares total variation;

[9]MSE is the sum of squares of the error divided by its degrees of freedom;

[10]SSE is the sum of all squared deviations.

As mentioned earlier, the mean square values can be used to to estimate the variance for each component of the model. Equations (53-56) can be used to solve for the individual variances components [105]. In these equations, the expected value is the long term average.

$$E(MSO) \approx \sigma_{\epsilon}^2 + n\sigma_{\gamma}^2 + nk\sigma_{\beta}^2 \tag{53}$$

$$E\ (MSP) \approx \sigma_{\epsilon}^2 + n\sigma_{\gamma}^2 + nc\sigma_{\alpha}^2 \tag{54}$$

$$E\ (MSI) \approx \sigma_{\epsilon}^2 + n\sigma_{\gamma}^2 \tag{55}$$

$$E\ (MSE) = \sigma_{\epsilon}^2 \tag{56}$$

III. Experimental Methodology and Determination of Parameters Affecting Visual Assessments

1. Introduction

An objective color difference formula that accurately represents average perceptual assessments of observers is a desirable tool for color quality control of various materials, and it is critical for effective electronic communication of colorimetric data for color management in a product supply chain. Existing formulas are based on several different sets of perceptual data that have been established under various experimental conditions, using samples representing a diverse range of substrates and different groups of observers. Currently, the $CMC_{(2:1)}$ color difference formula is the most widely used standardized color difference model, and is incorporated into standards by the International Organization for Standardization (ISO) [115], American Society for Testing and Materials (ASTM) [116], Society for Automotive Engineers (SAE) [117], and the American Association of Textile Chemists and Colorists (AATCC) [7] among others.

In 2001, the International Commission on Illumination (CIE) recommended the CIEDE2000 formula [118]. Luo et al. reported accuracy of prediction for several formulas using average data from a large visual dataset that combined four separate experimental datasets. The combined dataset was used in the development of the CIEDE2000 formula. Using the PF/3 performance metric, a value of 32.6 for the CIEDE2000 formula vs. 37.9 for $CMC_{(1:1)}$ was reported [8].

While the new formula produced an improvement for the combined dataset, it is not known if significant further improvement is possible for either a specific application type or for general applicability. In fact, four subsequent independent field tests of CIEDE2000 vs. $CMC_{(2:1)}$ based on textile samples resulted in a similar level of accuracy for the two formulas [48, 119-121]. However, one recent independent study by our group showed significant improvement in performance for CIEDE2000 compared to CMC for 67 sample pairs around one CIE recommended blue color center [94]. To date, no data have been produced to provide a definitive answer to the disparity between the theoretical performance and the field-tested performance. However, our hypothesis is that the differences are mainly due to a large inter-observer variability, differences in visual assessment protocols, and insufficient fit of the formulas to the mean observer data.

There are many variables that affect the degree of accuracy in assessing color difference formulas. Visual assessment of color, even when conditions are closely controlled, is variable due to the subjective experiences of even well-trained expert assessors. Hence, it is critical to ascertain the level of variability in observers and to maintain highly controlled conditions of observation.

The study reported here is part of a larger project funded by the U.S Department of Commerce, through the National Textile Center, with the primary goal of optimizing color difference accuracy by identifying and minimizing the variables in visual assessment of small color difference of textile materials and establishing the experimental conditions that provide the optimum level of intra- and inter-observer variability. The best experimental conditions (for textile samples) were first established, and highly controlled replication experiments were then performed under identical experimental conditions in different regions of the world (US, Europe, Asia, and Latin America) the results of which are reported in section IV.

2. Assessment of the Type of Observer: Comparison of Naïve and Expert Observers in the Assessment of Small Color Differences

Many studies of visual assessment of color difference involve color normal observers that are inexperienced in any industrial color difference assessment. The experience of observers in the development of visual datasets is potentially significant. Hence, the specific issue addressed in the study reported here is: Are visual color difference data produced by naïve observers on average different from those of experts? Do naïve and expert observers differ in intra- and inter-observer variability in perceptual color difference assessments? Ultimately, is it ok to use naïve observers? The following sections attempt to answer these questions.

For this study, a naïve observer is defined as a color normal observer with no prior experience of commercial pass/fail color difference assessments. An expert is defined as a color normal observer whose employment involves, or has involved, commercial shade matching in the textile industry.

Our hypothesis was that the repetitions of assessment for naïve observers produce results that are not statistically different to those of expert assessors.

2.1. Experimental

2.1.1. Samples

Visual assessments were made using a subset of 31 sample pairs (a standard and a batch) dyed with disperse dyes on unbrightened spun polyester knitted fabric. Figure 39 shows the location of the entire set of colored samples in a CIE a^*b^* plane. Figure 40 shows the appearance of the standards used. In the selection of samples the ratios of ΔL^*_{ab}, ΔC^*_{ab} and ΔH^*_{ab} to ΔE^*_{ab} were considered. Samples were selected such that differences were mainly due to lightness, chroma, hue or a combination of these attributes. Colorimetric data of all samples can be found in appendix A.

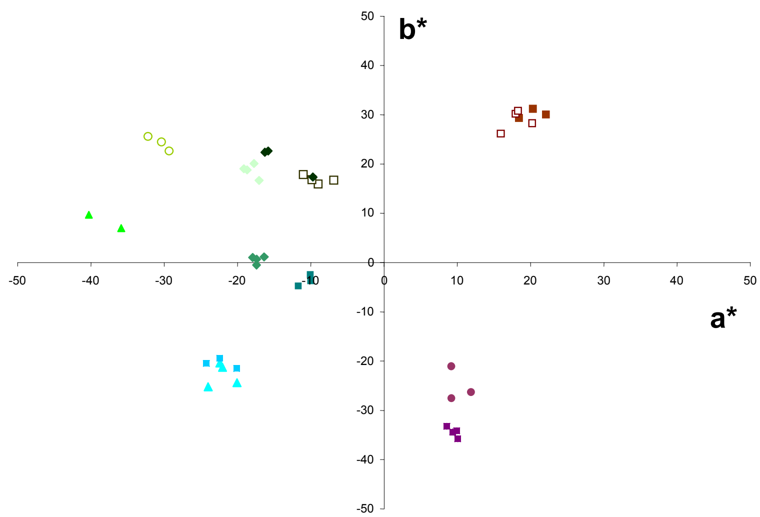


Figure 39. Location of dyed samples in the CIE a^*b^* plane.

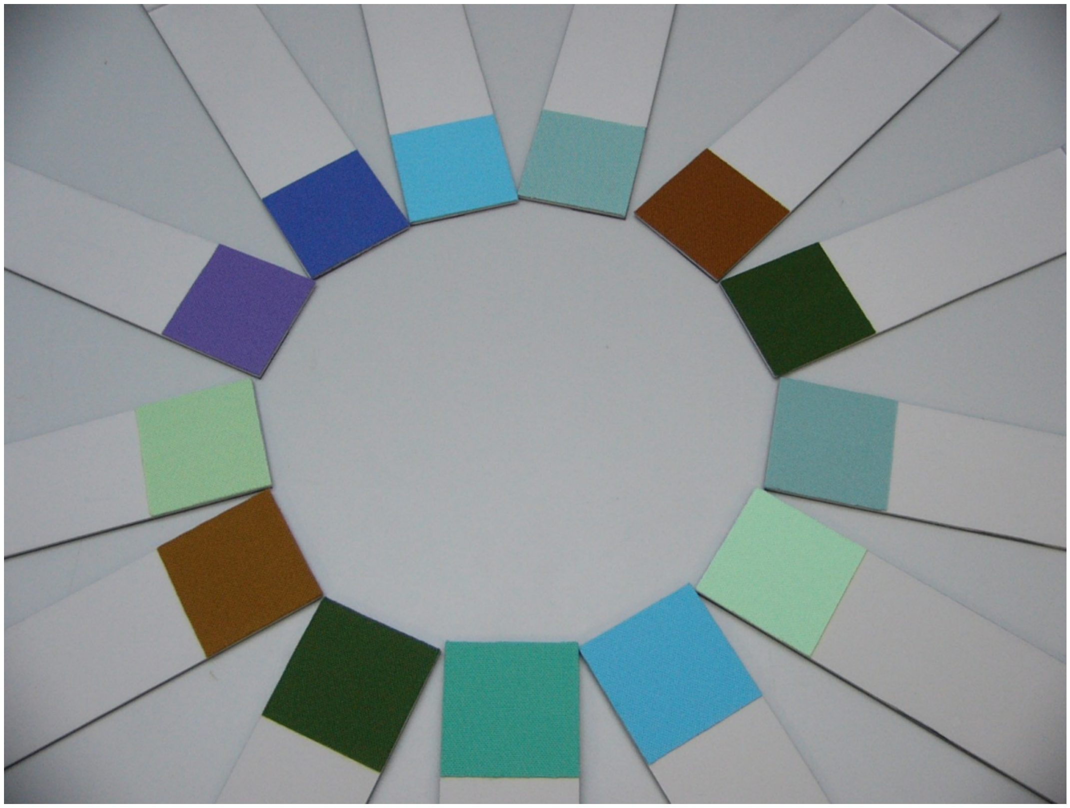


Figure 40. Appearance of color standards used for visual assessments.

Each sample was cut to precise 2” × 2” dimensions and mounted onto custom manufactured plastic holders. The sample mountings used precision cut PVC as a backing, and all the components were uniformly spray painted to a L^* of 74, which is approximately equivalent to Munsell N 7.25. Each sample mounting could slide in a slot on a custom designed display easel. With this setup, sharp dividing lines with no shadows were produced.

The reflectance of all samples was measured using a Datacolor SF600X spectrophotometer with specular component excluded (SPEX) and UV included, and colorimetric data were calculated using the CIE illuminant D_{65} and the CIE 10 degree supplemental standard observer. Each measurement was based on an average of four readings. The 31 sample pairs used in the experiment had an average ΔE^*_{ab} of 2.87, with a range of 0.76-8.34.

2.1.2. Sample Viewing

The samples were illuminated at 45° and viewed perpendicular to the plane of the samples. A custom built 45° easel was specially designed for the visual assessment. The easel was built from PVC with 15"x15" dimensions. It included a front panel that could hold two samples: a standard and a batch. The panel includes railings that allow the colored samples to slide, where samples were divided by a sharp line with no gap as shown in Figure 41. In this setup a standard AATCC Gray Scale was used directly underneath the samples for the assessment of color differences. The easel was located in a Macbeth SpectraLight III viewing booth, illuminated with a filtered tungsten daylight simulating lamp with a correlated color temperature of 6500 ± 200 K and constant illuminance of approximately 1400 lx in the middle of the display board. All extraneous light was eliminated and the illumination conditions were carefully controlled during the experiment in order to minimize variability.

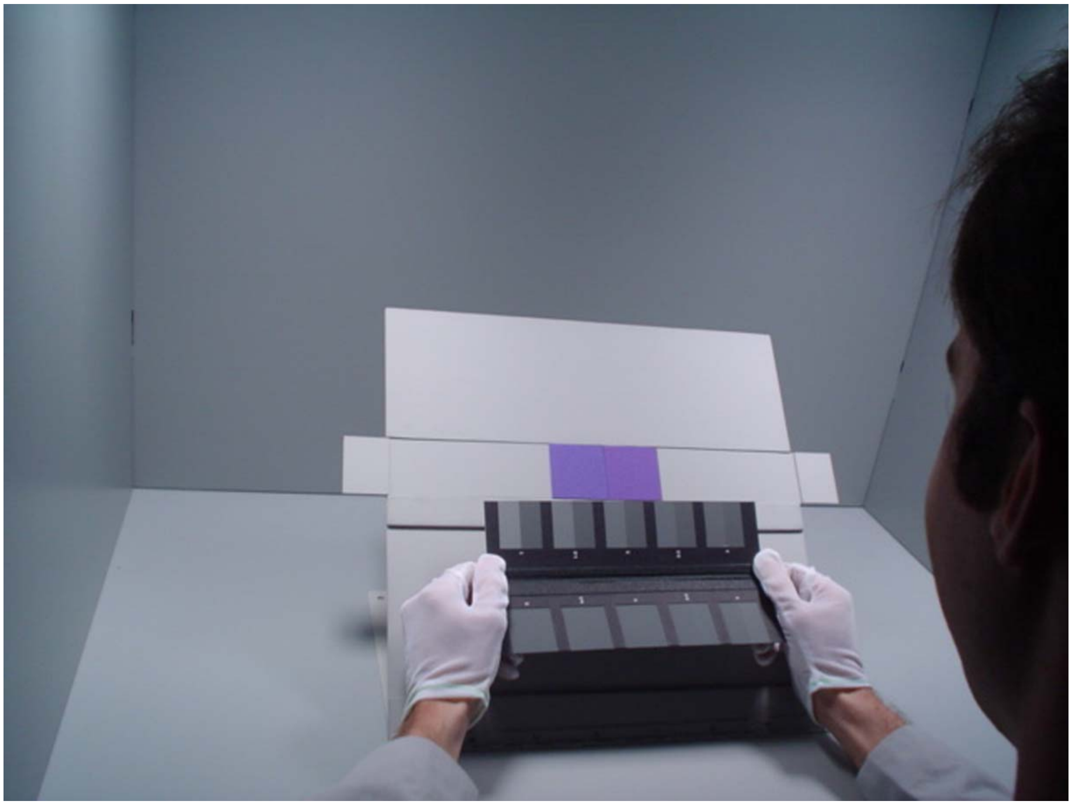


Figure 41. The experimental set up for visual assessment of color difference using an AATCC gray scale.

Each observer wore a mid-gray laboratory coat and a pair of gloves to minimize sample degradation. The samples were placed by the experimenter who also wore a mid-gray laboratory coat. At the beginning of the experiment, the observer was asked to view the empty illuminated viewing booth for 2 minutes to adapt to the light source, during which time the experiment was explained.

2.1.3. Psychophysical Method

An American Association of Textiles Chemists and Colorists (AATCC) Gray Scale for visual assessment of change of shade [119] was used as a guide for assessing the perceptual differences in color. A rating of 5 represents no color difference between samples and the color difference increases geometrically when moving from 5 to 1 in nine steps. For each sample pair the subject was asked: "Which gray scale difference is in closest agreement with the difference between the displayed sample pair? The result can be between two steps, such as 3-4." "A fundamental assumption is made that the total color difference can be so evaluated in terms of an equivalent lightness difference" [62] .

Fifty observers participated in the current experiment; 25 naïve (mostly students of North Carolina State University, tested for normal color vision using the Neitz test [30], of which 11 were females and 14 were males) and 25 expert observers (industrial colorists from the U.S. textile industry, including 10 females and 15 males). Each observer sat in front of the viewing booth, so that he/she could move the reference gray scale freely. Each naïve observer assessed all samples three times on separate days. Due to geographical constraints, each expert observer assessed the sample set once. The same viewing booth, sample sets and sample presentation, were used in each case. The samples were presented in a random order to each observer in each trial.

2.2. Analysis of Data

2.2.1. Gray Scale Transformations to Visual Difference

A total of 3,100 assessments were made using 31 sample pairs. For the analysis, the raw data in grade units, G, were transformed to a visual difference, ΔV, for each pair using Equation (57) [12]. The equation was obtained by fitting a third-degree polynomial equation between measured ΔE^*_{ab} and the AATCC gray scale grades

$$\Delta V = -0.180G^3 + 2.421G^2 - 12.312G + 23.749 \quad (57)$$

The polynomial fitting was evaluated using statistical analysis software (SAS). Initially a fourth-degree was used, however, the results obtained from SAS indicated that the use of a fourth degree polynomial equation to convert gray scale ratings to visual data produces over-fitting. Therefore, a third degree polynomial equation was used for the analysis. SAS results can be found in Appendix B. The average visual response from all observers was calculated for each sample pair. These values were then used to calculate observer repeatability and accuracy. Repeatability refers to the percentage of agreement between assessments for each naïve observer. Accuracy is the percentage of agreement between each observer's assessments for all samples and the average visual response from all observers for all samples in a pair-wise comparison.

In addition, observer assessments were used to compare the performance of naïve and expert assessors.

2.2.2. Comparing Performance of Naïve and Expert Assessors

Paired t-tests were used to evaluate statistical significance of differences between trials conducted by the naïve observers. Also, paired t-tests were used to compare average visual results for each of the three identical naive assessments and the average experts' assessments. These paired t-tests are based on an assumption that the underlying population of measures follows a normal (Gaussian) distribution. Analyses were performed using STATCRUNCH® [120] and SAS® software.

2.2.3. Analysis of Variance

An analysis of error variance was conducted to identify systematic and random errors. For the analysis of each method, the two sources of error were considered separately. A detailed examination of this methodology is given in section II.

2.2.3.1. Systematic Errors: Intra-Group Variability

Systematic errors can be used to represent accuracy. Accuracy refers to the degree to which measurements Y tend to agree with a true value, $\overline{Y_m}$ [113], as shown in Equation 58.

$$Accuracy = \overline{Y}_m - Y \tag{58}$$

Accuracy can also be thought of as intra-group variability. Assuming that the average perceived color difference for each pair is the true value, the intra-group variability is the degree to which each observer's assessment of each pair agrees with the true value of such pair. In other words, intra-group variability represents the degree of deviation of an observer's response compared to the average response of the group. For the analysis, the intra-group variability was calculated separately for each observer in each trial. A measure of intra-group variability is obtained using Equation 59.

$$Intra - group\ sum\ of\ squares = \sum_{i,j,k} \left(DV_{ijk} - \overline{DV}_{i\bullet\bullet} \right)^2 \tag{59}$$

Here the subscript *i* represents the pair number (*i* =1 to 31), *j* represents the observer (*j* =1 to 25), *k* is the trial number (*k* =1 to 3), and K is the number of pairs. In order to obtain an objective assessment, intra-group sum of squares can be converted to standard deviation using Equation 60:

$$Intra - group\ std.\ deviation = \sqrt{\frac{\sum \left(DV_{ijk} - \overline{DV}_{i\bullet\bullet} \right)^2}{K}} \tag{60}$$

$\overline{DV}_{i\bullet\bullet}$ represents the average perceived color difference from all observers for a pair in a particular trial. The intra-group standard deviation was calculated for the naïve and the expert observers using Equation 60. Here Y in equation 58 is defined by DV_{ijk}, which represents the perceived color difference of a pair for an observer in a particular trial and $\overline{Y}_m$ is determined by $\overline{DV}_{i\bullet\bullet}$.

2.2.3.2. Random Errors: Variance Component Analysis

Random errors are due to parameters that cannot be predicted, and in this case include observer variation, the effect of each trial, and the effect of the method used. To assess the role of random errors the analysis of variance with random effects, also known as variance components analysis, was used. The two factors used in the design of this study were: the observers and the pairs. The effect of observers is treated as a random effect. That is, the observers used here are only a sample of all possible observers. As a result our interest was not in the particular value they gave but instead the amount of variability they introduced into the results. The variance components analysis is used to estimate the amount of variability introduced into the measures from various factors.

The total variation among the visual differences DV is obtained using Equation 61:

$$SST = \sum (DV_{ijk} - \overline{\overline{DV}})^2 \qquad (61)$$

The abbreviation SST represents "sum of squares total" variation and $\overline{\overline{DV}}$ is the grand average of the results. The summation extends over 2325 assessments made by naïve observers. The sum of squares total can be decomposed into various components to estimate their effects, in this case: pairs, observers as well as their interaction. Table 5 shows the variance components as well as their definition used in this analysis. Each entry in the column "mean square" is a numeric measure of variability. The variance components model states that this mean square is created by summing an appropriate combination of the variances: σ^2_{pair}, $\sigma^2_{observer}$, $\sigma^2_{interaction}$, and σ^2_{error}. As a result, an algebraic combination of mean squares can be used to estimate the variance of components themselves. A more detailed explanation of the variance components methodology is given elsewhere [113, 114].

Table 5. Variance components analysis.

Source	DF	Sum of Squares	Mean Square
Pair	K-1	[1]SSP = nc $\sum_{i=1}^{k}\left(\overline{DV}_i - \overline{\overline{DV}}\right)^2$	[2] $MSP = \frac{SSP}{K-1}$
Observer	c-1	[3]SSO = nK $\sum_{j=1}^{c}\left(\overline{DV}_j - \overline{\overline{DV}}\right)^2$	[4] $MSO = \frac{SSO}{c-1}$
Pair × Observer	(c-1)(K-1)	[5]SSI = [6]SSM- SSO-SSP	[7] $MSI = \frac{SSI}{(c-1)(K-1)}$
Error	Ck(n-1)	[8]SST- SSM	[9] $MSMet = \frac{^{10}SSMet}{cK(n-1)}$

K is the number of sample pairs and i represents the pair number; i = 1-31;

c is the number of observers and j represents the observer number; j = 1-25;

n is the number of trials;

[1]SSP is the sum of all squared deviations of the average of the n pairs from the grand mean;

[2]MSP is the sum of squares for the pairs divided by its degrees of freedom;

[3]SSO is the sum of all squared deviations of the average of the j observers from the grand mean;

[4]MSO is the sum of squares for the observers divided by its degrees of freedom;

[5]SSI is the sum of all squared deviations for the interaction;

[6]SSM is the sum of all squared deviations of the average of the n measurements made by the j^{th} observer on the pair i^{th};

[7]MSI is the sum of squares for the interaction divided by its degrees of freedom

[8]SST is the sum of square total variation;

[9]MSMet is the sum of squares of the visual assessment method divided by its degrees of freedom;

[10]SSMet is the sum of all squared deviations.

As mentioned earlier, the mean square values can be split to estimate the variance for each component of the model. Equations 62-64 were used to solve for the individual variance components:

$$MSO = \sigma^2_{error} + n\sigma^2_{interaction} + nK\sigma^2_{observer} \quad (62)$$

$$MSP = \sigma^2_{error} + n\sigma^2_{interaction} + nc\sigma^2_{pair} \quad (63)$$

$$MSI = \sigma^2_{error} + n\sigma^2_{interaction} \quad (64)$$

It was impractical to repeat assessments for the expert observers. As a result not all of the variance components estimated for the naïve observers could be estimated for the experts. Equations 65 and 66 were used to calculate the variance components for the experts:

$$MSO = \sigma^2_{error} + nK\sigma^2_{observer} \quad (65)$$

$$MSP = \sigma^2_{error} + nc\sigma^2_{pair} \quad (66)$$

2.2.4. Intra-Observer Variability

The individual observer variability among the trials was calculated by subtracting the grand average of assessments made by observer, *j*, from the total sum of squares of the deviation of assessments made by the j^{th} observer on the i^{th} pair for all trials, as shown in Equation 67.

$$Intra-observer\ sum\ of\ squares = \sum_{k=1}^{3}\sum_{i=1}^{31}(DV_j - \overline{\overline{DV_j}})^2 \qquad (67)$$

Intra-observer sum of squares can be converted to standard deviation using Equation 68.

$$Intra-observer\ standard\ deviation = \sqrt{\frac{\sum_{k=1}^{3}\sum_{i=1}^{31}(DV_j - \overline{\overline{DV_j}})^2}{x}} \qquad (68)$$

The subscript *i* represents the pair number (*i* = 1 to 31), *j* denotes the observer (*j* = 1 to 25), *k* (*k* = 1 to 3) is the trial, and *x* is the total number of observations made by an observer.

In addition to the above statistical techniques, PF/3, and STRESS functions were used to analyze and compare the results. A detailed explanation of these methodologies is given in section II. PF/3 and STRESS values were calculated using Excel.

2.3. Results and Discussion

The raw data in gray scale units was transformed to visual differences according to the methodology explained in section 2.2.1. Figure 42 shows the average visual difference results for the three trials carried out by the naïve observers and those for expert observers.

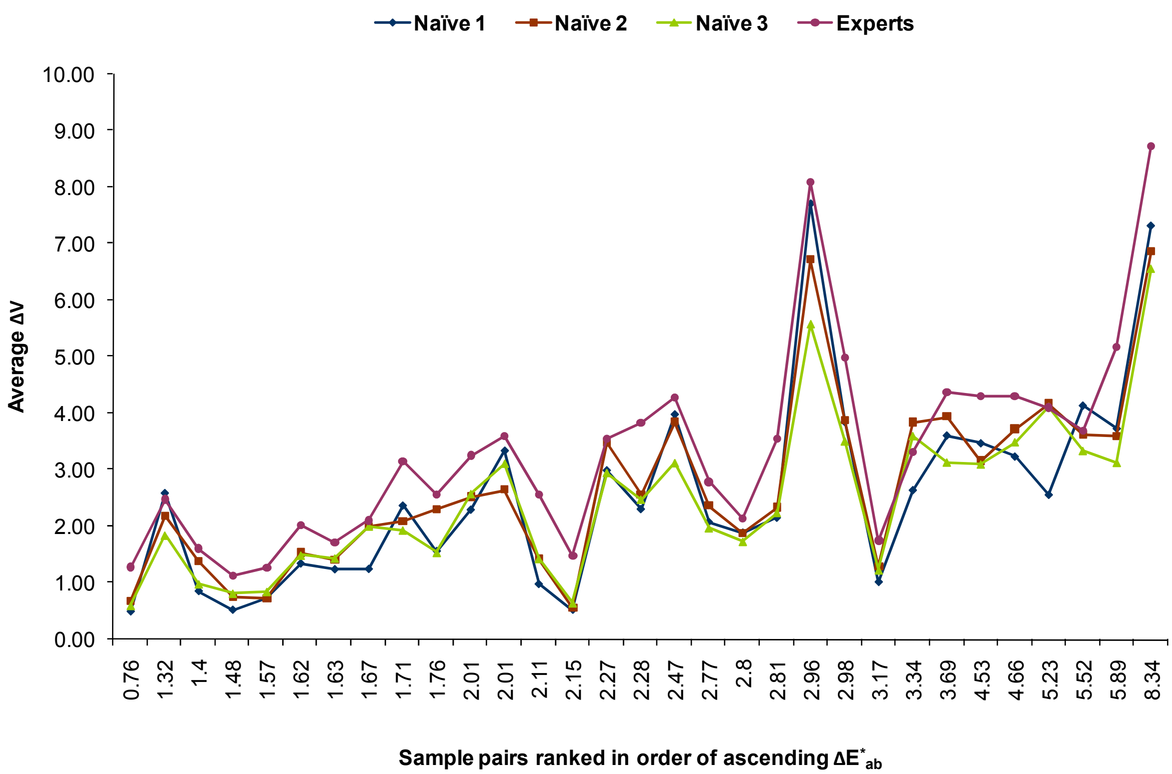


Figure 42. Graph of ΔV for the visual assessments of each sample pair, ranked in ascending order of ΔE^*_{ab}.

A paired two tailed t-test, results of which are summarized in Table 6, was used to evaluate statistical difference between each trial among the naïve observers. Results of a paired t-test between the second and the third trial conducted by naïve observers show that a statistically significant difference exists between the trials. This, however, may be due to the inconsistency of observers' assessments between trials. No statistical difference was found between the first and the second or the first and the third trials.

Table 6. Summary of t-test statistics for assessments carried out by naïve assessors.

Group	**t**	**P**	**Significance**
Trial 1 vs. Trial 2	-1.69	0.1016	No significant difference at α = 0.05
Trial 2 vs. Trial 3	3.94	0.0004	Significant difference at α = 0.05
Trial 1 vs. Trial 3	0.62	0.5428	No significant difference at α = 0.05

Table 7. Summary statistics for the naïve observer trials vs. expert observers.

Group	**t**	**P**	**Significance**
Trial 1 vs. Experts	-9.29	<0.0001	Statistically significant difference at α = 0.05
Trial 2 vs. Experts	-6.64	<0.0001	Statistically significant difference at α = 0.05
Trial 3 vs. Experts	-7.67	<0.0001	Statistically significant difference at α = 0.05

Due to geographic and availability constraints, the expert observers performed the experiment only once. Since both naïve and expert observer groups used the same sample set, i.e. the two sets of data were dependent, paired t-tests were performed to compare the average data for each of the three naïve observer trials and the average experts' assessments.

Table 7 shows the results of this comparison based on paired t-tests which point to a significant difference at a 95% confidence level. The average visual difference for each pair was compared against each trial and against the experts' ratings. The average perceived visual gray scale rating for the expert observers was approximately 9% lower than that for the average of the three trials carried out by the naïve observers, demonstrating that the visual differences among pairs were perceived to be larger by expert assessors. The calculation was carried out using equation 69 where $\overline{GE_i}$ represents the average gray scale rating by expert observers for a particular pair, $\overline{GN_i}$ represents the average gray scale rating by naive observers for a particular pair, and K represents the number of sample pairs.

$$\% = \frac{\sum_{i=1}\left(1-\left(\frac{\overline{GE_i}}{\overline{GN_i}}\right)\right)\times 100}{K} \tag{69}$$

In terms of visual difference, the disparity between naïve and expert assessors can be significantly larger approaching as high as 150% and with an average of approximately 40%. The disparity is due to the arrangement of gray pairs in the current scale. The gray pair arrangement in the standard AATCC is not perceptually linear. The arrangement is geometric and therefore the scale is perceptually a geometric scale.

A geometric scale is one in which the size of the contrast (lightness difference) between gray pairs in the scale approximately doubles as the number of each pair progressively increases. So a 9% difference in gray scale rating could represent a large visual difference in on end of the scale. This suggests that for the samples used in this study, expert observers on average assessed color differences to be perceptually larger compared to naïve observers. It also indicates that the long-term experience of observers in making color difference judgments may play a significant role in visual assessments used for the development of datasets utilized in color difference formulas.

2.3.1. Systematic Errors: Intra-Group Variation (Observer Accuracy)

To determine the variability of individual observers in comparison to the group average in each trial, intra-group variability was calculated. Assuming that the average perceived color difference for each pair in each group represents the true value, the intra-group variability, or accuracy, is the degree to which each observer's assessments of each pair agrees with the true value of such pair in that trial. Table 8 summarizes the mean standard deviation as an indicator of the intra-group variation among observers for each trial. It might be suggested that experienced experts would, as a group, exhibit less intra-group variability than an identically-sized group of observers with no such experience.

Since experts employed in this study represented different sectors of the color industry the higher intra-group variation for experts may be due to an inherent bias resulting from varying professional constraints and requirements when judging color differences, previous knowledge regarding color discrimination, the period of professional experience, and perhaps even the conscious or subconscious psychological pressure to produce test results that show the expert as a 'strict' and highly discerning judge of color difference.

Results of this experiment show that naïve observers on average exhibited smaller intra-group variability, expressed by mean standard deviation, compared to that of the experts.

Table 8. Summary of intra-group standard deviation for naïve and expert observers.

Observer	**Naïve Trial 1**	**Naïve Trial 2**	**Naïve Trial 3**	**Experts**
Mean	1.57	1.67	1.46	1.90
Minimum	0.75	0.65	0.89	0.83
Maximum	3.02	5.57	3.87	3.65
Range	2.27	4.92	2.98	2.82

2.3.2. Random Errors: Variance Component Analysis

2.3.2.1. Variance Component Analysis for Naive Observers

Table 9 includes the variance components analysis for the naïve observers. Each entry in the column is a numeric measure of variability. An algebraic combination of the mean squares, using Equations 62-64, can be used to estimate the variance of components themselves. These estimates are summarized in Table 10.

Table 9. Variance components analysis for naïve observers.

Source	DF	Sum of Squares	Mean Square
Pair	30	5275.33	175.84
Observer	24	1788.10	74.50
Pair × Observer	720	2548.16	3.54
Error	1550	2662.57	1.72

Table 10. Variance components estimates for naïve observers.

Source	Estimated Value
Variance due to pairs	2.30
Variance due to observers	0.76
Variance due to interaction	0.62
Variance due to error	1.72

The highest variance shown in Table 10 is due to pairs. This is expected since the color differences of all the sample pairs presented to the observers ranged from 0.76-8.34 ΔE_{ab}^*. While the variance due to observers is higher than that for the effect of repeated assessments (shown as interaction in Table 10), it is smaller than the variance due to the method, which essentially comprises all other variables in the system. Hence, in addition to the choice of observers, the choice of visual assessment method (including display system, choice of sample pairs, and the number of observers, repeats and samples) should be a primary concern when conducting color difference assessment experiments.

2.3.2.2. Variance Component Analysis for Expert Observers

Table 11 includes the variance components analysis for the expert observers. A similar analysis, using Equations 65 and 66, was employed to estimate the effect of components in overall variance obtained for expert observers, as shown in Table 12. Since the expert observers did not repeat the assessment no term for repeated assessment (i.e. interaction) is shown as in Tables 9 and 10.

Table 11. Variance components analysis for expert observers.

Source	DF	Sum of Squares	Mean Square
Pair	30	2366.24	78.87
Observer	24	902.01	37.58
Error	720	2023.78	2.81

Table 12. Variance components estimates for expert observers.

Variance Type	Estimated Value
Variance due to pairs	3.04
Variance due to observers	1.12
Variance due to error	2.81

For a better comparison between expert observers and naïve observers, the data from the naïve observers was analyzed using the model employed for the expert observers, which does not include the interaction term. Tables 13 and 14 include the results found.

Table 13. Variance components analysis for naive observers.

Source	DF	Sum of Squares	Mean Square
Pair	30	5275.33	175.84
Observer	24	1788.10	74.50
Error	2270	5210.74	2.30

Table 14. Variance components estimates for naive observers.

Variance Type	Estimated Value
Variance due to pairs	2.31
Variance due to observers	0.78
Variance due to error	2.30

The estimated values are higher in all cases for expert observers for the visual assessment methodology employed in this study than naïve observers. However, the observer' variance estimate for both group of observer was not statistically different according to F statistics. On the other hand, variance due to error is statistically significant based on F statistics. The error component contains all the unknown parameters that were not included in the variance components model. These parameters may influence other factors, such as observer, within the model.

2.3.3. Intra-Observer Variability (Observer Consistency)

Equation 68 was used to calculate intra-observer variability, which is the deviation of an observer in a trial involving 31 assessments in comparison with his/her own average results from all trials.

Table 15 summarizes the intra-observer variability found among naïve using the grand average from all trials (93 assessments).

Table 15. Summary of intra-observer variability expressed by standard deviations.

Naïve observer	Grand Average
Mean	0.98
Minimum	0.35
Maximum	2.46
Range	2.11

Ideally all observers would repeat their assessment at different times with a standard deviation of 0. However, studies involving human experience, including color difference perception, usually demonstrate high standard deviation in repeatability among observers. Results obtained in this study indicate that the range of variation in the repetition of the assessment was large. Further comparisons of these results can be found in section 5.

2.3.4. Observer Accuracy using PF/3 Metric

In addition to conventional statistical methods, PF/3 metric was used to assess observer accuracy. PF/3 values were calculated using the grand mean visual results from all observers for each pair compared to each observer's assessments. It was assumed that the grand mean visual results were the true value for each pair. An accurate observer is one that agrees closely with the mean (true) visual results from all observers [12].

PF/3 for accuracy was calculated for the naïve observers in each trial and a mean value was obtained. Table 16 summarizes the results of PF/3 for accuracy.

Table 16. Summary of naïve observers' variation in terms of PF/3 for accuracy.

Observer	**Trial 1**	**Trial 2**	**Trial 3**	**Mean Naïve Trials**	**Experts**
Mean	73.52	67.44	60.20	67.05	60.31
Minimum	35.36	29.58	38.04	34.33	29.63
Maximum	108.56	137.93	104.09	113.17	81.79
Range	73.20	108.35	66.05	78.84	52.16

The data indicate that expert observers have a smaller range of variation compared to all trials using naïve observers. These results approximately reflect those reported in Table 8 for intra-group variation. The mean PF/3 for both sets of observers, regardless of prior knowledge or experience in dealing with such assessments, is higher than those reported by Guan and Luo (PF/3=40) [12][20], Chou et al (PF/3=35) [47], and Kuo and Luo [46] which ranged from 35-40. While these studies used somewhat different experimental methodologies, the reasons for the discrepancy in PF/3 values data are not clear. However, since the experimental method reported here was deliberately controlled to minimize common variables (such as observer sample handling, poor mounting and juxtaposition of samples, etc.) it is unlikely that the method used is a primary factor in producing relatively large PF/3 values. A possible cause of the r the discrepancy in the data could be the inclusion of outliers in this study.

On the other hand, the results obtained in this study are in good agreement with the reported values by Mangine (PF/3=64) [121] and Lee (PF/3=57.6) [94] . Mangine's study employed a different methodology, but the color centers and the dyed substrates used for the construction of samples were the same as those used in this study. In Lee's work a single color center was investigated using the same fabric used in the present study and a similar visual assessment methodology was also employed.

2.3.5. PF/3 Observer Repeatability (Observer Consistency)

In order to assess observer repeatability between trials, the PF/3 measure was calculated for naïve observers and the mean values for the group as well as each observer were obtained. A summary of results is shown in Table 17.

Table 17. Summary of PF/3 of repeatability for naïve observers.

Observer	Trials 1&2	Trials 2&3	Mean
Mean	82.09	63.89	72.99
Minimum	36.07	30.45	33.26
Maximum	176.33	88.79	128.81
Range	140.26	58.35	95.55

The mean PF/3 values for each trial are consistently high and the repeatability among naïve observers is low. However, results show that observers' repeatability improves between trials 2 and 3 compared to that between trials 1 and 2 as evidenced by the smaller PF/3 values.

Similar to PF/3 results reported for accuracy the repeatability PF/3 results obtained in this study are significantly higher than those reported by Chou et al. (PF/3 = 27 and 41) [47]; Cui et al (PF/3=37) [38], Xu et al. and (PF/3=26) [45] and Xin et al[43] (PF/3=22.1). In contrast, the results obtained in this study nearly correspond to those obtained by Mangine (PF/3 = 64.8) [121] and Lee (PF/3 = 66.9)[94]. Using a paired comparison psychophysical test method Mangine et al. [121] reported that repetition of assessments for naïve observers produced more consistent results which she labeled 'observer training effect'. In order to establish whether repeat observations affect perceptual assessments a statistically robust methodology should be employed to ascertain the minimum number of repetitions required whereby an increase in repeating the test does not result in a significant change between assessments. However, the data shown in the experiments reported here do not show a definitive evidence of learning or training. Also, analysis of the literature in relation to small color difference perception shows little, if any, evidence supporting a mechanism leading to a cognitive improvement of the perceptual sensation due to repeated presentation of textile samples with small color differences. Whilst this was not tested, it may be argued, that observer consistency improves with repeated assessments possibly due to a better understanding of the experimental protocols.

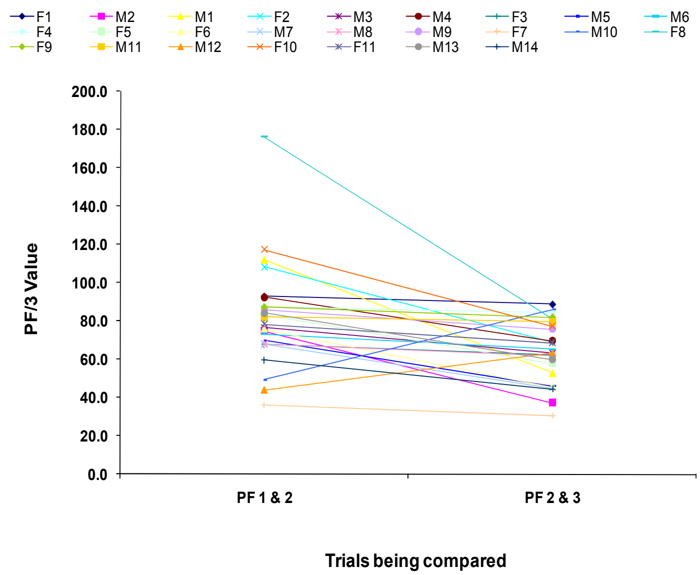


Figure 43. Comparison of the performance of individual observers in terms of PF/3 repeatability among trials.

2.3.6. Performance of Color Difference Formulas against each Visual Dataset

The performance of color difference equations was tested against the averaged data for each of the naïve observer trials as well as the expert observer visual data. Figure 44 shows the consistent reduction in PF/3 with each additional trial for the naïve observers. In the case of trial three, the PF/3 data between naïve observers and experts is comparable. This demonstrates that replications may be required in assessments involving naïve observers; although this could also apply to experts as they were not subjected to multiple assessments.

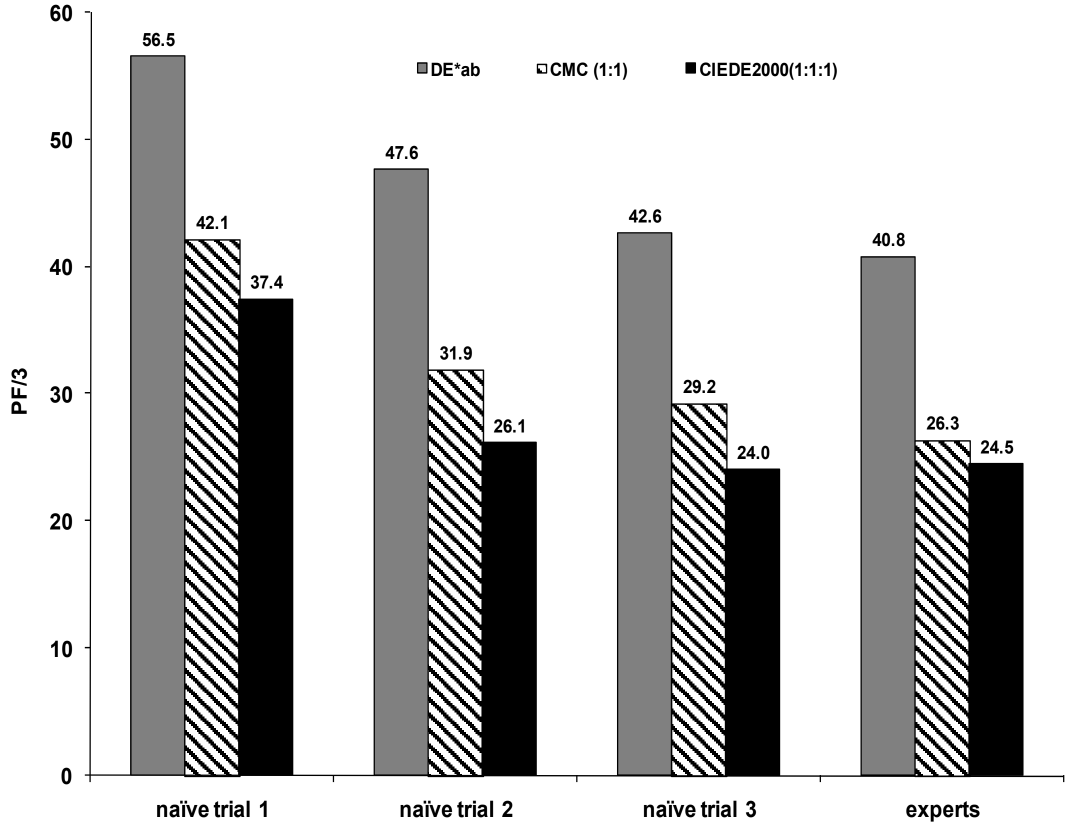


Figure 44. Graph of PF/3 for ΔE^*_{ab}, $CMC_{(1:1)}$, and CIEDE2000 $_{(1:1:1)}$ for the average naïve observers (three trials) and expert observers.

2.3.7. Performance of Color Difference Formulas against each Visual Dataset using the STRESS Function

STRESS function described in section 6.1.5 of chapter II can be used to assess the significance of variation between models as well as evaluate the performance of two independent groups of observers: in this case naïve and expert observers.

Tables 18, 19, 20, and, 21 show the results of the STRESS function and the associated F values for naïve and expert observers when compared against ΔE^*_{ab}, CMC, and CIEDE2000 at K_L or *l* setting of 1 since this methodology tested perceptibility.

Table 18. The STRESS values for color difference models.

Observer	ΔE^*_{ab}	$CMC_{(1:1)}$	$CIEDE2000_{(1:1:1)}$
Avg Naïve	0.37	0.28	0.23
Naïve Trial 1	0.43	0.35	0.30
Naïve Trial 2	0.36	0.27	0.22
Naïve Trial 3	0.34	0.25	0.20
Expert	0.35	0.26	0.22

Table 19. F values between trials against each other based on ΔE^*_{ab}.

Group	F
	ΔE^*_{ab}
Naïve T1 & Naïve T2	1.37
Naïve T2 & Naïve T3	1.18
Naïve T1 & Naïve T3	1.61

Table 20. F values between naïve and expert observers for different equations.

F-value	ΔE^*_{ab}
Avg Naïve/ Expert	1.08
Naïve Trial 1/ Expert	1.45
Naïve Trial 2/Expert	1.06
Naïve Trial 3/Expert	0.90

For the methodology employed in this study the calculated critical F values are $F_C = 0.48$ and $1/F_C = 2.08$ [120]. Table 19 shows that based on ΔE^*_{ab}, no significant difference exists between the responses of each trial within the naïve observers. Table 20 shows that based on, ΔE^*_{ab}, no significant difference exists between observer groups since the calculated F values fall between F_C and $1/F_C$. In addition since the calculated F values are close to unity, responses may be considered fairly equivalent. Table 21 compares the performance of color difference equations for both sets of observer groups against each other.

Table 21. F values between different equations for each set of observers.

	ΔE^*_{ab} /$CMC_{(1:1)}$	$CMC_{(1:1)}$/$CIEDE2000_{(1:1:1)}$	ΔE^*_{ab} /$CIEDE2000_{(1:1:1)}$
Avg Naïve	1.71	1.54	2.64
Experts	1.92	1.37	2.62

As can be seen, no significant difference exists between ΔE^*_{ab} and $CMC_{(1:1)}$ or between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ for the two groups of observers. A significant difference, however, is seen in the comparison of ΔE^*_{ab} versus $CIEDE2000_{(1:1:1)}$ equations and since calculated F values are greater than 1/Fc (2.08), it can be concluded that ΔE^*_{ab} color difference prediction model performs significantly poorer than $CIEDE2000_{(1:1:1)}$ for both sets of observer groups.

2.4. Conclusions

In this study several statistical methods, including, PF/3, t-test and the recently proposed STRESS function were used to compare the inter and intra- observer variability in visual assessment of colored samples for a group of naïve as well as a group of expert color assessors. The objectives were to test whether statistically significant differences exist between naïve and expert observers as well as among replicated trials in the assessment of small color differences.

Results of a paired t-test, shown in Table 6, demonstrate that, for the psychophysical method employed in this study, the repetition of assessments among naïve observers may affect observer performance as evidenced by a statistically significant difference between the second and the third trial assessments. However, the comparison of the visual data from trial to trial fitted into ΔE^*_{ab} color difference model and using the STRESS function shows no significant differences between naïve trials.

In terms of PF/3 for accuracy naïve observer trials show higher values than those of experts as shown in Table 16. Again the mean PF/3 for accuracy for both sets of observers, regardless of prior knowledge or experience in dealing with such assessments, is higher than those reported previously. The mean PF/3 values for repeatability were also higher than those reported in the literature for as yet unknown reasons.

While there could be a number of experts with superior color discrimination ability, as with any observer group, this general observation could be due to experts' better understanding of protocols employed for the assessment of color differences rather than a superior perceptual ability in determining color differences.

Results of the intra-observer variability based on PF/3, however, indicate that the repetition of the assessment on average improves observer repeatability as shown in Table 17. The apparent improvement in repeatability maybe due to learning the procedure required for the assessment of small color differences.

A comparison of average ΔV obtained from each naïve observer trial and that of expert assessors based on paired t-tests showed significant differences between observer groups in all trials at 95% confidence level as demonstrated in Table 7.

Results of variance components analysis to determine the root causes of variability show that for both groups of observers the highest variance is due to sample pairs. This was expected since the range of color differences of sample pairs presented to the observers was large (0.76-8.34 ΔE^*_{ab}). The role of observer variability in the overall variability was not as pronounced as shown in Tables 12 and 14. The variability due to observer component between two observer groups was not statistically significantly different. This indicates that there are probably no inherent differences between individual observers in each group.

In both observer groups the variance due to the method, defined as a combination of all other unknown factors, was higher than that for observers as shown in Tables 12 and 14. Discarding large variance due to sample variability, for both observer groups the highest variance was due to error which included all the unknown parameters. The difference between the error component values according to the F statistics is statistically significant which implies unknown factors affect the visual assessment between both groups of observers.

The performance of ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ color difference equations based on naïve and expert assessors' visual data was evaluated using the PF/3 and STRESS metrics. The PF/3 values obtained for the models in this study were comparable to those reported in the literature. In general, for the controlled psychophysical method used in this study, judgments made by naïve observers on average tracked well those made by expert observers. Experts, however, on average produced a 9% positive bias in gray scale ratings compared to naïve observers. Results also show a consistent reduction in naïve observer values from trial one to three and that by the third trial naïve observer PF/3 values were comparable to those based on expert assessments. This could mean that replications may be required in assessments involving naïve observers although this observation cannot be excluded for expert observers as they did not repeat the assessments.

However, the comparison of the visual data fitted into ΔE^*_{ab} color difference model and using the STRESS function shows no significant differences between naïve and expert observer responses as shown in Tables 19 and 20. Since naïve observers are by definition a form of representation of the general population, providing a randomly selected large group of observers are employed and the observers are provided with an opportunity to become familiar with the methodology employed in visual assessments, their responses can be used in the development of visual datasets and color difference prediction models. This may also avoid the potentially problematic varied commercial color difference experience of experts in their assessment of color.

It can be also concluded that a good working protocol is to use a significantly large number of observers and sample set, a highly controlled visual methodology, and conduct statistical analysis measures on repetitions and test for any significant change in repeatability of assessments.

3. Effect of Variables on Visual Assessment of Small Color Differences

3.1. Visual Assessment using Pair Comparison Method

Visual perception of color is affected by many factors as was previously discussed in chapter II.

The visual judgment of a color difference between a pair of colored objects can be completed using different psychophysical methods and each method may influence the magnitude of judgment obtained [9].

In this part of the study, we evaluated the effectiveness of pair comparison technique as a psychophysical method for the assessment of perceived color difference of textile samples. For these assessments, an anchor gray pair was used as a guide to evaluate the perceptual differences among colored samples.

3.1.1. Experimental

3.1.1.1. Samples

Visual assessments were made using the same set of 31 sample pairs (a standard and a batch) dyed with disperse dyes on unbrightened spun polyester knitted fabric used in section 2.1.1 of chapter III.

3.1.1.2. Sample Viewing

The sample viewing conditions used in this part of the study were the same as the conditions used in section 2.1.2 of chapter III but, with minor modifications as described in 3.1.1.3 below.

Each observer wore a mid-gray laboratory coat and a pair of gray gloves to minimize damaging samples. The samples were placed by the experimenter who also wore a mid-gray laboratory coat. At the beginning of the experiment, the observer was asked to view the empty illuminated viewing booth for 2 minutes to adapt to the light source, during which time the experiment was explained.

3.1.1.3. Psychophysical Method

An anchor gray pair with a color difference of ΔE^*_{ab} 3.9 was used as a guide to assess the perceptual differences among colored samples. The color pairs were the same size as the anchor pair. The standard and sample were placed against one another so that there was no space between them. The color pairs were presented above the anchor neutral difference pair with no space in between. Figure 45 shows the setting for this testing method.

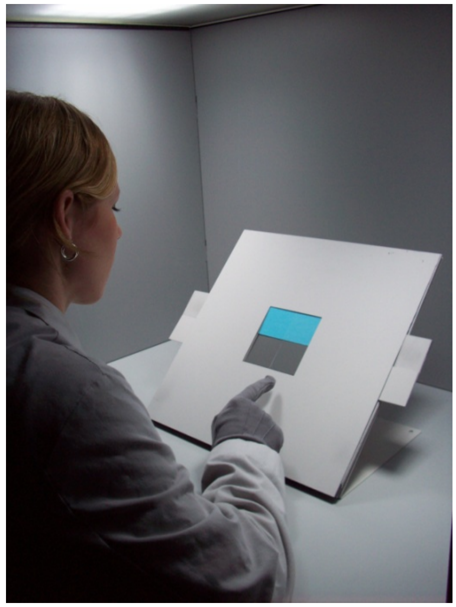


Figure 45. The experimental set up for visual assessment of color difference using the pair comparison method.

Each observer was asked: "Which pair exhibits a larger color difference?" by comparing the standard pair and a sample pair [12]. Each color pair was presented randomly and the judgment was recorded. Each visual assessment session took approximately 20 minutes.

3.1.2. Analysis of Data

3.1.2.1. Transformations to Visual Difference

A total of 2,325 assessments were made using 31 sample pairs. Visual results or ΔV were obtained for each trial of the experiment and for the total number of observations. To estimate ΔV, the visual probability *P,* for each sample pair was calculated first.

Visual probability is the percentage of observations judging the sample pair having a larger color difference than the standard pair. Equation 70 shows the methodology to calculate *P*.

$$P_i = S_i / N_i \qquad (70)$$

Where:

i = Color pair 1 to 31

S_i = Number of observations judging the *ith* color pair to be larger than the standard pair

N_i = Number of observations for the *ith* color pair

The CIE in the 1978 guidelines proposed a relationship between a color difference stimulus and *P* [122]. This relationship was used in Luo and Rigg's study [37]. The visual difference of a colored pair can be estimated using equation 71.

$$P_i = 1/\left[1 + \exp(\alpha - \Delta V_i)\right] \qquad (71)$$

Where α corresponds to a ΔV value having a probability of 50% ($P = 0.5$) and the magnitude depends upon the scale of the standard. Equation 71 can be rearranged as Equation 72 and it allows converting visual probability to a linear color difference scale.

$$\Delta V_i = \alpha - \ln[(1/P_i) - 1] \quad (72)$$

However, when applying equation 73, it is important to notice that samples having a *P* value less than 0.05 or greater than 0.95 have to be discarded due to the sigmoid (S-shaped) relationship between *P* and ΔV[122]. These samples are those that were judged to have larger or smaller color difference by the majority of the observers and they are not practical for the visual data [12]. Therefore, it is important to select colored sample pairs that are close in color difference to the standard.

3.1.2.2. PF/3 Metric

To test the degree of repeatability from each trial, the PF/3 measure, described in detail in section 6.1.3 of chapter II, was used. In addition, PF/3 was used to determine the performance of color difference formulas against each trial in the visual dataset.

3.1.2.3. STRESS Function

In addition to PF/3, the STRESS function proposed by Garcia et. al [110], and described in section 6.1.5 of chapter II, was used to determine the statistical significance of any difference in the performance of various formulas.

3.1.3. Results and Discussion

3.1.3.1. Visual Probability

First the raw data was transformed to visual probability *P*. The visual probability for each pair was compared in each trial. Figure 46 shows *P* for each sample pair in each of the trials.

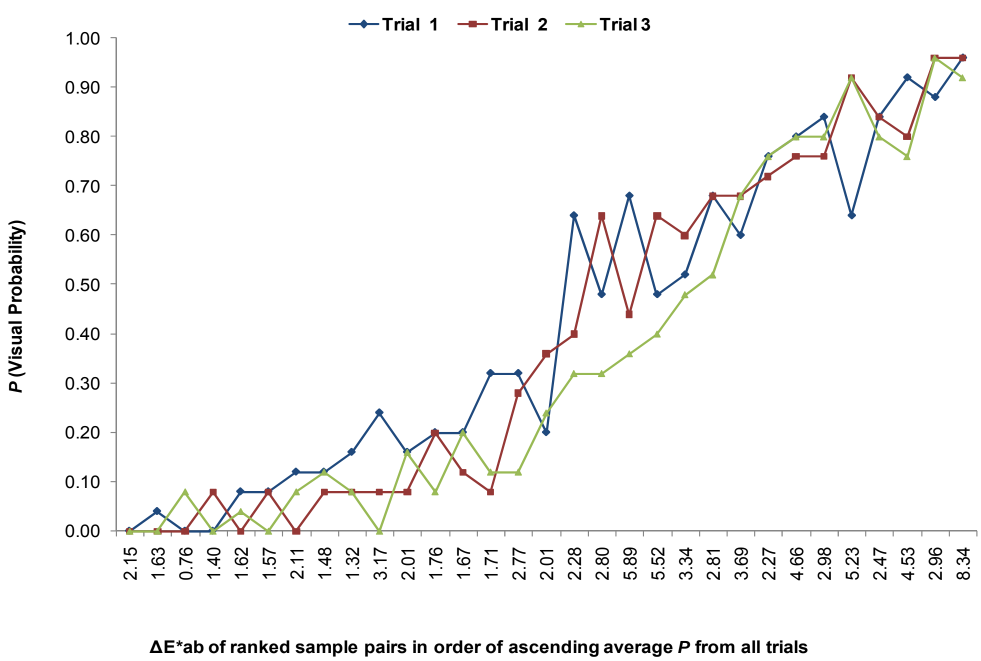


Figure 46. Graph of sample pairs ranked in order of ascending visual probability in each trial.

Figure 46 shows that the visual probability for each pair is different among trials. As previously mentioned, samples with a *P* value smaller than 0.05 or greater than 0.95 should be discarded. It can be seen that the number of samples to be discarded is different in each trial. Figures 47 and 48 and 49 show the correlation between *P* values among three trials. The correlation is higher between trial 2 and 3 indicating that visual probability (*P)* for each of the colored samples was more repeatable than those *P* values obtained between trials 1 and 2 and 1 and 3.

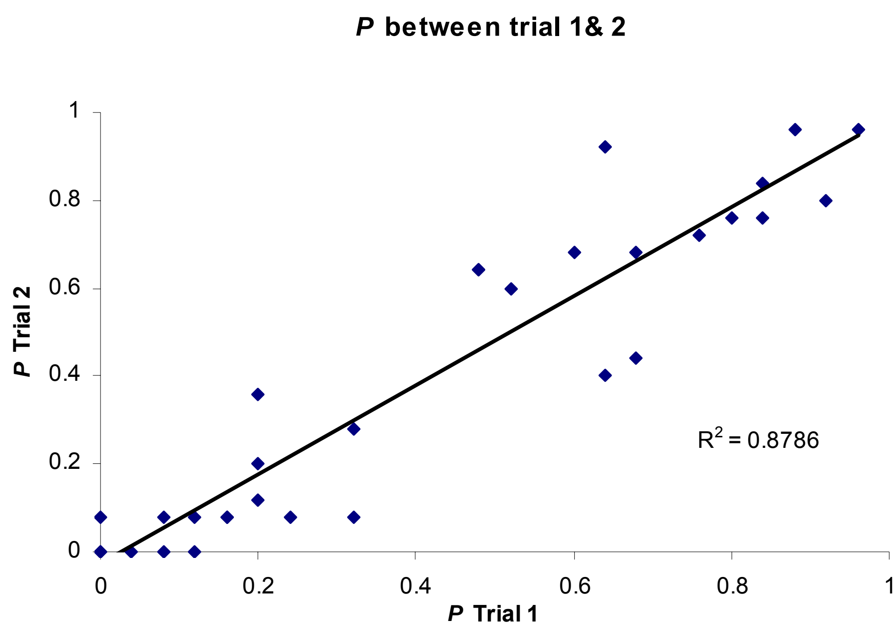


Figure 47. Correlation of visual probability between trials 1 and 2.

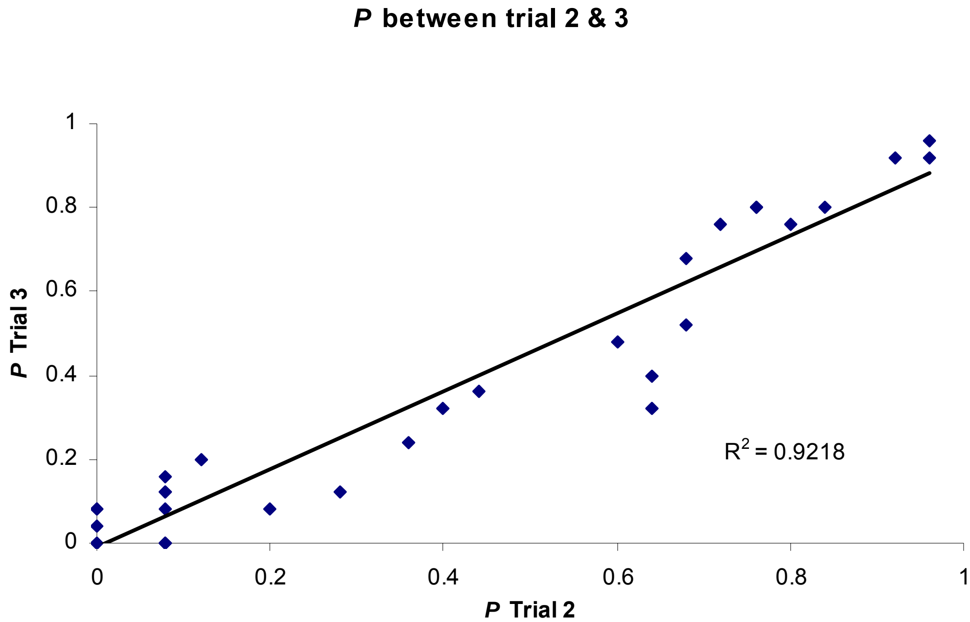


Figure 48. Graph of correlation in visual probability between trials 2 and 3.

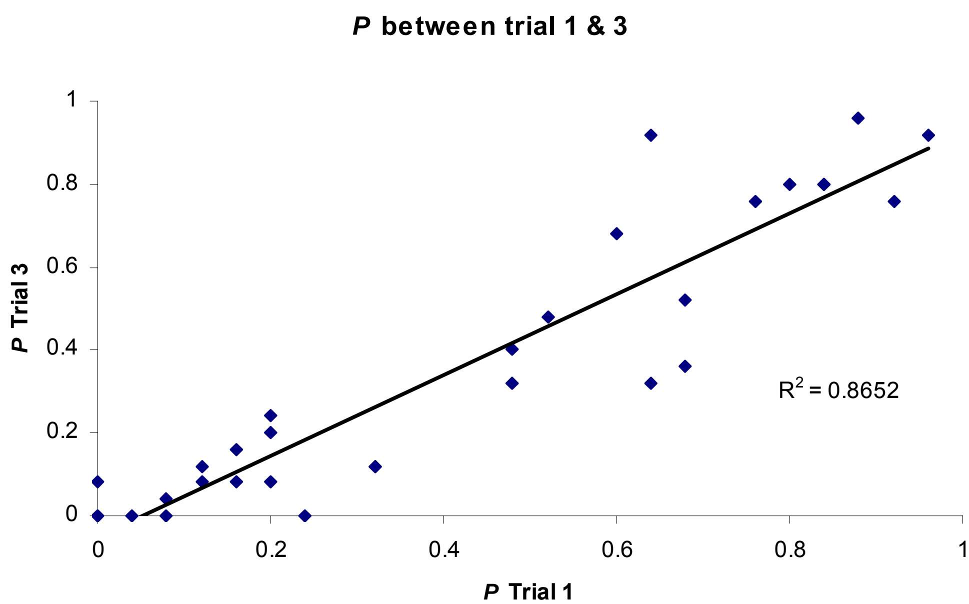


Figure 49. Correlation of visual probability between trials 1 and 3.

3.1.3.2. Visual Probability Converted to Visual Difference ΔV

The CIE guidelines suggest a method of calculation α based on the scale of the standard. It is important to keep in mind that α is a constant and its value depends on a particular dataset. In this study the value of α was selected to be 3.9. Using α =3.9 and Equation 72, the visual difference was calculated for each pair in each trial. The sample pairs having a P value less than 0.05 or greater than 0.95 were discarded.

In addition, all the data was combined for further comparisons with other methods. In trial 1, a total of 5 pairs were discarded. In trials 2 and 3, a total of 7 pairs were discarded in each case. The pairs discarded in each trial were not always the same; only 2 pairs were common in all trials. For more details regarding the samples refer to appendix 2. However, for any meaningful comparisons of the visual data obtained from trials, it was necessary to remove all the discarded pairs for each trial and therefore a total of 10 pairs were removed out of a total of 31 sample pairs. As already stated in order for the pair comparison method to yield useful results the standard should be close to the average color differences of the samples. However, due to the large range of color differences amongst samples in this study, the pair comparison psychophysical assessment method would probably not be a useful metric for the assessment of differences. In addition, the selection of an anchor pair representing the color difference of the majority of samples within the group is crucial and for samples with a wide range of color difference a selection would be unfeasible.

3.1.3.3. Agreement between Trials

To determine the repeatability of the assessments made by the group, the PF/3 measure was used to compare the visual difference obtained for each pair in each trial as shown in Table 22.

Table 22. Comparison of visual data among trials.

Trials Compared	PF/3 Values
Trial 1 & 2	25.57
Trial 2 & 3	22.19
Trial 1 & 3	24.54

A low PF/3 value indicates good agreement between two datasets. In this study, a high agreement was found between trials 2 and 3. These results suggest the possibility of an improvement in the observers' consistency, as a group, when judging small color differences.

3.1.3.4. Performance of Color Difference Formulas against Each Visual Dataset

To test the practical relevance of data obtained using this experimental methodology, the PF/3 of three color difference formulas were tested against the visual data from each trial. In addition, the observations from three trials were combined and compared to each formula in order to determine the performance of combined visual data obtained with this psychophysical method. The data were tested using (l:c) or (K_L, K_c) weighing functions of 1 in CMC and CIEDE2000 respectively since this methodology tested perceptibility.

Figure 50 shows that regardless of the color difference model, the performance of each model decreased with the number of trials. In addition, the graph shows that $CMC_{(1:1)}$ performed better than $CIEDE2000_{(1:1:1)}$ using this psychophysical methodology.

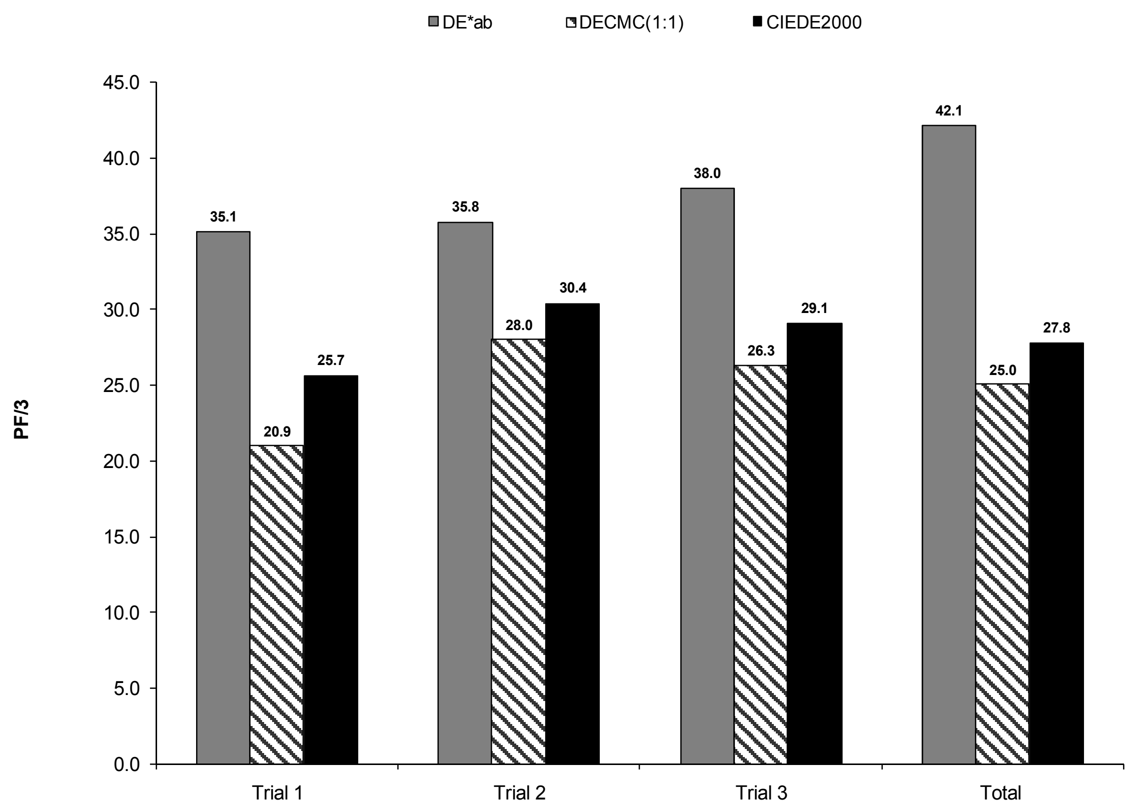


Figure 50. PF/3 for ΔE^*_{ab}, $CMC_{(1:1)}$, and $CIEDE2000_{(1:1:1)}$ against trials 1 to 3 and the combined dataset.

3.1.3.5. Performance of Color Difference Formulas against Each Visual Dataset Using STRESS Function

To test the statistical significance of the differences found in the performance among the three color difference formulas, the Standardized Residual Sum of Squares (*STRESS)* function was used. Each formula was evaluated using combined data. Tables 23 and 24 show the calculated STRESS and F values for each trial as well as the combined data when compared against ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ equations at K_L or l setting of 1. As shown in section 6.1.5 of chapter II, Equation 46 can be used to assess the significance of variation between models.

Table 23. The STRESS values for color difference models.

	ΔE^*_{ab}	$CMC_{(1:1)}$	$CIEDE2000_{(1:1:1)}$
Trial 1	0.3390	0.2029	0.2319
Trial 2	0.3237	0.2254	0.2371
Trial 3	0.3532	0.2369	0.2500
Combined	0.3570	0.2204	0.2491

For the methodology employed in this study in each individual trial the DF was 20, and the critical F values of $F_C = 0.405$ and $1/F_C = 2.46$ were calculated by STATCRUNCH® [120]. Since for the combined dataset the DF from total sample range changes, DF = 25, different critical F values of $F_C = 0.45$ and $1/F_C = 2.23$ are obtained. This difference is due to the fact that fewer samples were discarded from the combined data. Table 24 compares the performance of color difference equations for the entire dataset.

Table 24. F values between different equations for each set of data.

	ΔE^*_{ab} /$CMC_{(1:1)}$	$CMC_{(1:1)}$/$CIEDE2000_{(1:1:1)}$	ΔE^*_{ab} /$CIEDE2000_{(1:1:1)}$
Combined data	2.62	0.78	2.05

For the combined observations, no significant difference exists between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ or ΔE^*_{ab} and $CIEDE2000_{(1:1:1)}$. A significant difference, however, is seen in the comparison of ΔE^*_{ab} versus $CMC_{(1:1)}$ equations since the calculated F value is greater than 1/Fc (2.23). In this case also it can be concluded that ΔE^*_{ab} is significantly poorer than $CMC_{(1:1)}$ as a color difference prediction.

3.1.4. Conclusions

In this study, two statistical methods namely PF/3 and the recently proposed STRESS function were used to analyze the variability in the visual assessment of colored samples for a group of naïve assessors. The objective was to evaluate the use of pair comparison as a psychophysical method in the visual judgment of small color differences of textile samples used in this study.

The methodology employed in this study required the use of visual probability, *P,* in order to determine a visual difference [122]. Sample pairs with a *P* value smaller than 0.05 or greater than 0.95 is are discarded as they constitute judgments that for the majority of observers have larger or smaller color differences than that of the anchor pair [12].

In this study, it was found that in each trial different samples as well as a different number of sample pairs had to be discarded. Therefore, samples discarded from each trial were removed for the analysis of results of all trials to provide a means of comparison between trials. A total of 10 samples were thus discarded out of 31 samples. As such it has to be emphasized that the selection of sample pairs for a pair comparison psychophysical testing method should correspond with the magnitude of color difference of the anchor pair to yield useful information and avoid loss of data. In addition, the methodology employed in this study does not allow making any inferences in relation to observer variability among trials. However, in a previous study [121] observer accuracy and repeatability were reported using a mathematical transformation that is not recommended in the CIE guidelines [122].

Results of PF/3 of agreement between trials demonstrate that, for the psychophysical method employed in this study, the correlation of assessments among trials is high as demonstrated for a lower PF/3 value (Table 22). It also indicates that in the repeatability of the assessment as a group, does not improve with the repetition of trials.

PF/3 was used to compare each trial and the combined observations against ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ color difference formulas. The results indicate that for the data obtained in this study the PF/3 value based on $CMC_{(1:1)}$ formula is smaller compared to that of $CIEDE2000_{(1:1:1)}$ model as shown in Figure 50. In addition, the PF/3 values obtained all color difference models are approximately in agreement with those reported previously using different visual datasets [8].

Additionally, the significance of difference in performance of ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models against visual data was tested based on the STRESS function. Comparisons show that no significant difference exists between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ and ΔE^*_{ab} and $CIEDE2000_{(1:1:1)}$. A significant difference, however, is seen in the comparison of ΔE^*_{ab} versus $CMC_{(1:1)}$ equations. Since the calculated F value for ΔE^*_{ab} /CMC is greater than 1/Fc (2.23), it can be concluded that ΔE^*_{ab} is significantly poorer than $CMC_{(1:1)}$ as a color difference prediction model for the first trial as well as the combined visual dataset.

3.2. Visual Assessment Using a Jumbo Gray Scale

This study was carried out to evaluate the role of physical size of the scale in visual assessment of small color differences. A custom made scale using AATCC Gray Scale papers was constructed.

3.2.1. Experimental

3.2.1.1. Samples

Visual assessments were made using the same set of 31 sample pairs (a standard and a batch) dyed with disperse dyes on unbrightened spun polyester knitted fabric described in section 2.1.1 of this chapter.

3.2.1.2. Sample Viewing

The sample viewing conditions used in this part of the study were exactly the same as those described in section 2.1.2 of this chapter. Each observer wore a mid-gray laboratory coat and a pair of gray gloves to minimize sample degradation. The samples were placed by the experimenter who also wore a mid-gray laboratory coat. At the beginning of the experiment, the observer was asked to view the empty illuminated viewing booth for 2 minutes to adapt to the light source, during which time the experiment was explained.

3.2.1.3. Psychophysical Method

A custom made scale using standard AATCC gray scale papers was used for the assessment of perceived magnitude of small color differences. The paper used for the construction of the scale was supplied by X-rite. The standard gray papers were precision cut to 2” × 2” and mounted into PVC panel as a backing, and all the components were uniformly spray painted to L* of 74, which is approximately equivalent to Munsell N 7.25. Due to the size of the pairs, the scale was divided into three parts as shown in Figure 51. Given the previous experience when using a gray scale method, an identical gray scale was constructed in which the position of the standard of the each of the pairs in the scale was reversed. This small addition was used to avoid observers needing to rotate the scale 180 degrees when assessing some of the test pairs.

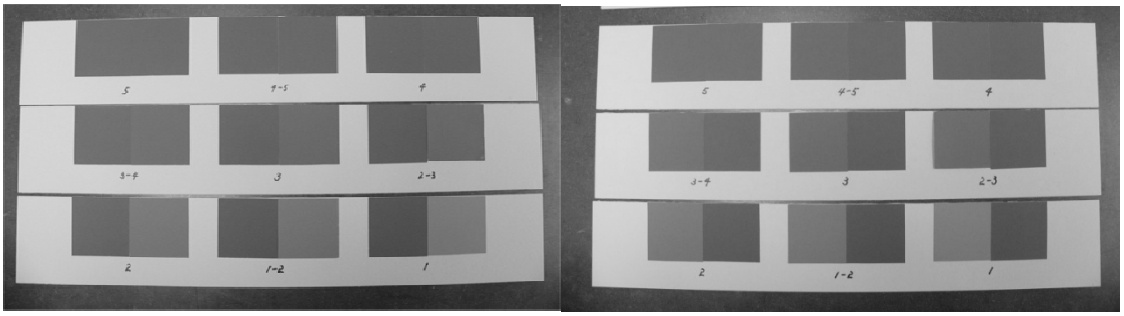


Figure 51. AATCC custom made gray scale panels.

As in the naïve vs. expert observer study, the two larger scales were used as a guide to assess the perceptual differences in color. For each sample pair the subject was asked: "Which gray scale difference is in closest agreement with the difference between the displayed sample pair?" Observers were allowed to select intermediate values, such as 3.4, expressed in decimal points. The answer was recorded. In addition, the observers were encouraged to use either one of the two gray scales since they were identical. However, the selection of the scale was not recorded. Figure 52 shows the test method being performed.

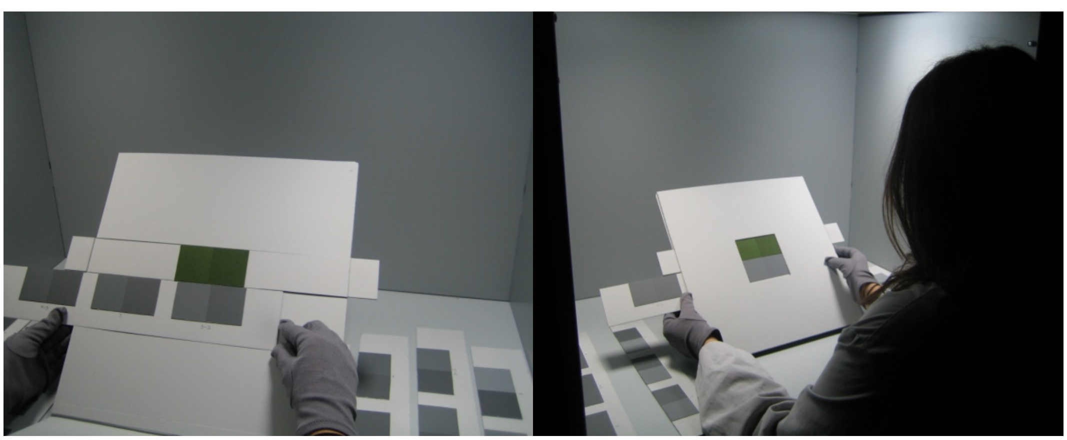


Figure 52. Experiment set up using the Jumbo gray scale.

Twenty five naïve observers (10 females and 15 males), mostly students of North Carolina State University, were tested for normal color vision using the Neitz test. [30]. Each observer sat in front of the viewing booth, so that he/she could move the reference gray scale freely. Each observer assessed all samples three times on separate days. The same viewing booth, sample sets, and sample presentation, was used in each trial. The samples were presented in a random order to each observer in each trial.

3.2.2. Analysis of Data

3.2.2.1. Gray Scale Transformations to Visual Difference

A total of 2,325 assessments were made using 31 sample pairs. For the analysis, the raw data in grade units, G, were transformed to a visual difference, ΔV, for each pair using Equation 73.

$$DV = -0.32G^3 + 3.76G^2 - 16.02G + 26.44 \tag{73}$$

The equation was obtained by fitting a third degree polynomial between measured ΔE^*_{ab} and AATCC scale grades. The average visual response from all observers was calculated for each sample pair. These values were used in the statistical analysis.

3.2.2.2. Analysis of Variance

The methodology employed in this analysis is described in section 2.2.3. An analysis of error variance was conducted to identify systematic and random errors. For the analysis of each method, the two sources of error were considered separately.

For the analysis of random errors, a variance components analysis was carried out. For the analysis of systematic errors, intra-observer standard deviation and intra-group standard deviation were calculated.

In addition to the above statistical techniques, PF/3, and STRESS functions were used to analyze and compare the results. A detailed explanation of these methodologies is also given in sections 6.1.3 and 6.1.5 of chapter II. PF/3 and STRESS values were calculated using Excel.

3.2.3. Results and Discussion

The raw data in gray scale units was transformed to visual differences according to the methodology explained in section 3.2.2.1. Figure 53 shows the average visual difference results for the three trials carried out by the observers.

Figure 53 shows there is close agreement among for the average of three trials in the perceived color difference of each pair.

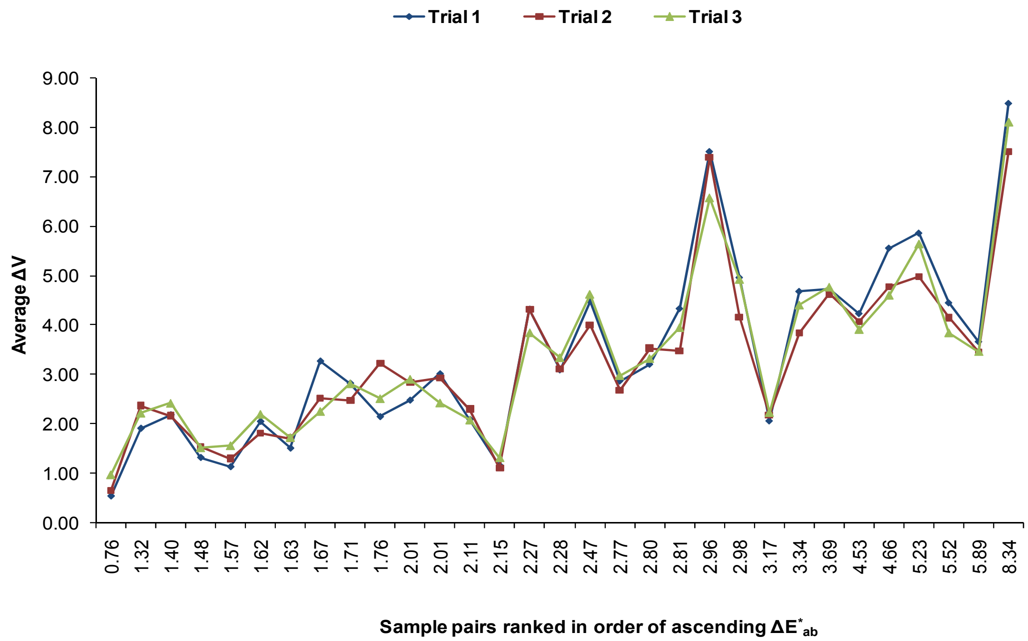


Figure 53. Graph of DV for the visual assessments of each sample pair, ranked in ascending order of ΔE_{ab}.

Table 25. Summary of t-test statistics for assessments carried out by naïve assessors.

Group	t	P	Significance
Trial 1 vs. Trial 2	1.91	0.0658	No significant difference at α =0.05
Trial 2 vs. Trial 3	-1.01	0.3221	No significant difference at α =0.05
Trial 1 vs. Trial 3	1.19	0.2449	No significant difference at α =0.05

A paired t-test, results of which are summarized in Table 25, was used to evaluate statistical difference between each trial among the observers.

No statistical difference was found between trials.

3.2.3.1. Variance Analysis

3.2.3.1.1. Systematic Errors: Intra-Group Variation

To determine the variability of individual observers in comparison to the group average in each trial, intra-group variability was calculated. The average perceived color difference for each pair in each group was assumed to represent the true value. Intra-group variability is the degree to which each observer's assessments of each pair agrees with the true value of the pair in that trial. Table 26 summarizes the mean standard deviation of the group of observers as an indicator of the intra-group variation among observers for each trial.

Table 26. Summary of intra-group standard deviation for observers.

Observer	Trial 1	Trial 2	Trial 3
Mean	1.80	1.70	1.68
Minimum	0.94	0.98	1.02
Maximum	4.24	2.79	3.17
Range	3.30	1.81	2.29

The results indicate that the mean intra-group standard deviation decreases with the number of trials. However, the range as well as the minimum and maximum intra-group standard deviation varied along the trials with no noticeable relationship among trials.

3.2.3.1.2. Random Errors: Variance Component Analysis

Table 27 includes the variance components analysis for the observers. Each entry in the column is a numeric measure of variability. An algebraic combination of the mean squares, using Equations 62-64 from section 2.2.3.2 can be used to estimate the variance of components themselves. These estimates are summarized in Table 28. The complete analysis can be found in appendix F.

Table 27. Variance components analysis for assessments using the Jumbo gray scale.

Source	DF	Sum of Squares	Mean Square
Pair	30	6223.33	207.44
Observer	24	1287.79	53.66
Pair × Observer	720	3231.49	4.49
Error	1550	3242.79	2.09

Table 28. Variance components estimates for assessments using the Jumbo gray scale.

Source	Estimated Value
Variance due to pairs	2.70
Variance due to observers	0.53
Variance due to interaction	0.79
Variance due to error	2.09

The highest variance shown in Table 28 is due to pairs. This is expected since the ΔE_{ab}^* range for the samples presented to the observers was from 0.76-8.34.

While the variance due to observers is smaller than that for the effect of repeated assessments (shown as interaction in Table 28), it is smaller than the variance due to the error, which essentially comprises all other variables in the system.

3.2.3.2. Intra-Observer Variability

Equation 68 from section 2.2.4 was used to calculate intra-observer variability to calculate the deviation of an observer in a trial involving 31 assessments in comparison with his/her own average results from all trials. Table 29 summarizes the intra-observer variability found among observers using this experimental methodology.

Table 29. Summary of intra-group standard deviation for observers in all trials.

Observer	Grand Average
Mean	1.09
Minimum	0.33
Maximum	2.48
Range	2.15

Under ideal conditions, observers would reproduce their assessment at different times with a standard deviation of zero. However, studies involving human experience, including color difference perception, usually demonstrate high standard deviation in repeatability among observers. Results obtained in this study indicate that the repetition of the assessment has a large range of variation. Further comparisons of these results can be found in chapter 5

3.2.3.3. PF/3 Observer Accuracy

PF/3 metric was also used to assess observer accuracy. PF/3 values were calculated using the grand mean visual results from all observers for each pair compared to each observer's assessments in each trial. It was assumed that the grand mean visual results were the true value for each pair. An accurate observer is one that agrees closely with the mean visual results from all observers [12]. PF/3 for accuracy was calculated for observers in each trial and a mean value was obtained. Table 30 summarizes the results of PF/3 for accuracy.

Table 30. Summary of observers' variation in terms of PF/3 for accuracy.

Observer	Trial 1	Trial 2	Trial 3	Mean Trials
Mean	63.30	62.59	58.39	61.43
Minimum	31.79	31.57	30.74	34.97
Maximum	126.97	109.19	110.54	102.52
Range	95.17	77.62	79.80	67.55

Similar to the method previously described, the mean PF/3, is higher than those reported by Guan and Luo (PF/3=40) [12], Cui et al (PF/3=37) [38] Chou et al(PF/3=35) [47], Xu et al. and (PF/3=33) [45]. While these studies used somewhat different experimental methodologies, the reasons for the discrepancy in PF/3 data are not clear.

However, since the experimental method reported here was controlled to minimize common variables (such as observer sample handling, sample mounting and juxtaposition, etc.) it is unlikely that the method used is a primary factor in obtaining large PF/3 values. On the other hand, the results obtained in this study are in good agreement with the reported values by Mangine (PF/3=64) [121] and Lee [94](PF/3=57. Mangine's study employed a different methodology, but the color centers and the dyed substrates used for the construction of samples were the same as those used in this study. In Lee's work a single color center was investigated using the same fabric used in the present study and a similar visual assessment methodology was also employed.

3.2.3.4. PF/3 Observer Repeatability

PF/3 of repeatability was calculated for the observers and the mean values for the group as well as each observer were obtained. A summary of results is shown in Table 31.

Table 31. Summary of PF/3 of repeatability for naïve observers.

Observer	**Trial 1&2**	**Trial 2 &3**	**Mean**
Mean	72.59	68.75	70.67
Minimum	25.70	22.81	24.26
Maximum	159.26	121.86	133.87
Range	133.56	99.05	109.61

The mean PF/3 values between trials are consistently high and the repeatability among observers is low. Possible reasons for the low repeatability are the inclusion of outliers in the statistical analysis. As in the PF/3 results reported for accuracy, the PF/3 repeatability results obtained in this study are significantly higher than those reported by Chou et al. (PF/3= 27 and 41) [47], Cui et al. (PF/3=37) [38], Xu et al. and (PF/3=26) [45] and Xin et al.[43] (PF/3=22.1). In contrast, the results obtained in this study are similar to the results obtained by Mangine (PF/3=64.8) [121] and Lee (PF/3=66.93) [94] which are closely related to the present research methodology.

Results show that on average observers' repeatability improves with increasing the number of trials for this method as evidenced by a smaller PF/3 range between trials 2 and 3 compared to that for trials 1 and 2. However, as shown in Figure 54, this is probably due to improved repeatability of the most inconsistent observers within the group, although a few observers show a reverse trend.

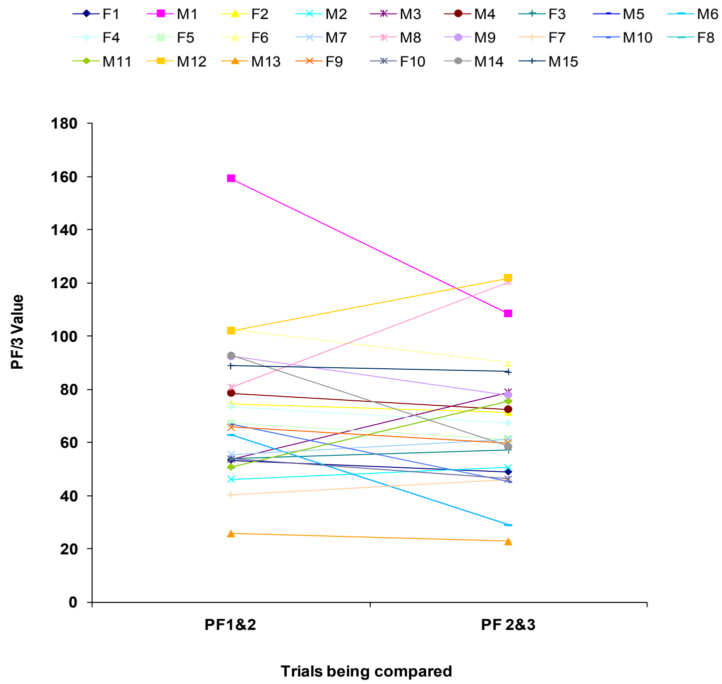


Figure 54. Comparison of the performance of individual observers in terms of PF/3 repeatability among trials.

3.2.3.5. Performance of Color Difference Formulas against Each Visual Dataset

The performance of color difference equations was tested against the average visual data for each of the observer trials. Two color difference models, namely CMC and CIEDE2000 equations, were evaluated at K_L or *l* setting of 1.Figure 55 shows a consistent reduction in PF/3 with each additional trial made by the observers.

As expected, the ΔE^*_{ab} equation performs rather poorly in comparison with the $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models.

In addition, $CIEDE2000_{(1:1:1)}$ performs better in comparison with the other two formulas as evidenced by the lower PF/3 value, which is in agreement with the reported values by Cui et al.[38]. This improvement is thought to be due to the use of a highly controlled visual assessment procedure, and especially the viewing conditions employed, which reduces the sources of variability amongst observers.

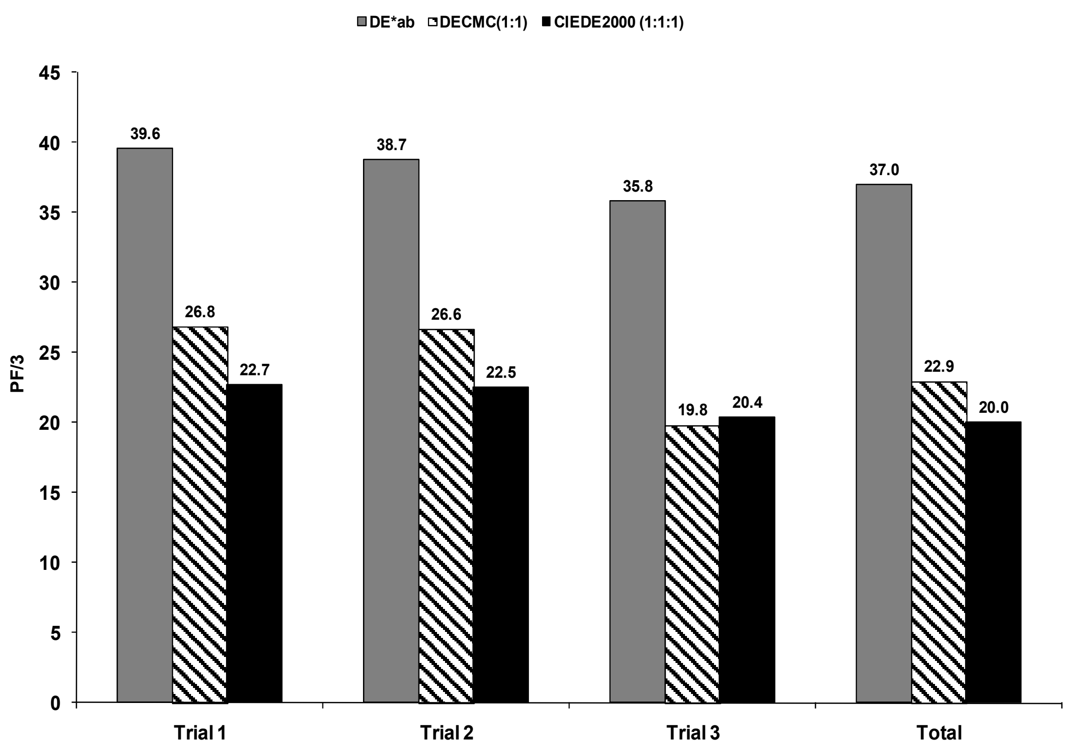


Figure 55. Graph of PF/3 ΔE^*_{ab}, $CMC_{(1:1)}$, and CIEDE2000 $_{(1:1:1)}$ for the average observers (three trials)

3.2.3.6. Performance of Color Difference Formulas against Each Visual Dataset using STRESS

In order to determine the significance of variation between models, the newly developed statistical metric, STRESS, was used [110] Tables 32, 33, and 34 show the results of calculating the STRESS function and F values for the observers when compared against ΔE^*_{ab}, CMC and CIEDE2000 equations at K_L or *l* setting of 1. Equation 46, from section 6.1.5 of chapter II was used to assess the significance of variation between models.

Table 32. Summary of the calculated STRESS values for color difference models.

Observer	ΔE^*_{ab}	$CMC_{(1:1)}$	$CIEDE2000_{(1:1:1)}$
Total	0.33	0.21	0.17
Trial 1	0.33	0.22	0.18
Trial 2	0.34	0.25	0.19
Trial 3	0.32	0.19	0.17

Table 33. F values between trials against based ΔE^*_{ab}.

Observer	ΔE^*_{ab}
Trial 1 vs. Trial 2	0.93
Trial 2 vs. Trial 3	1.13
Trial 1 vs. Trial 3	1.05

Table 34. Summary of the calculated F values between different equations for the average set of observations.

Observer	ΔE^*_{ab} /$CMC_{(1:1)}$	$CMC_{(1:1)}$/$CIEDE2000_{(1:1:1)}$	ΔE^*_{ab} /$CIEDE2000_{(1:1:1)}$
Total	2.36	1.56	3.68

For the methodology employed in this study the calculated critical F values, obtained from STATCRUNCH [120], are FC = 0.48 and 1/FC = 2.08. Table 33 compares the variability of a given trial against another trial using the ΔE^*_{ab} formula in order to determine the significance of difference between trials. Table 34 compares the performance of color difference equations based on the average observations from three trials. If the calculated F is bigger than $1/F_C$ or smaller than F_C, the difference between two given models is significant. In this case, the comparison between ΔE^*_{ab} and $CMC_{(1:1)}$ shows that for the total number of observations, ΔE^*_{ab} performs significantly poorer than $CMC_{(1:1)}$. In addition, a significant difference is seen in the comparison of ΔE^*_{ab} versus $CIEDE2000_{(1:1:1)}$ equations and since the calculated F values are greater than 1/Fc (2.08), ΔE^*_{ab} performs statistically poorer than $CIEDE2000_{(1:1:1)}$ as a color difference prediction model.

However, no significant difference between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was observed since the calculated F-values are between 1 and $1/F_C$. Additional detail regarding interpreting calculated F-values is given in section 6 of chapter II.

3.2.4. Conclusions

The present study was carried out to test the variability among observers and the performance of different models when using a custom made scale for the assessment of small color differences. The scale was designed to match exactly the size of the colored samples.

Several statistical methods, including paired t-test, PF/3 and STRESS function were used to compare inter- and intra-observer variability in visual assessment of colored samples for a group of observers.

Results of a paired t-test demonstrate that, for the psychophysical method employed in this study, the average perceived color difference for the group did not change throughout the repeated assessments. No statistical significance was found among trials.

PF/3 results show that, on average, observers' repeatability improves with increasing the number of trials for this method as evidenced by a smaller PF/3 range between trials 2 and 3 compared to that for trials 1 and 2. However, this is probably due to improved repeatability of the most inconsistent observers within the group, although a few observers showed a reverse trend. The mean PF/3 values for repeatability are somewhat higher than some of those from published results which employ different assessment methodologies (Table 31).

The mean PF/3 values for observer accuracy, and the intra-group standard deviation among trials, do not change significantly from trial to trial (Table 30). Again the mean PF/3 values obtained in this study are higher than those reported previously for unknown reasons [8].

In order to determine the root causes of variability for observers, a variance components analysis was conducted. Results show that the highest variance is due to sample pairs. This was expected since the range of color differences of sample pairs presented to the observers was 0.76-8.34 ΔE^*_{ab} units.

It was also shown that the effect of observer is smaller than that due to repetition of assessments (Table 28). The variance due to the error, which is a combination of all other unknown factors, is the second highest source of variation (Table 28)

The visual data obtained from observers was also compared against ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ color difference formulas and $CMC_{(1:1)}$ and outperformed ΔE^*_{ab}. According to the PF/3 metric $CIEDE2000_{(1:1:1)}$ shows higher agreement with visual data than $CMC_{(1:1)}$ for the samples tested (Fig 55). The PF/3 values obtained for the color difference models tested are in general agreement with those reported previously.

The significance of variation among the color difference formulas when tested against the observer visual data was also compared using the STRESS function. Results show a significant difference between ΔE^*_{ab} and $CMC_{(1:1)}$ and between ΔE^*_{ab} and $CIEDE2000_{(1:1:1)}$ and show that ΔE^*_{ab} is significantly poorer than both $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ equations. However, no significant difference between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was found.

Results of further comparisons with other methodologies employed in this dissertation will be discussed in chapter 5.

3.3. Visual Assessment Using Jumbo Scale with Gap

This experiment was carried out to evaluate the role of a two inch separation between colored sample pairs and the Jumbo scale for the assessment of small color differences. In the experiment conducted in section 3.2. The method involved placing the Jumbo gray scale pairs directly adjacent to the colored samples pairs being assessed. This has potential to introduced unwanted crispening [9, 18, 43] effects, thereby leading to increased variability in visual data. Hence, the hypothesis for the experiment described in this section was that a significant gap between the gray scale and the sample pairs would lead to a reduction in inter and intra-observer variability. In addition, in practice, viewing conditions including sample presentation often diverge from recommended standards; therefore we evaluated how intra- and inter-observer variability might be affected by the methodological conditions evaluated in the present section. The custom made scale built for the methodology employed in the previous section was used in the current study.

3.3.1. Experimental

3.3.1.1. Samples

Visual assessments were made using the same set of 31 sample pairs (a standard and a batch) dyed with disperse dyes on unbrightened spun polyester knitted fabric used in the studies reported previously and described in section 2.1.1.

3.3.1.2. Sample Viewing

The sample viewing conditions used in this part of the study were exactly the same as those described in section 2.1.2.

3.3.1.3. Psychophysical Method

The custom made scale described in section 3.2 was used as a guide for the assessment of small color differences. The scale was constructed using the AATCC papers provided by X-rite Inc. The standard gray papers were cut to 2"x 2" dimensions corresponding to the size of the samples used in this study. In addition to the construction method previously described in section 3.2, a piece of PVC backing measuring 1" x 18" was attached to the back panel of the gray scale which, when placed on the custom made sample stand, created a 2' gap between colored pairs and the gray scale as shown in Figure 56.

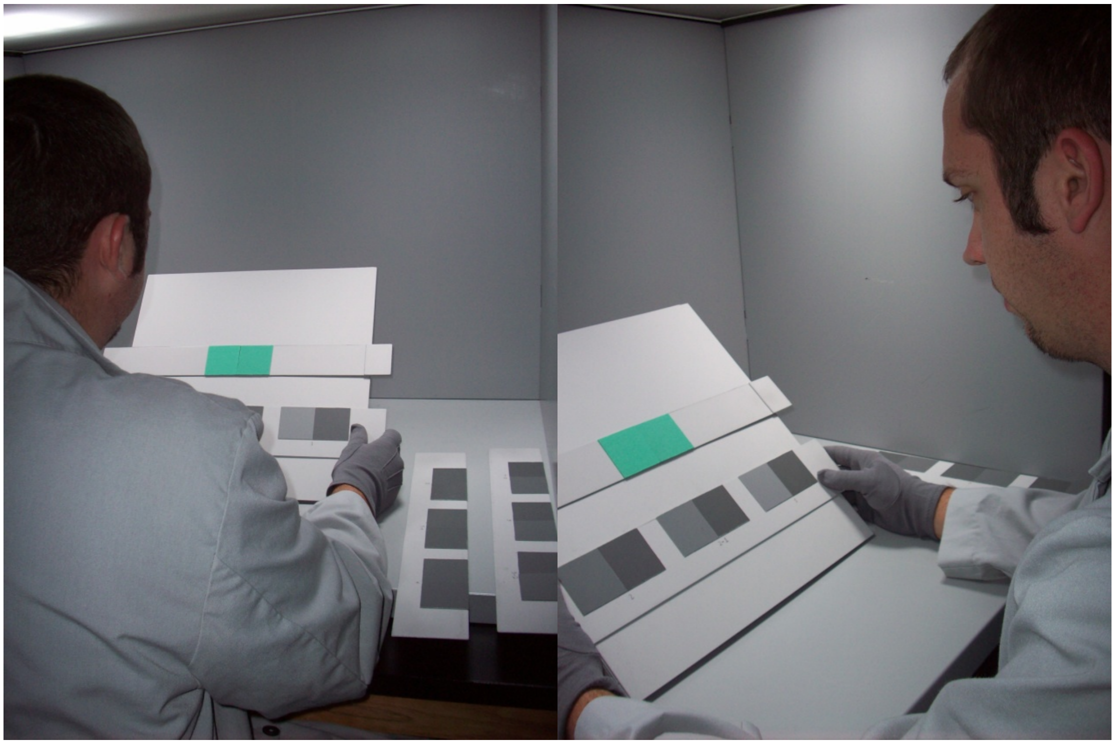


Figure 56. Experimental set up for the Jumbo Scale experiment with two inch gap..

The size of the gap between samples and the gray scale corresponded with that of the height and width of the colored samples and individual gray scale papers. Two sets of scales were prepared to avoid observers needing to rotate the scale 180 degrees when assessing some of the test pairs. The scale was used to assess the perceptual differences in color. For each sample pair the subject was asked to perform the same task described in section 3 of chapter III. Observers were allowed to select intermediate values, such as 3-4, expressed in decimal points. The answer was recorded. Figure 56 shows the setting for this testing method.

Twenty five naïve observers (13 females and 12 males) participated in this study. The observers were mostly students of North Carolina State University, tested for normal color vision using the Neitz test [30], participated in this study. Each observer sat in front of the viewing booth, so that he/she could move the reference gray scale freely. Each observer assessed all samples three times on separate days. The same viewing booth, sample sets, and sample presentation was used in each trial.

3.3.2. Analysis of Data

3.3.2.1. Gray Scale Transformations to Visual Difference

A total of 2,325 assessments were made using 31 sample pairs. For the analysis, the raw data in grade units, G, were transformed to a visual difference, ΔV, for each pair using Equation 74.

$$\mathrm{DV} = -0.35G^3 + 3.98G^2 - 16.53G + 26.98 \tag{74}$$

The equation was obtained by fitting a third-degree polynomial equation between measured ΔE^*_{ab} and the scale grades. The average visual response from all observers was calculated for each sample pair. These values were used to carry out the statistical analysis.

3.3.2.2. Analysis of Variance

The methodology employed in this analysis is described in section 6 of chapter II. An analysis of error variance was conducted to identify systematic and random errors. For the analysis of each method, the two sources of error were considered separately.

For the analysis of random errors, a variance components analysis was carried out. For the analysis of systematic errors, intra-observer standard deviation and intra-group standard deviation were calculated.

In addition to the above statistical techniques, PF/3, and STRESS functions were used to analyze and compare the results. A detailed explanation of these methodologies is also given in sections 6.1.3 and 6.1.5.

3.3.3. Results and Discussion

The raw data in gray scale units was transformed to visual differences according to the methodology explained earlier. Figure 57 shows close agreement between the average visual difference results for each of the three trials.

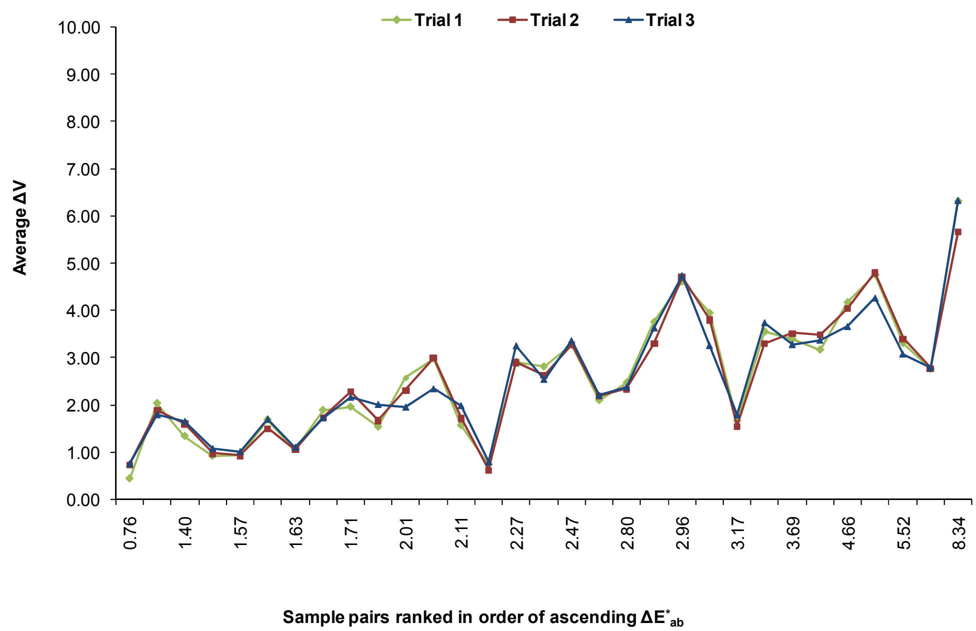


Figure 57. Graph of DV for the visual assessments of each sample pair, ranked in ascending order of DV.

A paired t-test, results of which are summarized in Table 35, was used to evaluate statistical difference between each trial among the observers. No statistical difference was found between trials.

Table 35. Summary of t-test statistics for assessments carried out by naïve assessors.

Group	t	P	Significance
Trial 1 vs. Trial 2	0.84	0.4079	No significant difference at α =0.05
Trial 2 vs. Trial 3	-0.03	0.9782	No significant difference at α =0.05
Trial 1 vs. Trial 3	0.55	0.584	No significant difference at α =0.05

3.3.3.1. Variance Analysis

3.3.3.1.1. Systematic Errors: Intra-Group Variation

To determine the variability of individual observers in comparison to the group average in each trial, intra-group variability was calculated. The average perceived color difference for each pair in each group was assumed to represent the true value. Hence, the intra-group variability, is the degree to which each observer's assessments of each pair agrees with the true value of such pair in that trial. Table 36 summarizes the mean standard deviation of the group of observers as an indicator of the intra-group variation among observers for each trial.

Table 36. Summary of intra-group standard deviation for observers.

Observer	Trial 1	Trial 2	Trial 3
Mean	1.53	1.45	1.45
Minimum	0.48	0.75	0.76
Maximum	3.92	3.88	3.39
Range	3.44	3.14	2.63

The results indicate that the range in intra group standard deviation decreases with the number of trials and the means for trial 2 and 3 are identical, and both are lower than the mean for trial 1. However, the minimum value increases from trial 1 to 3.

3.3.3.1.2. Random Errors: Variance Component Analysis

Table 37 includes the variance components analysis for the observers. Each entry in the column is a numeric measure of variability. An algebraic combination of the mean squares, using Equations 62-64 from section 2.2.3.2, was used to estimate the variance of components themselves. These estimates are summarized in Table 38.

Table 37. Variance components analysis for assessments using the Jumbo gray scale with gap.

Source	**DF**	**Sum of Squares**	**Mean Square**
Pair	30	3654.40	121.81
Observer	24	2072.53	86.36
Pair × Observer	720	2510.45	3.49
Method	1550	1885.13	1.22

Table 38. Variance components estimates for assessments using the Jumbo gray scale with a gap.

Source	**Estimated Value**
Variance due to pairs	1.58
Variance due to observers	0.89
Variance due to interaction	0.76
Variance due to method	1.22

The highest variance shown is Table 38 is due to pairs. This is expected since the ΔE_{ab}^{*} range for the samples presented to the observers was from 0.76-8.34.

While the variance due to observers is slightly higher than that for the effect of repeated assessments (shown as interaction in Table 38), it is smaller than the variance due to the error, which essentially comprises all other variables in the system.

3.3.3.2. Intra-Observer Variability

Equation 68 from section 2.2.4 from Chapter III was used to calculate intra-observer variability, which is the deviation of an observer in a trial involving 31 assessments in comparison with his/her own average results from all trials. Table 39 summarizes the intra-observer variability found among observers using this experimental methodology.

Table 39. Summary of intra-observer standard deviation for observers in all trials.

Observer	Grand Average
Mean	0.77
Minimum	0.28
Maximum	1.81
Range	1.53

Under ideal conditions, observers would reproduce their assessment at different times with a standard deviation of zero. However, studies involving human experience, including color difference perception, usually demonstrate high standard deviation in repeatability among observers.

These values are used to compare the intra-observer standard deviation for observers based on the use of the Jumbo scale with a 2” gap when compared to other methods as shown in section 5.

3.3.3.3. PF/3 Observer Accuracy

PF/3 metric was used to assess observer accuracy. PF/3 values were calculated using the grand mean visual results from all observers for each pair compared to each observer’s assessments in each trial. It was assumed that the grand mean visual results are the true value for each pair. An accurate observer is one that agrees closely with the mean visual results from all observers [12]. PF/3 for accuracy was calculated for observers in each trial and a mean value was obtained. Table 40 summarizes the results of PF/3 for accuracy.

Table 40. Summary of observers’ variation in terms of PF/3 for accuracy.

Observer	Trial 1	Trial 2	Trial 3	Mean Trials
Mean	62.88	62.47	62.90	62.69
Minimum	21.67	34.41	35.92	32.82
Maximum	117.73	125.72	107.48	115.36
Range	96.06	91.32	71.56	82.54

Similar to method previously described, the mean PF/3, is higher than those reported by Guan and Luo (PF/3=40) [12], Cui et al (PF/3=37) [38] Chou et al(PF/3=35) [47], Xu et al. and (PF/3=33)[45].

While these studies used somewhat different experimental methodologies, the reasons for the discrepancy in PF/3 values data are not clear. However, since the experimental method reported here was deliberately controlled to minimize common variables (such as observer sample handling, poor mounting and juxtaposition of samples, etc.) it is unlikely that the method used is a primary factor in obtaining large PF/3 values.

Again, the results obtained in this study are in good agreement with the reported values by Mangine's (PF/3=64) [121] and Lee's (PF/3=57) [94]. Mangine's study employed a different methodology, but the color centers and samples are similar to those used in this study. The samples used in this study correspond to areas of the color space not explored in previous studies. Lee's data were obtained using the same texture fabric of the present study and the same visual assessment methodology.

3.3.3.4. PF/3 Observer Repeatability

PF/3 of repeatability was calculated for the observers and the mean values for the group as well as each observer were obtained. A summary of results is shown in Table 41.

Table 41. Summary of PF/3 of repeatability for naïve observers.

Observer	**Trial 1&2**	**Trial 2 &3**
Mean	61.87	59.37
Minimum	20.91	31.71
Maximum	102.74	103.15
Range	81.83	71.44

The mean PF/3 values between trials are consistently high and the repeatability among observers is low. As in the PF/3 results reported for accuracy, the PF/3 repeatability results obtained in this study are significantly higher than those reported by Chou et al. (PF/3 = 27 and 41) [47], Cui et al (PF/3 = 37) [38], Xu et al. and (PF/3 = 26) [45] and Xin et al[43] (PF/3 = 22.1). However, the discrepancy with the reported values is not a high as the values obtained with other methodologies described previously.

According to table 41 above, PF/3 calculations show on average observer repeatability slightly improves with increasing the number of trials for this method. However, the reduction is unlikely to be practically significant as shown in Figure 58, it can be noticed that repeatability remains nearly constant through trials with the exception of few observers.

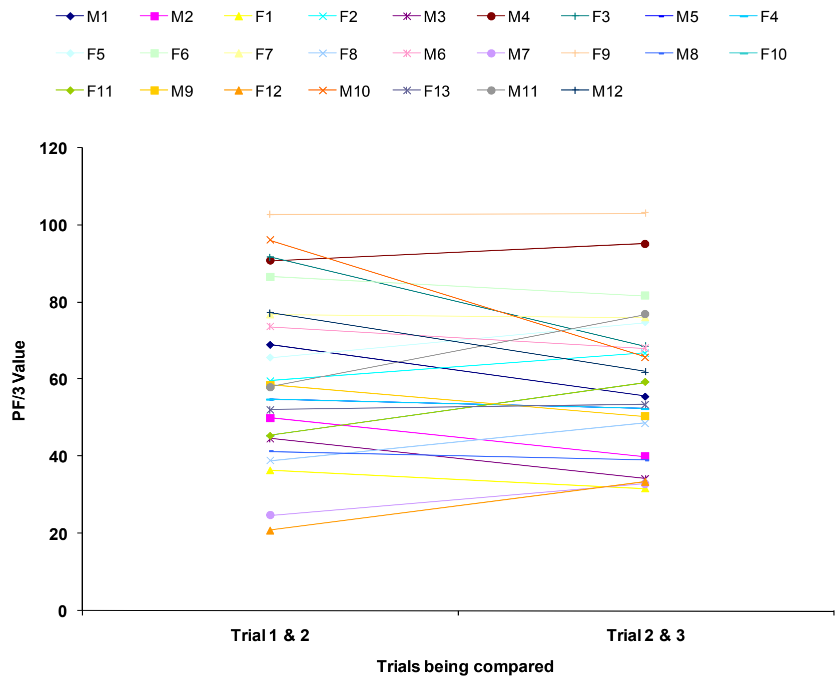


Figure 58. Comparison of the performance of individual observers in terms of PF/3 repeatability among trials.

3.3.3.5. Performance of Color Difference Formulas against Each Visual Dataset

The performance of color difference equations was tested against the average visual data for each of the observer trials. This was done to test the likely practical relevance of the variability in the visual data, since a primary goal of development of a robust visual assessment protocol is to establish visual data that will in turn be used in the development of a new and improved color difference formula.

Two color difference models, namely CMC and CIEDE2000 equations, were evaluated at K_L or l setting of 1.

Figure 59 shows that the performance of individual models does not change significantly among the trials or based on the use of the grand average of all three trials. However, the ΔE^*_{ab} equation performs rather poorly in comparison with the $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models, but the $CIEDE2000_{(1:1:1)}$ equation performs slightly better than $CMC_{(1:1)}$ with a higher agreement.

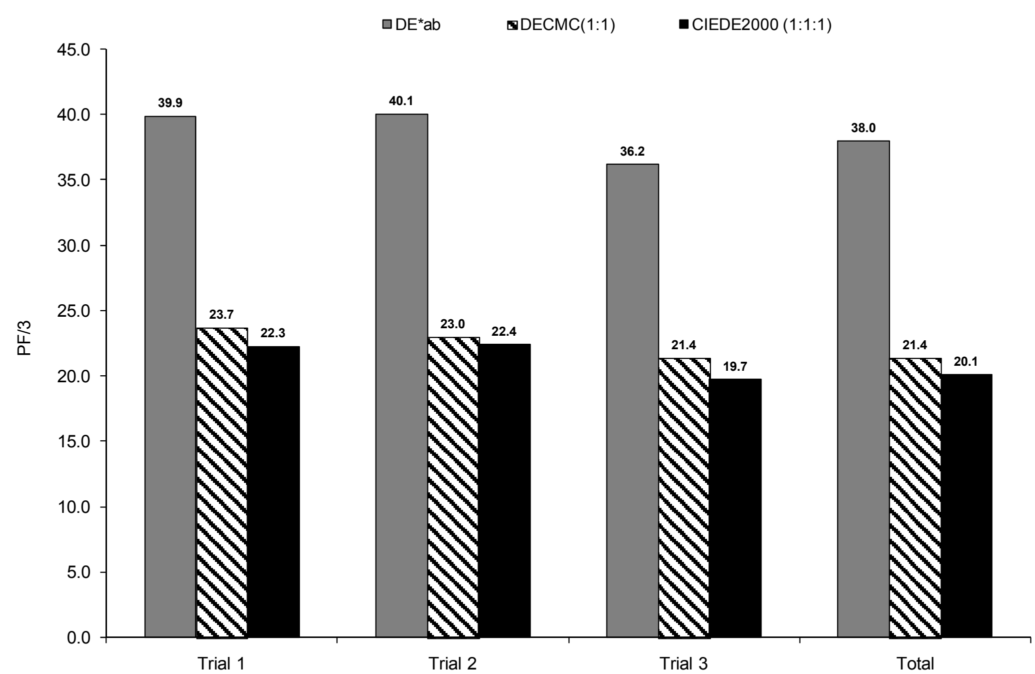


Figure 59. Graph of PF/3 for ΔE^*_{ab}, $CMC_{(1:1)}$, and $CIEDE2000_{(1:1:1)}$ for the average observers (three trials).

3.3.3.6. Performance of Color Difference Formulas against Each Visual Dataset using STRESS

In order to establish the significance of variation between models, the newly developed statistical metric, STRESS, was used. Tables 42, 43, and 44 show the results of calculating the STRESS function and F values for the observers when compared against ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ equations at K_L or l setting of 1. Equation 46, from section 6.1.5 of chapter II, can be used to assess the significance of variation between models.

Table 42. The STRESS values for color difference models.

Observer	ΔE^*_{ab}	$CMC_{(1:1)}$	$CIEDE2000_{(1:1:1)}$
Total	0.31	0.18	0.16
Trial 1	0.32	0.18	0.16
Trial 2	0.32	0.19	0.18
Trial 3	0.31	0.19	0.16

Table 43. F values between trials against ΔE^*_{ab} formula.

Observer	ΔE^*_{ab}
Trial 1 vs. Trial 2	0.99
Trial 2 vs. Trial 3	1.07
Trial 1 vs. Trial 3	1.06

Table 44. F values between different equations for the average set of observations.

Observer	ΔE^*_{ab} /$CMC_{(1:1)}$	$CMC_{(1:1)}$/$CIEDE2000_{(1:1:1)}$	ΔE^*_{ab} /$CIEDE2000_{(1:1:1)}$
Total	3.00	1.26	3.78

For the methodology employed in this study the calculated critical F values, obtained from STATCRUNCH, for the data are $F_C = 0.48$ and $1/F_C = 2.08$ [120]. Table 43 compares the variability of a given trial against another trial using ΔE^*_{ab} formula in order to determine the significance of difference between trials. The results show that no significant difference exits between trials. Table 44 compares the performance of color difference equations based on the average observations from three trials. If the calculated F is bigger than $1/F_C$ or smaller than F_C, the difference between two given models is significant. In this case, the comparison between ΔE^*_{ab} and $CMC_{(1:1)}$ shows that for the total number of observations, ΔE^*_{ab} performs significantly poorer than $CMC_{(1:1)}$.

In addition, a significant difference is seen in the comparison of ΔE^*_{ab} versus $CIEDE2000_{(1:1:1)}$ equation and since the calculated F value is greater than $1/F_C$ (2.08), it can be concluded that ΔE^*_{ab} performs significantly poorer than $CIEDE2000_{(1:1:1)}$ as a color difference prediction model.

On the other hand, no significant difference between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was observed since the calculated F-values are between 1 and $1/F_C$. Additional detail regarding interpreting calculated F-values is given in section 6 of chapter II.

3.3.4. Conclusions

The present study was carried out to test the variability among observers and the performance of different models when using a 2” gap between the Jumbo scale and the displayed samples in the assessment of small color differences.

PF/3 results indicate that on average observers’ repeatability slightly improves with increasing the number of trials for this method as evidenced by a smaller PF/3 range between trials 2 and 3 compared to that for trials 1 and 2. However, this is probably due to improved repeatability of the most inconsistent observers within the group, although a few observers showed a reverse trend. The mean PF/3 values for repeatability are somewhat higher than those from published results which employ different assessment methodologies (Table 41). The mean PF/3 values for observer accuracy, and the intra-group standard deviation among trials, do not change significantly from trial to trial (Tables 36 and 40). However, their range of variation decreases with each successive trial. Again the mean PF/3 values obtained in this study are higher than those reported previously for unknown reasons (Table 40).

In order to determine the root causes of variability for observers, a variance components analysis was conducted. Results show that the highest variance is due to sample pairs. This was expected since the range of color differences of sample pairs presented to the observers was 0.76-8.34 ΔE^*_{ab} units.

It was also shown that the effect of observer is higher than that due to repetition of assessments (Table 38). Indeed, among the components variability due to the repetition of assessment is the lowest and has no pronounced effect on the overall variability (Table 38). The variance due to the error, which is a combination of all other unknown factors, is the second highest source of variation (Table 38).

The visual data obtained from observers was also compared against ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ color difference formulas and $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ outperformed ΔE^*_{ab} for this visual dataset. According to the PF/3 metric $CIEDE2000_{(1:1:1)}$ shows higher agreement with visual data than $CMC_{(1:1)}$ for the samples tested (Fig 59), although STRESS showed no significant difference between these two formulas. The PF/3 values obtained for the color difference models tested are in general agreement with those reported previously.

However, the STRESS function showed ΔE^*_{ab} performed significantly poorer than both $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ equations.

4. Development of a Linear Gray Scale for Visual Assessment of Small Color Differences

The visual assessment of small color differences between two objects is of critical concern to all industries in which the color of a given product must be controlled to ensure consistency and quality. Large industry sectors that require color quality control of their products include but are not limited to, textiles, plastics, cosmetics, paint, printing, packaging, photography and imaging. A number of national and international standard methods exist today that enable quantification of the visual difference between two objects. These national and international standard methods have been developed by institutions such as the International Organization for Standardization [100, 123] (ISO), the American Society for Testing and Materials (ASTM)[62, 123, 124], the American Association of Textile Chemists and Colorists (AATCC)[36, 63], the British Standards Institute (BSI), the Deutsches Institut für Normung (DIN)[123], the Japanese Standards Institute (JSA)[125], and others[123, 126]. Some of these standard methods are used worldwide by various industries with the goal of ensuring consistency in quality control data from one assessment to another, and between, for example, a customer and one or more suppliers. Unfortunately, large variability is obtained in practically all assessments that involve human sensory judgments, which limits the usefulness of these methods. In the case of visual assessment of materials, in most cases certain variables must be controlled carefully.

These variables include the brightness and color temperature of the lighting, the color of the surround, angle of illumination and viewing, and the reference method by which magnitude estimation is made.

Most standard visual methods that quantify the magnitude of color difference perception employ a gray scale. A gray scale is used by trained observers to select a numerical color difference between a defined 'standard' color and 'trial'. Figure 60 shows a photograph that depicts a typical example of the use of a current standard AATCC Gray Scale for Color Change [36]. The AATCC Gray Scale for Color Change is a standard scale for the assessment of change in color, usually of textile materials, following a certain chemical, mechanical or standard test. In the Gray Scale for Color Change, the reference sample in each pair of gray samples is the same and is of medium lightness ($Y = 12 \pm 1$ corresponding approximately to $L^* = 41$ +/- 2 for a D65/10^o observer). The ISO scale is very similar in construction to the AATTC standard gray scale shown in Figure 60.

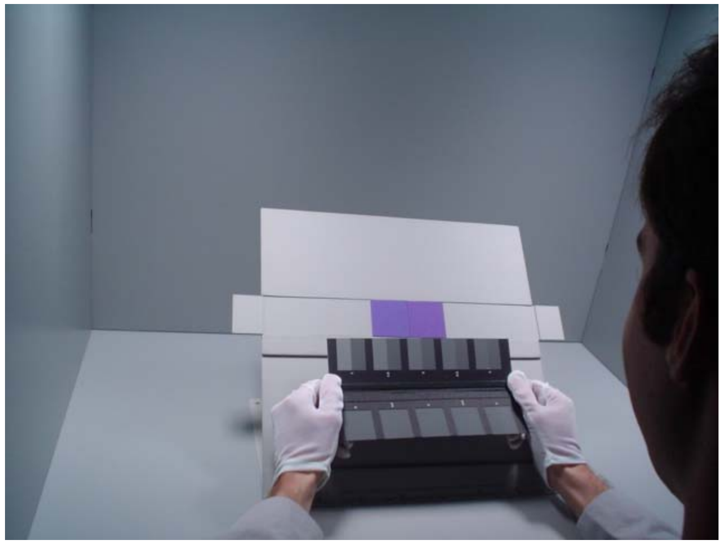


Figure 60. Photograph of a typical AATCC gray scale for color change**.**

The numerical rating selected when using a standard gray scale corresponds to the designated number of the pair of grays in the scale. Each pair represents a set visual contrast, which is compared to the contrast between a ‘standard’ and a ‘trial’ sample during visual assessments. A selection of, ‘2’, for instance, corresponds to a perceived difference in the objects being assessed identical in magnitude to the difference in gray pair number two in the scale. Where the contrast of the test pair is judged to be in between two pairs of grays in the scale, the assessor may interpolate and select a number between two numbers designated in the scale. For example, a value of 2.5 would correspond to a visual difference perceived to bisect pair 2 and pair 3 in the scale.

The arrangement of gray pairs in the current scales used in the AATCC and ISO standards is not perceptually linear. In other words the size of the difference between gray pairs does not follow a linear pattern. The arrangement is geometric and therefore these scales are perceptually geometric scales. A geometric scale is one in which the size of the contrast (lightness difference) between gray pairs in the scale approximately doubles as the number of each pair progressively increases. The contrast relationship amongst pairs in the scale can be graphically demonstrated in terms of the overall ΔE^*_{ab} color difference values between samples [36]. Figure 61 illustrates the ΔE^*_{ab}(using illuminant D65 and 10^o CIE supplemental standard observer) versus gray scale ratings for the AATCC Gray Scale for Color Change as well as the ISO Gray Scale for Color Change [100], illustrating quantitatively the geometric nature of these standard scales.

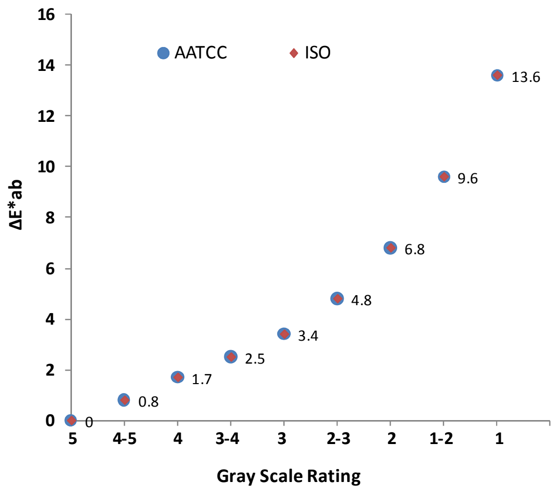


Figure 61. ΔE^*_{ab} (illuminant D65, CIE 10^o supplemental standard observer) versus gray scale ratings for the AATCC gray scale for color change as well as the ISO gray scale for color change.

Scales such as the ISO and AATCC Gray Scale for Color Change are used throughout the world as part of standard methods and evaluation procedures for the quantitative assessment of visual difference between two colored samples, such as textile materials. It is well-known that standard methods involving visual assessment usually produce highly variable results in terms of repeatability and reproducibility. This variability limits the utility of the methods employing these scales and reduces confidence in the data produced.

Hence, any modification to a standard method that enhances reliability of the data produced is of value to the industries that employ the use of these scales. This study describes the development of a new perceptually linear scale which when used in a manner similar to that of the ISO or AATCC Gray Scale for Color Change results in a significant reduction in assessment variability.

The scale can be used for the assessment of perceived magnitude of color difference of chromatic or achromatic stimuli.

4.1. Development of a Perceptually Linear Gray Scale

In a perceptually linear gray scale the difference in contrast between gray pairs increases in a perceptually linear manner. In other words, and using arbitrary units to illustrate the point, the increase in contrast between the gray pairs may progress in 0.5 units i.e.: 0.5, 1.0, 1.5, 2.0…., or in other selected units such as 1, 2, 3, 4…., or 2, 4, 6, 8…. Hence, the contrast (lightness difference) between all adjacent samples in the scale remains constant. In order to develop such a scale for the assessment of small color differences a relatively large number of gray samples with small perceptual differences in lightness were needed. These samples were then shown to a group of color normal observers to select pairs of progressively increasing lightness that on average had equal perceptual differences when compared to an anchor pair - in which the contrast did not vary. While other methods may be developed to establish a perceptually linear scale, one simple method of developing such a scale is described below.

4.1.1. Experimental

4.1.1.1. Selection of Observers

A total of 25 adult observers participated in this part of the study, 13 females and 12 males. Observers ranged from 23 to 56 years of age, and were of a variety of ethnicities. All observers were found to be color normal following a Neitz color vision test [30], and all wore a medium gray laboratory coat throughout the experiment.

4.1.1.2. Samples

For the development of a perceptually linear gray scale, a set of custom-made gray samples was manufactured, from which 31 samples were selected. The samples were developed using PP 684 hi-hide flat deep base flat interior latex paint from Porter Paints [127]. In addition, up to five different pigments were added in varying proportions to modify the lightness of the paint while keeping the sample perceptually achromatic with C^* kept below 1.5 in most cases. The pigments used were universal black, yellow oxide, red oxide, amber, and white. Each paint sample was then applied as a uniform coating on a Leneta un-lacquered plain white paper. The paper had a reflectance value of 80.91% and weighted approximately 18 grams per sheet. An EZ Coater EC-200 (supplied by ChemInstruments) using a rod size wire # 30 bar and a speed of approximately 303 ipm (768cm/m) was used to coat the papers. Figure 62 shows the coating process used to produce the samples.

Figure 62. EZ Coater EC-200 used for the coating and production of gray samples.

After preparation each sample was mounted onto a PVC plastic backing with a thickness of approximately 2 mm and the sample and backing were precision cut to 2”×2” dimensions. Using a Datacolor Spectraflash SF600X spectrophotometer with a large area view (30 mm), and including ultraviolet light and excluding the specular component (SPEX) the reflectance of each sample was measured. Four measurements were made of each sample and the data were averaged. Appendix K shows the reflectance data (D65/10^{o}) for all samples used in the development of the scale.

Each gray sample was prepared to have slight visual difference in lightness when compared to its nearest neighbor. The darkest sample was selected as the ‘standard’. Table 45 shows a summary of colorimetric data for each sample based on D65/10^{o} CIE illuminant/observer conditions, additional colorimetric data based on A and U30 can be found in appendix I. The lightness range was 9 units varying approximately from L* = 41 to L* = 50.

The average ΔE^*_{ab} (D65/10°) between adjacent sample pairs ordered from lowest to highest measured L^*, was 0.34 with a range of 0.09-0.64. The ΔE^*_{ab} between the standard gray and the first sample was 0.64 and 8.74 between the standard and the last sample in the 31 samples used for the development of the scale.

Table 45. Summary of colorimetric data for all gray samples used in the development of the perceptually linear scale.

	D65/ 2° Observer					D65/ 10° Observer				
Sample	**L***	**a***	**b***	**C***	**h°**	**L***	**a***	**b***	**C***	**h°**
Standard	*41.54*	*-0.68*	*-0.50*	*0.85*	*216.16*	*41.55*	*-0.65*	*-0.46*	*0.79*	*215.58*
1	42.16	-0.70	-0.36	0.79	207.22	42.17	-0.66	-0.32	0.73	205.78
2	42.33	-0.75	-0.35	0.83	204.90	42.34	-0.70	-0.30	0.77	203.30
3	42.58	-0.78	-0.63	1.00	219.22	42.60	-0.74	-0.59	0.95	218.21
4	43.00	-0.80	-0.64	1.02	218.96	43.02	-0.76	-0.59	0.97	217.91
5	43.22	-0.75	-0.60	0.96	218.54	43.24	-0.72	-0.56	0.91	217.56
6	43.48	-0.78	-0.64	1.00	219.37	43.50	-0.74	-0.59	0.95	218.56
7	43.86	-0.75	-0.62	0.98	219.51	43.88	-0.72	-0.58	0.92	218.69
8	44.37	-0.79	-0.57	0.98	215.82	44.39	-0.75	-0.53	0.92	214.99
9	44.73	-0.70	-0.52	0.87	216.70	44.75	-0.67	-0.48	0.82	215.64
10	44.99	-0.82	-0.61	1.02	216.54	45.01	-0.78	-0.56	0.96	215.59
11	45.22	-0.80	-0.47	0.92	210.54	45.24	-0.75	-0.43	0.86	209.82
12	45.38	-0.89	-1.07	1.39	230.35	45.42	-0.89	-1.01	1.35	228.78
13	45.61	-0.89	-1.09	1.41	230.63	45.64	-0.89	-1.03	1.36	229.08
14	46.13	-0.92	-1.14	1.46	231.16	46.17	-0.92	-1.08	1.42	229.43
15	46.37	-0.91	-1.11	1.44	230.72	46.41	-0.91	-1.05	1.39	229.00
16	46.79	-0.90	-1.02	1.36	228.57	46.83	-0.89	-0.96	1.31	227.31
17	47.04	-0.80	-0.77	1.11	223.97	47.07	-0.78	-0.72	1.07	222.60
18	47.26	-0.80	-1.03	1.31	232.04	47.29	-0.80	-0.98	1.26	230.75
19	47.59	-0.84	-1.02	1.32	230.66	47.62	-0.83	-0.97	1.28	229.42
20	47.68	-0.84	-1.05	1.35	231.32	47.72	-0.84	-1.00	1.30	229.97
21	47.90	-0.84	-1.03	1.33	230.88	47.93	-0.83	-0.98	1.28	229.75
22	48.39	-0.82	-1.23	1.48	236.35	48.43	-0.83	-1.17	1.44	234.61
23	48.40	-0.83	-1.43	1.66	239.75	48.44	-0.86	-1.37	1.62	237.74
24	48.59	-0.83	-1.66	1.86	243.27	48.65	-0.88	-1.59	1.82	240.98
25	48.69	-0.88	-1.39	1.65	237.60	48.73	-0.90	-1.33	1.61	235.73

Table 45. (Continued)

	D65/ 2° Observer					D65/ 10° Observer				
Sample	**L***	**a***	**b***	**C***	**h°**	**L***	**a***	**b***	**C***	**h°**
26	48.98	-0.84	-1.23	1.49	235.42	49.02	-0.86	-1.16	1.45	233.64
27	49.06	-0.85	-1.18	1.46	234.36	49.10	-0.86	-1.12	1.41	232.73
28	49.25	-0.87	-1.11	1.41	232.03	49.29	-0.87	-1.05	1.36	230.46
29	49.45	-0.85	-0.74	1.13	221.17	49.47	-0.82	-0.69	1.08	219.91
30	49.77	-0.88	-0.73	1.15	219.87	49.8	-0.85	-0.68	1.09	218.60
31	50.25	-0.89	-0.81	1.21	222.20	50.28	-0.87	-0.75	1.15	220.92

4.1.1.3. Visual Assessment Procedure

All samples were placed inside a Macbeth SpectraLight III viewing booth (X-Rite, Inc.) in a 'U' shaped pattern from darkest (left) to lightest (right), as shown in Figure 63, and were illuminated with a filtered tungsten calibrated daylight simulating lamp with a color temperature of 6500 ± 200 K. The observer could see all the samples in the U-pattern at all times. A duplicate set of gray samples was prepared and kept by the experimenter, out of sight of the observer, and these were used to insert a sample into the lower left portion of the E-shaped sample holder shown in Figure 63 and schematically in greater detail in Figure 64.

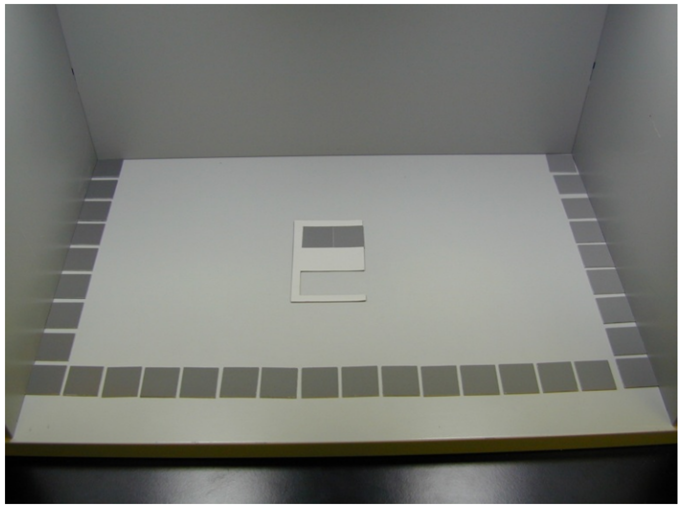


Figure 63. Photograph showing the placement of the gray samples in a ‘U’ shaped pattern inside a Macbeth SpectraLight III viewing booth, and an ‘E’ shaped sample holder in the center.

An anchor pair with a color difference of approximately ΔE^*_{ab} 0.98 was used as a constant visual guide. This pair was placed inside the upper opening of an ‘E’ shaped holder, as shown in Figure 63. The E shaped holder was painted neutral gray to approximately Munsell N 7.25. The distance between the anchor pair and the samples below was set at 2” which was identical to the size of gray samples. Using the duplicate set of gray samples, the experimenter placed the darkest sample (standard from Table 45) in the left hand side of the lower opening of the E shaped holder. Each observer was then asked to place a sample in the lower opening to match the perceived color difference of the anchor pair. No time constraints were imposed and observers were allowed to try as many gray samples from the U-shaped sample arrangement as desired. The instructions given to the observers can be found in Appendix J.

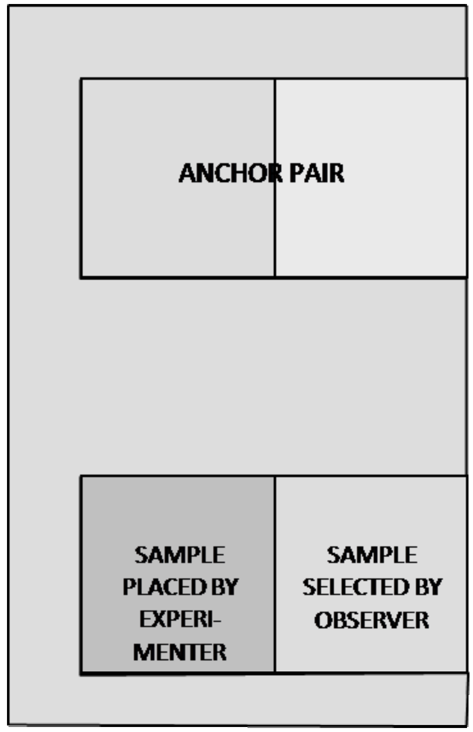


Figure 64. Schematic arrangement of the placement of gray samples in an 'E' shaped sample holder.

Observers were asked to validate their selections at the end of the assessment. All selected samples were placed in front of the observer to determine whether any sample did not belong to a perceptually linear color difference progression within the set. The sequence involved the standard for all observers as the point of departure. The observers were then allowed to replace the non-conforming sample(s) with one from the set that, in their judgment, produced the best linear progression. Each observer repeated the assessment twice on separate days. Validations as well as any changes made were recorded. The results are shown in Table 46.

Table 46. Selections of 10 samples by observers in the two assessments used in the development of perceptually linear gray scale.

	Trial One										Trial Two									
Observer	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
F1	3	6	8	11	14	18	21	24	27	31	3	7	10	14	18	21	23	25	28	31
M1	3	7	8	10	11	12	13	15	17	19	2	3	5	8	10	12	16	18	23	24
M2	2	3	6	8	12	13	14	16	19	21	1	2	3	5	8	10	11	14	15	16
F2	4	8	11	13	14	17	23	26	28	31	2	5	7	11	13	15	18	22	24	27
F3	3	8	11	14	17	20	24	26	29	31	3	6	8	11	15	18	21	23	26	31
M3	3	7	9	13	17	20	23	25	28	30	2	5	9	11	15	19	22	26	28	31
M4	1	3	5	6	9	11	13	14	15	17	2	4	5	7	9	10	11	12	13	14
M5	2	5	9	13	16	19	21	23	27	29	3	6	8	10	13	15	18	20	23	26
M6	3	7	10	12	16	19	21	22	26	30	2	4	8	10	13	16	20	23	26	30
F4	3	5	7	9	12	15	19	22	26	29	4	7	9	11	12	18	21	23	27	29
F5	2	5	9	13	16	19	22	23	27	29	3	7	9	11	14	17	19	22	25	28
F6	4	8	12	14	16	19	23	24	26	27	3	6	8	10	13	16	19	22	26	29
F7	3	9	13	14	15	17	18	20	21	24	2	4	7	10	13	15	17	20	22	24
F8	4	6	9	12	15	18	21	23	24	29	2	4	6	10	13	14	18	22	24	29
F9	1	4	6	9	12	14	16	20	21	30	1	4	7	9	12	14	16	20	22	27
F10	2	4	8	11	12	16	21	24	28	31	1	4	7	10	12	16	21	26	28	31
M7	2	4	7	9	11	17	18	21	25	28	2	6	9	12	15	17	21	23	26	28
M8	3	7	10	13	14	18	23	25	30	31	1	4	8	11	16	17	20	25	26	31
M9	1	4	7	8	10	12	15	18	22	28	2	5	8	11	13	16	19	22	25	29
F11	4	8	13	17	22	24	25	27	29	31	4	6	9	12	15	17	19	23	27	30
M10	1	3	7	10	13	18	22	25	26	27	2	4	8	10	13	16	18	22	27	29
M11	5	10	11	13	15	16	20	23	26	29	4	7	9	13	15	17	20	22	23	26
M12	3	4	8	11	13	17	19	22	24	26	2	4	8	10	12	14	17	19	22	25
F12	3	4	7	10	14	16	18	22	24	26	3	6	8	9	11	16	18	21	22	28
F13	2	6	10	12	15	18	21	23	25	29	3	5	8	10	12	16	20	22	28	30

4.1.2. Data Analysis

Table 47 shows the range of samples selected by observers together with their corresponding color differences expressed in ΔE^*_{ab}.

Table 47. Summary of ranges and ΔE^*_{ab} of samples selected for each selection.

Sample Pairs	**Range**	**Gray Sample Range Selected**	**ΔE^*_{ab} Range**
Selection 1	4	1-5	0.64-1.69
Selection 2	8	2-10	0.81-3.46
Selection 3	10	3-13	1.06-4.14
Selection 4	12	5-17	1.69-5.53
Selection 5	14	8-22	2.84-6.92
Selection 6	14	10-24	3.46-7.19
Selection 7	14	11-25	3.69-7.24
Selection 8	15	12-27	3.91-7.58
Selection 9	17	13-30	4.14-8.25
Selection 10	17	14-31	4.67-8.74

The data from two assessments was used and an average response for each selection was calculated for each observer as well as all observers. Table 48 shows the average, median, and mode for the 10 selections in the two assessments, and the final samples selection validated by the observers in each of the assessments. Median was used for the selection of the final set. In cases where median was a fraction, the closest value to the average was selected.

Table 48. Summary statistics for the 10 selections for all observers.

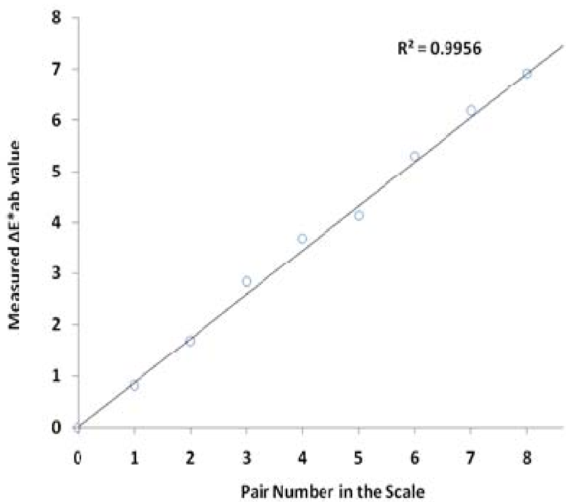
Final Order	Average	Mode	Median	Sample Selected
1	2.52	2.5	2.5	2
2	5.40	5	5.5	5
3	8.24	9	8.5	8
4	10.82	12	11	11
5	13.52	16	13.5	13
6	16.30	16	16.5	16
7	19.14	20.5	20	20
8	21.80	22.5	22.5	22
9	24.52	26	26	26
10	27.52	31	28.5	28

For validation purposes, and to test the reproducibility of the scale, samples selected by observers were reproduced. The largest ΔE^*_{ab} difference between the original and reproduced samples was 0.14 (for the standard) with a range of 0.01-0.14 as shown in Table 49. In the original design of the linear scale, 10 pairs were considered and therefore sample 28 was not included in the final perceptual linear scale and thus a total of 9 samples were selected.

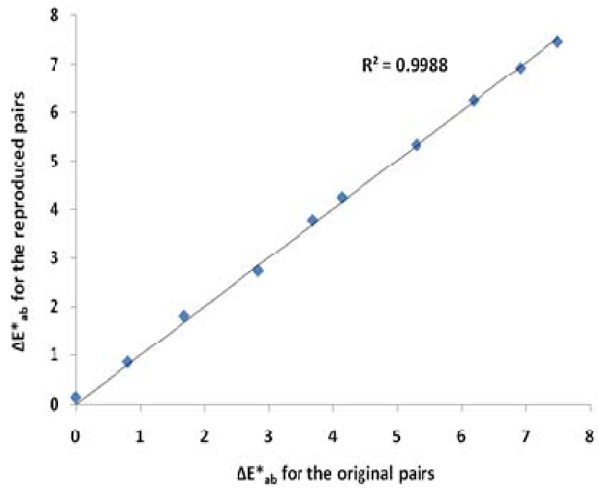
Table 49. ΔE^*_{ab} (D65/10^o) for the reproduction of samples for the final scale.

Selected Sample	ΔE^*_{ab}	ΔE^*_{ab} (reproduced scale)	Difference
Std.	0.00	0.14	0.14
2	0.81	0.86	0.05
5	1.69	1.82	0.13
8	2.84	2.74	0.10
11	3.69	3.79	0.10
13	4.14	4.25	0.11
16	5.30	5.35	0.05
20	6.19	6.25	0.06
22	6.92	6.93	0.01
26	7.50	7.47	0.03

Figure 65a shows the ΔE^*_{ab} values for the original and reproduced samples in the final selection (1-9), against the standard for each set. A correlation value of 0.999 was obtained between the original and the reproduced set, which demonstrates the ability to effectively reproduce the scale. The perceptual linearity of the scale can be seen in Figure 65b. The measured ΔE^*_{ab} of selected set of samples against the pair number produces a linear correlation coefficient value of 0.996. Considering the inherent observer and production variability involved, the progression of color differences for the set of samples selected is highly perceptually linear. Figure 66 shows a photograph of the final set of samples in a prototype scale.



a.



b.

Figure 65a-b. ΔE^*_{ab} Values between the gray sample and the standard within the new scale (a-left), and the correlation of the ΔE^*_{ab} of each pair before and after reproduction of the samples (b-right).

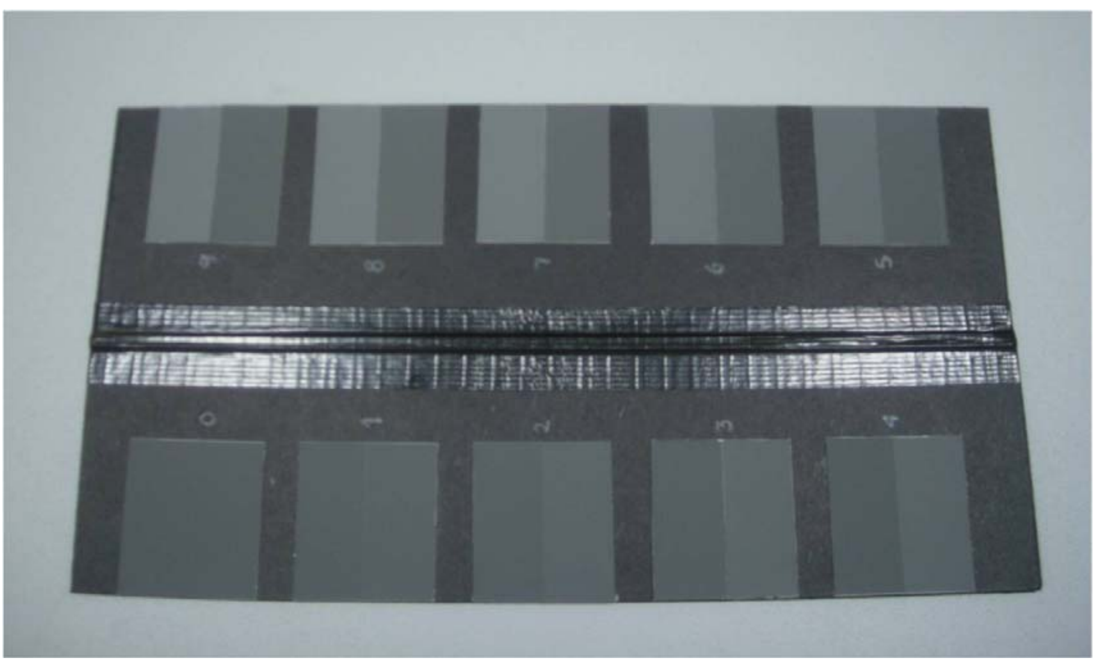


Figure 66. Final arrangement of the gray pairs in the prototype linear scale.

4.1.3. Conclusions

A perceptually linear gray scale for the assessment of perceived magnitude of color difference of chromatic or achromatic stimuli has been developed. The perceptually linear gray scale comprises ten discrete pairs of gray samples that are mounted onto a suitable support structure in a defined order of linearly increasing contrast.

The key aspect of the present study is that this scale has been purposely developed to be perceptually linear as tested under simulated daylight (D65) and a neutral gray background (Munsell N 7.25).

In other words, the gray pairs in the scale are ordered from having no lightness contrast to having very high contrast in a perceptually linear manner, as determined by a statistically valid group of color normal observers. The size and actual physical orientation of the samples as well as the scale may be modified to suit a particular application.

The scale may also be used as a quality control tool in the production and reproduction of colored goods, and as a reference scale in the development of reliable visual datasets for use in colorimetry and color science applications in general.

The perceptually linear scale may be incorporated into visual color difference assessment protocols to improve the repeatability and reliability of the results.

Important and practical examples of the utility of the scale include, but are not limited to, the visual assessment of change in color for:

- dyed fabric before and after a standard washing, rubbing or light fastness test.
- automotive paint coatings before and after exposure to light in a standard test.
- the reproduction of color of cosmetic products compared to a defined color standard.
- the reproduction of color of food and beverage products compared to a defined color standard.
- the reproduction of color of printed packaging compared to a defined color standard.
- the reproduction of color of a photographic print compared to a defined color standard.
- the reproduction of color of a textile sample compared to a defined color standard.
- the reproduction of color of plastics and films compared to a defined color standard.
- the reproduction of colored images on a color projection device such as monitors compared to a defined color standard.

These tests are currently required by companies requiring high quality color control of their products. This scale may also be used to train observers in the assessment of color difference attributes, such as lightness, chroma and hue. Other uses for the scale include application in experimental methods aimed at the development of robust visual datasets for establishment of new mathematical color spaces and/or color difference models.

4.2. Validation of the Perceptually Linear Scale

For the validation of the final linear scale, each selected sample was paired with the standard sample in order of increasing lightness contrast (i.e. increasing ΔE^*_{ab}). An additional pair was prepared comprising two identical standards representing no color difference which was used as the first pair in the scale.

Each observer was asked the following questions:

1- Is this a natural sequence of samples from dark to light? In other words, is there a pair that is misplaced in the sequence? If not, which pair should you switch? ____________.

2- Does the sequence represent a perceptually linear change in contrast? In other words, is the perceptual change in contrast from one pair to the other the same among all pairs in the sequence? If not, which pair should I switch? _____________.

Eighty four percent of observers agreed that the presented linear scale represented a natural sequence from dark to light. While 14 of the 25 observers answered "yes" to the second question, the remaining observers showed no consistency in which pair or pairs needed to be modified. The observers were then asked to assess each sample in the scale by examining each adjacent pair only and very good agreement was obtained. Table 50 summarizes the high percent agreement amongst observers for adjacent pairs in the sequence. The results obtained from this assessment validate the perceptual linearity of the final scale.

Table 50. Agreement amongst observers for adjacent pairs in the sequence.

	Std.	**Std. & 2**	**2 & 5**	**5 & 8**	**8 & 11**	**11 & 13**	**13 & 16**	**16 & 20**	**20 & 22**	**22 & 26**
Total Yes	25	24	24	25	24	23	24	22	24	21
Total No	0	1	1	0	1	2	1	3	1	4
% Yes	100	96	96	100	96	92	96	88	96	84
% No	0	4	4	0	4	8	4	12	4	16

4.3. Visual Assessment of Small Color Differences Using a Novel Perceptually Linear Gray Scale

In practice, most standard visual methods that quantify the magnitude of color difference perception employ gray scales that are perceptually non-linear. This study was carried out to evaluate the application of a novel perceptually linear gray scale in visual assessment of small color differences. This new scale contains ten discrete gray pairs mounted onto a suitable support structure in a defined order of linearly increasing contrast. The development of the scale was discussed in a previous section. The hypothesis was that the use of this scale would enhance the reliability of perceptual small color difference assessments between two stimuli when compared to data obtained using perceptually non-linear scales and reduce inter and intra observer variability. To test this hypothesis the perceptually linear scale was used for the assessment of small color differences and intra- and inter-observer variability data were obtained.

4.3.1. Experimental

4.3.1.1. Samples

Visual assessments were made using the same set of 31 sample pairs (a standard and a batch) dyed with disperse dyes on unbrightened spun polyester knitted fabric used in other parts of this study. Detailed information regarding samples is given in section 2.1.1.

4.3.1.2. Sample Viewing

The sample viewing conditions used in this part of the study were exactly the same as those described in section 2.1.2. Each observer wore a mid-gray laboratory coat and a pair of gray gloves to minimize sample degradation. The samples were placed by the experimenter who also wore a mid-gray laboratory coat. At the beginning of the experiment, the observer was asked to view the empty illuminated viewing booth for 2 minutes to adapt to the light source, during which time the experiment was explained.

4.3.1.3. Psychophysical Method

A novel perceptually linear gray scale was used for the assessment of perceived magnitude of small color differences. The scale was constructed using the methodology described in section 3.1. Visual assessment results using AATCC gray scales discussed in sections 3.2 and 3.3 suggested that the presence of a gap between the scale and the displayed samples decreased intra and inter-observer variability. Therefore a PVC backing measuring 1” x 18” was attached to the back panel of the gray scale which, when placed on the custom made sample stand, created a 2” gap between colored pairs and the gray scale as shown in Figure 67.

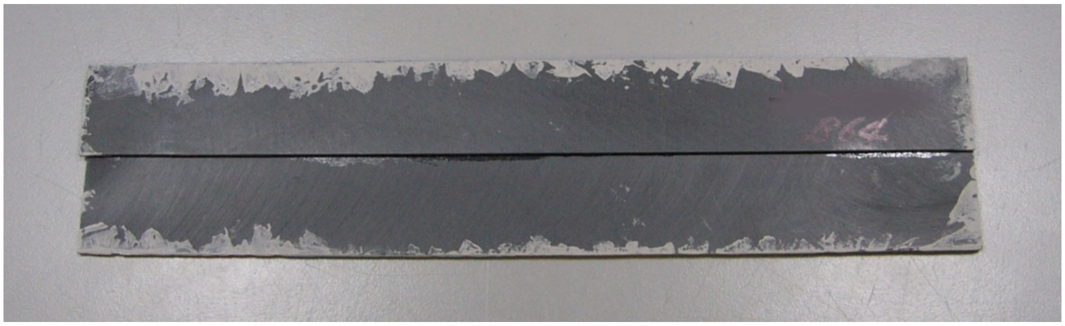


Figure 67. Reverse side of the gray scale.

The size of the gap between samples and the gray scale corresponded with that of the colored samples. Two sets of scales were prepared to facilitate observer assessment without the need to rotate the scale. A full description of the methodology is given in section 3.3. Comparison of results obtained from the different psychophysical methods evaluated in this study is given in section 5.

The scale was used to assess the perceptual differences in color. A rating of 0 represents no color difference between samples and a rating of 9 represents the largest contrast between gray pairs. The contrast between remaining gray pairs increases perceptually linearly in nine steps. For each sample pair the subject was asked: “Which gray scale difference is in closest agreement with the difference between the displayed sample pair? The result can be between two steps, such as 1-2.”

Twenty five naïve observers (13 females and 12 males) mostly students of North Carolina State University, tested for normal color vision using the Neitz test [30], participated in this study. Each observer sat in front of the viewing booth, so that he/she could move the reference gray scale freely. Each observer assessed all samples three times on separate days. The same viewing booth, sample sets, and sample presentation was used in each trial. Figure 68 shows the experimental set up for the present methodology.

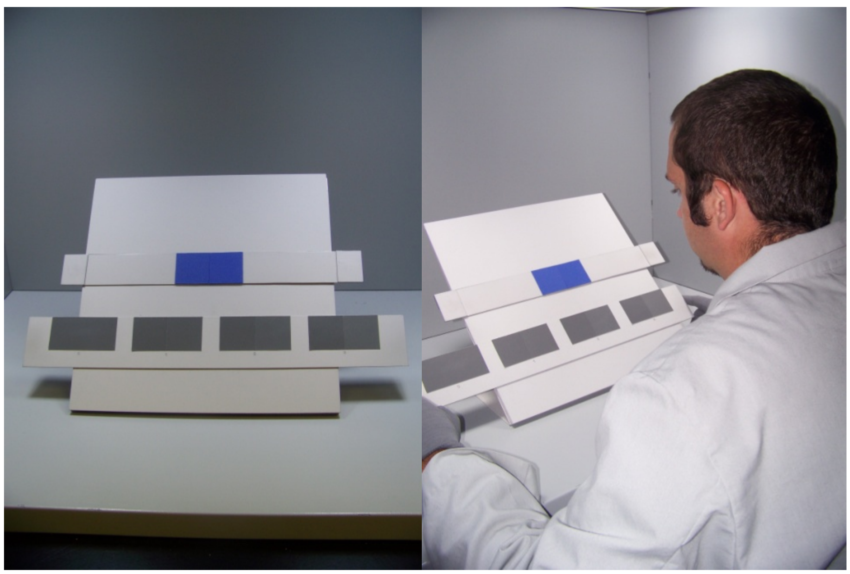


Figure 68. Experiment set up.

4.3.2. Data Analysis

4.3.2.1. Gray Scale Transformations to Visual Difference

A total of 2,325 assessments were made using 31 sample pairs. For the analysis, the raw data in grade units, G, were transformed to a visual difference, ΔV, for each pair using Equation 75.

$$DV = 0.8418\,G + 0.1718 \qquad (75)$$

The equation was obtained by fitting a linear regression between measured ΔE^{*}_{ab} and the novel perceptually linear gray scale grades. The average visual response from all observers was calculated for each sample pair. These values were used in the statistical analysis.

4.3.2.2. Analysis of Variance

The methodology employed in this analysis is described in section 2.2.3 from Chapter III. An analysis of error variance was conducted to identify systematic and random errors. For the analysis of each method, the two sources of error were considered separately.

For the analysis of random errors, a variance components analysis was carried out. For the analysis of systematic errors, intra-observer standard deviation and intra-group standard deviation were calculated.

In addition to the above statistical techniques, PF/3, and STRESS functions were used to analyze and compare the results. A detailed explanation of these methodologies is also given in section II.

4.3.3. Results and Discussion

The raw data in gray scale units was transformed to visual differences according to the methodology explained in section 4.3.2.1. Figure 69 shows a close agreement among three trials in the overall perceived color difference of each pair.

A paired t-test, results of which are summarized in Table 51, was used to evaluate statistical difference among the trials. No statistical difference was found between the first and the second trial; however, a statistically significant difference exists between the second and the third trial as well as the first and the third trial.

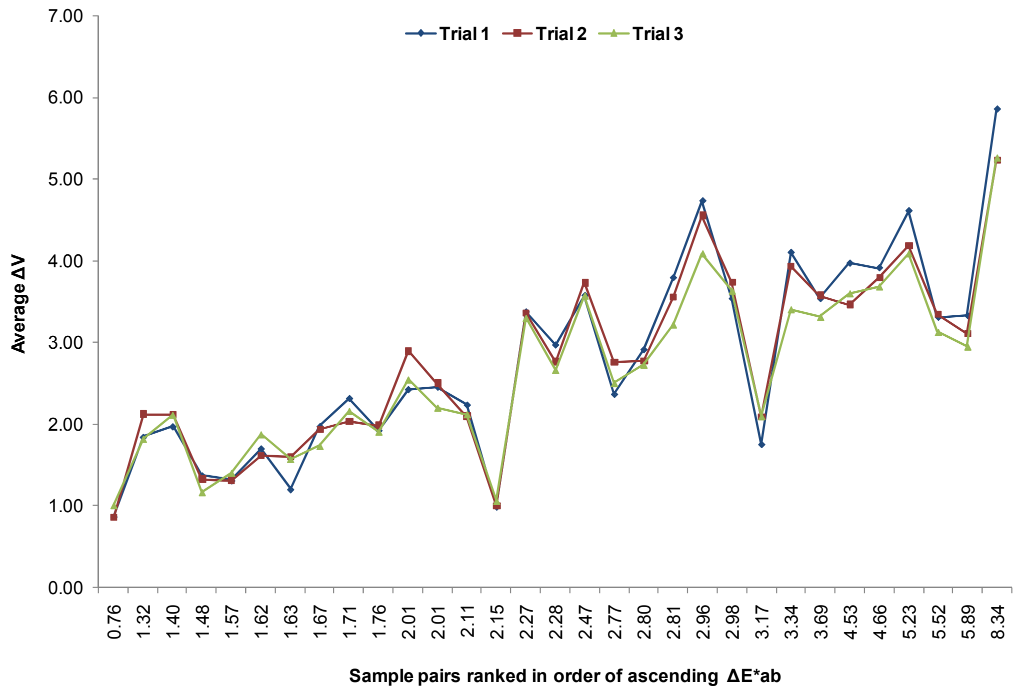


Figure 69. Graph of ΔV for the visual assessments of each sample pair, ranked in ascending order of ΔE^{*}_{ab}.

Table 51. Summary of t-test statistics for assessments carried out by observers.

Group	**t**	**P**	**Significance**
Trial 1 vs. Trial 2	0.63	0.5311	No significant difference at α = 0.05
Trial 2 vs. Trial 3	3.39	0.0020	Significant difference at α = 0.05
Trial 1 vs. Trial 3	2.79	0.0092	Significant difference at α = 0.05

4.3.3.1. Variance Analysis

4.3.3.1.1. Systematic Errors: Intra-Group Variation

To determine the variability of individual observers in comparison to the group average in each trial, intra-group variability was calculated. The average perceived color difference for each pair in each group is assumed to represent the true value, the intra-group variability, is the degree to which each observer's assessments of each pair agrees with the true value of such pair in that trial. Table 52 summarizes the mean standard deviation of the group of observers as an indicator of the intra-group variation among observers for each trial.

Table 52. Summary of intra-group standard deviation for observers.

Observer	Trial 1	Trial 2	Trial 3
Mean	1.11	1.24	1.27
Minimum	0.70	0.78	0.87
Maximum	1.70	2.26	1.92
Range	1.00	1.48	1.05

The results indicate that the mean as well as the minimum intra-group standard deviation increases with the number of trials. However, the range of intra-group standard deviation varied along the trials with no noticeable relationship among trials.

4.3.3.1.2. Random Errors: Variance Component Analysis

Table 53 includes the variance components analysis for the observers. Each entry in the column is a numeric measure of variability. An algebraic combination of the mean squares, using Equations 62-64 from section 2.2.3.2 from Chapter III, was used to estimate the variance of components themselves. These estimates are summarized in Table 53.

Table 53. Variance component analysis for assessments using the perceptually linear gray scale.

Source	DF	Sum of Squares	Mean Square
Pair	30	2721.55	90.72
Observer	24	1160.72	48.36
Pair × Observer	720	1032.16	1.43
Error	1550	880.61	0.57

Table 54. Variance component estimates for assessments using the perceptually linear gray scale.

Source	Estimated Value
Variance due to pairs	1.19
Variance due to observers	0.50
Variance due to interaction	0.29
Variance due to error	0.57

The highest variance shown in Table 54 is due to pairs. This is expected since the ΔE^*_{ab} range for the samples presented to the observers was from 0.76-8.34.

While the variance due to observers is higher than that for the effect of repeated assessments (shown as interaction in Table 54), it is close to the variance due to the error, which essentially comprises all other variables in the system.

4.3.3.2. Intra-Observer Variability (Observer Consistency)

Equation 68 from section 2.2.4 was used to calculate intra-observer variability, which is the deviation of an observer in a trial involving 31 assessments in comparison with his/her own average results from all trials. Table 55 summarizes the intra-observer variability found among observers using this experimental methodology. In addition the total variability for each observer was calculated using the grand average from all trials (93 assessments).

Table 55. Summary of intra-observer standard deviation for observers in all trials.

Observer	Average of Three Trials
Mean	0.62
Minimum	0.33
Maximum	1.22
Range	0.89

Under ideal conditions, observers would reproduce their assessment at different times with a standard deviation of zero. However, studies involving human experience, including color difference perception, usually demonstrate high standard deviation in reproducibility.

These values are used to compare the intra-observer standard deviation for observers based on the use of the linear scale when compared to other methods as shown in section 5.

4.3.3.3. Observer Accuracy Using PF/3 Metric

PF/3 metric was also used to assess observer accuracy. PF/3 values were calculated using the grand mean visual results from all observers for each pair compared to each observer's assessments in each trial. It is assumed that the grand mean visual results are the true value for each pair. An accurate observer is one that agrees closely with the mean visual results from all observers in all trials [12]. PF/3 for accuracy was calculated for observers in each trial and a mean value was obtained. Table 56 summarizes the results of PF/3 for accuracy.

Table 56. Summary of observers' variation in terms of PF/3 for accuracy.

Observer	**Trial 1**	**Trial 2**	**Trial 3**	**Mean Trials**
Mean	50.40	47.47	49.39	49.09
Minimum	30.24	20.07	23.78	27.75
Maximum	90.05	84.00	89.92	83.30
Range	59.81	63.93	66.14	55.55

Similar to the methodologies previously described, the mean PF/3, is slightly higher than those reported by Guan and Luo (PF/3 = 40) [12], Cui et al (PF/3 = 37) [38] Chou et al (PF/3 = 35) [47], Xu et al. and (PF/3 = 33) [45].

However, compared to other methods examined, values based on the linear gray scale method are in closer agreement with the reported values.

4.3.3.4. PF/3 Observer Repeatability

PF/3 of repeatability was calculated for the observers and the mean values for the group as well as each observer were obtained. A summary of results is shown in Table 57.

Table 57. Summary of PF/3 of repeatability for naïve observers.

Observer	Trial 1&2	Trial 2 &3	Mean
Mean	53.82	51.03	52.43
Minimum	20.15	24.03	23.37
Maximum	109.00	82.66	85.27
Range	88.85	58.62	61.90

The mean PF/3 values between trials are consistently high and the repeatability among observers is low. As in the PF/3 results reported for accuracy, the PF/3 repeatability results obtained in this study are significantly higher than those reported by Chou et al. (PF/3 = 27 and 41) [47], Cui et al (PF/3 = 37) [38], Xu et al. and (PF/3 = 26) [45] and Xin et al[43] (PF/3 = 22.1). However, the discrepancy with the reported values is not a high as the values obtained with other methodologies described previously.

Results show that on average observers' repeatability improves with increasing the number of trials for this method as evidenced by a smaller PF/3 range between trials 2 and 3 compared to that for trials 1 and 2. However, as shown in Figure 70, this is probably due to improved repeatability of the most inconsistent observers within the group, although a few observers show a reverse trend.

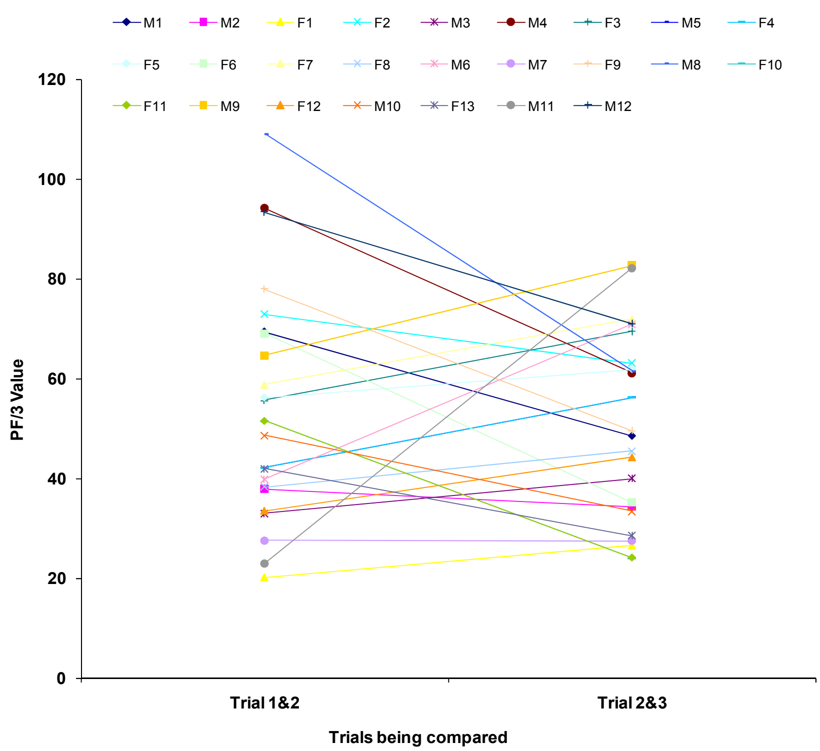


Figure 70. Comparison of the performance of individual observers in terms of PF/3 repeatability among trials.

4.3.3.5. Performance of Color Difference Formulas against Each Visual Dataset

The performance of color difference equations was tested against the average visual data for each of the observer trials. Two color difference models, namely CMC and CIEDE2000 equations, were evaluated at K_L or l setting of 1.Figure 71 shows that the performance of individual models does not change significantly among the trials or based on the use of the grand average of three trials (Total). However, the ΔE^*_{ab} equation performs rather poorly in comparison with the $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models but the $CMC_{(1:1)}$ equation performs slightly better than $CIEDE2000_{(1:1:1)}$ with a agreement.

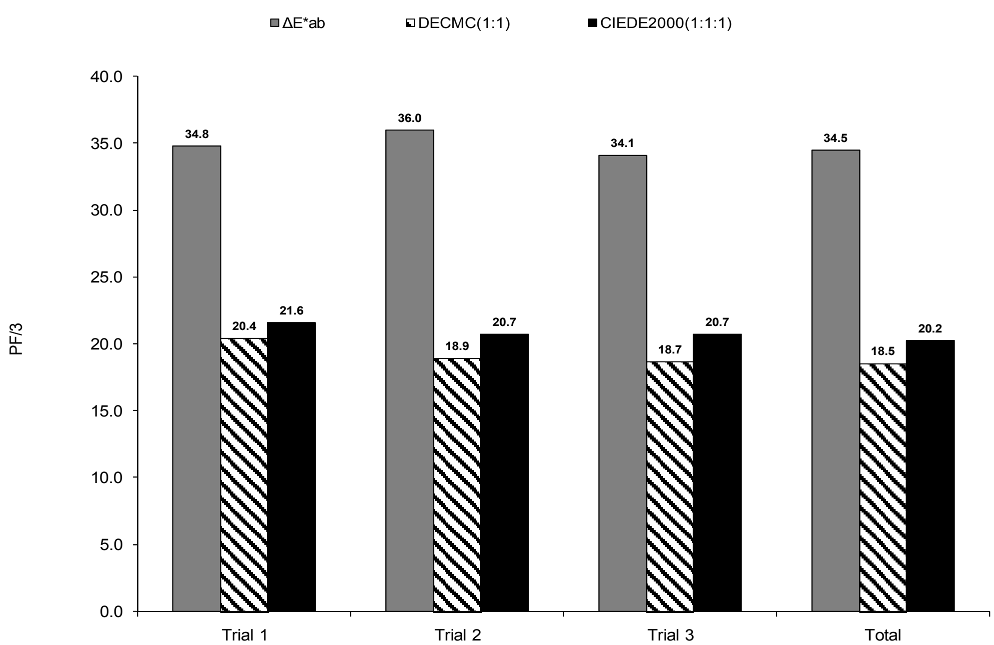


Figure 71. Graph of PF/3 for ΔE^*_{ab}, $CMC_{(1:1)}$, and CIEDE2000 $_{(1:1:1)}$ for the average observers (three trials and total).

4.3.3.6. Performance of Color Difference Formulas against Each Visual Dataset using STRESS

In order to establish the significance of variation between models, the newly developed statistical metric, STRESS, was used. Tables 58, 59, and 60 show the results of calculating the STRESS function and F values for the observers when compared against ΔE^{*}_{ab}, CMC and CIEDE2000 equations at K_L or l setting of 1. Equation 46, from section 6.1.5 of chapter II, can be used to assess the significance of variation between models.

Table 58. The STRESS values for color difference models.

Observer	ΔE^*_{ab}	$CMC_{(1:1)}$	$CIEDE2000_{(1:1:1)}$
Average Trials	0.3128	0.1761	0.1766
Trial 1	0.3069	0.1861	0.1771
Trial 2	0.3285	0.1820	0.1867
Trial 3	0.3103	0.1748	0.1811

Table 59. F values between trials against ΔE^{*}_{ab} formula.

Observer	ΔE^*_{ab}
Trial 1 vs. Trial 2	0.873
Trial 2 vs. Trial 3	1.121
Trial 1 vs. Trial 3	0.978

Table 60. F values between different equations for the average set of observations.

Observer	ΔE^*_{ab} / $CMC_{(1:1)}$	$CMC_{(1:1)}$/$CIEDE2000_{(1:1:1)}$	ΔE^*_{ab}/$CIEDE2000_{(1:1:1)}$
Average Trials	3.15	0.99	3.14

For the methodology employed in this study the calculated critical F values, obtained from STATCRUNCH, for the data are F_C = 0.48 and $1/F_C$ = 2.08 [120]. Table 59 compares the variability of a given trial against another trial using ΔE^*_{ab} formula in order to determine the significance of difference between trials. The results suggest that there are not significant differences between trials. Table 60 compares the performance of color difference equations based on the average observations from three trials. If the calculated F is bigger than $1/F_C$ or smaller than F_C, the difference between two given models is significant. In this case, the comparison between ΔE^*_{ab} and $CMC_{(1:1)}$ shows that for the total number of observations, ΔE^*_{ab} performs significantly poorer than $CMC_{(1:1)}$. In addition, a significant difference is seen in the comparison of ΔE^*_{ab} versus $CIEDE2000_{(1:1:1)}$ equation and since the calculated F value is greater than $1/F_C$ (2.08), it can be concluded that ΔE^*_{ab} performs significantly poorer than $CIEDE2000_{(1:1:1)}$ as a color difference prediction model.

On the other hand, no significant difference between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was observed since the calculated F-values are between 1 and $1/F_C$. Additional detail regarding interpreting calculated F-values is given in section 6 of chapter II.

4.3.4. Conclusions

The present study was carried out to test the variability among observers and the performance of different models when using a novel perceptually linear scale for the assessment of small color differences.

The scale was specifically designed to progressively increase in lightness contrast in a perceptually linear manner under illuminant D65 as a primary reference.

Several statistical methods, including paired t-test, PF/3 and the STRESS functions were used to compare inter and intra-observer variability in visual assessment of colored samples for a group of observers.

PF/3 results show that on average observers' repeatability improves with increasing the number of trials for this method as evidenced by a smaller PF/3 range between trials 2 and 3 compared to that for trials 1 and 2. However, this is probably due to improved repeatability of the most inconsistent observers within the group, although a few observers showed a reverse trend. The mean PF/3 values for repeatability are somewhat higher than those from published results which employ different assessment methodologies (Table 57).

The mean PF/3 values for observer accuracy, and the intra-group standard deviation among trials, do not change significantly from trial to trial (Tables 52 and 56). These results suggest that individual observers' agreement with the mean 'true' is low. Again the mean PF/3 values obtained in this study are higher than those reported previously for unknown reasons (Table 56).

In order to determine the root causes of variability for observers, a variance components analysis was conducted.

Results show that the highest variance is due to sample pairs. This was expected since the range of color differences of sample pairs presented to the observers was 0.76-8.34 ΔE^*_{ab} units. It was also shown that the effect of observer is higher than that due to repetition of assessments (Table 54). Indeed, among the components variability due to the repetition of assessment is the lowest and has no pronounced effect on the overall variability (Table 54). The variance due to the error, which is a combination of all other unknown factors, is the second highest source of variation (Table 54). The analyses show that the intra-observer and inter-observer variability remain fairly constant through the repeated assessments when using the novel perceptually linear scale.

The visual data obtained from observers was also compared against ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ color difference formulas and $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ outperformed ΔE^*_{ab}. According to the PF/3 metric $CMC_{(1:1)}$ shows higher agreement with visual data than $CIEDE2000_{(1:1:1)}$ for the samples tested (Fig 71). The PF/3 values obtained for the color difference models tested are in general agreement with those reported previously.

The significance of variation among the color difference formulas when tested against the observer visual data was also compared using the STRESS function. Results show a significant difference between ΔE^*_{ab} and $CMC_{(1:1)}$ and between ΔE^*_{ab} and $CIEDE2000_{(1:1:1)}$ and show that ΔE^*_{ab} is significantly poorer than both $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ equations. However, no significant difference between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was found.

5. Comparison of Various Psychophysical Methods

The development of an objective color difference formula that accurately represents the average perceptual assessments of observers would be a desirable tool for color quality control of textile materials. Existing formulas are based on several different sets of perceptual data that have been established under various experimental conditions. For instance, visual datasets used for the development of color difference formulas contain a number of variables that include:

• substrates (e.g., wool fabric, sewing thread, and paint chips);

• light sources (e.g., different daylight simulators);

• observer groups (different numbers and expertise i.e., 'naive' no previous experience and some 'experts');

• methods of judgment (e.g., geometric gray scale or anchor pair);

• methods of data analysis (PF/3, Probit, correlation coefficient, etc.).

These factors may affect the degree of accuracy in assessing color differences and may generate statistically significant variability in the dataset produced which in turn may limit the effectiveness of the color difference model upon which the visual dataset is based.

There are numerous variables that affect observer variability. Visual assessment of color, even when conditions are closely controlled, is variable due to the subjective experiences of even well trained expert assessors. Our hypothesis is that one of the main causes of variability lies in the variability among visual assessment protocols used for the establishment and testing of color difference formulas. Therefore, it is critical to ascertain the level of variability in observers and to maintain highly controlled conditions of observation.

To test the hypothesis, a set of experiments was devised to determine the optimum psychophysical experimental method in terms of intra- and inter observer variability. Results from a subset of experiments were first required to determine the psychophysical method resulting in the least variability. These experiments, and results generated, were described in sections 2-4.

This section includes a summary and a comparison of intra- and inter-observer variability results based on the use of naïve vs. expert assessors (using the AATCC standard scale), Jumbo gray scale, Jumbo gray scale with gap, as well as the linear scale developed in this study.

Finally, the highly controlled experimental conditions (for textile samples) that generated the least variability amongst observers are described. This methodology was then used to obtain replicated visual data from different groups of observes in the USA, Latin America (Colombia), Europe (Czech Republic), and Asia (observers of Chinese origin residing in the USA).

5.1. Average Perceived Color Difference Using Different Experimental Methods

Figure 72 illustrates the average perceived visual difference for several sample pairs based on different experimental methodologies. The values for each pair represent the average of observations from twenty-five observers participating in three trials. It was assumed that the mean visual results were the true value for each pair in each particular visual assessment methodology studied. The average visual difference for each pair from each study was compared against each other.

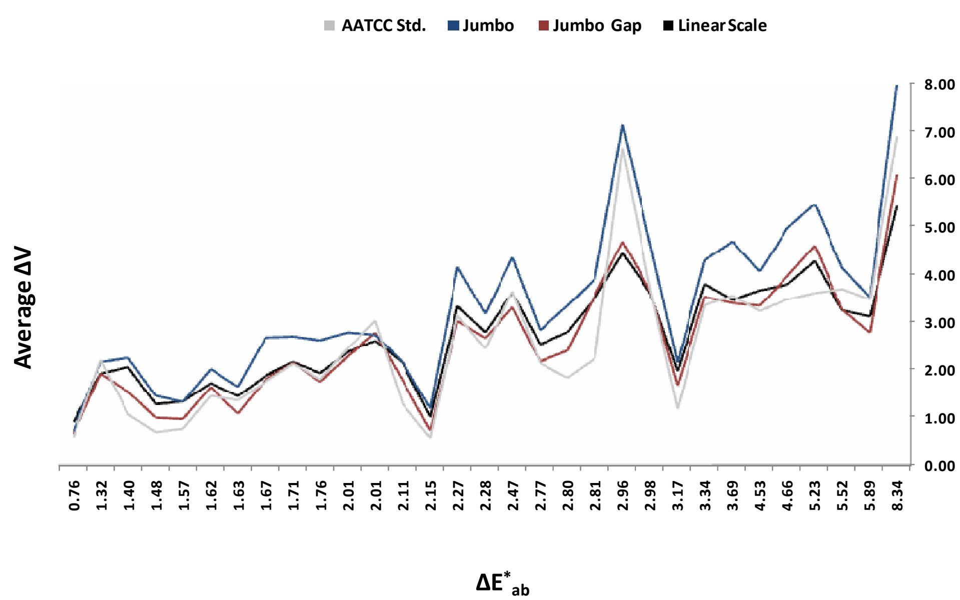


Figure 72. Graph of ΔV of each sample pair, ranked in ascending order of ΔE^*_{ab} for each visual assessment methodology studied.

The Figure shows that for the samples used in this study, observers on average perceived a larger color difference for each pair when using the specially designed Jumbo scale. On the other hand, observers for the most part perceived a smaller color difference between pairs when using the standard AATCC scale. The assessments using the Jumbo scale with a 2” gap between samples and the scale as well as the novel linear scale also employing a 2” gap between samples and the scale seem to produce approximately similar perceived color differences. This supports previous studies that indicate the type of psychophysical methodology employed has a significant role in the magnitude of visual assessments [9, 12, 32].

Figure 73 shows the variance for visual differences amongst observers in each method.

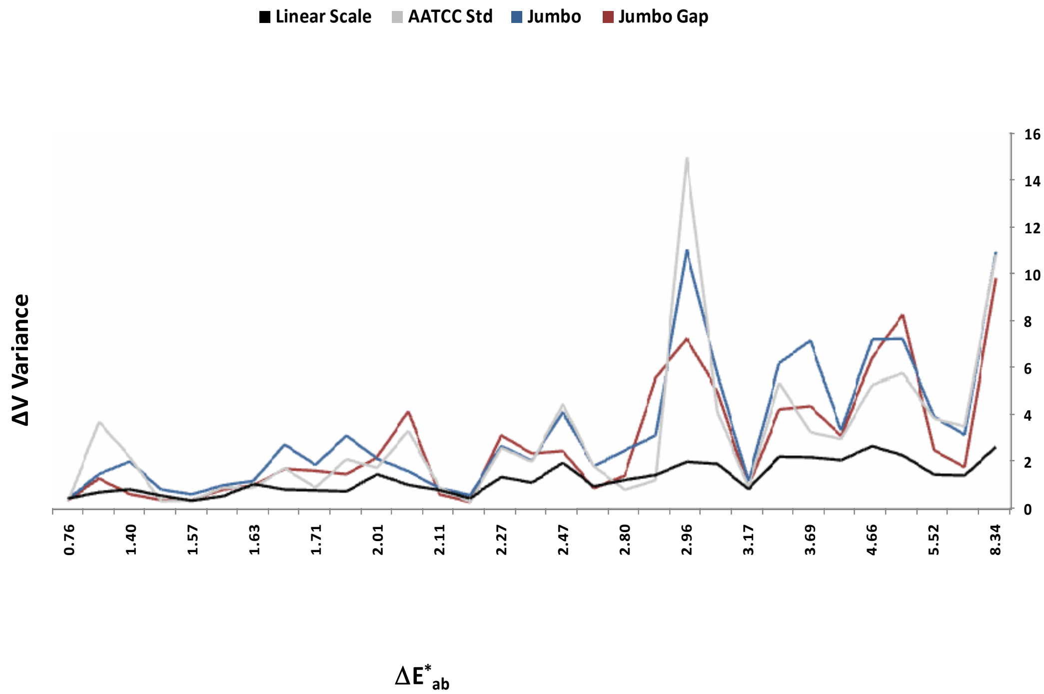


Figure 73. Graph of variance in ΔV of each sample pair, ranked in ascending order of ΔE^*_{ab} for each visual assessment methodology studied.

The figure shows clearly that the variance in the assessment of each of the colored pairs amongst observers decreases when using the novel linear scale developed in this study. The patterns of the visual assessment variance, based on other methodologies shown, are approximately similar and, on average, the magnitudes are larger than those based on the linear scale.

5.2. Intra-Group Variability

Table 61 shows a summary of results from all methodologies employed in this study. The Table shows the mean of individual observers' standard deviation within a group with respect to the overall average of that group. In general, for all the methodologies studied, the range as well as the mean of the intra-group variability changes in magnitude across the trials, which suggest no plausible improvement from trial to trial. In addition, the results show that the novel linear scale produces the lowest intra-group variability (mean and range) among all the methodologies. The experimental Jumbo scale, developed specifically to match the image size of the samples used in this study, generated the largest variability among all methods studied. The separation of samples from the scale improved the agreement amongst observers when the Jumbo scale was used.

Table 61. Summary of intra-group standard deviation for each visual assessment methodology studied.

Method of visual assessment	Summary	Trial 1	Trial 2	Trial 3
Standard AATCC scale	Mean	1.57	1.67	1.46
	Minimum	0.75	0.65	0.89
	Maximum	3.02	5.57	3.87
	Range	2.27	4.92	2.98
Jumbo scale	Mean	1.80	1.67	1.66
	Minimum	0.94	0.94	0.80
	Maximum	4.10	2.83	3.18
	Range	3.06	1.89	2.38
Jumbo scale with gap	Mean	1.53	1.45	1.45
	Minimum	0.48	0.75	0.76
	Maximum	3.93	3.89	3.39
	Range	3.45	3.14	2.63
Novel linear scale	Mean	1.11	1.24	1.27
	Minimum	0.70	0.78	0.87
	Maximum	1.70	2.26	1.91
	Range	1.00	1.47	1.05

5.3. Intra-Observer Variability

Table 62 shows the summary of intra-observer variability results obtained based on the methods studied.

The results show that the use of the Jumbo scale without a gap produces the largest intra-observer variability. This is evidenced by not only the mean but also the range of values obtained within the group of observers.

On the other hand, it is evident that the use of the novel linear scale produces the smallest intra-observer variability among all the methodologies tested in this study. Similar to the intra-group variability, the use of a 2” gap between samples and the scale reduces the variability of observers within trials. The standard AATCC scale shows a large intra-observer variability range when compared to other methods.

Table 62. Summary of intra-observer variability expressed by standard deviations for each visual assessment methodology studied.

Method of visual assessment	Summary	Grand Average
Standard AATCC scale	**Mean**	0.98
	Minimum	0.35
	Maximum	2.46
	Range	2.11
Jumbo scale	**Mean**	1.09
	Minimum	0.33
	Maximum	2.48
	Range	2.15
Jumbo scale with gap	**Mean**	0.77
	Minimum	0.28
	Maximum	1.81
	Range	1.53
Novel linear scale	**Mean**	0.62
	Minimum	0.33
	Maximum	1.22
	Range	0.89

5.4. Variance Component Analysis

Table 63 shows the parameters, i.e. observers; pairs; interaction; and the error: due to the variation that could not be accounted for by other parameters, as well as a summary of the level of variation due to these parameters in the variance component analysis model.

Table 63. Summary of variance component estimates deviations for each visual assessment methodology studied.

Method of visual assessment	**Estimated value of types of variance**			
	pairs	**observers**	**interaction**	**error**
Standard AATCC scale	2.29	0.76	0.61	1.72
Jumbo scale	2.71	0.53	0.80	2.09
Jumbo scale with gap	1.58	0.89	0.76	1.22
Novel linear scale	1.19	0.50	0.29	0.57

It can be seen that the variance due to pairs is the highest within a methodology. As mentioned previously, this is expected since the range of samples' color difference is large. The visual assessments based on the use of the Jumbo scale have the largest variability for nearly all parameters including the 'error'. It can also be seen that the use of a 2" gap in the psychophysical method reduces the variability in all the parameters compared to when a gap was not used. In addition, the visual assessment methodology based on the novel linear scale shows a considerable reduction in variability.

In terms of the observer parameter variability decreases in comparison to other methods tested, however, the magnitude of the reduction is statistically insignificant as determined using the F statistics. The variability due to the interaction parameter also decreases by more than 50% in comparison to all methods studied which is statistically significant. Most importantly, the variation due to unknown factors, or error, decreases by more than 53% in comparison to the other methodologies studied and this reduction is statistically significant.

5.5. PF/3 Accuracy and PF/3 Repeatability

An ‘accurate’ observer is one that agrees closely with the mean (true) visual results of all observers. A low PF/3 of accuracy value means that there is a good agreement between a person’s responses and the true visual results. A ‘repeatable’ observer is one that agrees closely with his/her previous visual assessments. In other words, a low PF/3 repeatability value means that there is a good agreement between responses of an individual obtained in different trials. Previous studies based on the use of a geometric scale have reported lower PF/3 accuracy and repeatability values compared to those obtained in this study as shown in Table 64. However, the results obtained in this study are in good agreement with those reported by Mangine and Lee [94, 121] . In Mangine’s study a different methodology was employed, but the color centers and the dyed substrates used for the construction of samples were the same as those used in this study.

On the other hand, the PF/3 accuracy as well as PF/3 repeatability results based on the use of a linear scale are in closer agreement with previously reported values.

It can be argued that the use of the novel linear scale reduces the PF/3 accuracy and PF/3 repeatability among observers in the visual assessment of small color differences in comparison to other methodologies examined. A systematic assessment of the correlation between PF/3 accuracy and PF/3 repeatability is shown in Section 5.5.1 of this chapter.

Table 64. Summary of PF/3 of repeatability and accuracy for each visual assessment methodology studied.

Visual Assessment Method	Summary	PF/3 Accuracy Average	PF/3 Repeatability Average
Standard AATCC scale	Mean	67.05	72.99
	Minimum	34.33	33.26
	Maximum	113.17	128.81
	Range	78.84	95.55
Jumbo scale	Mean	61.43	70.67
	Minimum	34.97	24.26
	Maximum	102.52	133.87
	Range	67.55	109.61
Jumbo scale with gap	Mean	62.69	60.62
	Minimum	32.82	27.21
	Maximum	115.36	102.94
	Range	82.54	75.73
Novel linear scale	Mean	49.09	52.43
	Minimum	27.75	23.37
	Maximum	83.30	85.27
	Range	55.55	61.90

5.5.1. Correlation between PF/3 Accuracy and PF/3 Repeatability

In order to assess whether there was a statistical relationship among the PF/3 accuracy and PF/3 repeatability results of individual observers, and to determine whether a high correlation between the values existed, a Pearson's correlation technique was employed. A Pearson's correlation coefficient specifies the strength and direction of a linear relationship between two independent variables [111]. A summary of the results for each visual assessment methodology studied is shown in Table 65. The correlation coefficients were calculated using STATCRUNCH[120]®.

Table 65. Pearson's correlation coefficient between PF/3 repeatability and PF/3 accuracy for each visual assessment methodology studied.

Visual Assessment Method	**Correlation coefficient between PF/3 repeatability and accuracy**
Standard AATCC scale	0.80
Jumbo scale	0.83
Jumbo scale with gap	0.86
Novel linear scale	0.76

The results indicate a strong correlation among variables. Although correlation coefficients based on various methods are approximately 0.80, the highest correlation is for values obtained from the Jumbo scale with a 2" gap. While this cannot be fully generalized, this implies that an observer exhibiting high repeatability would also show high accuracy in their assessments.

5.6. PF/3 of Agreement among Visual Results from Different Methodologies

The PF/3 metric can be used to determine the agreement between the results obtained from any two psychophysical methods and was therefore employed to establish the degree of agreement between visual assessments based on the methods examined in this study. Table 66 summarizes the results.

Table 66. PF/3 of agreement among visual results from different methodologies.

Methods of visual assessment	**PF/3**	**PF/3 (1)**
Standard AATCC scale vs. Jumbo scale	33.04	30.31
Standard AATCC scale vs. Jumbo scale with gap	21.42	21.35
Standard AATCC scale vs. Linear scale	29.66	29.16
Jumbo scale vs. Jumbo scale with gap	22.24	24.81
Jumbo scale vs. Linear scale	20.59	22.62
Jumbo scale with gap vs. Linear scale	13.24	13.07

(1) In the calculation of PF/3 the order of X and Y is reversed

A low PF/3 value indicates good agreement between the visual results obtained from two methodologies. Table 66 shows the highest agreement was between results based on the Jumbo scale employing a 2” gap and the novel linear scale, which also employed a 2” gap between the colored pairs and the scale. The size of gray scale pairs in both methods was identical to the size of the colored samples being assessed; however, in contrast to the perceptually linear scale the Jumbo scale was geometric.

On the other hand, Table 66 shows the lowest agreement between the Standard AATCC and the Jumbo scale. These results are not surprising since the psychophysical methodologies differed significantly. One of the shortcomings of the PF/3 metric is that it is asymmetric and therefore the reversal of the variables being compared generates different values. Although the differences are not large, this indicates additional care has to be exercised in the interpretation of PF/3 values.

5.7. Performance of Existing Color Difference Models Using PF/3 Metric

The performance of ΔE^*_{ab}, $CMC_{(1:1)}$, and $CIEDE2000_{(1:1:1)}$ color difference equations, in terms of PF/3, was tested against the averaged data for each of the methodologies examined in this study. Table 67 summarizes the results.

Table 67. PF/3 for ΔE^*_{ab}, $CMC_{(1:1)}$, and $CIEDE2000_{(1:1:1)}$ for each visual assessment methodology studied.

Visual Assessment Method	**PF/3**		
	ΔE^*_{ab}	$CMC_{(1:1)}$	$CIEDE2000_{(1:1:1)}$
Standard AATCC scale	47.63	32.78	27.38
Jumbo scale	36.97	22.92	20.01
Jumbo scale with gap	38.00	21.37	20.09
Novel linear scale	34.47	18.55	20.25
Pair Comparison	42.12	25.00	27.79

The results are in good agreement with previously reported values from studies using similar methodology [8].

It can be seen that for all the psychophysical methods examined ΔE^*_{ab} performs worst among the models tested. The best performance amongst all methodologies examined in this study is based on the use of the linear scale where $CMC_{(1:1)}$ gives the highest agreement with the visual data. The $CIEDE2000_{(1:1:1)}$ equation performs well in the majority of methodologies examined, however, $CMC_{(1:1)}$ also outperforms $CIEDE2000_{(1:1:1)}$ when the pair comparison method was used.

5.8. Performance of Color Difference Models Using STRESS

The STRESS function was used to determine whether the visual assessment results based on various psychophysical methods employed were statistically different. Table 68 compares a given method against another using ΔE^*_{ab} data as reference. As shown, no significant differences are observed amongst data generated from various methodologies employed. Table 69 compares the performance of ΔE^*_{ab}, $CMC_{(1:1)}$, and $CIEDE2000_{(1:1:1)}$ color difference equations based on the average observations from each psychophysical method. The critical values ($F_C = 0.48$ and $1/F_C = 2.08$) were obtained from STATCRUNCH [120]. If the calculated F is bigger than $1/F_C$ or smaller than F_C, the difference between two given models is significant. Table 69 shows that for the majority of cases, ΔE^*_{ab} performs significantly poorer than $CMC_{(1:1)}$ with the exception of the standard AATCC method where the difference between models is statistically insignificant.

In addition, ΔE^{*}_{ab} performs significantly poorer than CIEDE2000$_{(1:1:1)}$ in all cases. On the other hand, no significant differences between CMC$_{(1:1)}$ and CIEDE2000$_{(1:1:1)}$ were observed for any of the methods tested since the calculated F-values are between 1 and $1/F_C$.

Table 68. F values between visual assessment methodologies based on ΔE^{*}_{ab}

Methods Compared	**ΔE^{*}_{ab}**
Standard AATCC scale vs. Jumbo scale	1.26
Standard AATCC scale vs. Jumbo scale with gap	1.40
Standard AATCC scale vs. linear scale	1.38
Jumbo scale vs. Jumbo scale with gap	1.11
Jumbo scale vs. linear scale	1.10
Jumbo scale with gap vs. linear scale	0.99

Table 69. F values between different equations for each visual assessment methodology studied.

Method	**ΔE^{*}_{ab} /CMC$_{(1:1)}$**	**CMC$_{(1:1)}$/CIEDE2000$_{(1:1:1)}$**	**ΔE^{*}_{ab}/CIEDE2000$_{(1:1:1)}$**
Standard AATCC	1.71	1.54	2.64
Jumbo Scale	2.36	1.56	3.68
Jumbo with Gap	3.00	1.26	3.78
Linear Scale	3.15	0.99	3.14

$F_C = 0.48$ and $1/F_C = 2.08$

5.9. Conclusions

This study was conducted to determine an experimental methodology that produced the least variability amongst observers when all other conditions were kept identical.

The methodology would then be used to run an international replication experiment to assess the repeatability of results and ascertain the degree of variability amongst observer groups (if any) based on the methodology developed. The conclusions are summarized below.

The visual assessments based on the use of the Jumbo scale on average produced larger apparent visual color differences for all pairs, whereas assessments based on the use of the Jumbo scale with a 2' gap between samples as well the novel linear scale, also employing a 2' gap, produces similar perceived color differences. In contrast, for the most part smaller apparent color differences for each pair were obtained when the standard AATCC scale was employed. However, the use of the STRESS function to determine the significance of variation amongst results shows that these differences are not statistically significant.

An important observation based on the analysis of results using various methods was that the *variance* in the assessment of each of the colored pairs amongst observers decreased when the novel linear scale developed in this study was used. The patterns of the visual assessment's variance based on other methodologies are approximately similar and, their magnitudes are larger than those of the linear scale.

For all the methods studied, the magnitude as well as the range of the intra-group variability among trials does not follow a pattern suggesting no plausible improvement based on the repetition of trials. The use of the Jumbo scale resulted in the largest variability among methods while the separation of sample from the scale improved observers' agreement.

In addition, the results show that the use of the perceptually linear scale results in the lowest intra-group variability (mean as well as range) among all the methodologies examined.

The use of the Jumbo scale without a gap produces the largest intra-observer variability. Similar to the intra-group variability, the use of a 2” gap between samples and the scale reduces the variability of observers within trials. The standard AATCC scale shows a large intra-observer variability range when compared to other methods. On the other hand, it is evident that the use of the novel linear scale produces the smallest intra-observer variability among all the methodologies tested in this study.

The variance component analysis showed that the variance due to pairs is the highest within a methodology. This was expected since the range of samples’ color difference is large. The visual assessments based on the use of the Jumbo scale have the largest variability for nearly all parameters including the ‘error’. The use of a 2” gap in the psychophysical method reduces the variability in all the parameters compared to when a gap was not used.

The visual assessment methodology based on the novel linear scale shows a considerable reduction in variability. In terms of the observer parameter variability decreases in comparison to other methods tested, however, the magnitude of the reduction is statistically insignificant as determined using F statistics. The variability due to the interaction parameter also decreases by more than 50% in comparison to all methods studied, which is statistically significant.

The variation due to unknown factors, or error, decreases by more than 53% in comparison to the other methodologies studied and this reduction is statistically significant.

Previous studies based on the use of a geometric scale have reported lower PF/3 accuracy and repeatability values compared to those obtained in this study as shown in Table 64. However, the results obtained in this study are in good agreement with those reported by Mangine and Lee [94, 121]. On the other hand, the PF/3 accuracy as well as PF/3 repeatability results based on the use of a linear scale are in closer agreement with previously reported values. It can be argued that the use of the novel linear scale reduces the PF/3 accuracy and PF/3 repeatability among observers in the visual assessment of small color differences in comparison to other methodologies examined. It was also shown, that a high Pearson's correlation coefficient (~0.8) is obtained between PF/3 accuracy and PF/3 repeatability values for individual observers. This implies that repeatable observers are also essentially accurate.

The PF/3 metric was used to determine the agreement between the results obtained from any two psychophysical methods. The highest agreement was between results based on the Jumbo scale employing a 2" gap and the novel linear scale, which also employed a 2" gap between the colored pairs and the scale. The size of gray scale pairs in both methods was identical to the size of the colored samples being assessed. The lowest agreement was found between the Standard AATCC and the Jumbo scale. These results are not surprising since the psychophysical methodologies differed significantly.

The performance of ΔE^*_{ab}, $CMC_{(1:1)}$, and $CIEDE2000_{(1:1:1)}$ color difference equations, in terms of PF/3 and STRESS, was tested against the averaged data for each of the methodologies examined in this study. The results showed that for all the psychophysical methods examined ΔE^*_{ab} performs worst among the models tested and is significantly poorer than $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$. $CMC_{(1:1)}$ gives the highest agreement with the visual data based on the use of the linear scale and the pair comparison method. The $CIEDE2000_{(1:1:1)}$ equation performs best for the remaining methodologies examined. However, a comparison of $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ based on the STRESS function showed that the differences between these models are not statistically significant.

Based on the comparison of results from different experimental methodologies studied, the novel linear scale produced the least variability amongst observers when all other conditions were kept identical.

The linear gray scale was used in the international replication experiment to assess the repeatability of results and ascertain the degree of variability amongst observer groups (if any) based on the methodology developed. This is explained in detail in Section 1, Chapter IV.

IV. Inter and Intra-Observer Variability in an Internationally Replicated Study of Assessing Small Color Differences among Textile Samples

Arguably, the most important area of color management in textiles is the accurate assessment of differences in color between two-textile materials (e.g., a standard and a laboratory or production trial). There is significant variability in judging small color differences between colored samples among human observers. The range and magnitude of variability among observers based on a systematic study, however, has not been reported [9]. Existing visual data, used for the development of color difference formulas, include a variety of experimental settings that prevents an accurate determination of observer variation specific to assessment conditions.

A key aspect of this work was to establish a reliable method for visual assessment of textile samples containing small color differences that produced the smallest variability among a statistically valid set of observers. The statistical variability was established based on results obtained from several psychophysical experiments described in previous sections. It was shown that the use of a novel perceptually linear gray scale in the assessment of small color differences of textiles decreased variability within a set of observers. The experimental conditions based on the use of the linear gray scale were first evaluated in the U.S. with a group of observers.

Once the procedure was established the experiment was repeated under identical experimental conditions with observers in Europe, Latin America, USA (RIT), as well as Chinese observers residing in the USA. The aim was to ascertain, for the first time, the variability in visual data using observers from different regions of the world under highly controlled conditions of observation. This variability may then be used to determine the maximum possible performance of a predictive color difference formula.

1. Experimental

1.1. Samples

Visual assessments were made using the entire set of 69 sample pairs (including 13 standards and 56 batch samples) dyed with disperse dyes on unbrightened spun polyester knitted fabric. Figure 74 shows the location of the entire set of colored samples in a CIE a^*b^* plane. Figure 75 shows the appearance of samples used. Colorimetric data and reflectance data of all samples can be found in Appendices M and N.

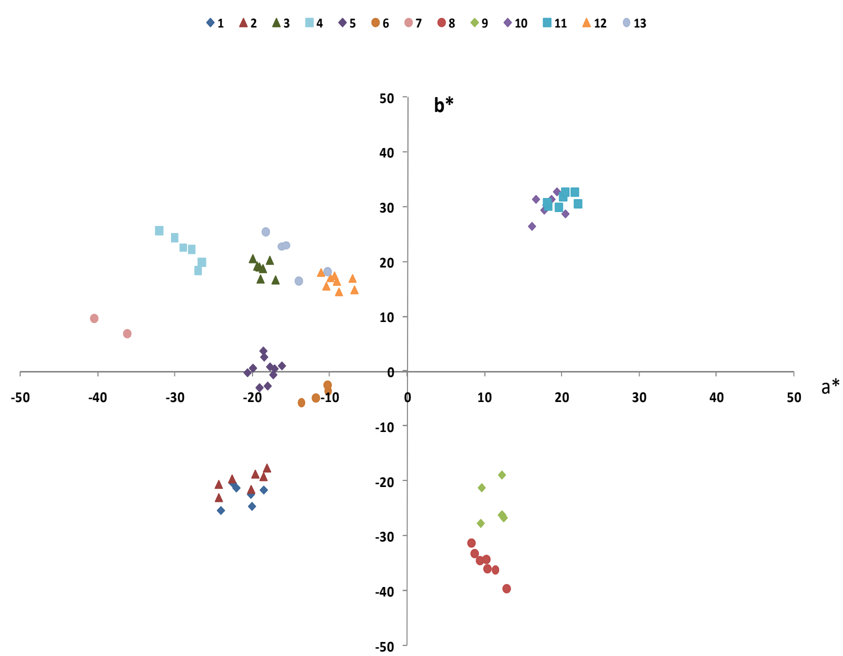


Figure 74. Location of dyed samples in the CIE a^*b^* plane

Figure 75. Appearance of a sample pair used in visual assessments

Each sample was cut to precise 2” × 2” dimensions and mounted onto custom manufactured plastic holders. The sample mountings used and all the components were uniformly spray painted to a L^* of 74, which is approximately equivalent to Munsell N 7.25 to match the interior of a Macbeth SpetraLight III viewing booth. Each sample mounting could slide in a slot on a custom designed display easel, with this setup; sharp dividing lines with no shadows were produced. The reflectance of all samples was measured using a Datacolor SF600X spectrophotometer with specular component excluded (SPEX) and UV included, and colorimetric data were calculated using the CIE illuminant D_{65} and the CIE 10 degree supplemental standard observer. Each measurement was based on an average of four readings. The 69 sample pairs used in the experiment had an average ΔE^*_{ab} of 3.37, with a range of 0.56-7.57.

1.2. Sample Viewing

The sample viewing conditions used in this part of the study were exactly the same as those described in section 2.1.2 from chapter III. Each observer wore a mid-gray laboratory coat and a pair of gray gloves to minimize sample degradation. The samples were placed by the experimenter who also wore a mid-gray laboratory coat. At the beginning of the experiment, the observer was asked to view the empty illuminated viewing booth for 2 minutes to adapt to the light source, during which time the experiment was explained.

1.3. Psychophysical Method

A novel perceptually linear gray scale was used for the assessment of perceived magnitude of small color differences. The scale was constructed using the methodology described in section 3.3 from part III. The identical visual assessment procedure utilized in this study included all experimental conditions described in section 4 from chapter III. Two sets of scales were prepared to facilitate observer assessment without the need to rotate the scale. A full description of the methodology is given in section 4.3 of chapter III.

The scale was used to assess the perceptual differences in color. A rating of 0 represents no color difference between samples and a rating of 9 represents the largest contrast between gray pairs.

The contrast between remaining gray pairs increases perceptually linearly in nine steps. For each sample pair the subject was asked: "Which gray scale difference is in closest agreement with the difference between the displayed sample pair? The result can be between two steps, such as 1-2."

A hundred observers divided into 4 panels representing different regions of the world, as shown in Figure 76, participated in the study.

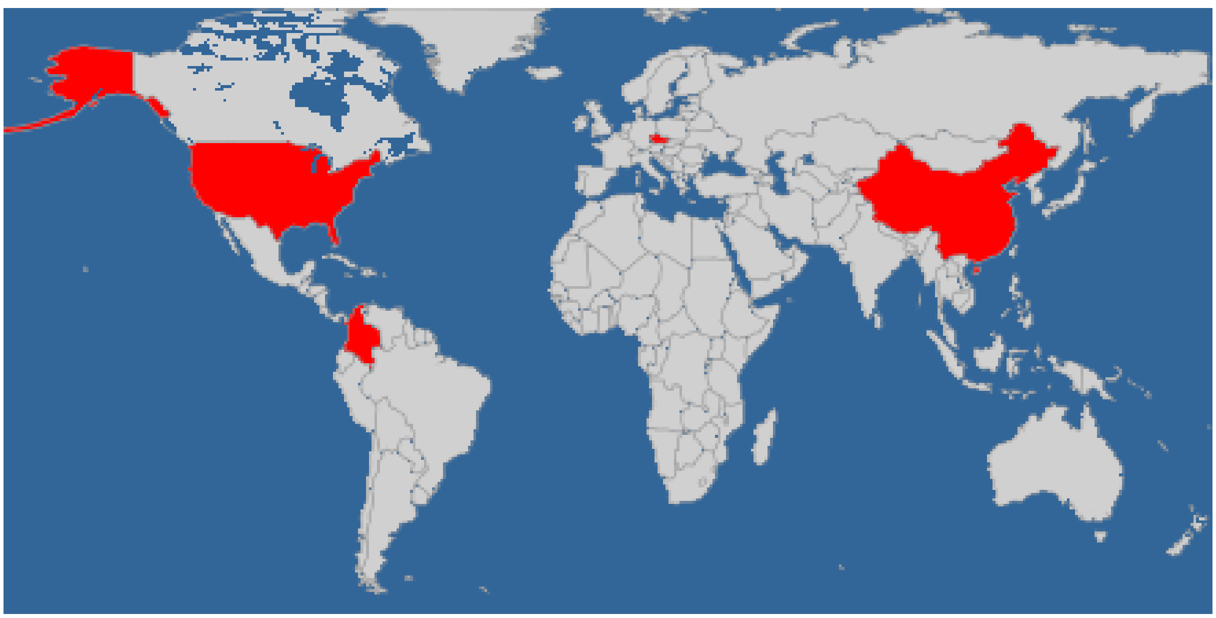


Figure 76. Global regions represented amongst observers participating in this study.

A breakdown of observer groups according to regional representation is given below.

South America: Twenty-five observers participated from Colombia (mostly students of LaSalle College in Bogotá) of which 19 were female and 6 males with an average age of 21 ranging from 16 to 44. All the participants were native Colombians.

Europe: Twenty-five observers participated from the Czech Republic (mostly students of Technical University of Liberec, Liberec) of which 21 were female and 4 males with an average age of 28 ranging from 11 to 64. All the participants were native Czechs and some of the observers had experience in assessing color.

North America: Twenty-five observers participated from the USA (mostly students, faculty and staff from the Color Science group of the Rochester Institute of Technology (RIT), Rochester, New York) of which 10 were female and 15 males with an average age of 34 ranging from 22 to 67. The participants in this panel, however, were from different cultural, ethnic and geographical backgrounds. In addition, some of the observers in this panel also had experience in assessing color and color differences.

Asia: Twenty-five observers represented Asia. These observers were native Chinese residents in the USA and were students, and staff at North Carolina State University of which 13 were female and 12 males with an average age of 30 ranging from 22 to 50.

The observers were tested for normal color vision using the Neitz test [30]. The observer sat in front of the viewing booth, so that he/she could move the reference gray scale freely. Each observer assessed all samples three times on separate days. The same viewing booth, sample sets, and sample presentation was used in each trial and each location. Figure 77 shows the visual assessment setting in each of the locations.

Figure 77. Examples of observers from various locations tested in this study.

2. Data Analysis

2.1. Gray Scale Transformations to Visual Difference

A total of 20,700 assessments were made using 69 sample pairs. Each location accounted for 5175 assessments. For the analysis, the raw data in grade units, G, were transformed to visual difference, ΔV, for each pair using Equation 76.

$$\Delta V = 0.8302G + 0.1122 \quad (76)$$

The equation was obtained by fitting a linear regression between measured ΔE^*_{ab} and the perceptually linear gray scale grades. The average visual response from all observers in each location was calculated for each sample pair. These values were used in the statistical analysis. Results from each location were then analyzed separately.

2.2. Analysis of Variance

The methodology employed in this analysis is described in section II. An analysis of error variance was conducted to identify systematic and random errors. For the analysis of each method, the two sources of error were considered separately.

For the analysis of random errors, a variance components analysis was carried out. For the analysis of systematic errors, intra-observer standard deviation and intra-group standard deviation were calculated.

In addition to the above statistical techniques, PF/3, and STRESS functions were used to analyze and compare the results. A detailed explanation of these methodologies is given in section II.

3. Results and Discussion

3.1. Variance of Pairs

Figure 78 shows the variance in the visual assessment for each colored pair for each observer panel. It can be noticed that variability among observer panels is distinctly different from each other for all pairs. However, no relationship was found between the degree of pair variation and the colorimetric properties of the sample pair assessed.

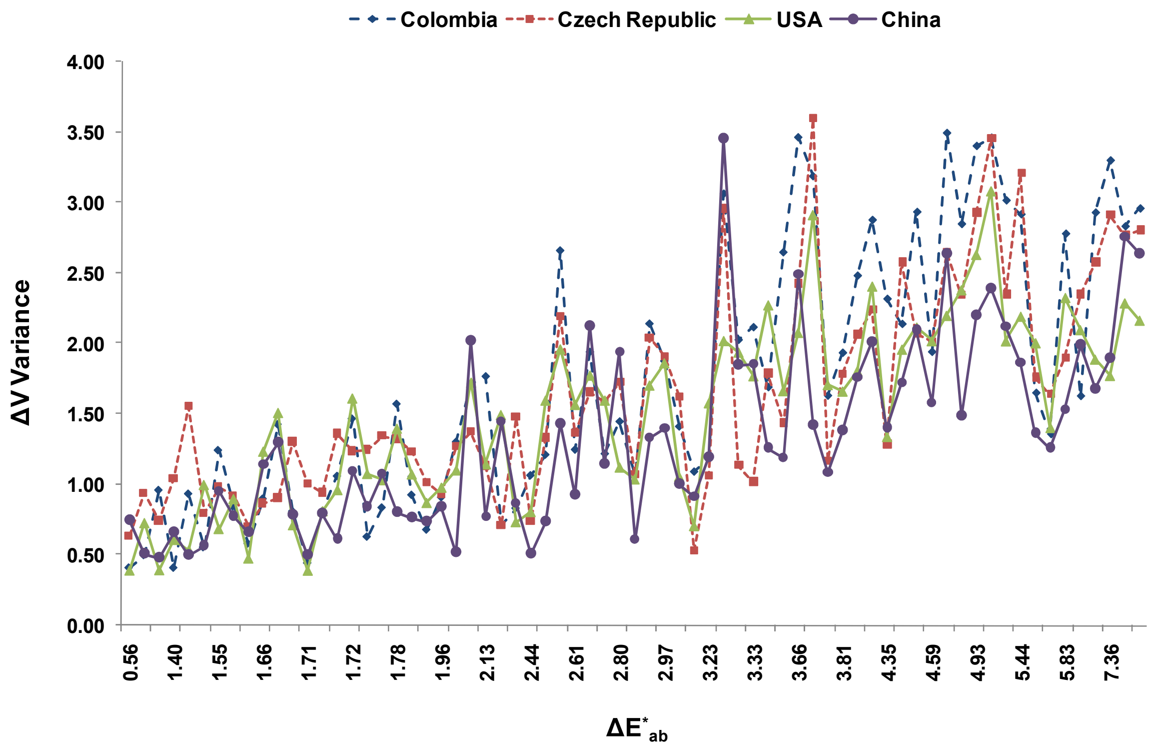


Figure 78. Variance of each colored pair in order ascending ΔE for each observer panel.

3.2. Systematic Errors

3.2.1. Intra-group variation (Observer Accuracy)

To determine the variability of individual observers in comparison to the group average in each trial, intra-group variability was calculated.

Assuming that the average perceived color difference for each pair in each group represents the true value, the intra-group variability, or accuracy, is the degree to which each observer's assessments of each pair agrees with the true value of such pair in that trial. Table 70 summarizes the mean standard deviation as an indicator of the intra-group variation among observers for each trial in each location.

Table 70. Summary of intra-group standard deviation for the observers representing different locations.

Panel of Observers	Observer Summary	Trial 1	Trial 2	Trial 3
Colombia	**Mean**	1.21	1.21	1.33
	Minimum	0.92	0.63	0.75
	Maximum	1.73	2.29	2.73
	Range	0.81	1.66	1.97
Czech Republic	**Mean**	1.09	1.31	1.22
	Minimum	0.72	0.70	0.70
	Maximum	2.22	2.39	1.91
	Range	1.49	1.70	1.21
USA (RIT)	**Mean**	1.14	1.11	1.10
	Minimum	0.70	0.59	0.55
	Maximum	2.50	3.32	2.64
	Range	1.80	2.73	2.09
China	**Mean**	1.13	1.10	1.09
	Minimum	0.68	0.57	0.65
	Maximum	1.89	1.92	1.93
	Range	1.21	1.35	1.28

The intra-group variation range for Colombian observers increases with the number of trials and the mean in the third trial is also increased.

On the other hand, the mean intra-group standard deviation for observers representing China and the USA slightly decreases with the number of trials although the change in the range does not follow a pattern. For observers representing Europe no patterns can be observed based on the repetition of the assessment. The results obtained in these locations are in good agreement with those from observers employed during the development of the linear scale described in section check Chapter III. Overall, although there are differences amongst various observer panels, the mean variation amongst all observer panels is comparable.

3.3. Random Errors

3.3.1. Variance Component Analysis

Table 71 includes the variance components estimates for all the observers for each location. Variance components analysis was carried out for each location separately. Each entry in the column is a numeric measure of variability. An algebraic combination of the mean squares, using Equations 62-64 of section 2.2.3.2 from Chapter III, can be used to estimate the variance of components themselves.

Table 71. The variance component estimates for all the observers for each location.

	Estimated value of variance components			
Panel of Observes	**Pairs**	**Observers**	**Interaction**	**Error**
Colombia	1.17	0.45	0.28	1.03
Czech Republic	1.10	0.49	0.21	0.96
USA	1.12	0.61	0.26	0.67
China	1.04	0.48	0.27	0.60

For all locations, the highest variance shown in Table 71 is due to pairs. This is expected since the color differences of all the sample pairs presented to the observers ranged from 0.56-7.57 ΔE^*_{ab}. While the variance due to observers is higher than that for the effect of interaction it is smaller than the variance due to the 'error', which essentially comprises all other influential factors in the system of which some including observer mood, fatigue, stress, time of day, etc. could influence other parameters including the observer.

Table 72 shows the significance of the difference in estimates among observer panels. The results in Table 71 show that Colombian observers have the lowest variance estimate and USA observers have the highest. However, based on F-statistics the differences in observers' variance are not statistically significant as shown in Table 72. With the exception of responses from Czech observers, the variance due to interaction (observer-colored pair) is comparable among all observer panels. Results from F statistics show that the interaction factor for Czech observers is significantly different from that of other observer panels.

Table 72. Summary of the significance of difference between variance estimates for each observer panel.

Observers	**Significance of Variance (p-value)**			
	Pairs	Observers	Interaction	Error
Colombian vs. Czech	0.40	0.42	< 0.01	0.02
Colombian vs. USA	0.43	0.23	0.07	< 0.01
Colombian vs. Chinese	0.31	0.44	0.23	< 0.01
Czech vs. USA	0.47	0.30	< 0.01	< 0.01
Czech vs. Chinese	0.41	0.48	< 0.01	< 0.01
USA vs. Chinese	0.38	0.28	0.22	< 0.01

Table 71 shows that the Chinese observers had the lowest variance due to error while the Colombian observers had the highest. According to F statistics, all observer panels significantly differ in terms of the 'error' component. As explained previously, the error component contains all the unknown parameters that were not included in the variance components model. These parameters may influence other factors, such as observer, within the model. This indicates that there are probably no inherent differences between individual observers in each group. This study did not include a genetic assessment of participating observers to determine whether differences due to genetics contributed to the overall variability and as such, no further assessment in relation to individual observer's perception of color could be obtained. Therefore, an accurate analysis of the significance of variation of error between observer panels requires a more in depth analysis of psychological, physiological, temporal, spatial and geographical factors that might influence observer assessments. This was beyond the scope of this study.

3.4. Intra-Observer Variability (Observer Consistency)

Equation 68 was used to calculate intra-observer variability, which is the standard deviation of an observer in a trial involving 69 assessments in comparison with his/her own average results from all trials. The total variability for each observer was calculated using the grand average from all trials (207 assessments).The results are shown in Table 73.

Table 73. Summary of intra-observer variability expressed by standard deviations.

Observer Summary	Panel of Observers			
	Colombia	Czech Republic	USA	China
Mean	0.80	0.75	0.66	0.60
Minimum	0.40	0.38	0.31	0.32
Maximum	1.34	1.42	1.16	0.95
Range	0.94	1.04	0.85	0.63

Studies involving human experience, including color difference perception, usually demonstrate high standard deviation in repeatability among observers. Results obtained in this study indicate that the panel of observers from China and the USA had the lowest mean intra-observer standard deviation. On the other hand, Colombian observers showed the highest mean intra-observer standard deviation.

In addition, the results from this study indicate that Chinese observers had the lowest overall intra-observer standard deviation, which means they had a higher agreement with themselves. The Czech observers had the highest overall intra-observer standard deviation, which means that the observers within that group had lower agreement with themselves.

In general, the mean and range of intra-observer standard deviation obtained from the panel of observers representing China and the USA were almost identical to those based on the use of the linear gray scale prior to running the international experiment (see section 4.3 of Chapter III). The mean and range of intra-observer deviation obtained from observers in Colombia and the Czech Republic are slightly higher.

However, the intra-observer variability results based on the use of other gray scale techniques prior to the international experiment are higher in comparison with the values reported here (see section 5 Chapter III). This further supports that the use of the linear scale results in more consistency amongst observers as well as within observer responses in comparison with other methods.

As can be seen the range of variation among the mean standard deviation as well as other parameters shown in Table 73 is small, with the largest variation of 0.50 due to maximum standard deviations. This indicates that overall observations in terms of intra-observer variability among groups are comparable.

3.5. Observer Accuracy using PF/3 Metric

In addition to conventional statistical methods, PF/3 metric was used to assess observer accuracy. PF/3 values were calculated using the grand mean visual results from all observers for each pair, compared to each observer's assessments. PF/3 of accuracy was calculated individually for each group of observers. A low PF/3 of accuracy value denotes a good agreement between a person's responses and the 'true' response. Table 74 summarizes the PF/3 of accuracy for observers from each location.

Table 74. Summary of the PF/3 of accuracy for observer panels compared to their own average.

Observer Summary	Panel of Observers				
	Colombian	**Czech**	**USA**	**Chinese**	**Range**
Mean	82.88	67.20	52.28	57.63	30.60
Minimum	49.05	30.39	25.71	30.25	23.34
Maximum	152.71	120.23	86.02	108.62	66.69
Range	103.66	89.85	60.32	78.37	43.34

The data indicates poor accuracy among all observer panels with observers from the USA displaying relatively the highest accuracy, followed by Chinese observers, Czech and finally Colombian observers who showed the poorest accuracy. Results from the US panel are close to those of a panel of observers tested at North Carolina State University prior to the international experiment (see section 5 Chapter III). This may be due to the varied ethnic background of observers in both panels and the fact that some of the observers in these panels had prior experience in the visual assessment of color.

In general, the PF/3 values shown here (PF/3 ≈ 52-82) are higher than those reported previously, e.g. Guan and Luo (PF/3 = 40), Chou et al (PF/3 = 35), and Kuo and Luo (PF/3= 35-40)[12, 46, 47]. A possible justification for the apparent discrepancy in the data could be the inclusion of outliers in this study. It should be noted that researchers in the area have recommended removing observers who are significantly different from the average of the group. However, since the main objective of this study was to examine observer variability, it was decided to leave all the data points obtained to have a realistic indication of variability.

Pf/3 of accuracy was also calculated using the grand mean visual results (visual assessments from all locations) for each pair. Table 75 summarizes the results of accuracy calculated using the grand average of visual results from all locations.

Table 75. Summary of the PF/3 of accuracy for observer panels compared to grand average visual assessments from all locations.

Observer Summary	Panel of Observers				
	Colombian	**Czech**	**USA**	**Chinese**	**Range**
Mean	87.36	67.38	52.30	57.97	35.06
Minimum	45.86	32.53	26.91	31.36	18.95
Maximum	162.30	117.75	87.58	106.96	74.72
Range	116.44	85.22	60.67	75.60	55.77

The data indicates that the mean accuracy among all observer panels increases when individual results are compared with the average assessments from all observer panels. It is also interesting to note that the USA observers' results do not change significantly when compared against their own or the grand average results as shown in Tables 74 and 75. This may be due to the fact that the US observer panel comprised a variety of ethnicities and age groups, while other panels were ethnically biased. Since the grand average is a more accurate representation of the mixed global population it might therefore correlate with the results of the US panel better than any of the other panels.

3.6. PF/3 Observer Repeatability

In order to assess observer repeatability between trials the PF/3 measure was calculated for each observer in each observer panel. The mean value for each observer panel was also obtained. A summary of results from each location studied is shown in Table 76.

Table 76. Summary of PF/3 repeatability results between Trials from various observer panels.

Observer Summary	**Panel of Observers**							
	Colombian		**Czech**		**USA**		**Chinese**	
Trials	**1 & 2**	**2 & 3**	**1 & 2**	**2 & 3**	**1 & 2**	**2 & 3**	**1 & 2**	**2 & 3**
Mean	106.39	91.43	90.55	83.73	62.35	50.71	66.44	56.32
Minimum	36.16	32.93	35.11	26.77	35.35	29.11	21.95	23.67
Maximum	182.63	155.68	221.34	273.66	113.34	88.14	103.93	112.82
Range	146.46	122.75	186.23	246.89	77.99	59.03	81.98	89.15

The results show that, in general, the mean PF/3 values for all observer panels are consistently high indicating that observer's repeatability is poor. Colombian observers showed the lowest repeatability, followed by Czech observers, while the Chinese and the USA observers showed approximately similar repeatability. In addition, the results indicate that, for all observer panels, the mean repeatability improves between trials 2 and 3 compared to trials 1 and 2 as evidenced by the smaller PF/3 values.

In general, the mean PF/3 repeatability values of all observer panels are higher than those reported previously by Chou et al.[47] (PF/3 = 27 and 41); Cui et al (PF/3 = 37) [38], Xu et al. and (PF/3 = 26)[45] and Xin et al[43] (PF/3 = 22.1). Once again, a possible justification for the discrepancy in the data could be explained by the inclusion of outliers since all the answers obtained in this study were used in the statistical analysis.

3.7. Performance of Color Difference Formulas against Visual Datasets from each Observer Panel

The performance of color difference equations was tested against the average visual data for each of the observer panels. Two color difference models, namely CMC and CIEDE2000 equations, were evaluated at K_L or *l* setting of 1.

In addition, Figures 79 and Table 77 show that the performance of individual models does not change significantly among the trials or compared to the grand average of three trials for each observer panel. The PF/3 data among all the locations is comparable. The results are in good agreement with previously reported values from similar studies.

The ΔE^*_{ab} equation performs rather poorly in comparison with the $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models for all observer panels which is not unexpected.

The performance of $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ equations for the visual data obtained from the Chinese, USA and Czech panel of observers is comparable. Although, $CIEDE2000_{(1:1:1)}$ performs slightly better than $CMC_{(1:1)}$ for the Colombian observers, the improvement is marginal.

Table 77. Summary of PF/3 of ΔE^*_{ab}, $CMC_{(1:1)}$, and $CIEDE2000_{(1:1:1)}$ equations compared against visual data obtained from various observer panels.

Panel of Observers	**PF/3 metric for color difference models**		
	ΔE^*_{ab}	$CMC_{(1:1)}$	$CIEDE2000_{(1:1:1)}$
Colombia	32.08	23.22	22.90
Czech Republic	30.15	20.00	20.82
USA	29.93	19.54	20.14
China	29.83	20.92	21.66
Range	2.65	3.92	3.12

The accuracy of prediction of average perceived difference based on the formulas evaluated in this study is higher than those reported previously [8] The RIT-Dupont dataset is the closest in agreement with the results obtained in this study [8, 9].

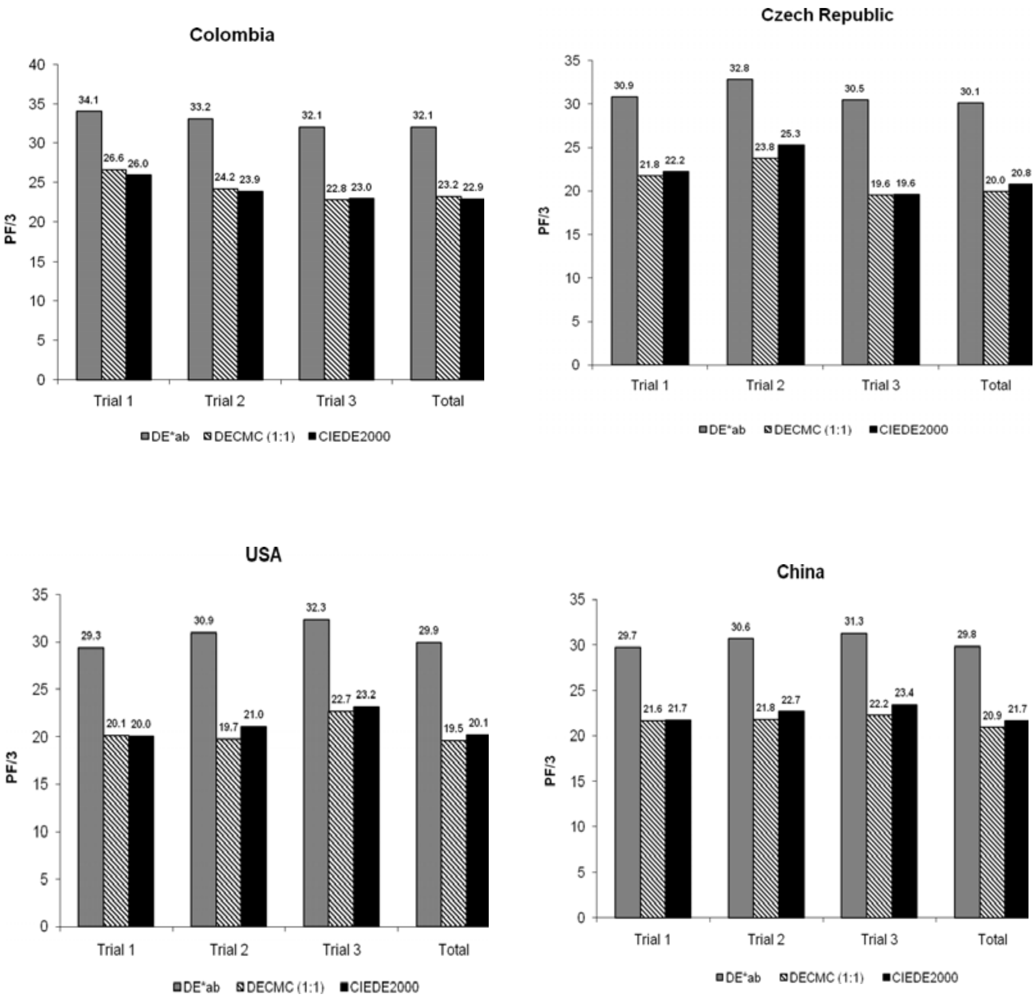


Figure 79.Graph of PF/3 for ΔE^*_{ab}, $CMC_{(1:1)}$, and CIEDE2000 $_{(1:1:1)}$ each panel of observers.

3.8. PF/3 of Agreement among Visual Results from Different Observer Panels

PF/3 metric was used to determine agreement among the results obtained from different observer panels. Table 78 summarizes the results.

Table 78. PF/3 of agreement between average visual data for each panel compared against other observer panels.

Panel of Observers	PF/3 Values of Agreement
Colombia vs. Czech Republic	19.38
Colombia vs. USA	21.18
Colombia vs. China	19.75
Colombia vs. Total Population	14.61
Czech Republic vs. USA	6.84
Czech Republic vs. China	9.67
Czech Republic vs. Total Population	5.82
USA vs. China	9.38
USA vs. Total Population	6.81
China vs. Total Population	6.82

A lower PF/3 value indicates good agreement between the visual results obtained from two observer panels. Table 78 shows the highest agreement is between the Czech Republic and the USA panels where the agreement between the datasets is high, followed by the agreement between the Chinese panel and the USA observers as well as that with the Czech observers.

On the other hand, the agreement between Colombian observers and other observer panels is lower. The reasons for this discrepancy are unknown. A possible reason is that Colombian observers had no prior experience in any psychophysical experiment of this nature while some of the observers in the remaining panels were experienced in studies pertaining to visual assessments.

3.9. Performance of Color Difference Formulas against Each Visual Dataset using STRESS

In order to establish the significance of variation between models, the newly developed statistical metric, STRESS, was used. Table 79, show the STRESS function values for ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ equations at K_L or l setting of 1 for all observer panels. Table 80 shows the F values comparing the significance of difference between two observer panels based on ΔE^*_{ab}. Table 81 shows F values comparing the significance of difference between predictive color difference equations for each set of observers respectively. Equation 46, from section 6.1.5 of chapter II can be used to assess the significance of variation between models.

Table 79. The STRESS values for color difference models using the average of three trials for each observer panel.

Observer Panels	**ΔE^*_{ab}**	**$CMC_{(1:1)}$**	**CIEDE2000$_{(1:1:1)}$**
Colombia	0.2984	0.2394	0.2162
Czech Republic	0.2736	0.2076	0.1951
USA	0.2715	0.1981	0.1855
China	0.2607	0.1967	0.1855

For the methodology employed in this study the calculated critical F values, obtained from STATCRUNCH, are $F_C = 0.62$ and $1/F_C = 1.61$ [120]). If the calculated F is bigger than $1/F_C$ or smaller than F_C, the difference between two given models is significant. Table 80 compares the variability of a given panel of observers against another panel of observers using ΔE^*_{ab} formula in order to determine the significance of difference between observer panels. The results suggest that there are not significant differences between observer panels.

Table 80. F values based on a comparison of observer panels against ΔE^*_{ab} formula.

Observer Panels	**ΔE^*_{ab}**
Colombia vs. Czech Republic	1.19
Colombia vs. USA	1.15
Colombia vs. China	1.25
Czech Republic vs. USA	0.97
Czech Republic vs. China	1.05
USA vs. China	1.09

$F_C = 0.62$ and $1/F_C = 1.61$

Table 81 compares the performance of color difference equations based on the average observations from three trials for each observer panel. In this case, the comparison between ΔE^{*}_{ab} and $CMC_{(1:1)}$ shows that for the total number of observations, ΔE^{*}_{ab} performs significantly poorer than $CMC_{(1:1)}$ for visual assessments obtained from Colombia, Czech, USA, and Chinese observers.

Table 81. F values between different equations for each set of observers.

Observer Panels	**ΔE^{*}_{ab} /$CMC_{(1:1)}$**	**$CMC_{(1:1)}$/$CIEDE2000_{(1:1:1)}$**	**ΔE^{*}_{ab} /$CIEDE2000_{(1:1:1)}$**
Colombia	1.93	1.17	2.25
Czech Republic	2.34	1.00	2.34
USA	2.48	1.01	2.51
China	2.25	1.00	2.24

$F_C = 0.62$ and $1/F_C = 1.61$

In addition, a significant difference is seen in the comparison of ΔE^{*}_{ab} versus $CIEDE2000_{(1:1:1)}$ equation and it can be concluded that ΔE^{*}_{ab} performs significantly poorer than $CIEDE2000_{(1:1:1)}$ as a color difference prediction model for responses obtained from all observer panels.

On the other hand, no significant difference between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was observed since the calculated F-values are between F_C and $1/F_C$. The results show that there are no significant differences between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models based on the responses obtained from various observer panels.

3.10. Confidence intervals for each pair for the observers panels

Confidence intervals were calculated for the average assessments obtained for each pair from each observer panel. Due to the large number of colored pairs (69) studied, samples were divided into groups of 10 pairs and plotted in order of ascending ΔE^{*}_{ab} against visual difference as shown in Figures 80-86. Plots show the confidence intervals for each pair from each observer panel.

Figure 80 shows that with the exception of three pairs circled in red, assessed by Colombian observers, observations from various panels overlapped at a margin of 95% confidence. Figures 81, 82, and 86 also show overlapping assessments among observer panels with the exception of only one pair in each graph. These "isolated" assessments were again obtained from the Colombian observers. Figures 83, 84, and 85, show overlapping assessments among observer panels for each of the pairs assessed with no "isolated" responses. Therefore, on average a good agreement amongst observations was obtained.

In general, judgments by a given panel do not follow a pattern in relation to other observer panels. These results, in addition to previous results discussed in this chapter, show that the overall average perceived color difference for each colored pair is within the margin of error for observer panels representing Czech Republic, USA, and China and with some exceptions Colombia.

Interestingly, for the most part Colombian observers judged samples to have smaller color differences compared to assessments obtained from the remaining observer panels. Nevertheless, the results from Colombian observers remain in good agreement with those of other panels. A possible reason for this observation could be that Colombian observers had no prior experience in any psychophysical experiment of this nature while some of the observers in the remaining panels were experienced in studies pertaining to visual assessments.

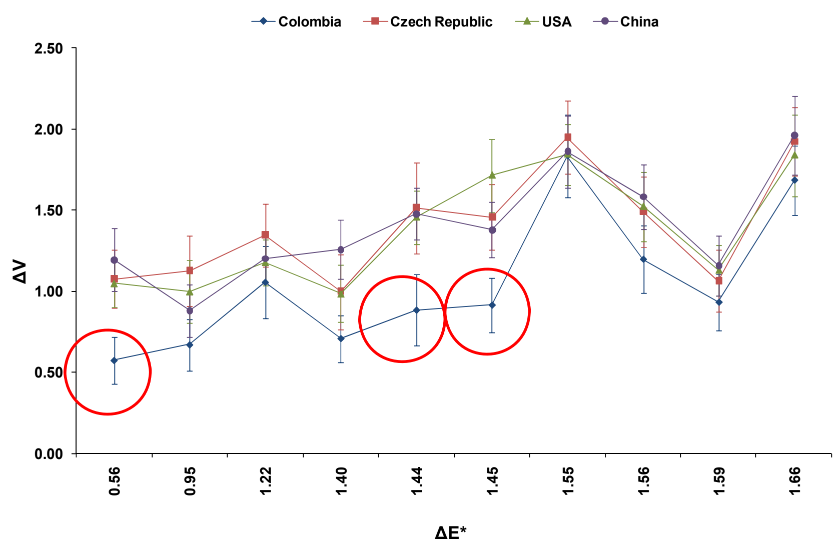


Figure 80. Confidence intervals for each pair (ΔE^{*}_{ab} range 0.56-1.66) for the observers panels.

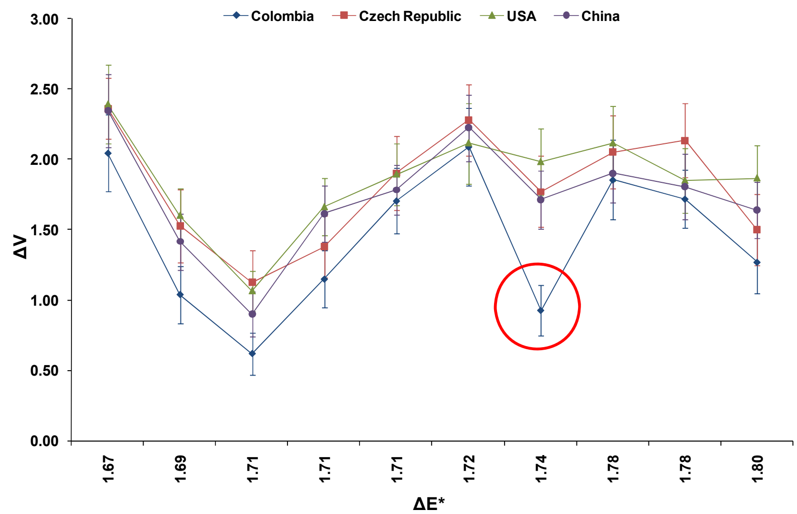


Figure 81. Confidence intervals for each pair (ΔE^*_{ab} range 1.67-1.80) for the observers panels.

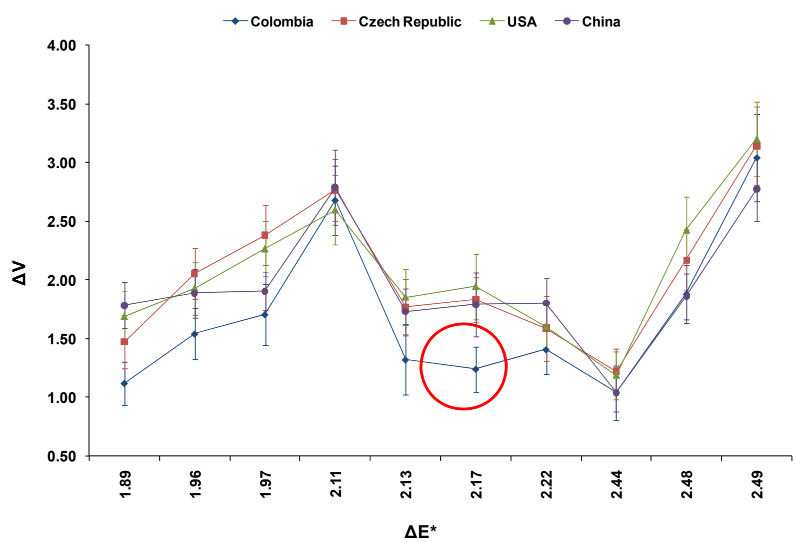


Figure 82. Confidence intervals for each pair (ΔE^*_{ab} range 1.89-2.49) for the observers panels.

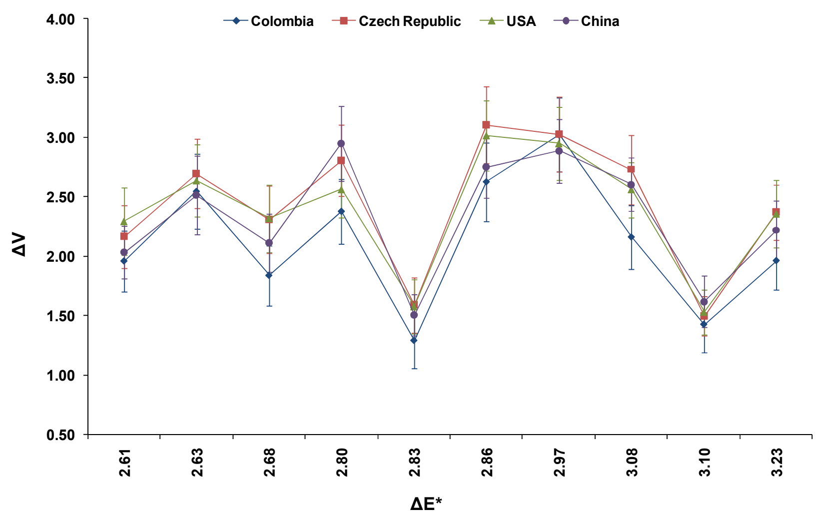


Figure 83. Confidence intervals for each pair (ΔE^*_{ab} range 2.61-3.23) for the observers panels.

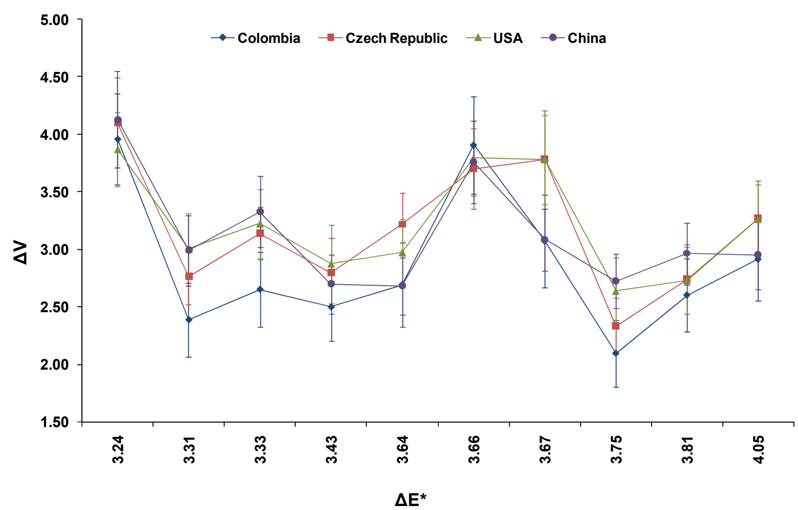


Figure 84. Confidence intervals for each pair (ΔE^*_{ab} range 3.24-4.05) for the observers panels.

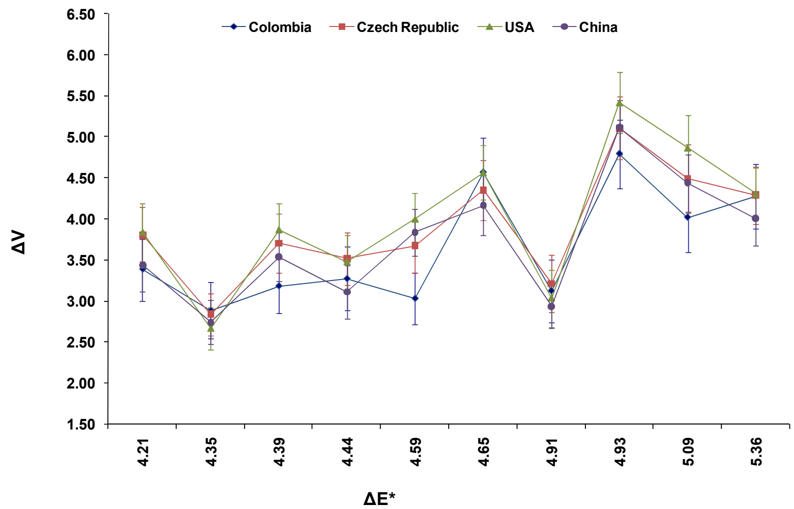


Figure 85. Confidence intervals for each pair (ΔE^*_{ab} range 4.21-5.36) for the observers panels.

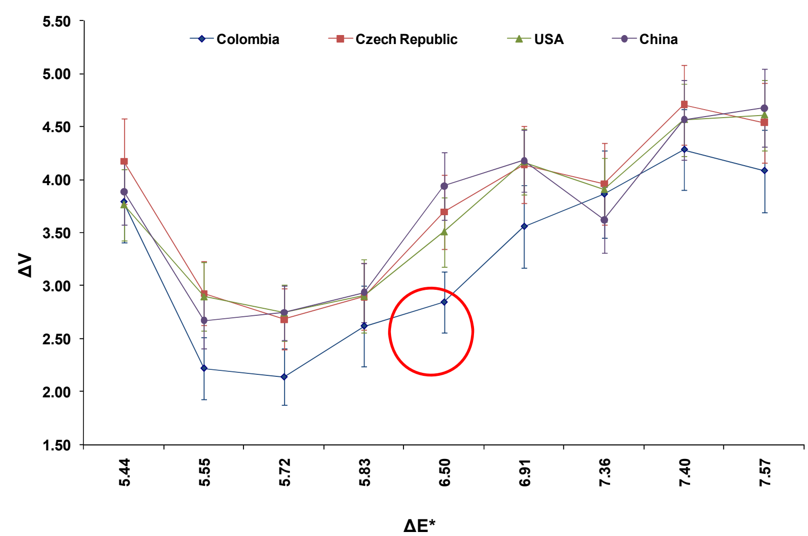


Figure 86. Confidence intervals for each pair (ΔE^*_{ab} range 5.44-7.57) for the observers panels.

3.11. Confidence Intervals for each Pair for all the Observers' Panels Combined

Confidence intervals were calculated for the average assessments obtained for each pair from a total of 300 observations. The number of observations included all the assessments from all the observer panels. Table 82 shows the results for all the 69 sample pairs.

Table 82. Confidence Intervals for each pair for all the observers'' panels combined.

Pair	**ΔV**	**confidence**	**L. limit**	**U. limit**
1	3.344	0.172	3.172	3.516
2	3.905	0.181	3.724	4.085
3	1.982	0.128	1.854	2.109
4	3.078	0.170	2.908	3.248
5	3.617	0.175	3.442	3.792
6	3.100	0.161	2.939	3.262
7	2.759	0.147	2.611	2.906
8	1.669	0.126	1.543	1.795
9	2.716	0.150	2.566	2.866
10	2.229	0.128	2.101	2.356
11	4.222	0.174	4.048	4.396
12	0.927	0.089	0.838	1.016
13	1.519	0.102	1.417	1.621
14	1.878	0.118	1.760	1.996
15	1.703	0.122	1.582	1.825
16	1.367	0.102	1.265	1.468
17	1.071	0.088	0.983	1.160
18	1.123	0.100	1.023	1.223
19	2.841	0.165	2.676	3.007
20	3.430	0.192	3.238	3.622
21	4.456	0.201	4.254	4.657
22	3.840	0.178	3.662	4.017
23	1.448	0.106	1.342	1.554

Table 82. (Continued)

Pair	ΔV	confidence	L. limit	U. limit
24	1.602	0.112	1.489	1.714
25	1.854	0.110	1.744	1.964
26	1.872	0.111	1.761	1.983
27	1.452	0.106	1.346	1.557
28	2.891	0.150	2.740	3.041
29	3.791	0.182	3.609	3.974
30	4.414	0.188	4.226	4.602
31	2.785	0.151	2.634	2.937
32	2.288	0.129	2.160	2.417
33	1.196	0.091	1.104	1.287
34	5.110	0.190	4.920	5.300
35	2.680	0.150	2.530	2.830
36	2.446	0.136	2.310	2.582
37	3.499	0.166	3.333	3.666
38	3.634	0.160	3.474	3.795
39	0.987	0.095	0.892	1.083
40	3.086	0.149	2.936	3.235
41	1.518	0.106	1.411	1.624
42	2.580	0.137	2.443	2.717
43	4.533	0.185	4.348	4.718
44	4.016	0.191	3.825	4.207
45	1.333	0.110	1.223	1.442
46	1.599	0.119	1.480	1.718
47	1.492	0.110	1.381	1.602
48	2.781	0.142	2.639	2.923
49	2.179	0.131	2.048	2.310
50	2.972	0.150	2.823	3.122
51	1.397	0.110	1.287	1.506
52	1.569	0.115	1.454	1.684
53	1.822	0.113	1.709	1.935
54	2.114	0.128	1.986	2.242
55	2.085	0.127	1.958	2.212
56	3.041	0.163	2.878	3.203

Table 82. (Continued)

Pair	ΔV	confidence	L. limit	U. limit
57	2.876	0.153	2.723	3.029
58	0.919	0.094	0.825	1.013
59	2.146	0.134	2.011	2.280
60	2.676	0.143	2.533	2.818
61	1.851	0.115	1.736	1.967
62	2.709	0.147	2.562	2.857
63	2.516	0.129	2.387	2.645
64	2.065	0.119	1.945	2.184
65	2.599	0.154	2.445	2.754
66	0.973	0.087	0.885	1.060
67	4.014	0.172	3.842	4.186
68	4.479	0.185	4.294	4.664
69	3.574	0.166	3.408	3.740

3.12. Performance of Color Difference Formulas against Visual Datasets from all observers' panels

The performance of color difference equations was tested against the grand mean visual results (visual assessments from all locations). Two color difference models, namely CMC and CIEDE2000 equations, were evaluated at K_L or *l* setting of 1.

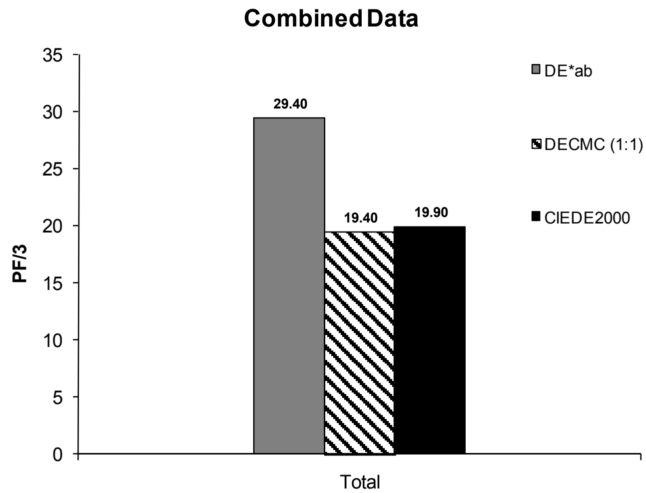


Figure 87. Graph of PF/3 for ΔE^*_{ab}, $CMC_{(1:1)}$, and $CIEDE2000_{(1:1:1)}$ for the combined data.

The ΔE^*_{ab} equation performs rather poorly in comparison with the $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models using the grand mean visual from all observer panels, which is not unexpected. $CMC_{(1:1)}$ performs about the same as $CIEDE2000_{(1:1:1)}$ for the grand mean visual results from all observers.

In addition, the performance of color difference equations was tested against the confidence Interval's lower and upper values for each of the colored pairs. Table 83 shows the PF/3 values for each of the formulas evaluated. The results show a variation in PF/3 of less than a unit with a 95% confidence level.

Table 83. PF/3 values for different equations using the lower and upper limit of the confidence interval from the combined data

	ΔE*ab	**$CMC_{(1:1)}$**	**$CIEDE2000_{(1:1:1)}$**
Lower limit	29.37	19.21	19.56
Upper limit	29.54	19.67	20.29
Range	0.17	0.46	0.73

3.13. Performance of Color Difference Formulas against the grand mean visual from all observer panels using STRESS

In order to establish the significance of variation between models, the newly developed statistical metric, STRESS, was used. Table 84 shows the STRESS function values for ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ equations at K_L or l setting of 1 for the grand mean visual results from all observer panels. Table 85 shows F values comparing the significance of difference between predictive color difference equations for all the evaluated models. Equation 46 from section 6.1.5, chapter II, can be used to assess the significance of variation between models.

Table 84. STRESS values for color difference models using the average from the combined data of all observer panels.

STRESS	**ΔE^*_{ab}**	**$CMC_{(1:1)}$**	**$CIEDE2000_{(1:1:1)}$**
Total population	0.2548	0.1652	0.1610

For the methodology employed in this study the calculated critical F values, obtained from STATCRUNCH, are $F_C = 0.62$ and $1/F_C = 1.61$ [120]. If the calculated F is bigger than $1/F_C$ or smaller than F_C, the difference between two given models is significant.

Table 85. F values between different equations using the average from the combined data of all observer panels

	ΔE^*_{ab} / $CMC_{(1:1)}$	$CMC_{(1:1)}$ / $CIEDE2000_{(1:1:1)}$	ΔE^*_{ab} / $CIEDE2000_{(1:1:1)}$
Total population	2.38	1.05	2.50

Table 85 compares the performance of color difference equations based on the grand average visual results from all the observers' panels. In this case, the comparison between ΔE^*_{ab} and $CMC_{(1:1)}$ shows that for grand average, ΔE^*_{ab} performs significantly poorer than $CMC_{(1:1)}$.

In addition, a significant difference is seen in the comparison of ΔE^*_{ab} versus $CIEDE2000_{(1:1:1)}$ equation and it can be concluded that ΔE^*_{ab} performs significantly poorer than $CIEDE2000_{(1:1:1)}$ as a color difference prediction model for responses obtained from all observer panels.

On the other hand, no significant difference between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was observed since the calculated F-values are between F_C and $1/F_C$. The results show that there are no significant differences between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models based on the responses obtained from all observers' panels. Additional detail regarding interpreting calculated F-values is given in section II.

In addition, the performance of color difference equations was tested against the confidence Interval's lower and upper values for each of the colored pairs. Table 86 shows the STRESS values for each of the formulas evaluated. The results show a variation in STRESS of less than 0.1 with a 95% confidence level

Table 86. STRESS values for different equations using the lower and upper limit of the confidence interval from the combined data

	ΔE*ab	**CMC $_{(1:1)}$**	**CIEDE2000$_{(1:1:1)}$**
Lower limit	0.2539	0.1640	0.1518
Upper limit	0.2559	0.1668	0.1643
Range	0.0020	0.0028	0.0125

.

4. Conclusions

The present study was carried out to systematic test the range and magnitude of variability among different panels of observers. For the first time, the variability in visual data using observers from different regions of the world under highly controlled conditions of observation was evaluated. The visual assessment was carried out using a controlled method for visual assessment of small color differences in textile samples based on the use of a novel perceptually linear gray scale. The experimental conditions were first evaluated in the U.S with a group of observers. The visual assessment was carried out under identical experimental conditions with observers from Colombia, Czech Republic, USA (RIT), and a group of Chinese students studying in the USA.

Several statistical methods, PF/3 and the STRESS functions were used to compare inter and intra- observer variability in visual assessment of colored samples for each group.

The results show that the average assessments incorporating the confidence intervals from observer panels for each of the pairs evaluated overlaps among panels with some exceptions for the Colombian observers. Therefore, on average a good agreement amongst observations was obtained. The judgments by a given panel do not vary systematically from those in other observer panels. Interestingly, for the most part Colombian observers judged samples to have smaller color differences compared to assessments obtained from other observer panels.

A possible reason for this observation could be that Colombian observers had no prior experience in any psychophysical experiment of this nature while some of the observers in the remaining panels had some experience in studies pertaining to visual assessments. Nevertheless, the results from Colombian observers are in good agreement with those of other panels. In addition, PF/3 metric was used to determine the agreement in visual data among observer panels. Results show that with the exception of Colombian observers all panels are in good agreement with the highest agreement between the Czech Republic and the USA panels, and the lowest between Colombian observers and the USA observers. However, it can be argued that all panels demonstrate remarkable agreement considering the perceptual nature of the study. In addition, it can be seen that the agreement between individual panels and the total observer population increases for all observer panels. This suggests that observer panels based on a mixed ethnic makeup may be better suited for investigations that involve human perceptions. Additionally, results of a STRESS function based on ΔE^*_{ab} formula show that there are no significant differences between observer panels.

In a separate observation it was noted that variability among observer panels in relation to the assessment of individual color pairs is distinctly different. Due to the large number of colored samples used and the nature of the assessments, further investigation to elucidate the underlying reasons for this apparent deviation was not conducted. In addition, no correlation between the observed variation and the magnitude of measured difference for the sample pairs was found.

Although there are differences in variability for each pair amongst observer panels, the standard deviation of responses for an observer in any panel from the mean response of that panel is not significantly different amongst panels.

The variance component analysis showed that, for all panels, the highest variance was due to pairs (See Table 70). This was expected since the color differences of all the sample pairs ranged from 0.56-7.57 ΔE^*_{ab}. While the variance due to observers is higher than that for the effect of interaction it is smaller than the variance due to the 'error', which essentially comprises all other influential factors in the system. The results in Table 70 also show that Colombian observers have the lowest observer variance estimate and USA observers have the highest. However, based on F-statistics the differences in observers' variances are not statistically significant. With the exception of responses from Czech observers, the variance due to interaction (observer-colored pair combination) was comparable among all observer panels. F statistics showed that the interaction factor for Czech observers is significantly different from that of other observer panels. Chinese observers had the lowest variance due to error while the Colombian observers had the highest. According to F statistics all observer panels significantly differ in terms of the 'error' component. The error component contains all the unknown parameters that were not included in the variance components model. These parameters may also influence other factors, such as the observer component within the model.

This indicates that an accurate analysis of the significance of variation of error between observer panels requires a more in depth analysis of psychological, physiological, temporal, spatial and geographical factors that might influence observer assessments. This was beyond the scope of this study.

Results from intra-observer standard deviation show that Chinese and the USA observers had the lowest mean values while Colombian observers had the highest. Low intra-observer variability implies a high degree of agreement amongst responses of individual observers. Colombian observers therefore had the lowest agreement with themselves while the Chinese observers had the highest, although all values are relatively low.

A comparison of the intra-observer variations with those based on the use of other gray scale techniques, prior to the international experiment, shows that intra-observer variability for all observer panels is within the range of variability based on the linear gray scale as shown in section 5 of Chapter III. This further supports that the use of the linear gray scale improves observer consistency in comparison with other methods.

PF/3 Accuracy (PF/3 ≈ 52- 82) and Repeatability (PF/3 ≈ 52- 87) values obtained in this study were higher than those reported previously, e.g. Guan and Luo[12] (PF/3 = 40), Chou et al (PF/3 = 35)[47], and Kuo and Luo[46](PF/3= 35-40). A possible justification for the apparent increase in the PF/3 values could be the inclusion of outliers in this study.

Since the main objective of this study was to examine observer variability it was decided to process all the data points obtained to generate a realistic representation of variability and this can reduce accuracy and repeatability among observers. In terms of accuracy the ranking for observer panels was USA > China > Czech Rep > Colombia. Interestingly the results from the US panel are close to those based on an observer panel tested at North Carolina State University prior to commencing the international replication experiment (see section 5.5 Chapter III).

In terms of repeatability, Colombian observers showed the lowest repeatability, followed by Czech observers, while the Chinese and the USA observers showed approximately similar performances. In addition, the results showed that for all observer panels the mean repeatability improves between trials 2 and 3 compared to trials 1 and 2 as evidenced by the smaller PF/3 values. In general, the mean PF/3 repeatability values of all observer panels are higher than those reported previously by Chou et al. (PF/3 = 27 and 41)[47]; Cui et al (PF/3 = 37) [38], Xu et al. and (PF/3 = 26)[45] and Xin et al [43](PF/3 = 22.1). Once again, a possible justification for the discrepancy in the data could be explained by the inclusion of outliers in the statistical analysis.

The performance of color difference equations was tested against the average visual data for each of the observer panels using PF/3 and STRESS functions. Two color difference models, namely CMC and CIEDE2000 equations, were evaluated at K_L or l setting of 1. The PF/3 values for all panels are within a range of 4 units.

Results showed that ΔE^*_{ab} equation performs rather poorly in comparison with the $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models for all observer panels which is not unexpected. The performance of $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ equations for the visual data obtained from the Chinese, USA and Czech panel of observers is within a range of one PF/3 unit. Although, $CIEDE2000_{(1:1:1)}$ performs slightly better than CMC for the Colombian observers, the improvement is marginal (<1 unit). A comparison of ΔE^*_{ab} vs. $CMC_{(1:1)}$ and ΔE^*_{ab} vs. $CIEDE2000_{(1:1:1)}$, based on STRESS function showed that ΔE^*_{ab} performs significantly poorer than $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ for all observer panels. No significant difference between $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models was observed based on the responses obtained from various observer panels. In addition, when the visual data from all observer panels was combined the performance of color difference models, based on PF/3 and STRESS, against the combined dataset did not change significantly compared to that based on individual panels. Moreover, the inclusion of upper and lower confidence levels on to the visual data does not change the PF/3 values significantly (<1 unit) while the change in STRESS values is also not significant (<0.1)

V. Summary of Conclusions

This study was conducted to investigate the intra and inter- observer variability in the visual assessment of small color differences. A key aspect of the study was the development of a methodology with the objective to produce the least variability amongst observers when all other conditions were kept constant. The methodology was used to assess, for the first time, the repeatability of results and ascertain the degree of variability among observer panels representing geographical locations on four continents. The conclusions are summarized below.

1. Role of Observer: Expert vs. Naïve Observers

An experiment based on the use of the standard AATCC gray scale was employed to determine the intra and inter- observer variability for a group of naïve as well as a group of expert color assessors.

According to various statistical methods, PF/3 and STRESS functions results showed that for the samples used in this study, judgments made by naïve observers on average tracked well those made by expert observers. Variance component analysis of observers showed no statistical significance between naïve and experts. However, experts on average produced a 9% positive bias in gray scale ratings compared to naïve observers. Although PF/3 data indicated potential learning effect of naïve observers when comparisons were made with color difference formulas, STRESS analysis showed no evidence of learning effect in the naïve observers.

The ‘error’ term in variance component analysis was statistically different between the two groups of observers, which implied that unknown factors affect observer groups differently.

Since naïve observers are by definition a form of representation of the general of the general population, providing a randomly selected large group of observes are employed and the observers are provided with an opportunity to become familiar with the methodology employed for visual assessments, their response can be used in the development of visual datasets.

2. Development of a Linear Gray Scale for Visual Assessment of Small Color Differences

A robust perceptually linear gray scale for the assessment of perceived magnitude of color difference of chromatic or achromatic stimuli was developed comprising ten discrete pairs of gray samples. The perceptual linearity of the scale was validated under simulated daylight (6500K and a neutral gray background (Munsell N 7.25). The performance of the scale was assessed relative to standard and custom-built geometric scales, and was used in the development of visual datasets during the international replication study.

3. Observer Variability in Small Color Difference Assessments Using Various Psychophysical Methods

A subset of experiments was devised to determine the optimum psychophysical experimental method in terms of intra and inter- observer variability.

3.1. Pair Comparison Method

The selection of sample pairs in a pair comparison psychophysical testing method should correspond with the magnitude of color difference of the anchor pair to yield useful information and avoid loss of data. Results of visual assessments based on pair comparison revealed that in each trial a different number of sample pairs had to be discarded due to their high or low visual probability. Ten samples were thus discarded from of a total of 31 samples used during the initial testing phase. Useful inferences in relation to observer variability among trials could not be obtained because the visual probability is calculated by the pair rather than by the observer. Consequently, this psychophysical method was not continued.

3.2. Gray Scale Methods

Three aspects of visual assessments based on gray scales were examined: type of scale (geometric vs. linear), size (standard AATCC vs. Jumbo) and sample separation (hairline vs. 2” gap).

3.2.1. Type of Scale: Standard Geometric Scale (AATCC) vs. Novel Perceptually Linear Scale

The variance in the assessment of each of the colored pairs as well as the variability amongst observers was substantially lower using the linear scale than the standard AATCC scale. The visual assessment based on the linear scale also showed a considerable reduction in the estimates of variability between the two methods. Most importantly, the 'error' term was decreased by more than 53% when the linear scale was employed. These findings were confirmed when the performance of color difference models were tested against the visual data. The performance of ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ color difference models was tested using PF/3 and STRESS metrics based on the average visual data obtained using the two scales. Results based on the linear scale yield the highest agreement between color difference models and the visual data obtained. The results also showed, in both cases, that ΔE^*_{ab} performed worst among the models, however, when the standard AATCC scale was used the difference between ΔE^*_{ab} and $CMC_{(1:1)}$ was not significant. No significant difference between the performances of $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was observed in both methods.

For the experiment conducted, various statistical methods, including PF/3 and STRESS, showed that observers on average perceived a smaller color difference between pairs using the standard AATCC gray scale when compared to the linear scale.

In addition, although PF/3 Accuracy and Repeatability values for both methodologies were higher in comparison to previously reported values, the results obtained based on the use of the linear scale were lower (i.e. more accurate and repeatable) in comparison to those based on the standard AATCC scale.

3.2.2. Size of Scale: Standard AATCC Geometric Scale vs. Jumbo Gray Scale

A custom geometric scale, comprising 10 gray pairs, 2” x 4” in size, was developed and tested. Results showed that for the samples used in this study, observers on average perceived a larger color difference between pairs when the larger “Jumbo” gray scale was used compared to the smaller, standard AATCC scale. The variance in the assessment of each of the colored pairs amongst observers was found to be approximately similar for both methodologies. However, in comparison to the linear gray scale the magnitude of variation was higher.

The Jumbo scale generated the largest intra-group as well as intra-observer variability between both methods and amongst all methodologies studied. In addition, the assessment based on the Jumbo scale showed a significant increase in the estimates of variability between the two methods.

While PF/3 Accuracy and Repeatability values for both methodologies were higher in comparison to reported values of other research groups that used different methods, the results obtained based on the use of the AATCC standard gray scale were higher in comparison to those based on the Jumbo gray scale. Based on PF/3 metric, the disagreement between visual assessments of the two methods was approximately 22, indicating reasonably good agreement.

The performance of ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ color difference models was tested using PF/3 and STRESS metrics based on the average visual data obtained using the two scales. Results based on the Jumbo scale yield a higher agreement between color difference models and the visual data obtained. The results also showed, in both cases, that ΔE^*_{ab} performed statistically poorer than $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$. No significant difference between the performances of $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was observed when comparing the two methods.

3.2.3. Sample Separation: Jumbo Gray Scale vs. Jumbo Gray Scale with Gap

The custom Jumbo scale, comprising 10 gray pairs, 2” x 4” in size, was used to test the potential effect of having sample separation between the pairs versus juxtaposition (i.e. hairline gap vs. 2” gap). Results showed that for the samples used in this study, observers on average perceived a larger color difference between pairs when the larger Jumbo gray scale was used compared to the standard scale.

The variance in the assessment of each of the colored pairs amongst observers was found to be approximately similar for both methodologies. However, in comparison to the linear gray scale the magnitude of variation was higher in both cases.

The Jumbo scale generated the largest intra-group as well as intra-observer variability between both methods and amongst all methodologies studied. In addition, the assessment based on the Jumbo scale showed a significant increase in the estimates of variability between two methods.

PF/3 Accuracy and Repeatability values for both methodologies were higher in comparison to data reported by other researchers using different visual assessment methods; accuracy and repeatability for the Jumbo scale with gap were lower than those based on the Jumbo gray scale with no gap. Hence, attempts to minimize the potential for crispening effects by providing a 2" gap between sample pairs actually led to reduced accuracy and repeatability. Based on PF/3 metric, the disagreement between visual assessments of the two methods was approximately 23, indicating reasonably good agreement.

The performance of ΔE^*_{ab}, $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ color difference models were tested using PF/3 and STRESS metrics based on the average visual data obtained using the two scales. Results based on the separations studies (hairline vs. gap) yielded approximately the same agreement between color difference models and the visual data obtained.

The results also showed, in both cases, that ΔE^*_{ab} performed statistically poorer than $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$. No significant difference between the performances of $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ was observed using either method.

4. Replication Experiment

A replication study was carried out to test systematically and for the first time, the inter and intra-observer variability in visual data among four panels of observers representing different regions of the world, Colombia, Czech Republic, USA, and a group of Chinese students in the USA, when using a highly controlled experimental method.

The visual assessments employed the perceptually linear scale. Sixty nine polyester knitted samples around 13 color centers were assessed three times by 25 observers in each location.

Based on various statistical methods, including PF/3 and STRESS functions, results showed that for the samples used in this study, on average, all observer groups demonstrated good agreement considering the perceptual nature of the study. However, variability among observer panels for each colored pair was not consistent.

For most sample pairs, Colombian observers judged samples to have a smaller color difference compared to results obtained from other observer panels. However, the agreement between each observer panel and the total observer population increased for all observer panels.

This suggests that visual data obtained from observer panels based on a mixed ethnic makeup may be more suitable for investigations that involve human perceptions.

In addition, the degree of variation (intra-group standard deviation) among responses from any panel compared to the mean response of that panel was not significantly different between the panels. Furthermore, variance component analysis showed no statistically difference amongst observers' estimates for each observer group. Additionally, according to F-statistics, all observer panels were significantly different in terms of the error. The error component includes all the unknown parameters not included in the variance component model. This indicates that a complete understanding of the significance of variation due to method error requires further experimentation.

4.1. Performance of Existing Models Based on Results of Replication Study

The performance of $\Delta E^{*}_{ab,}$ $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ color difference models was tested using PF/3 and STRESS metrics based on the average visual data obtained from each of the observer panels and for the combined data. The purpose of testing the visual data against existing color difference models was to show whether or not the visual data obtained for each geographic group was statistically different from each other in practical performance, i.e. to understand the potential impact of variability on conventional color difference modeling. PF/3 values for all panels were within a range of four units, indicated close agreement.

The results also showed that ΔE^*_{ab} performed statistically poorer than $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ for all the visual data. No significant difference between the performances of $CMC_{(1:1)}$ and $CIEDE2000_{(1:1:1)}$ models was observed based on the responses obtained from various observer panels. In addition, when the visual data from all observer panels was combined the performance of color difference models, based on STRESS and PF/3, against the combined dataset did not change significantly compared to that based on individual panels. Moreover, the inclusion of upper and lower confidence levels on to the visual data does not change the PF/3 values significantly (<1 unit) while the change in STRESS values was also insignificant. (< 0.1).

Overall, for the highly controlled visual method developed, which requires a perceptually linear scale, controlled lighting, surround and illumination and observer viewing angles, and textile color difference samples precision cut and consistently displayed - good agreement was found in regard to intra- and inter-observer variability between naïve observers from four different continents. When combined with the finding that naïve observers are appropriate to use in small color difference assessment of textiles, the methodology developed is appropriate for use in the development of a comprehensive small color difference visual dataset using color centers throughout global color space and color difference pairs varying in lightness, chroma and hue.

VI. Future Work

Results obtained in this study highlight the need to conduct additional investigation of a number of parameters and attributes that influence the degree of intra and inter- observer variability in the assessment of small color differences. Specifically the following recommendations would lead to increasing our current understanding of the factors that influence human visual perception:

- A large set of dyed textiles samples dispersed around color centers that are comprehensively distributed throughout color space should be established to examine the magnitude of intra and inter-observer variability among observers as a function of location of samples using the best experimental method developed in this work. The visual method could then be repeated using all key light sources. The data obtained from this study would be the single largest visual dataset, with presumably close to the minimum observer variability practicably possible for textile materials. These data should then be used to establish an optimized color difference formula for each light source.
- The intra-and inter observer variability in visual data obtained from a group of expert observers should be tested using the perceptually linear scale under the highly controlled experimental methodology established, and the experiments should be conducted with repeat assessments to enable comprehensive comparison with naïve observers who conducted the same experiment.

- In order to establish whether repeated observations affect perceptual assessment of small color differences, a statistically robust methodology should be employed to ascertain the minimum number of repetitions whereby a further increase in repeating the test would not result in a significant change in assessments. This work could also be undertaken in tandem with investigating the minimum number of observers required to be statistically robust.
- Further independent research is required to validate the demonstrated reduction in intra and inter observer variability of visual data, demonstrated in this study based on the use of a perceptually linear gray scale is repeatable for magnitude estimation of small color differences of samples in CIE recommended color centers.
- Although the perpetual linearity of the scale was validated under simulated daylight (6500 K) conditions, validation and verification of this scale should be established using incandescent and all commercially important fluorescent lamps.
- A perceptual study can be developed to assess the perceived magnitude of differences amongst adjacent pairs in a standard AATCC gray scale and establish whether naïve observers perceive a geometric increase in contrast among sample pairs.
- A fundamental understanding of the significance of variation of error between observer panels requires a more in-depth analysis of psychological, physiological, temporal, spatial and geographical factors that might influence observer assessments.

Although, this was beyond the scope of this study, the results obtained indicate that additional parameters could be identified for inclusion in the variance component analysis.

- In this study variance due to sample pairs was high as the ΔE_{ab}^{*} range of samples assessed was between 0.76-8.34. In order to reduce the role of sample types in the assessment of intra and inter observer variability a study involving the assessment of just noticeable differences among different panels of observers would be required.
- It would be prudent to repeat parts of the inter- and intra-observer variability study with a focus on determining the effect, if any, of the diversity of observer population in terms of gender and age.
- A limited set of observer panels representing different regions of the world was studied. Additional panels of observers representing other regions of the world, such as Africa, Australia, and a panel of continental European observers could be examined to increase the confidence in the visual datasets.
- Observer panels could be genetically tested to establish whether differences among individuals are due to genetic variations.
- The same samples used in this study should be assessed by 25 naïve observers at least three times using AATCC Evaluation Procedure 9 without mounting the samples in a controlled way.

That is, observers would assess the samples in the same manner that is commonly undertaken in commercial visual assessment protocols for textile materials. This experiment would enable comparison of the effect of controlling the sample size and position on inter and intra observer variability, and would demonstrate if commercial practice should be modified to more controlled sample presentation in order to reduce variability.

- A more comprehensive pair comparison assessment method can be employed to compare inter and intra observer variability amongst observers based on this psychophysical testing method. The experiment could be repeated with varying color difference of the anchor pair employed to test the importance of the magnitude of color difference of anchor pair on inter and intra observer variability.
- The perceptually linear gray scale should be employed in a controlled experiment to determine the significance of chromatic adaptation on intra and inter- observer variability using light sources that have poor color rendering properties. The linear gray scale was custom built to a size representing the sample size of the colored pairs studied. In order to assess the suitability of this scale in practical assessment of chromatic and achromatic stimuli, further analysis of its size, number of pairs, background color, and mounting conditions should be conducted. In particular, the scale should be manufactured to a practically useful size similar to the current AATCC and ISO gray scales, and the smaller linear scale should be tested in comparison to the current standards scales.

VII. References

1. Graystone, J., *Integrating Colour Delivery Skills*. in *The Colour Delivery Challenge*. May 3-4 of 2000. Leeds, UK.

2. Eckes, G., *Six Sigma for Everyone*.2003, Hoboken, New Jersey: John Wiley and Sons, Inc.

3. Butts, K., *"Why Don't my Numbers Match Yours"?. in Color Management and Communication Workshop,* 2007. Raleigh, NC.

4. Parrot, K., *Instrumental Colour Quality Control: "Getting the Best from your System"*, in *Colour Science Volume 3: Colour Physics*, Gilchrist, A., and. Nobbs, J.H. 2001, University of Leeds Department of Colour Chemistry: Leeds, UK.

5. Sanger, A., *Creativity: Getting Color Right*. in *Color Management and Communication Workshop,* 2007. Raleigh, NC.

6. Koksal, G., Smith, W., and Smith, B., *A System Analysis of Textile Operations.* Textile Chemist & Colorist, 1992. **24**(10): p. 30.

7. American Association of Textile Chemists and Colorists, *AATCC Test Method 173-1998. CMC: Calculation of Small Color Differences for Acceptability*, in *AATCC Technical Manual*, 2005, AATCC: Research Triangle Park ,NC. p. 311.

8. Luo, M.R., Cui G., and Rigg, B., *The Development of the CIE 2000 Colour-Difference Formula: CIEDE2000.* Color Research & Application, 2001. **26**(5): p. 340.

9. Kuehni, R.G., *Colour Difference Formulas: Accurate Enough for Objective Colour Quality Control of textilesTextiles?* Coloration Technology, 2003. **119**(3): p. 164.

10. Kuehni, R.G., *Determination of Unique Hues Using Munsell Color Chips.* Color Research & Application, 2001. **26**: p. 61.

11. Hinks, D., Cárdenas, L.,Kuehi, R.G., and Shamey, R., *Unique-Hue Stimulus Selection Using Munsell Color Chips.* Journal of the Optical Society of America, 2007. **24**(10): 3371 .

12. Guan, S.S. and Luo, M.R., *Investigation of Parametric Effects Using Small Colour Differences.* Color Research & Application, 1999. **24**(5): p. 331.

13. Guan, S.S. and Luo, M.R., *Investigation of Parametric Effects Using Large Colour Differences.* Color Research & Application, 1999. **24**(5): p. 356.

14. Xin, J.H., Lam, C.C., and Luo, M.R., *Investigation of Parametric Effects Using Medium Colour-Difference Pairs.* Color Research & Application, 2001. **26**(5): p. 376.

15. Kuehni, R.G., *Color Space and its Divisions*, 2003, Hoboken, New Jersey: John Wiley & Sons, Inc. 408.

16. Kuehni, R.G. *Color : An Introduction to Practice and Principles*, 2005, Hoboken, NJ: John. Wiley & Sons Inc. 199.

17. Coren, S., Ward, L., and Enns, J., *Brightness and Color*, in *Sensation and Perception*, 2003, John Wiley & Sons, Inc.: Danvers, MA. 608.

18. Fairchild, M.D., *Color Appearance Models*, 2nd edition. 2005. Wiley-IS&T: Chichester, UK. 385.

19. McDonald, R.D., *Colour Physics for Industry 2nd edition*, 1997, Bradford: Society of Dyers and Colourists. 534.

20. Fletcher, R. and Voke, J., *Defective Colour Vision : Fundamentals, Diagnosis, and Management*,1985, Bristol ; Boston: A. Hilger. 608.

21. Szaflarski, D., *How We See: The First Steps of Human Vision*. Access Excellence Classic Collection. 2009. [September 2004]. Available from: http://www.accessexcellence.org/AE/AEC/CC/vision_background.php

22. Schwartz, S.H., *Visual Perception : A Clinical Orientation*, 1994, Norwalk, CN.: Appleton & Lange. 384.

23. Chang, R., *Physical Chemistry for the Chemical and Biological Sciences*, 2000, Sausalito, CA: University Science Books. 1018.

24. Casiday, R., and Frey, R., *"I Have Seen the Light!" Vision and Light-Induced Molecular Changes*. 2000. [September 2004]. Available from: http://www.chemistry.wustl.edu/~edudev/LabTutorials/Vision/Vision.html

25. Hubel, D., *Color Vision*, in *Eye, Brain, and Vision* 2nd edition, 1995, W H Freeman & Co: New York, NY. 256.

26. Dalton, J., *Extraordinary Facts Relating to the Vision of Colours: With Observations.* Memoirs of the Literary and Philosophical Society of Manchester. **5**: p. 28.

27. McIntyre, D., *Colour Blindness: Causes and Effects*. 2002, Chester, UK: Dalton Publishing. 112.

28. Lee, D. and Honson, M., *Chromatic Variation of Ishihara Diagnostic Plates.* Color Research & Application, 2003. **28**(4): p. 267.

29. Mollon, J.D., Pokorny, J., and Knoblauch, K., *Normal and Defective Colour Vision*. 2003, Oxford University Press: New York. 460.

30. Neitz, J., *The Neitz Test of Color Vision*, 2001, Western Psychological Services: Los Angeles, CA.

31. Coren, S., Ward L., and Enns J., *Psychophysics*, in *Sensation and Perception*, 2003, John Wiley & Sons, Inc.: Danvers, MA. 608.

32. Montag, E.D., and Wilber, D.C., *A Comparison of Constant Stimuli and Gray-Scale Methods of Color Difference Scaling.* Color Research & Application, 2003. **28**(1): p. 36.

33. Kim, D.H., *The Influence of Parametric Effects on the Appearance of Small Colour Differences*. 1997, PhD thesis. University of Leeds.

34. Witt, K., and Doring, G., *Parametric Variations in a Threshold Color-Difference Ellipsoid for Green Painted Samples.* Color Research & Application, 1983. **8**(3): p. 153.

35. Xu, H., Yaguchi, H., and Shiori S., *Correlation Between Visual and Colorimetric Scales Ranging from Threshold to Large Color Difference.* Color Research & Application, 2002. **27**(5): p. 349.

36. American Association of Textile Chemists and Colorists, *Evaluation Procedure 9-Visual Assessment of the Change in Color*, in *AATCC Technical Manual*, 2005, AATCC: Research Triangle Park, NC. p.396.

37. Luo, M.R. and Rigg, B., *Chromatic-Discrimination Ellipses for Surface Colours.* Color Research and Application, 1986. **11**(1): p. 25.

38. Cui, G., Luo, R., Rigg, B., and Li, W., *Colour-Difference Evaluation Using CRT Colours. Part I: Data Gathering and Testing Colour Difference Formulae.* Color Research & Application, 2001. **26**(5): p. 394.

39. Noriega, L.A., Heptinstall, A.R., Luo, M.R. and Westland, S., *The Perception of Achromatic Differences*, in Proceedings of Colour Science, 1998. Leeds, UK. p. 195

40. Witt, K., *Linearity and Additivity of Small Color Differences.* Color Research and Application, 1995. **20**(1): p. 36.

41. Witt, K., *Geometric Relations Between Scales of Small Color Differences.* Color Research and Application, 1999. **24**(2): p. 78

42. Westland, S., Graham, C., Addison, S., Sharrot, P., and Rigg, B., *Effect of Sleeve Colour and Background Colour on Change in Colour Assessments.* Coloration Technology, 2001. **117**: p. 123.

43. Xin, J.H., Lam, C.C., and Luo, M.R., *Evaluation of the Crispening Effect Using CRT-Displayed Colour Samples.* Color Research & Application, 2004. **29**(5): p. 374

44. Lam, Y.M. and. Xin, J.H, *Evaluation of the Quality of Different D65 Simulators for Visual Assessment.* Color Research and Application, 2002. **27**(4): p. 243.

45. Xu, H., Luo, M.R.and Rigg, B., *Evaluation of Daylight Simulators. Part 2: Assessment of the Quality of Daylight Simulators Using Actual Metameric Pairs.* Coloration Technology, 2003. **119**(5): p. 253.

46. Kuo, W.G. and Luo, M.R., *Methods for Quantifying Metamerism. Part 1 -- Visual Assessment.* Journal of the Society of Dyers & Colourists, 1996. **112**(12): p. 312.

47. Chou, W., Lin, H., Luo, M.R., Westland, S., Rigg, B., and Nobbs, J., *Performance of Lightness Difference Formulae.* Coloration Technology, 2001. **117**: p. 19.

48. Gibert, J.M., Daga, J.M., Gilabert, E.J., and Valldeperas, J., *Evaluation of Colour Difference Formulae.* Coloration Technology, 2005. **121**(3): p. 147.

49. Kim, D.H., Cho, E.K.and Kim, J.P., *Evaluation of CIELAB-Based Colour-Difference Formula Using a New Dataset.* Color Research and Application, 2001. **26**(5): p. 369.

50. Rich, D.C. and Billmeyer, F.W.Jr., *Small and Moderate Color Differences IV. Color-Difference-Perceptibility Ellipses in Surface-Color Space.* Color Research & Application, 1983. **8**(1): p. 31.

51. Stroka, D., Brockes, A.,and Pafthausen, W., *Influence of Experimental Parameters on the Evaluation of Color-Difference Ellipsoids.* Color Research and Application, 1983. **8**(3): p. 189.

52. Qiao, Y., Berns, R.S., Reniff, L., and Montag, E., *Visual Determination of Hue Suprathreshold Color-Difference Tolerances.* Color Research and Application, 1998. **23**(5): p. 302-313.

53. Kuehni, R.G. and Marcus, R.T., *.An Experiment in Visual Scalling of Small Color Differences.* Color Research and Application, 1979. **4**(2): p. 83.

54. Alder, C., Chaing, K.P., Chong, T.F., Coates, E., Khalili, A.A., and Rigg, B., *Uniform Chromaticity Scales- New Experimental Data.* Journal of the Society of Dyers & Colourists, 1982. **98**: p. 14.

55. Cheung, M. and Rigg, B., *Colour-Difference Ellipsoids for five CIE Colour Centres.* Color Research and Application, 1986. **11**(3): p. 185.

56. Coates, E., Fong, Y., and Rigg, B., *Uniform Lightness Scales.* Journal of the Society of Dyers & Colourists, 1981. **97**: p. 179.

57. Witt, K., *Three-Dimensional Threshold of Color-Difference Perceptibility in Painted Samples: Variability of Observers in Four CIE Color Regions.* Color Research & Application, 1987. **12**(3): p. 128.

58. Berns, R.S., *Deriving Instrumental Tolerances from Pass-Fail and Colorimetric Data.* Color Research and Application, 1996. **21**(6): p. 459.

59. Mangine, H., Jakes K., and Noel, C., *.A Preliminary Comparison of CIE Color Differences to Textile Color Acceptability Using Average Observers.* Color Research & Application, 2005. **30**(4): p. 288.

60. Steen, D. and Dupont, D., *.Defining a Practical Method of Ascertaining Textile Color Acceptability.* Color Research & Application, 2002. **27**(6): p. 391.

61. Committee E12 on Color and Appearance, American Society for Testing and Materials, *E 1499-97 Standard Guide for Selection, Evaluation, and Training of Observers*, in *ASTM Standards on Color and Appearance.* 2004, ASTM International: West Conshohocken, PA. p. 534.

62. Committee E12 on Color and Appearance, American Society for Testing and Materials, *D 2616-95 Standard Test Method for Evaluation of visual Color Difference With a Gray Scale*, in *ASTM Standars on Color Appearance* 2004, ASTM International: West Conshohocken, PA. p. 123.

63. American Association of Textile Chemists and Colorists, *Evaluation Procedure 1 Gray Scale for Color Change,* in *AATCC Technical Manual*, 2005, AATCC: Research Triangle Park ,NC. p. 379.

64. Wyszecki, G. *Current Developments in Colorimetry*. in *Proceedings of AIC Colour '73*(American Institute of Chemists). 1973. Hilger, London.. p. 21.

65. Berns, R.S., *Billmeyer and Saltzman's Principles of Color Technology*, 3rd edition. 2000, John Wiley & Sons: New York,304.

66. CIE, *Colorimetry Technical Report 15, 3rd. edition.* 2004, CIE: Vienna, Austria.

67. Stiles, W.S. and. Burch, J.M., *N.P.L. Colour Matching Investigation: Final Report.* Journal of Modern Optics, 1959. **6**(1): p. 1.

68. Speranskaya, N.I., *Determination of Spectrum Color Coordinates for Twenty-Seven Normal Observers.* Optics and Spectroscopy, 1959. **7**: p. 424.

69. North, A. and Fairchild, M.D., *Measuring Color Matching Functions. Part I.* Color Research & Application, 1993. **18**(3): p. 155.

70. North, A. and Fairchild, M.D., *Measuring Color Matching Functions. Part II. New data for Assessing Observer Metamerism.* Color Research & Application, 1993. **18**(3): p. 163.

71. Rich, D.C. and Jalijali, J., *Effects of Observer Metamerism in the Determination of Human Color Matching Functions.* Color Research & Application, 1995. **20**(1): p. 29.

72. Randall, D., *Instruments for the Measurement of Color*, in *Color Technology in the Textile Industry*, 1997, American Association of Textile Chemists and Colorists: Research Triangle Park, NC. 196

73. Christment, A., *Color and Colorimetry*, 1998, Paris, France: Editions 3C Conseil. 30.

74. Committee E12 on Color and Appearance, American Society for Testing and Materials, *E 805-94 Standard Practice for Identification of Instrumental Methods of Color or Color Difference Measurement Materials*. 2004, ASTM International: West Conshohocken, PA.

75. American Association of Textile Chemists and Colorists, *Evaluation Procedure 6-Instrumental Color Measurement*, in *AATCC Technical Manual*, 2005, AATCC: Research Triangle Park,NC. p. 384.

76. American Association of Textile Cemists and Colorists, *Evaluation Procedure 7-Instrumental Assessment of the Change in Color of a Test Specimen*, in *AATCC Technical Manual*, 2005, AATCC: Research Triangle Park ,NC. p. 392.

77. Conelly, R., *Good Sample Presentation: How to Get Color Measurement Results that Make Sense*, in *Color Technology in the Textile Industry*, 1997, American Association of Textile Chemists and Colorists: Research Triangle Park, NC. 196

78. Laidlaw, A.C., *Care and Feeding of Color Measuring Instrumentation: How toImplement a System for Maintaining its Integrity*, in *Color Technology in the Textile Industry*, 1997, American Association of Textile Chemists and Colorists: Research Triangle Park, NC. 196

79. Pointer, M.R. and Attridge, G.G., *Some Aspects of the Visual Scaling of Large Colour Differences.* Color Research & Application, 1997. **22**(5): p. 298.

80. Hering, E., *Outlines of a Theory of the Light Sense*, 1964,Cambridge, MA: Harvard University Press. 344

81. Cunthasaksiri, P., Shinoda, H., and Ikeda, M., *Recognized Visual Space of Illumination: A New Account of Center-Surround Simultaneous Color Contrast.* Color Research & Application, 2004. **29**(4): p. 255.

82. Wu, R.C., Wardman, R.H., and. Luo, M.R, *A Comparison of Lightness and Contrast Effects in CRT and Surface Colours.* Color Research & Application, 2005. **30**(1): p. 13.

83. Wu, R.C. and Wardman, R.H., *Proposed Modificiation to the CIECAM02 Colour Appearance Model to Include the Simultaneous Contrast Effects.* Color Research & Application, 2007. **32**(2): p. 121

84. Luo, M.R., *A Proposed Colour Inconstancy Index by CMC*, in *Colour Science Volume 3: Colour Physics*, Gilchrist, A., and. Nobbs, J.H. 2001, University of Leeds Department of Colour Chemistry: Leeds, UK.

85. Luo, M.R. and Hunt, R.W.G., *A Chromatic Adaptation Transform and a Colour Inconstancy Index.* Color Research & Application, 1998. **23**(3): p. 154.

86. Kuo, W.G. and Luo, M.R., *Methods for Quantifying Metamerism. Part 2 -- Instrumental Methods.* Journal of the Society of Dyers & Colourists, 1996. **112**(12): p. 354.

87. Alfvin, R. and Fairchild, M.D., *Observer Variability in Metameric Color Matches using Color Reproduction.* Color Research & Application, 1997. **22**(3): p. 174.

88. Diaz, J.A., Chiron, A., and Vienot, F., *Tracing a Metameric Match to Individual Variations of Color Vision.* Color Research & Application, 1998. **23**(6): p.379.

89. Wyzecki, G., *The Degree of Color Metamerism and its Specifications*, in *Color Technology in the Textile Industry*, 1997, American Association of Textile Chemists and Colorists: Research Triangle Park, NC. 196

90. Li, Z. and Berns, R.S., *Comparison of Methods of Parameric Correction for Evaluating Metamerism.* Color Research & Application, 2007. **32**(4): p. 293.

91. Luo, M.R., *Development of Colour-Difference Formulae.* Review of Progress in Coloration, 2002. **32**: p. 28.

92. Committee E12 on Color and Appearance, American Society for Testing and Materials, *E 284-99a Standard Terminology of Appearance.* 2004, ASTM International: West Conshohocken, PA. p. 294

93. Berns, R.S., Alman, D., Reniff, L., Snyder, G.D., and Balonon-Rosen, M. R., *Visual Determination of Suprathreshold Color-Difference Tolerances Using Probit Analysis.* Color Research & Application, 1991. **16**(5) : p. 297.

94. Lee, S.G., *Assessment of Metrics in Color Spaces*. 2007, Masters thesis. North Carolina State University

95. Melgosa, Hita, E., Poza, A.J., Alman, D.H., Berns, R.S., *Suprathreshold Color-Difference Ellipsoids for Surface Colors.* Color Research and Application, 1997. **22**(3): p. 148.

96. Melgosa, M., Hita, E., Romero, J., and Jiménez del Barco, L., *Color-Discrimination Thresholds Translated from the CIE(x.y,Y) Space to the CIE 1976 (L*a*b*).* Color Research and Application, 1984. **19**(1): p. 10.-

97. Pridmore, R.W. and Melgosa, M., *Effect of Luminance of Samples on Color Discrimination Ellipses: Analysis and Prediction of Data.* Color Research & Application, 2005. **30**(3): p. 186.

98. Kuehni, R.G., *Advances in Color Difference Formulas.* Color Research & Application, 1982. **7**(1): p. 19.

99. McDonald, R., *Industrial Pass/Fail Colour Matching Part 1- Preparation of Visual Color Matching Data.* Journal of the Society of Dyers & Colourists, 1980. **96**: p. 372.

100. International Organization for Standardization, *ISO 105-A02 Textiles Tests for Colour Fastness -- Part A02: Grey Scale for Assessing Change in Colour.* 4th edition. 1993. ISO: Geneva, Switzerland.

101. Luo, M.R. and Rigg, B., *BFD (l:c) Color-Difference Formula. Part 2. Performance of the Formula.* Journal of the Society of Dyers & Colourists, 1987. **103**(3): p. 126.

102. Luo, M.R. and Rigg, B., *BFD (1:c) Color Difference Formula. Part 1. Development of the Formula.* Journal of the Society of Dyers & Colourists, 1987. **103**(2): p. 86.

103. CIE, *Industrial Colour-Difference Evaluation CIE 116,* 1995, CIE: Viena, Austria.

104. Melgosa, M. and Huertas, R., *Relative Significance of the Terms in CIEDE2000 and CIE 94 Color-Difference Formulas.* Journal of the Optical Society of America, 2004. **21**(12): p. 2269.

105. Schultze, W., *The Usefulness of Colour-Differences Formulae for Fixing Colour Tolerances* in *Color Metrics Proceedings* of the Helmholtz Memorial Symposium, 1971. Vos, J.J., Friele, L.F. and Walraven, P.L. (Association Internationale de la Couleur (AIC): Soesterberg, Holland. p.254.

106. Xin, J.H., Shen, H.L., and Lam, C.C., *Investigation of Texture Effect on Visual Colour Difference Evaluation.* Color Research and Application, 2005. **30**(5): p. 341.

107. Xu, H. and Yaguchi, H., *Visual Evaluation at Scale of Threshold to Suprathreshold Color Difference.* Color Research and Application, 2005. **30**(3): p. 198.

108. Cui, G., Luo, M.R., Rigg, B., and Roesler, G., *Uniform Colour Spaces Based on the DIN99 Colour-Difference Formula.* Color Research and Application, 2002. **27**(4): p. 282.

109. Luo, M.R., Cui, G., and Li, C., *Uniform Color Spaces Based on CIECAM02 Colour Appearance Model.* Color Research and Application, 2006. **31**(4): p. 320.

110. Garcia, P., Huertas, R., Melgosa, M., and Cui, G., *Measurement of the Relationship Between Perceived and Computed Color Differences.* Journal of the Optical Society of America, 2007. **24**(7): p. 1823.

111. Agresti, A. and Finlay, B., *Statistical Methods for the Social Sciences* 3rd. edition, 1997, Upper Saddle River, NJ: Prentice Hall. 643.

112. Rao, P.V., *Statistical Research Methods in the Life Sciences*, 1998, Pacific Groove, CA: Duxbury Press. 889.

113. Farnum, N.R., *Modern Statistical Quality Control and Improvement*, 1993, Belmont, California: Duxbury Press. 512.

114. Lyman Ott, R. and Longnecker, M., *An Introduction to Statistical Methods and Data Analysis*, 5th edition, 2001, Pacific Grove, California: Duxbury Press. 1184.

115. International Organization for Standardization, *ISO 105-J03 Textiles Tests for Colour Fastness -- Part J03: Calculation of Colour Differences,.* 4th edition. 1993. ISO: Geneva, Switzerland.

I116. American Society for Testing and Materials, *D2244-07: Standard Practice for Calculation of Color Tolerances and Color Differences from Instrumentally Measured Color Coordinates*.2001, ASTM International: West Conshohocken, PA.

117. Textile and Flexible Plastics Committee, Society of Automotive Engineers, *SAE J1767: Instrumental Color Difference Measurements for Colorfastness of Automotive Interior Trim Materials.* 1995, SAE International: Warrendale, PA.

118. CIE, *Improvement to Industrial Colour-Difference Evaluation* Technical Report 142, 2001, CIE: Viena.

119. American Association of Textile Chemists and Colorists, *Evaluation Procedure 1 Gray Scale for Color Change,* in *AATCC Technical Manual*, 2007, AATCC: Research Triangle Park ,NC. p. 377.

120. *Statcrunch, Data Analysis on the Web 4th and 5th edition* Available from: www.statcrunch.com.

121. Mangine, H., *Variability in Experimental Color Matching Conditions: Effects of Observer, Daylight Simulators, and Color Incostancy*. 2005, PhD thesis. Ohio State University.

122. Robertson, A.R., *CIE Guidelines for Coordinated Research on Colour-Difference Evaluation.* Color Research and Application, 1978. **3**(3): p. 149.

123. Techincal Services Department Hunter Associates Laboratory, *Tests Methods for Color Measurement*. Application Note, 2006, **18** (4): p. 1.

124. American Society for Testing and Materials, *F 2109: Standard Test Method to Determine Color Change and Staining Caused by Aircraft Maintenance Chemicals upon Aircraft Cabin Interior Hard Surfaces*. 2001, ASTM International: West Conshohocken, PA.

125. Japanese Standards Association, *JIS L 0809: Instrumental Determination of Colour Fastness-Change in Colour and Staining*. 2001, JSI: Akasaka, Minato-ku, Japan.

126. Standards Australia, *AS 2001.4.A02: Methods of Test for Textiles-Colourfastness Tests-Gray Scale for Assessing Change in Colour.* 2001, AS: Sydney, Australia.

127. Porter Paints. 2007; Available from: www.porterpaints.com.

VII. Appendices

Appendix A. Colorimetric Data for the Subset of 31 Samples Used in the First Part of the Study.

Pair	L*	a*	b*	C*	h°	DL*	DH*	DC*	ΔE^*_{ab}	CMC (1:1)	CMC (2:1)	CIEDE2000 (1:1:1)	CIEDE2000 (2:1:1)
Standard I	61.53	-22.4	-20.37	30.27	222.28								
1	60.38	-24	-25.22	34.81	226.42	-1.14 D	4.54 B	2.35 B	5.23	2.91	2.79	2.47	2.32
2	60.15	-22.04	-21.24	30.61	223.94	-1.38 D	0.33 B	0.88 B	1.67	1.29	0.84	1.32	0.81
3	61.73	-20.05	-24.39	31.57	230.58	0.21 L	1.30 B	4.47 B	4.66	3.09	3.09	2.71	2.7
Standard II	62.18	-22.44	-19.42	29.68	220.87								
4	62.59	-24.24	-20.46	31.72	220.16	0.41 L	2.04 B	-0.38 G	2.11	1.1	1.06	0.97	0.92
5	63.45	-20.14	-21.48	29.44	226.84	1.27 L	-0.24 D	3.08 B	3.34	2.34	2.16	2.19	1.98
Standard III	71.26	-19.13	19.05	27	135.11								
6	71.1	-17.04	16.67	23.84	135.64	-0.17 D	-3.16 D	0.23 G	3.17	1.67	1.66	1.52	1.52
7	71.41	-17.73	20.11	26.81	131.4	0.15 L	-0.19 D	-1.74 Y	1.76	1.22	1.21	1.18	1.18
8	69.72	-18.68	18.81	26.5	134.81	-1.55 D	-0.49 D	-0.14 Y	1.63	1.23	0.66	1.22	0.65
Standard IV	75.65	-30.37	24.49	39.01	141.11								
9	75.73	-32.2	25.61	41.15	141.51	0.08 L	2.13 B	0.27 G	2.15	0.95	0.95	0.78	0.78
10	78.67	-29.32	22.65	37.05	142.31	3.01 L	-1.97 D	0.80 G	3.69	2.47	1.5	2.31	1.37
Standard V	63.02	-17.92	1.02	17.95	176.73								
11	62.67	-16.35	1.16	16.39	175.93	-0.35 D	-1.56 D	-0.24 Y	1.62	1.06	1.03	1.1	1.07
12	62.43	-17.39	-0.49	17.4	181.6	-0.59 D	-0.56 D	1.50 B	1.71	1.41	1.34	1.27	1.19
13	61.12	-17.38	0.68	17.39	177.75	-1.90 D	-0.56 D	0.32 B	2.01	1.62	0.9	1.68	0.92
Standard VI	60.9	-9.99	-3.72	10.66	200.44								
14	60.64	-11.7	-4.75	12.63	202.11	-0.26 D	1.97 B	0.34 B	2.01	1.65	1.64	1.59	1.57
15	60.57	-10.05	-2.45	10.34	193.72	-0.34 D	-0.32 D	-1.23 Y	1.32	1.34	1.32	1.08	1.05
Standard VII	57.54	-35.86	6.99	36.54	168.98								
16	59.41	-40.28	9.71	41.43	166.45	1.87 L	4.90 B	-1.72 Y	5.52	2.92	2.57	2.66	2.22
Standard VIII	37.72	9.34	-34.4	35.65	285.19								
17	38.23	8.57	-33.25	34.33	284.46	0.51 L	-1.32 D	-0.44 G	1.48	0.89	0.75	0.56	0.41
18	34.98	9.92	-34.15	35.56	286.2	-2.74 D	-0.09 D	0.63 R	2.81	3	1.56	2.36	1.27
19	37.39	10.05	-35.76	37.14	285.71	-0.33 D	1.50 B	0.33 R	1.57	0.81	0.75	0.5	0.43
Standard IX	42.17	11.87	-26.28	28.83	294.3								
20	42.63	9.18	-21.06	22.97	293.57	0.46 L	-5.86 D	-0.33 B	5.89	3.02	2.99	2.68	2.65
21	42.19	9.19	-27.54	29.04	288.46	0.01	0.20 B	-2.95 B	2.96	2.48	2.48	2.71	2.71

Appendix A. (Continued)

Pair	L*	a*	b*	C*	h°	DL*	DH*	DC*	ΔE*ab	CMC (1:1)	CMC (2:1)	CIEDE2000 (1:1:1)	CIEDE2000 (2:1:1)
Standard X	41.75	17.98	30.21	35.16	59.24								
22	41.78	15.94	26.17	30.64	58.66	0.04	-4.52 D	-0.33 R	4.53	2.11	2.11	1.84	1.84
23	41.97	16.52	31.1	34.77	54.44	-0.38 D	-0.39 D	-2.93 R	2.98	3.48	3.46	2.26	1.32
24	42.97	18.29	30.82	35.84	59.32	1.23 L	0.68 B	0.05 Y	1.4	1.28	0.7	1.15	0.62
Standard XI	37.96	20.32	31.22	37.25	56.94								
25	37.19	18.43	29.35	34.66	57.88	-0.77 D	-2.60 D	0.59 Y	2.77	1.59	1.42	1.27	1.14
26	36.73	22.1	30.05	37.31	53.66	-1.24 D	0.05 B	-2.13 R	2.47	2.88	2.64	1.9	1.67
Standard XII	31.06	-8.98	15.95	18.3	119.38								
27	31.01	-11.01	17.87	20.99	121.65	-0.05	2.69 B	0.78 G	2.8	1.84	1.83	1.74	1.74
28	30.82	-6.85	16.71	18.06	112.29	-0.24 D	-0.24 D	-2.25 Y	2.27	2	1.98	2.17	2.16
29	32.99	-9.82	16.81	19.47	120.3	1.93 L	1.17 B	0.30 G	2.28	2.48	1.42	1.71	1.08
Standard XIII	34.24	-16.26	22.39	27.67	125.99								
30	33.68	-15.82	22.66	27.63	124.93	-0.56 D	-0.03	-0.51 Y	0.76	0.73	0.48	0.58	0.42
31	33.16	-9.69	17.36	19.88	119.18	-1.07 D	-7.79 D	-2.78 Y	8.34	4.64	4.52	4.68	4.62

Appendix B. SAS Results of Polynomial Procedure to Obtain a Statistically Reliable Equation for Converting Gray Scale Rating to Visual Differences.

The GLM Procedure

Dependent Variable: difference

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	165.2357533	41.3089383	710.79	<.0001
Error	4	0.2324689	0.0581172		
Corrected Total	8	165.4682222			

R-Square	Coeff Var	Root MSE	difference Mean
0.998595	4.759106	0.241075	5.065556

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	152.2270817	152.2270817	2619.31	<.0001
grade2	1	12.2707161	12.2707161	211.14	0.0001
grade3	1	0.7238716	0.7238716	12.46	0.0242
grade4	1	0.0140840	0.0140840	0.24	0.6483

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	0.62572844	0.62572844	10.77	0.0305
grade2	1	0.11933497	0.11933497	2.05	0.2252
grade3	1	0.03610514	0.03610514	0.62	0.4747
grade4	1	0.01408397	0.01408397	0.24	0.6483

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	24.88722222	2.53097022	9.83	0.0006
grade	-14.37506087	4.38095909	-3.28	0.0305
grade2	3.65593240	2.55133060	1.43	0.2252
grade3	-0.47733230	0.60560399	-0.79	0.4747
grade4	0.02475524	0.05028717	0.49	0.6483

Appendix B. (Continued)

The GLM Procedure

Dependent Variable: difference

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	165.2216693	55.0738898	1116.88	<.0001
Error	5	0.2465529	0.0493106		
Corrected Total	8	165.4682222			

R-Square	Coeff Var	Root MSE	difference Mean
0.998510	4.383722	0.222060	5.065556

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	152.2270817	152.2270817	3087.11	<.0001
grade2	1	12.2707161	12.2707161	248.85	<.0001
grade3	1	0.7238716	0.7238716	14.68	0.0122

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	5.43212494	5.43212494	110.16	0.0001
grade2	1	1.58890629	1.58890629	32.22	0.0024
grade3	1	0.72387162	0.72387162	14.68	0.0122

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	23.74936508	0.94977391	25.01	<.0001
grade	-12.31153439	1.17299821	-10.50	0.0001
grade2	2.42082251	0.42646545	5.68	0.0024
grade3	-0.18026936	0.04705017	-3.83	0.0122

Appendix C. Variance Component Analysis: Naïve Observers.

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	774	9611.58835	12.41807	7.23	<.0001
Error	1550	2662.57369	1.71779		
Corrected Total	2324	12274.16203			

R-Square	Coeff Var	Root MSE	value Mean
0.783075	51.01888	1.310645	2.568941

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	30	5275.328663	175.844289	102.37	<.0001
observer	24	1788.097868	74.504078	43.37	<.0001
swatch*observer	720	2548.161814	3.539114	2.06	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	5275.328663	175.844289	102.37	<.0001
observer	24	1788.097868	74.504078	43.37	<.0001
swatch*observer	720	2548.161814	3.539114	2.06	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	5275.328663	175.844289	49.69	<.0001
observer	24	1788.097868	74.504078	21.05	<.0001
Error	720	2548.161814	3.539114		
Error: MS(swatch*observer)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch*observer	720	2548.161814	3.539114	2.06	<.0001
Error: MS(Error)	1550	2662.573686	1.717789		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	0.7631
swatch*observer	0.6071
Residual	1.7178

Appendix D. Variance Component Analysis: Expert Observers.

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	54	3268.248323	60.523117	21.53	<.0001
Error	720	2023.775024	2.810799		
Corrected Total	774	5292.023347			

R-Square	Coeff Var	Root MSE	value Mean
0.617580	50.31922	1.676544	3.331816

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	30	2366.235974	78.874532	28.06	<.0001
observer	24	902.012349	37.583848	13.37	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	2366.235974	78.874532	28.06	<.0001
observer	24	902.012349	37.583848	13.37	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	2366.235974	78.874532	28.06	<.0001
observer	24	902.012349	37.583848	13.37	<.0001
Error: MS(Error)	720	2023.775024	2.810799		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	1.1217
Residual	2.8108

Appendix E. Variance Component Analysis: Naïve Observers without Interaction.

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	54	7063.42653	130.80420	56.98	<.0001
Error	2270	5210.73550	2.29548		
Corrected Total	2324	12274.16203			

R-Square	Coeff Var	Root MSE	value Mean
0.575471	58.97698	1.515084	2.568941

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	30	5275.328663	175.844289	76.60	<.0001
observer	24	1788.097868	74.504078	32.46	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	5275.328663	175.844289	76.60	<.0001
observer	24	1788.097868	74.504078	32.46	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	5275.328663	175.844289	76.60	<.0001
observer	24	1788.097868	74.504078	32.46	<.0001
Error: MS(Error)	2270	5210.735500	2.295478		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	0.7764
Residual	2.2955

Appendix F. Variance Component Analysis: Jumbo Scale.

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	774	10742.61467	13.87935	6.63	<.0001
Error	1550	3242.78874	2.09212		
Corrected Total	2324	13985.40341			

R-Square	Coeff Var	Root MSE	value Mean
0.768130	43.30861	1.446417	3.339790

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	30	6223.333028	207.444434	99.16	<.0001
observer	24	1287.791320	53.657972	25.65	<.0001
swatch*observer	720	3231.490321	4.488181	2.15	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	6223.333028	207.444434	99.16	<.0001
observer	24	1287.791320	53.657972	25.65	<.0001
swatch*observer	720	3231.490321	4.488181	2.15	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	6223.333028	207.444434	46.22	<.0001
observer	24	1287.791320	53.657972	11.96	<.0001
Error	720	3231.490321	4.488181		
Error: MS(swatch*observer)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch*observer	720	3231.490321	4.488181	2.15	<.0001
Error: MS(Error)	1550	3242.788736	2.092122		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	0.5287
swatch*observer	0.7987
Residual	2.0921

Appendix G. Variance Component Analysis: Jumbo Scale Gap.

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	774	8237.37933	10.64261	8.75	<.0001
Error	1550	1885.12890	1.21621		
Corrected Total	2324	10122.50823			

R-Square	Coeff Var	Root MSE	value Mean
0.813769	42.54197	1.102820	2.592311

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	30	3654.402454	121.813415	100.16	<.0001
observer	24	2072.528216	86.355342	71.00	<.0001
swatch*observer	720	2510.448659	3.486734	2.87	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	3654.402454	121.813415	100.16	<.0001
observer	24	2072.528216	86.355342	71.00	<.0001
swatch*observer	720	2510.448659	3.486734	2.87	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	3654.402454	121.813415	34.94	<.0001
observer	24	2072.528216	86.355342	24.77	<.0001
Error	720	2510.448659	3.486734		
Error: MS(swatch*observer)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch*observer	720	2510.448659	3.486734	2.87	<.0001
Error: MS(Error)	1550	1885.128897	1.216212		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	0.8911
swatch*observer	0.7568
Residual	1.2162

Appendix H. Reflectance Data of all Samples Used in the Development of the Linear Scale Expressed in Decimal Fraction (D65/10^o).

Λ (nm)	1	2	3	4	5	6	7	8	9	10	11	12	13
400	0.113	0.115	0.115	0.119	0.121	0.122	0.124	0.126	0.128	0.131	0.132	0.133	0.137
410	0.119	0.122	0.122	0.126	0.129	0.130	0.132	0.135	0.138	0.140	0.142	0.143	0.148
420	0.122	0.125	0.126	0.129	0.132	0.133	0.135	0.138	0.141	0.143	0.145	0.146	0.151
430	0.123	0.127	0.127	0.131	0.134	0.135	0.137	0.139	0.143	0.145	0.147	0.148	0.152
440	0.125	0.128	0.129	0.132	0.135	0.136	0.138	0.141	0.144	0.146	0.149	0.150	0.154
450	0.125	0.128	0.130	0.132	0.135	0.137	0.139	0.141	0.144	0.147	0.149	0.150	0.154
460	0.125	0.128	0.129	0.132	0.135	0.136	0.138	0.141	0.144	0.147	0.149	0.150	0.154
470	0.124	0.128	0.129	0.132	0.135	0.136	0.138	0.140	0.144	0.146	0.148	0.150	0.153
480	0.124	0.127	0.128	0.131	0.134	0.135	0.137	0.139	0.143	0.145	0.147	0.149	0.152
490	0.123	0.127	0.128	0.130	0.133	0.135	0.136	0.139	0.143	0.145	0.147	0.148	0.151
500	0.123	0.127	0.128	0.130	0.133	0.134	0.136	0.139	0.142	0.145	0.147	0.148	0.151
510	0.123	0.127	0.128	0.130	0.133	0.135	0.136	0.139	0.142	0.145	0.147	0.148	0.151
520	0.123	0.127	0.128	0.130	0.133	0.134	0.136	0.139	0.142	0.145	0.147	0.148	0.150
530	0.123	0.127	0.129	0.130	0.133	0.135	0.137	0.139	0.143	0.145	0.147	0.149	0.150
540	0.123	0.127	0.128	0.130	0.133	0.134	0.136	0.139	0.142	0.145	0.147	0.148	0.150
550	0.123	0.127	0.128	0.130	0.133	0.134	0.136	0.139	0.142	0.144	0.147	0.148	0.149
560	0.123	0.127	0.128	0.130	0.132	0.134	0.136	0.138	0.142	0.144	0.146	0.148	0.149
570	0.123	0.127	0.128	0.129	0.132	0.134	0.135	0.138	0.142	0.144	0.146	0.148	0.148
580	0.122	0.126	0.127	0.129	0.131	0.133	0.135	0.137	0.141	0.144	0.145	0.147	0.147
590	0.121	0.126	0.127	0.128	0.131	0.132	0.134	0.137	0.140	0.143	0.144	0.146	0.147
600	0.121	0.125	0.126	0.127	0.130	0.131	0.133	0.136	0.139	0.142	0.143	0.145	0.145
610	0.119	0.124	0.125	0.126	0.128	0.130	0.132	0.134	0.138	0.141	0.142	0.144	0.144
620	0.118	0.123	0.124	0.125	0.127	0.129	0.130	0.133	0.137	0.139	0.141	0.143	0.142
630	0.117	0.122	0.122	0.123	0.126	0.128	0.129	0.132	0.135	0.138	0.139	0.141	0.141
640	0.116	0.120	0.121	0.122	0.125	0.126	0.128	0.131	0.134	0.137	0.138	0.140	0.139
650	0.116	0.120	0.120	0.121	0.124	0.126	0.127	0.130	0.133	0.136	0.137	0.139	0.138
660	0.115	0.119	0.119	0.120	0.123	0.124	0.126	0.128	0.132	0.135	0.136	0.138	0.137
670	0.113	0.117	0.118	0.119	0.122	0.123	0.124	0.127	0.130	0.133	0.134	0.136	0.135
680	0.112	0.117	0.117	0.118	0.120	0.122	0.123	0.126	0.129	0.132	0.133	0.135	0.134
690	0.111	0.115	0.116	0.116	0.119	0.121	0.122	0.125	0.128	0.131	0.132	0.134	0.133
700	0.111	0.115	0.115	0.116	0.118	0.120	0.121	0.124	0.127	0.130	0.131	0.133	0.131

Appendix H. (continued)

Λ (nm)	14	15	16	17	18	19	20	21	22	23	24	25	26
400	0.142	0.144	0.146	0.146	0.150	0.151	0.152	0.153	0.158	0.159	0.162	0.161	0.162
410	0.154	0.155	0.158	0.158	0.162	0.164	0.165	0.167	0.172	0.174	0.177	0.175	0.177
420	0.157	0.158	0.161	0.162	0.165	0.168	0.168	0.170	0.175	0.177	0.180	0.179	0.180
430	0.158	0.160	0.163	0.163	0.167	0.169	0.170	0.172	0.177	0.178	0.181	0.180	0.182
440	0.160	0.161	0.164	0.165	0.168	0.171	0.171	0.173	0.178	0.179	0.182	0.182	0.183
450	0.160	0.162	0.165	0.165	0.168	0.171	0.172	0.173	0.178	0.179	0.182	0.182	0.183
460	0.160	0.161	0.164	0.165	0.168	0.170	0.171	0.173	0.178	0.179	0.181	0.181	0.183
470	0.159	0.161	0.164	0.165	0.167	0.170	0.171	0.172	0.177	0.178	0.180	0.180	0.182
480	0.158	0.160	0.162	0.163	0.166	0.168	0.169	0.171	0.176	0.176	0.179	0.179	0.180
490	0.157	0.159	0.162	0.163	0.165	0.168	0.169	0.170	0.175	0.176	0.178	0.178	0.180
500	0.157	0.158	0.161	0.163	0.165	0.167	0.168	0.170	0.174	0.175	0.177	0.177	0.179
510	0.157	0.158	0.161	0.163	0.165	0.167	0.168	0.170	0.174	0.174	0.177	0.177	0.179
520	0.156	0.158	0.161	0.162	0.164	0.167	0.168	0.169	0.173	0.174	0.176	0.176	0.178
530	0.156	0.158	0.161	0.163	0.164	0.167	0.168	0.169	0.174	0.174	0.176	0.176	0.178
540	0.155	0.157	0.160	0.162	0.164	0.166	0.167	0.169	0.173	0.173	0.174	0.175	0.177
550	0.155	0.157	0.160	0.162	0.163	0.166	0.167	0.168	0.172	0.172	0.174	0.175	0.177
560	0.154	0.156	0.159	0.161	0.163	0.166	0.166	0.168	0.172	0.172	0.173	0.174	0.176
570	0.154	0.156	0.159	0.161	0.163	0.165	0.166	0.168	0.171	0.171	0.173	0.174	0.176
580	0.153	0.155	0.158	0.160	0.162	0.164	0.165	0.167	0.170	0.170	0.171	0.172	0.175
590	0.152	0.154	0.157	0.159	0.161	0.163	0.164	0.166	0.169	0.169	0.170	0.171	0.174
600	0.151	0.152	0.156	0.158	0.160	0.162	0.163	0.164	0.168	0.168	0.169	0.170	0.173
610	0.149	0.151	0.154	0.157	0.158	0.160	0.161	0.163	0.166	0.166	0.167	0.168	0.171
620	0.148	0.149	0.153	0.155	0.156	0.159	0.159	0.161	0.165	0.164	0.165	0.166	0.169
630	0.146	0.148	0.151	0.154	0.155	0.157	0.158	0.159	0.163	0.163	0.164	0.165	0.167
640	0.144	0.146	0.150	0.152	0.153	0.156	0.156	0.158	0.161	0.161	0.162	0.163	0.166
650	0.143	0.145	0.148	0.151	0.152	0.154	0.155	0.156	0.160	0.159	0.161	0.161	0.165
660	0.142	0.144	0.147	0.150	0.150	0.153	0.153	0.155	0.158	0.158	0.159	0.160	0.163
670	0.140	0.142	0.145	0.148	0.149	0.151	0.151	0.153	0.157	0.156	0.157	0.158	0.161
680	0.139	0.140	0.144	0.147	0.147	0.149	0.150	0.151	0.155	0.154	0.156	0.156	0.159
690	0.137	0.139	0.142	0.145	0.145	0.148	0.148	0.150	0.153	0.153	0.154	0.154	0.158
700	0.136	0.138	0.141	0.144	0.144	0.146	0.147	0.148	0.152	0.151	0.152	0.153	0.156

Appendix H. (Continued)

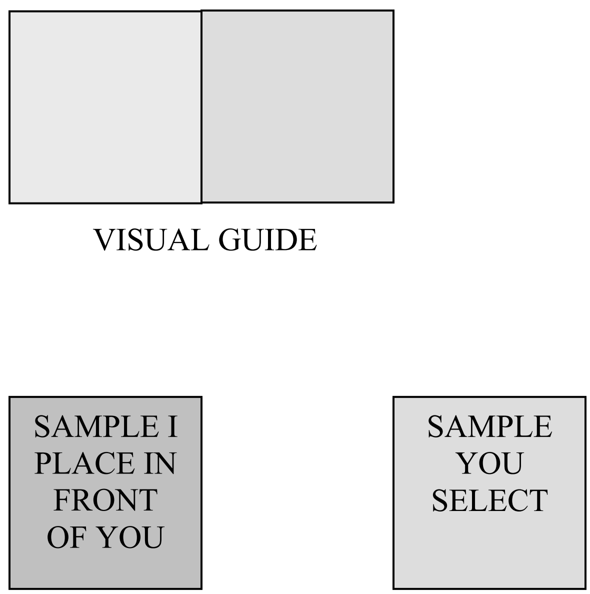
Λ (nm)	27	28	29	30	31
400	0.162	0.162	0.160	0.162	0.166
410	0.177	0.177	0.176	0.179	0.183
420	0.180	0.181	0.180	0.183	0.187
430	0.182	0.183	0.182	0.185	0.190
440	0.183	0.185	0.184	0.187	0.191
450	0.184	0.185	0.185	0.188	0.192
460	0.183	0.184	0.184	0.187	0.192
470	0.182	0.184	0.184	0.187	0.191
480	0.181	0.182	0.183	0.185	0.190
490	0.180	0.181	0.182	0.185	0.189
500	0.179	0.181	0.182	0.184	0.189
510	0.179	0.181	0.182	0.185	0.189
520	0.179	0.180	0.182	0.184	0.189
530	0.179	0.181	0.182	0.185	0.189
540	0.178	0.180	0.181	0.184	0.188
550	0.178	0.179	0.181	0.184	0.188
560	0.177	0.179	0.181	0.183	0.187
570	0.177	0.178	0.180	0.183	0.187
580	0.176	0.177	0.179	0.182	0.186
590	0.175	0.176	0.178	0.181	0.185
600	0.173	0.175	0.177	0.180	0.183
610	0.171	0.173	0.175	0.178	0.182
620	0.170	0.172	0.174	0.176	0.180
630	0.168	0.170	0.172	0.175	0.178
640	0.167	0.168	0.170	0.173	0.177
650	0.165	0.166	0.169	0.171	0.175
660	0.164	0.165	0.167	0.170	0.174
670	0.162	0.163	0.166	0.168	0.172
680	0.160	0.161	0.164	0.166	0.170
690	0.158	0.159	0.162	0.164	0.168
700	0.157	0.158	0.161	0.163	0.166

Appendix I. Summary of Colorimetric Data for all Gray Samples Used in the Development of the Perceptually Linear Scale Under Illuminant A and U-30.

	Illuminant A/2° Observer					Illuminant A/10° Observer					Illuminant U-30/2° Observer					Illuminant U-30/10° Observer				
Sample	L*	a*	b*	C*	h°	L*	a*	b*	C*	h°	L*	a*	b*	C*	h°	L*	a*	b*	C*	h°
Standard	41.44	-1.00	-0.68	1.21	213.99	41.44	-0.89	-0.67	1.11	216.65	41.47	-0.76	-0.62	0.98	219.26	41.48	-0.68	-0.61	0.91	221.89
1	42.07	-0.97	-0.54	1.11	209.24	42.07	-0.86	-0.52	1.01	211.32	42.10	-0.74	-0.44	0.86	210.58	42.11	-0.66	-0.42	0.78	212.41
2	42.23	-1.02	-0.54	1.15	208.01	42.24	-0.90	-0.52	1.04	209.90	42.27	-0.77	-0.43	0.88	209.49	42.27	-0.69	-0.41	0.80	211.12
3	42.46	-1.14	-0.84	1.42	216.39	42.47	-1.02	-0.82	1.32	218.83	42.50	-0.85	-0.77	1.15	222.05	42.50	-0.77	-0.76	1.08	224.56
4	42.88	-1.16	-0.85	1.44	216.33	42.89	-1.05	-0.84	1.34	218.73	42.92	-0.89	-0.78	1.18	221.20	42.92	-0.80	-0.77	1.11	223.71
5	43.11	-1.11	-0.80	1.37	215.95	43.12	-0.99	-0.79	1.27	218.42	43.15	-0.82	-0.72	1.09	221.18	43.15	-0.74	-0.71	1.03	223.65
6	43.36	-1.14	-0.84	1.42	216.39	43.37	-1.03	-0.83	1.32	218.92	43.39	-0.85	-0.77	1.14	222.13	43.39	-0.76	-0.76	1.07	224.68
7	43.74	-1.12	-0.82	1.39	216.38	43.75	-1.00	-0.81	1.29	218.94	43.78	-0.84	-0.75	1.12	221.92	43.79	-0.75	-0.74	1.06	224.49
8	44.26	-1.14	-0.78	1.38	214.37	44.27	-1.02	-0.76	1.28	216.81	44.30	-0.86	-0.69	1.10	219.03	44.30	-0.77	-0.68	1.03	221.44
9	44.63	-1.03	-0.71	1.25	214.69	44.64	-0.92	-0.70	1.15	217.18	44.66	-0.74	-0.64	0.98	220.61	44.66	-0.67	-0.62	0.91	222.88
10	44.87	-1.17	-0.82	1.43	214.98	44.88	-1.05	-0.80	1.33	217.36	44.91	-0.87	-0.74	1.15	220.29	44.91	-0.79	-0.73	1.07	222.69
11	45.11	-1.12	-0.67	1.30	211.11	45.12	-0.99	-0.66	1.19	213.51	45.16	-0.84	-0.57	1.02	214.18	45.16	-0.76	-0.56	0.94	216.43
12	45.22	-1.39	-1.33	1.92	223.60	45.24	-1.28	-1.31	1.83	225.82	45.26	-1.03	-1.28	1.64	231.14	45.27	-0.94	-1.27	1.58	233.41
13	45.44	-1.40	-1.34	1.94	223.77	45.46	-1.29	-1.33	1.85	225.99	45.48	-1.03	-1.31	1.66	231.85	45.49	-0.94	-1.30	1.61	234.11
14	45.96	-1.44	-1.40	2.01	224.26	45.98	-1.33	-1.39	1.92	226.39	46.01	-1.06	-1.37	1.73	232.22	46.01	-0.97	-1.36	1.67	234.41
15	46.21	-1.42	-1.37	1.98	224.00	46.23	-1.31	-1.36	1.89	226.15	46.25	-1.05	-1.34	1.70	231.96	46.26	-0.96	-1.33	1.64	234.13
16	46.64	-1.39	-1.27	1.88	222.38	46.65	-1.27	-1.26	1.79	224.73	46.68	-1.03	-1.22	1.60	229.91	46.69	-0.94	-1.21	1.54	232.28
17	46.91	-1.21	-1.00	1.57	219.52	46.93	-1.10	-0.98	1.47	221.88	46.95	-0.87	-0.94	1.28	227.14	46.96	-0.79	-0.92	1.21	229.40
18	47.11	-1.31	-1.26	1.82	223.87	47.13	-1.19	-1.25	1.73	226.38	47.15	-0.92	-1.24	1.55	233.41	47.16	-0.84	-1.23	1.49	235.70
19	47.44	-1.34	-1.26	1.84	223.20	47.45	-1.22	-1.25	1.75	225.69	47.48	-0.96	-1.23	1.56	231.97	47.49	-0.88	-1.22	1.50	234.30
20	47.53	-1.35	-1.29	1.87	223.65	47.55	-1.24	-1.28	1.78	226.1	47.57	-0.96	-1.26	1.59	232.60	47.58	-0.88	-1.25	1.53	234.88
21	47.75	-1.35	-1.27	1.86	223.18	47.76	-1.23	-1.26	1.76	225.73	47.79	-0.97	-1.25	1.58	232.20	47.80	-0.88	-1.24	1.52	234.56
22	48.23	-1.38	-1.48	2.02	226.90	48.25	-1.27	-1.47	1.95	229.18	48.27	-0.97	-1.49	1.78	236.72	48.28	-0.90	-1.48	1.73	238.76
23	48.22	-1.45	-1.69	2.23	229.20	48.24	-1.35	-1.68	2.16	231.31	48.26	-1.02	-1.71	2.00	239.12	48.27	-0.95	-1.71	1.96	240.99
24	48.40	-1.52	-1.92	2.45	231.66	48.43	-1.42	-1.93	2.39	233.57	48.44	-1.07	-1.98	2.25	241.65	48.45	-1.00	-1.98	2.22	243.33
25	48.51	-1.50	-1.65	2.23	227.84	48.53	-1.39	-1.65	2.15	230.00	48.55	-1.07	-1.67	1.98	237.41	48.56	-0.98	-1.67	1.94	239.41
26	48.81	-1.40	-1.47	2.03	226.52	48.83	-1.29	-1.47	1.95	228.74	48.85	-1.01	-1.47	1.78	235.48	48.86	-0.93	-1.46	1.73	237.58
27	48.90	-1.39	-1.43	2.00	225.77	48.92	-1.28	-1.43	1.92	228.07	48.94	-1.01	-1.41	1.74	234.49	48.95	-0.93	-1.41	1.68	236.66
28	49.09	-1.39	-1.36	1.94	224.36	49.11	-1.27	-1.35	1.86	226.69	49.13	-1.00	-1.33	1.66	233.12	49.14	-0.91	-1.32	1.60	235.29
29	49.31	-1.26	-0.98	1.60	217.82	49.33	-1.14	-0.96	1.49	220.22	49.36	-0.92	-0.89	1.28	224.02	49.36	-0.84	-0.88	1.21	226.28
30	49.64	-1.29	-0.98	1.62	217.15	49.65	-1.16	-0.96	1.51	219.48	49.68	-0.95	-0.89	1.30	222.92	49.69	-0.86	-0.87	1.22	225.17
31	50.11	-1.32	-1.06	1.69	218.60	50.13	-1.2	-1.04	1.59	220.94	50.15	-0.97	-0.99	1.38	225.45	50.16	-0.88	-0.97	1.31	227.72

Appendix J. Instructions for the Development of Gray Scale.

Thank you for agreeing to be an observer in this visual assessment. Please look at the samples in front of you for about two minutes so that your eyes are adapted to the experimental illumination setup. The purpose of this study is to develop a perceptually linear gray scale. The experiment is very simple. You will be shown a pair of samples with a perceptible color difference to guide your assessments. For each assessment I will place a sample below this pair. Your task is to find a sample from those in front of you that represents the same visual difference to that in the above pair when compared to the sample next to it. You will make a total of 9 selections.



Once you have completed the 9 selections your entire set will be shown to you. You can determine whether there are sample(s) that perceptually do not belong to the sequence of samples that in your opinion formed equal visual differences from one step to the next. You are allowed to change your selections until you are satisfied.

Do you understand the instructions? Are you ready to begin?

Appendix K. SAS Results of Polynomial Procedure for Obtaining a Statistically Reliable Equation to Convert Gray Scale Ratings to Visual Differences.

The GLM Procedure

Dependent Variable: difference

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	58.58747016	14.64686754	647.35	<.0001
Error	5	0.11312984	0.02262597		
Corrected Total	9	58.70060000			

R-Square	Coeff Var	Root MSE	difference Mean
0.998073	3.798467	0.150419	3.960000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	58.46427273	58.46427273	2583.95	<.0001
grade2	1	0.08051212	0.08051212	3.56	0.1179
grade3	1	0.04032168	0.04032168	1.78	0.2394
grade4	1	0.00236364	0.00236364	0.10	0.7596

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	0.25893860	0.25893860	11.44	0.0196
grade2	1	0.00000492	0.00000492	0.00	0.9888
grade3	1	0.00051386	0.00051386	0.02	0.8861
grade4	1	0.00236364	0.00236364	0.10	0.7596

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.0983216783	0.14560891	0.68	0.5295
grade	0.8545745921	0.25261268	3.38	0.0196
grade2	-.0018298368	0.12405085	-0.01	0.9888
grade3	0.0032051282	0.02126803	0.15	0.8861
grade4	-.0003787879	0.00117195	-0.32	0.7596

Appendix K. (Continued).

The GLM Procedure

Dependent Variable: difference

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	58.58510653	19.52836884	1014.52	<.0001
Error	6	0.11549347	0.01924891		
Corrected Total	9	58.70060000			

R-Square	Coeff Var	Root MSE	difference Mean
0.998032	3.503547	0.138740	3.960000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	58.46427273	58.46427273	3037.28	<.0001
grade2	1	0.08051212	0.08051212	4.18	0.0868
grade3	1	0.04032168	0.04032168	2.09	0.1980

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	0.72450442	0.72450442	37.64	0.0009
grade2	1	0.02179041	0.02179041	1.13	0.3283
grade3	1	0.04032168	0.04032168	2.09	0.1980

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.1146853147	0.12592377	0.91	0.3975
grade	0.7863927739	0.12818054	6.14	0.0009
grade2	0.0364277389	0.03423754	1.06	0.3283
grade3	-.0036130536	0.00249637	-1.45	0.1980

Appendix K. (Continued).

The GLM Procedure

Dependent Variable: difference

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	58.54478485	29.27239242	1315.06	<.0001
Error	7	0.15581515	0.02225931		
Corrected Total	9	58.70060000			

R-Square	Coeff Var	Root MSE	difference Mean
0.997346	3.767564	0.149196	3.960000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	58.46427273	58.46427273	2626.51	<.0001
grade2	1	0.08051212	0.08051212	3.62	0.0989

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	5.48613969	5.48613969	246.46	<.0001
grade2	1	0.08051212	0.08051212	3.62	0.0989

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.0236363636	0.11730430	0.20	0.8460
grade	0.9529545455	0.06070082	15.70	<.0001
grade2	-.0123484848	0.00649290	-1.90	0.0989

Appendix K. (Continued).

The GLM Procedure

Dependent Variable: difference

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	58.46427273	58.46427273	1979.10	<.0001
Error	8	0.23632727	0.02954091		
Corrected Total	9	58.70060000			

R-Square	Coeff Var	Root MSE	difference Mean
0.995974	4.340270	0.171875	3.960000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	58.46427273	58.46427273	1979.10	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	58.46427273	58.46427273	1979.10	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.1718181818	0.10102000	1.70	0.1274
grade	0.8418181818	0.01892278	44.49	<.0001

Appendix L. Variance Component Analysis: Perceptually Linear Scale.

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	774	4914.430995	6.349394	11.18	<.0001
Error	1550	880.606347	0.568133		
Corrected Total	2324	5795.037342			

R-Square	Coeff Var	Root MSE	value Mean
0.848041	27.70267	0.753746	2.720843

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	30	2721.545829	90.718194	159.68	<.0001
observer	24	1160.723391	48.363475	85.13	<.0001
swatch*observer	720	1032.161775	1.433558	2.52	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	2721.545829	90.718194	159.68	<.0001
observer	24	1160.723391	48.363475	85.13	<.0001
swatch*observer	720	1032.161775	1.433558	2.52	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	30	2721.545829	90.718194	63.28	<.0001
observer	24	1160.723391	48.363475	33.74	<.0001
Error	720	1032.161775	1.433558		

Error: MS(swatch*observer)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch*observer	720	1032.161775	1.433558	2.52	<.0001
Error: MS(Error)	1550	880.606347	0.568133		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	0.5046
swatch*observer	0.28
Residual	0.5681

Appendix M. Colorimetric Data for the 69 samples employed in the Replication Experiment.

Pair	L*	a*	b*	C*	hº	DL*	DC*	DH*	X	Y	Z	ΔE*ab	CMC (1:1)	CMC (2:1)	CIEDE2000 (1:1:1)	CIEDE2000 (2:1:1)
Standard I	61.96	-22.59	-20.26	30.34	221.90				23.36	30.36	49.64					
1	60.36	-18.64	-21.52	28.47	229.10	-1.60 D	-1.88 D	3.69 B	22.71	28.53	48.21	4.44	2.96	2.74	2.83	2.56
2	60.48	-24.10	-25.28	34.93	226.36	-1.48 D	4.58 B	2.53 B	21.63	28.66	52.04	5.44	3.09	2.90	2.67	2.43
3	60.48	-22.13	-21.14	30.60	223.68	-1.48 D	0.26 B	0.95 B	22.06	28.66	48.03	1.78	1.39	0.89	1.41	0.87
4	61.88	-20.14	-24.52	31.73	230.61	-0.08 D	1.39 B	4.71 B	23.83	30.27	53.72	4.91	3.25	3.25	2.85	2.85
5	59.10	-20.28	-22.31	30.15	227.73	-2.86 D	-0.20 D	3.08 B	21.19	27.13	46.91	4.21	3.14	2.39	3.13	2.26
Standard II	62.17	-22.66	-19.62	29.97	220.89				23.55	30.60	49.36					
6	61.84	-18.64	-19.21	26.77	225.87	-0.33 D	-3.20 D	2.46 B	24.14	30.22	48.44	4.05	2.32	2.31	2.20	2.18
7	62.10	-24.36	-23.03	33.52	223.39	-0.07 D	3.55 B	1.39 B	23.10	30.52	52.60	3.81	2.00	2.00	1.67	1.67
8	62.94	-24.37	-20.62	31.93	220.23	0.77 L	1.96 B	-0.35 G	23.90	31.51	51.63	2.13	1.18	1.05	1.08	0.92
9	63.65	-20.22	-21.53	29.53	226.80	1.48 L	-0.44 D	3.07 B	25.58	32.37	53.76	3.43	2.41	2.17	2.28	2.01
10	61.25	-19.69	-18.73	27.18	223.58	-0.92 D	-2.79 D	1.34 B	23.32	29.53	47.03	3.23	1.82	1.70	1.74	1.60
11	64.38	-18.20	-17.62	25.33	224.07	2.21 L	-4.64 D	1.53 B	26.83	33.27	51.12	5.36	3.11	2.69	3.03	2.56
Standard III	71.50	-19.18	19.06	27.04	135.17				34.79	42.92	30.71					
12	71.60	-20.04	20.53	28.69	134.30	0.10 L	1.65 B	-0.42 Y	34.67	43.07	29.81	1.71	0.91	0.91	0.78	0.78
13	71.39	-17.11	16.75	23.95	135.61	-0.11 D	-3.09 D	0.19 G	35.26	42.76	32.21	3.10	1.63	1.62	1.48	1.48
14	71.54	-17.84	20.24	26.98	131.40	0.04	-0.07 D	-1.78 Y	35.23	42.97	29.94	1.78	1.23	1.23	1.19	1.19
15	71.30	-19.04	16.91	25.46	138.40	-0.19 D	-1.58 D	1.48 G	34.59	42.63	32.00	2.17	1.32	1.32	1.20	1.19
16	72.91	-19.48	19.21	27.36	135.40	1.41 L	0.32 B	0.11 G	36.51	45.03	32.33	1.45	1.11	0.57	1.08	0.56
17	70.01	-18.73	18.76	26.51	134.97	-1.49 D	-0.53 D	-0.10 Y	33.08	40.76	29.15	1.59	1.19	0.64	1.17	0.63
Standard IV	75.92	-30.05	24.49	38.77	140.82				37.24	49.77	32.27					
18	76.20	-32.09	25.80	41.18	141.20	76.20	41.18	141.20	36.98	50.21	31.67	2.44	1.09	1.07	0.90	0.88
19	75.11	-26.52	19.93	33.17	143.07	75.11	33.17	143.07	37.25	48.46	34.61	5.83	2.66	2.61	2.36	2.30
20	78.87	-28.97	22.58	36.73	142.06	78.87	36.73	142.06	41.61	54.70	37.59	3.67	2.44	1.50	2.27	1.37
21	79.92	-27.87	22.22	35.64	141.43	79.92	35.64	141.43	43.49	56.55	39.36	5.09	3.31	2.05	3.07	1.85
22	73.32	-27.02	18.31	32.64	145.88	73.32	32.64	145.88	34.80	45.66	33.51	7.36	3.79	3.39	3.52	3.10
Standard V	62.64	-17.71	0.93	17.73	176.99				25.15	31.16	32.75					
23	62.50	-16.17	1.10	16.20	176.10	-0.14 D	-1.53 D	-0.26 Y	25.37	30.99	32.45	1.56	1.02	1.01	1.06	1.05

Appendix M. (Continued).

Pair	L*	a*	b*	C*	hº	DL*	DC*	DH*	X	Y	Z	ΔE*ab	CMC (1:1)	CMC (2:1)	CIEDE2000 (1:1:1)	CIEDE2000 (2:1:1)
24	62.72	-19.91	0.67	19.92	178.08	0.08 L	2.19 B	0.36 B	24.71	31.25	33.04	2.22	1.44	1.44	1.40	1.40
25	62.82	-18.46	2.73	18.66	171.59	0.18 L	0.92 B	-1.71 Y	25.15	31.37	31.67	1.96	1.58	1.58	1.37	1.36
26	62.61	-17.29	-0.56	17.30	181.86	-0.03	-0.43 D	1.49 B	25.21	31.12	33.81	1.55	1.30	1.30	1.14	1.14
27	61.07	-17.12	0.56	17.13	178.12	-1.57 D	-0.60 D	0.34 B	23.72	29.33	31.07	1.71	1.37	0.80	1.42	0.82
28	64.61	-18.57	3.87	18.97	168.22	1.97 L	1.23 B	-2.81 Y	26.98	33.56	33.08	3.64	2.99	2.64	2.71	2.30
29	61.89	-18.01	-2.63	18.20	188.32	-0.75 D	0.47 B	3.55 B	24.32	30.27	34.43	3.66	3.09	3.05	2.72	2.66
30	60.48	-19.06	-2.95	19.29	188.81	-2.16 D	1.56 B	3.81 B	22.73	28.66	32.87	4.65	3.82	3.50	3.50	3.10
31	63.75	-20.62	-0.17	20.62	180.47	1.12 L	2.89 B	1.16 B	25.59	32.50	35.00	3.31	2.29	2.15	2.19	2.03
Standard VI	61.32	-10.24	-3.63	10.86	199.54				25.57	29.61	34.45					
32	60.82	-11.86	-4.77	12.78	201.90	-0.50 D	1.92 B	0.49 B	24.68	29.04	34.65	2.04	1.67	1.64	1.59	1.54
33	60.88	-10.29	-2.50	10.59	193.64	-0.44 D	-0.28 D	-1.10 Y	25.11	29.11	33.04	1.22	1.22	1.18	1.00	0.95
34	56.84	-13.71	-5.65	14.83	202.41	-4.48 D	3.96 B	0.64 B	20.53	24.76	30.32	6.01	4.93	3.74	5.00	3.62
Standard VII	57.82	-36.18	6.85	36.82	169.28				17.01	25.77	23.42					
35	59.53	-40.56	9.80	41.73	166.41	1.71 L	4.90 B	-1.96 Y	17.56	27.61	23.42	5.55	2.90	2.61	2.62	2.26
Standard VIII	38.06	9.32	-34.41	35.65	285.16				10.80	10.12	27.88					
36	36.46	8.26	-31.19	32.27	284.83	-1.60 D	-3.39 D	-0.19 G	9.77	9.25	24.14	3.75	2.31	1.77	1.83	1.40
37	35.93	12.78	-39.49	41.51	287.94	-2.13 D	5.85 B	1.87 R	10.05	8.97	28.80	6.50	3.82	3.27	2.33	1.72
38	34.32	11.33	-36.15	37.88	287.41	-3.74 D	2.23 B	1.44 R	9.02	8.16	24.91	4.59	4.29	2.53	3.23	1.77
39	38.34	8.70	-33.19	34.31	284.69	0.28 L	-1.35 D	-0.29 G	10.87	10.28	27.40	1.40	0.72	0.67	0.45	0.40
40	34.84	10.15	-34.20	35.67	286.53	-3.21 D	0.02	0.85 R	9.15	8.42	24.27	3.33	3.51	1.85	2.79	1.53
41	37.33	10.28	-35.86	37.31	286.00	-0.73 D	1.65 B	0.54 R	10.50	9.72	28.00	1.89	1.17	0.95	0.75	0.53
Standard IX	42.35	12.28	-26.15	28.89	295.16				13.92	12.73	27.31					
42	42.90	9.55	-21.16	23.22	294.28	0.55 L	-5.68 D	-0.40 B	13.86	13.09	24.79	5.72	2.95	2.91	2.60	2.56
43	41.37	12.23	-18.82	22.44	303.03	-0.98 D	-6.45 D	3.50 R	13.25	12.09	21.88	7.40	4.49	4.41	4.87	4.81
44	42.23	9.42	-27.68	29.24	288.80	-0.12 D	0.34 B	-3.22 B	13.39	12.65	28.18	3.24	2.70	2.70	2.98	2.97
45	40.91	12.20	-26.16	28.87	295.01	-1.44 D	-0.02	-0.07 B	12.95	11.81	25.75	1.44	1.45	0.73	1.29	0.65
46	44.00	12.46	-26.66	29.43	295.04	1.65 L	0.53 B	-0.06 B	15.11	13.84	29.54	1.74	1.69	0.88	1.54	0.80

Appendix M. (Continued).

Pair	L*	a*	b*	C*	hº	DL*	DC*	DH*	X	Y	Z	ΔE*ab	CMC (1:1)	CMC (2:1)	CIEDE2000 (1:1:1)	CIEDE2000 (2:1:1)
Standard X	42.05	17.93	30.37	35.27	59.45				14.62	12.53	4.54					
47	41.97	19.30	32.85	38.10	59.57	-0.08 D	2.83 B	0.08 Y	14.79	12.48	4.05	2.83	1.31	1.30	1.07	1.07
48	42.20	16.08	26.44	30.94	58.69	0.15 L	-4.32 D	-0.43 R	14.43	12.63	5.41	4.35	2.06	2.05	1.78	1.77
49	42.23	16.60	31.45	35.57	62.17	0.18 L	0.30 B	1.68 Y	14.54	12.65	4.40	1.72	1.99	1.98	1.30	1.30
50	41.72	20.39	28.75	35.25	54.65	-0.33 D	-0.02	-2.95 R	14.79	12.32	4.75	2.97	3.48	3.47	2.27	2.25
51	40.68	17.66	29.43	34.32	59.04	-1.37 D	-0.94 D	-0.25 R	13.63	11.66	4.27	1.69	1.48	0.87	1.30	0.74
52	43.32	18.60	31.45	36.54	59.39	1.27 L	1.27 B	-0.04	15.66	13.38	4.77	1.80	1.41	0.87	1.26	0.76
Standard XI	38.02	20.10	31.91	37.71	57.80				12.27	10.10	3.08					
53	37.61	21.55	32.71	39.17	56.63	-0.41 D	1.46 B	-0.78 R	12.23	9.87	2.86	1.71	1.21	1.14	0.87	0.81
54	37.38	18.24	30.19	35.27	58.87	-0.64 D	-2.44 D	0.68 Y	11.61	9.74	3.17	2.61	1.51	1.39	1.19	1.10
55	37.03	18.08	30.86	35.76	59.64	-0.98 D	-1.94 D	1.18 Y	11.38	9.56	2.98	2.48	1.95	1.72	1.43	1.23
56	37.02	22.00	30.66	37.74	54.34	-1.00 D	0.03	-2.28 R	11.93	9.55	3.01	2.49	2.87	2.72	1.89	1.74
57	35.95	19.59	30.01	35.84	56.86	-2.07 D	-1.87 D	-0.60 R	10.95	8.98	2.83	2.86	2.48	1.56	1.94	1.21
58	38.53	20.35	32.66	38.49	58.07	0.52 L	0.78 B	0.18 Y	12.64	10.39	3.10	0.95	0.69	0.49	0.55	0.39
Standard XII	31.04	-9.10	16.46	18.81	118.94				5.51	6.67	3.62					
59	31.07	-11.18	18.15	21.31	121.65	0.03	2.50 B	0.95 G	5.35	6.68	3.36	2.68	1.77	1.77	1.70	1.70
60	30.88	-6.82	14.85	16.34	114.66	-0.16 D	-2.47 D	-1.31 Y	5.65	6.60	3.85	2.80	1.93	1.92	2.07	2.07
61	31.10	-10.48	15.55	18.75	123.99	0.06 L	-0.06 D	1.66 G	5.41	6.70	3.80	1.66	1.44	1.43	1.46	1.46
62	31.00	-7.07	17.02	18.43	112.56	-0.04	-0.38 D	-2.07 Y	5.67	6.65	3.52	2.11	1.81	1.81	1.99	1.99
63	28.71	-8.83	14.47	16.95	121.39	-2.33 D	-1.86 D	0.76 G	4.72	5.73	3.29	3.08	3.14	1.95	2.15	1.48
64	32.68	-9.95	17.14	19.82	120.14	1.64 L	1.01 B	0.41 G	6.06	7.39	4.00	1.97	2.12	1.23	1.48	0.97
65	33.42	-9.41	17.54	19.91	118.22	2.37 L	1.10 B	-0.24 Y	6.40	7.73	4.15	2.63	2.98	1.61	1.98	1.13
Standard XIII	34.25	-16.20	22.82	27.99	125.37				6.11	8.13	3.49					
66	34.04	-15.73	23.05	27.91	124.32	-0.21 D	-0.08 D	-0.51 Y	6.07	8.03	3.39	0.56	0.43	0.38	0.40	0.37
67	32.43	-14.07	16.51	21.69	130.44	-1.82 D	-6.30 D	2.18 G	5.60	7.28	4.03	6.91	4.14	3.73	3.61	3.38
68	32.71	-10.37	18.25	20.99	119.62	-1.54 D	-7.00 D	-2.43 Y	6.03	7.41	3.81	7.57	4.35	4.08	4.20	4.06
69	36.99	-18.31	25.53	31.42	125.65	2.73 L	3.43 B	0.15 G	7.03	9.53	3.83	4.39	3.59	2.36	2.71	1.87

APPENDIX N. Reflectance Data for the 69 Colored Samples Used in the Replication Study.

Pair	400	410	420	430	440	450	460	470	480	490	500	510	520	530	540	550	560	570	580	590	600
Standard I	38.17	41.13	43.53	45.09	45.93	46.62	47.27	47.78	48.00	48.08	47.52	45.91	43.11	39.21	34.97	30.58	25.98	22.02	19.06	16.95	15.10
1	37.13	41.00	43.90	45.92	45.95	45.49	45.60	44.75	43.65	44.00	44.15	41.91	39.04	36.49	33.15	28.92	24.46	20.79	18.01	16.06	14.39
2	35.93	40.43	44.33	47.35	48.47	48.14	49.55	51.37	51.55	51.67	50.77	47.34	42.43	37.07	31.97	27.04	22.12	18.13	15.35	13.54	12.03
3	35.82	39.57	42.61	44.84	45.14	45.15	45.57	45.27	44.92	45.79	46.35	44.55	41.26	37.45	33.32	28.79	23.93	19.80	16.91	14.98	13.36
4	43.06	46.13	48.30	49.43	49.86	50.30	51.07	51.67	51.71	50.81	48.62	45.69	42.22	37.91	33.62	29.61	25.50	21.66	18.60	16.33	14.57
5	35.01	38.90	42.00	44.30	44.51	44.33	44.49	43.88	43.03	43.50	43.72	41.54	38.34	35.22	31.53	27.16	22.61	18.86	16.15	14.31	12.76
Standard II	38.53	41.35	43.50	44.88	45.64	46.28	46.92	47.43	47.71	47.85	47.40	45.92	43.30	39.57	35.41	31.05	26.45	22.48	19.48	17.28	15.37
6	38.42	42.03	44.51	46.16	45.96	45.47	45.65	44.88	43.90	44.53	45.11	43.21	40.61	38.55	35.52	31.25	26.66	22.86	19.95	17.80	15.97
7	37.80	42.14	45.58	48.08	48.89	48.39	49.77	51.55	51.97	52.23	52.03	49.16	44.73	39.72	34.69	29.62	24.46	20.17	17.20	15.17	13.47
8	43.83	45.51	45.78	45.75	46.16	47.41	49.40	51.13	52.01	51.98	50.49	47.76	44.34	40.25	35.82	31.37	27.13	23.47	20.28	17.58	15.52
9	45.24	47.72	49.02	49.66	49.76	50.03	50.69	51.40	51.65	51.17	49.63	47.33	44.46	40.62	36.52	32.53	28.30	24.31	21.02	18.55	16.63
10	37.32	40.69	43.10	44.60	44.45	44.12	44.24	43.68	42.98	43.78	44.35	42.90	40.38	38.08	34.93	30.65	25.97	22.05	19.11	17.01	15.20
11	42.72	44.79	46.06	46.89	47.42	48.06	48.55	48.72	48.47	48.00	47.09	45.56	43.44	40.69	37.59	34.21	30.58	27.22	24.43	22.12	20.01
Standard III	25.06	23.78	23.22	23.46	24.17	25.87	28.58	32.18	37.50	42.87	47.74	50.23	50.58	50.32	49.38	48.36	46.37	42.62	38.83	36.54	35.50
12	24.01	22.70	22.20	22.48	23.18	24.93	27.67	31.37	36.93	42.68	48.04	50.82	51.40	50.94	49.85	48.63	46.47	42.62	38.80	36.49	35.45
13	26.95	25.61	25.02	25.18	25.89	27.54	30.10	33.47	38.28	42.93	47.00	48.96	49.24	49.11	48.45	47.81	46.24	42.85	39.24	37.03	36.01
14	24.49	23.20	22.64	22.90	23.54	25.22	27.86	31.36	36.55	41.77	46.55	49.05	49.69	49.63	49.04	48.44	46.92	43.47	39.79	37.54	36.50
15	26.40	25.09	24.55	24.70	25.48	27.19	29.87	33.41	38.53	43.58	48.06	50.24	50.38	50.01	48.93	47.81	45.71	41.94	38.17	35.90	34.88
16	26.29	25.01	24.49	24.78	25.51	27.28	30.09	33.83	39.35	44.98	50.14	52.79	53.32	52.87	51.80	50.63	48.46	44.56	40.68	38.34	37.28
17	23.90	22.58	22.02	22.22	22.91	24.56	27.16	30.60	35.66	40.71	45.24	47.51	47.92	47.67	46.84	45.99	44.19	40.60	36.92	34.71	33.69
Standard IV	34.66	29.87	26.10	23.04	21.31	21.44	23.98	32.37	53.01	71.42	78.24	74.56	67.46	59.47	52.82	47.82	43.91	41.21	39.39	38.35	38.17
18	33.83	29.07	25.23	22.23	20.55	20.70	23.26	31.72	52.75	72.16	79.98	76.55	69.22	60.90	54.03	48.72	44.34	41.07	38.90	37.71	37.53
19	37.73	33.13	29.14	25.98	24.11	24.20	26.71	34.95	53.85	68.67	73.13	69.74	63.61	56.83	51.32	47.07	43.45	40.64	38.77	37.81	37.73
20	38.62	34.64	31.03	27.94	26.00	26.20	28.77	37.33	58.96	77.47	83.62	79.71	72.62	64.82	58.31	53.23	49.04	45.93	43.69	42.26	41.89
21	40.99	36.52	32.54	29.29	27.39	27.58	30.51	39.82	61.26	78.60	84.43	80.98	74.11	66.44	60.06	55.13	50.97	47.88	45.83	44.71	44.57
22	36.63	32.06	28.22	25.14	23.35	23.44	25.92	33.96	52.28	66.52	69.72	67.00	60.80	53.99	48.42	44.10	40.41	37.60	35.73	34.78	34.71

APPENDIX N. (Continued).

Pair	610	620	630	640	650	660	670	680	690	700
Standard I	13.33	11.97	12.01	12.68	14.40	16.48	18.25	19.72	20.76	23.31
1	12.76	11.54	11.68	13.36	16.52	21.51	28.00	36.15	42.82	51.42
2	10.60	9.54	9.66	11.14	13.98	18.50	24.53	32.29	38.93	47.66
3	11.81	10.67	10.81	12.42	15.44	20.23	26.48	34.45	41.13	49.85
4	13.11	11.94	11.84	12.28	13.90	17.39	23.64	32.59	39.90	48.89
5	11.25	10.13	10.26	11.78	14.71	19.38	25.52	33.36	39.99	48.62
Standard II	13.53	12.13	12.16	12.83	14.50	16.49	18.02	19.26	20.19	22.67
6	14.15	12.83	13.03	14.98	18.54	24.04	30.98	39.48	46.23	54.88
7	11.83	10.65	10.85	12.60	15.83	20.89	27.41	35.56	42.29	51.13
8	13.92	12.45	11.28	10.51	10.42	10.79	11.88	12.94	13.66	15.6
9	15.04	13.75	13.60	13.99	15.66	19.29	25.95	35.39	42.87	51.81
10	13.43	12.13	12.35	14.23	17.67	23.04	29.87	38.25	44.96	53.65
11	18.00	16.34	15.59	15.72	16.08	16.38	16.58	16.72	17.15	19.12
Standard III	33.33	30.21	29.15	31.10	37.71	48.77	60.05	68.04	70.80	72.54
12	33.30	30.19	29.14	31.13	37.68	48.70	59.99	68.04	70.85	72.66
13	33.89	30.78	29.75	31.79	38.44	49.49	60.65	68.44	71.08	72.74
14	34.33	31.16	30.08	32.10	38.74	49.74	60.70	68.31	70.83	72.47
15	32.77	29.73	28.70	30.67	37.27	48.31	59.60	67.59	70.39	72.2
16	35.09	31.92	30.83	32.87	39.57	50.44	61.07	68.25	70.61	72.2
17	31.59	28.57	27.56	29.53	36.06	47.14	58.79	67.25	70.27	72.22
Standard IV	38.23	38.47	39.77	41.96	45.13	49.55	54.87	60.92	64.99	68.68
18	37.56	37.95	39.27	41.46	44.71	49.18	54.58	60.77	64.98	68.79
19	37.84	38.31	39.60	41.76	44.96	49.36	54.74	60.87	65.03	68.73
20	42.05	42.76	44.60	47.42	51.06	55.32	59.72	64.22	67.08	69.63
21	44.63	45.05	46.42	48.66	51.75	55.79	60.27	65.07	68.09	70.63
22	34.83	35.31	36.62	38.78	42.03	46.52	52.12	58.75	63.39	67.69

APPENDIX N. (Continued).

Pair	400	410	420	430	440	450	460	470	480	490	500	510	520	530	540	550	560	570	580	590	600
Standard V	36.83	34.46	31.33	28.41	26.42	26.24	27.48	32.38	41.74	45.80	44.22	40.39	36.67	33.71	31.67	30.22	29.44	29.31	28.97	27.62	25.39
23	36.77	34.32	31.21	28.27	26.26	26.06	27.27	32.07	41.09	44.82	43.16	39.41	35.84	33.06	31.14	29.88	29.38	29.43	29.31	28.24	26.08
24	36.59	34.35	31.34	28.48	26.53	26.40	27.69	32.74	42.51	46.99	45.51	41.56	37.67	34.53	32.28	30.58	29.49	28.99	28.21	26.61	24.30
25	35.76	33.18	30.04	27.10	25.17	25.03	26.35	31.46	41.42	46.11	44.74	40.85	37.06	34.02	31.92	30.43	29.66	29.56	29.26	27.94	25.70
26	37.57	35.44	32.50	29.58	27.56	27.39	28.59	33.39	42.29	45.83	44.04	40.25	36.57	33.69	31.65	30.19	29.37	29.19	28.79	27.40	25.15
27	35.39	33.01	29.93	27.04	25.11	24.89	26.04	30.67	39.50	43.25	41.61	37.93	34.39	31.63	29.70	28.36	27.74	27.63	27.39	26.13	23.98
28	37.15	34.40	31.11	28.09	26.11	26.10	27.48	32.97	43.68	48.81	47.51	43.48	39.49	36.26	34.05	32.53	31.80	31.73	31.51	30.17	27.90
29	39.70	36.03	32.33	29.10	27.01	26.90	28.56	34.73	46.43	51.33	49.08	43.93	38.65	34.01	30.41	27.51	25.03	23.23	21.92	20.94	20.27
30	38.01	34.36	30.68	27.52	25.51	25.48	27.11	33.27	45.11	50.29	47.99	42.60	37.19	32.42	28.72	25.75	23.24	21.46	20.18	19.23	18.59
31	39.74	35.92	32.09	28.85	26.81	26.65	28.75	35.65	49.17	55.72	53.78	48.08	42.18	36.90	32.83	29.56	26.80	24.80	23.38	22.31	21.62
Standard VI	35.71	35.95	34.84	33.87	32.38	31.59	31.49	30.67	30.06	30.71	31.72	31.79	31.88	32.65	33.55	33.63	32.64	30.90	28.81	26.44	23.96
32	34.31	34.94	34.06	33.42	32.21	31.86	31.94	31.42	31.01	31.87	32.99	32.84	32.58	32.82	32.94	32.52	30.94	28.96	27.02	24.99	22.81
33	33.64	33.85	32.76	32.00	30.70	30.28	30.33	29.81	29.44	30.35	31.61	31.58	31.63	32.14	32.87	32.74	31.49	29.81	28.09	26.23	24.13
34	34.75	32.18	28.60	26.24	25.70	23.60	26.22	32.92	36.60	36.78	35.30	33.12	30.64	28.11	25.77	23.62	21.65	20.11	19.00	18.28	17.78
Standard VII	17.56	17.37	16.58	17.43	18.19	19.14	22.72	25.80	27.64	32.18	39.78	42.49	40.35	36.20	31.99	27.95	24.11	20.88	18.13	15.74	13.61
35	24.43	22.08	18.52	16.20	16.11	14.16	17.44	28.80	40.85	45.81	46.26	44.19	40.90	36.97	32.98	29.06	25.24	21.98	19.20	16.76	14.58
Standard VIII	34.18	35.12	34.21	32.21	29.41	26.56	23.75	21.27	18.80	16.77	14.85	13.10	11.63	10.39	9.31	8.30	7.38	6.71	6.27	5.96	5.74
36	31.52	31.51	30.03	27.85	25.22	22.74	20.41	18.44	16.47	14.86	13.31	11.89	10.63	9.61	8.65	7.73	6.89	6.27	5.86	5.57	5.37
37	33.55	35.54	35.74	34.18	31.25	27.86	24.37	21.27	18.26	15.86	13.66	11.73	10.16	8.91	7.86	6.92	6.11	5.54	5.16	4.90	4.73
38	30.94	32.01	31.33	29.42	26.66	23.81	20.97	18.49	16.05	14.09	12.27	10.64	9.30	8.21	7.27	6.43	5.68	5.17	4.80	4.55	4.40
39	34.23	34.79	33.63	31.54	28.74	25.98	23.32	21.00	18.67	16.77	14.94	13.26	11.83	10.63	9.55	8.52	7.59	6.91	6.47	6.13	5.91
40	30.96	31.54	30.49	28.44	25.74	23.05	20.44	18.20	15.99	14.19	12.47	10.94	9.64	8.57	7.63	6.77	6.00	5.48	5.09	4.81	4.66
41	34.10	35.28	34.55	32.62	29.75	26.77	23.79	21.15	18.54	16.43	14.43	12.64	11.13	9.89	8.82	7.82	6.95	6.30	5.89	5.59	5.38
Standard IX	29.61	29.08	28.16	27.31	26.71	26.35	25.67	24.54	22.47	19.81	16.65	14.07	12.36	10.98	9.68	8.98	9.09	10.00	10.92	11.11	11.09
42	30.02	28.74	26.85	24.85	23.23	22.45	22.17	22.60	23.01	21.19	17.87	14.93	12.95	11.32	9.85	9.10	9.29	10.40	11.54	11.88	11.79
43	29.34	27.95	25.91	23.88	21.97	20.38	18.98	17.74	16.44	15.22	13.79	12.55	11.69	10.92	10.00	9.45	9.57	10.41	11.23	11.39	11.34
44	28.71	29.26	28.70	28.62	28.20	27.26	26.49	24.88	22.33	20.03	17.41	15.00	13.22	11.76	10.40	9.45	9.34	9.56	9.81	9.82	9.66
45	28.21	27.54	26.55	25.68	25.13	24.83	24.23	23.20	21.26	18.68	15.59	13.10	11.47	10.15	8.91	8.25	8.34	9.18	10.03	10.21	10.19
46	33.75	33.73	32.14	30.48	29.55	27.31	26.70	26.46	24.29	21.30	17.99	15.30	13.51	12.06	10.68	9.85	9.96	10.99	11.94	12.12	12.07

APPENDIX N. (Continued).

Pair	610	620	630	640	650	660	670	680	690	700
Standard V	22.90	20.56	18.80	17.59	16.77	15.98	15.21	14.89	15.05	16.68
23	23.59	21.24	19.45	18.23	17.42	16.59	15.81	15.49	15.65	17.32
24	21.82	19.49	17.77	16.60	15.79	15.02	14.29	14.02	14.12	15.69
25	23.20	20.81	19.06	17.82	17.00	16.20	15.42	15.12	15.24	16.90
26	22.67	20.33	18.60	17.38	16.58	15.80	15.02	14.73	14.86	16.50
27	21.58	19.31	17.62	16.45	15.67	14.92	14.17	13.95	14.02	15.63
28	25.29	22.79	20.94	19.65	18.78	17.92	17.09	16.74	16.93	18.67
29	20.37	20.91	23.42	27.80	34.21	42.57	51.67	60.19	64.90	68.42
30	18.69	19.24	21.61	25.88	32.11	40.46	49.85	59.04	64.39	68.49
31	21.71	22.27	24.82	29.31	35.81	44.23	53.28	61.58	66.10	69.47
Standard VI	21.49	19.21	17.56	16.40	15.60	14.83	14.06	13.81	13.91	15.53
32	20.53	18.36	16.75	15.64	14.88	14.14	13.41	13.11	13.23	14.82
33	21.83	19.60	17.94	16.78	15.97	15.19	14.41	14.11	14.22	15.84
34	17.56	17.50	18.18	19.75	22.37	26.31	31.83	39.61	46.79	55.49
Standard VII	11.73	10.13	9.01	8.26	7.77	7.26	6.84	6.70	6.80	7.83
35	12.64	10.98	9.78	8.97	8.47	7.93	7.50	7.34	7.43	8.54
Standard VIII	5.77	6.02	7.17	9.43	13.17	18.94	26.77	36.89	45.43	54.90
36	5.41	5.64	6.74	8.90	12.52	18.16	25.89	35.92	44.50	54.15
37	4.76	4.99	5.92	7.87	11.20	16.55	23.96	33.71	42.20	52.20
38	4.43	4.69	5.51	7.34	10.51	15.68	22.97	32.61	41.15	51.32
39	5.95	6.21	7.40	9.73	13.57	19.53	27.56	37.89	46.53	56.02
40	4.69	4.99	5.84	7.76	11.10	16.43	23.89	33.68	42.26	52.27
41	5.41	5.66	6.75	8.92	12.53	18.22	25.98	36.03	44.60	54.26
Standard IX	11.09	11.33	13.11	16.42	21.34	28.65	37.69	48.18	55.56	62.81
42	11.75	11.93	13.75	17.15	22.16	29.58	38.67	49.18	56.64	64.16
43	11.36	11.64	13.45	16.77	21.78	29.11	38.21	48.76	56.26	63.56
44	9.65	9.92	11.56	14.65	19.41	26.42	35.32	45.85	53.62	61.38
45	10.19	10.41	12.09	15.23	20.00	27.09	36.00	46.60	54.48	62.52
46	12.07	12.32	14.20	17.67	22.79	30.26	39.40	49.91	57.30	64.60

APPENDIX N. (Continued).

Pair	400	410	420	430	440	450	460	470	480	490	500	510	520	530	540	550	560	570	580	590	600
Standard X	5.80	5.36	4.74	4.47	4.10	3.99	3.95	3.91	4.02	4.38	5.15	5.58	6.20	7.68	10.39	13.28	15.13	15.75	16.31	17.05	17.95
47	5.13	4.70	4.21	3.94	3.69	3.55	3.53	3.49	3.61	3.90	4.57	4.98	5.57	7.00	9.80	13.08	15.43	16.33	16.92	17.63	18.50
48	7.01	6.39	5.72	5.35	4.89	4.76	4.70	4.66	4.72	5.22	6.10	6.59	7.22	8.67	11.11	13.40	14.58	14.97	15.46	16.26	17.22
49	5.57	5.10	4.58	4.27	3.99	3.86	3.82	3.79	3.90	4.26	5.01	5.45	6.07	7.66	10.54	13.83	16.05	16.71	17.04	17.40	17.93
50	6.15	5.59	4.99	4.69	4.30	4.18	4.13	4.09	4.18	4.57	5.34	5.75	6.30	7.63	9.98	12.31	13.66	14.27	15.16	16.41	17.87
51	5.38	4.95	4.45	4.19	3.88	3.76	3.71	3.67	3.80	4.10	4.79	5.19	5.76	7.13	9.64	12.35	14.06	14.63	15.15	15.82	16.66
52	6.11	5.60	4.98	4.66	4.31	4.19	4.14	4.11	4.17	4.63	5.45	5.93	6.58	8.15	10.98	14.06	16.06	16.82	17.48	18.34	19.34
Standard XI	5.59	4.33	3.56	3.13	2.86	2.71	2.61	2.55	2.56	2.59	2.71	2.85	3.27	4.08	5.65	8.22	11.94	15.96	18.98	20.06	19.62
53	4.84	3.76	3.12	2.84	2.65	2.53	2.47	2.43	2.46	2.48	2.59	2.72	3.06	3.8	5.19	7.55	11.11	15.2	18.58	20.28	20.01
54	6.11	4.59	3.77	3.26	2.96	2.77	2.65	2.59	2.61	2.63	2.76	2.92	3.38	4.22	5.79	8.31	11.88	15.58	18.17	18.87	18.2
55	4.91	3.91	3.34	2.99	2.77	2.64	2.57	2.51	2.54	2.57	2.68	2.82	3.19	3.97	5.53	8.14	11.82	15.53	18.02	18.6	17.86
56	5.53	4.16	3.43	3.05	2.81	2.66	2.56	2.5	2.51	2.53	2.63	2.77	3.11	3.8	5.15	7.37	10.64	14.26	17.2	18.76	18.74
57	4.93	3.75	3.16	2.83	2.63	2.51	2.44	2.4	2.42	2.44	2.53	2.65	3.01	3.6	4.88	7.07	10.36	14.05	16.94	18.02	17.63
58	5.57	4.29	3.57	3.13	2.9	2.72	2.62	2.57	2.6	2.63	2.76	2.92	3.39	4.26	5.86	8.48	12.22	16.3	19.43	20.63	20.19
Standard XII	4.27	3.77	3.34	3.11	2.98	2.91	3.04	3.50	4.07	4.76	5.68	6.15	6.65	7.62	8.86	9.36	8.81	7.25	5.97	5.37	5.03
59	3.68	3.20	2.90	2.75	2.66	2.62	2.73	3.27	4.10	5.02	6.11	6.65	7.15	8.03	9.03	9.13	8.59	7.03	5.77	5.19	4.88
60	4.75	4.24	3.74	3.47	3.30	3.21	3.33	3.57	3.98	4.43	5.21	5.63	6.14	7.16	8.57	9.32	8.93	7.47	6.17	5.56	5.22
61	4.50	3.92	3.45	3.21	3.09	3.01	3.16	3.71	4.39	5.18	6.16	6.63	7.09	7.97	8.98	9.09	8.56	7.00	5.74	5.18	4.86
62	4.13	3.68	3.28	3.07	2.95	2.88	3.02	3.32	3.84	4.36	5.17	5.62	6.14	7.17	8.65	9.44	9.05	7.62	6.32	5.68	5.31
63	3.76	3.33	3.00	2.84	2.74	2.68	2.82	3.14	3.68	4.23	5.02	5.42	5.83	6.63	7.62	7.97	7.43	6.09	5.00	4.51	4.21
64	4.80	4.17	3.67	3.39	3.25	3.16	3.33	3.90	4.58	5.40	6.43	6.97	7.51	8.53	9.82	10.30	9.65	7.97	6.59	5.93	5.55
65	3.27	2.78	2.67	2.58	2.59	2.68	3.11	3.97	7.51	11.43	12.12	10.86	9.55	8.41	7.38	6.62	6.43	6.44	6.52	6.47	6.35
Standard XIII	3.59	2.88	2.55	2.41	2.34	2.31	2.41	2.69	4.62	9.11	13.58	13.59	11.90	10.25	8.86	7.78	6.96	6.47	6.10	5.80	5.60
66	3.26	2.75	2.51	2.38	2.34	2.32	2.40	2.59	4.16	8.34	13.32	13.71	11.99	10.19	8.61	7.44	6.74	6.46	6.18	5.94	5.71
67	3.18	2.75	2.61	2.53	2.53	2.61	2.87	3.76	7.24	11.42	12.18	11.18	9.75	8.53	7.48	6.62	6.01	5.72	5.45	5.26	5.08
68	2.98	2.64	2.53	2.46	2.45	2.51	2.69	3.41	6.61	10.80	11.66	10.74	9.41	8.27	7.23	6.47	6.20	6.14	6.13	6.05	5.91
69	3.88	3.05	2.68	2.50	2.41	2.39	2.53	2.91	5.41	10.95	16.17	16.11	14.09	12.10	10.44	9.13	8.15	7.53	7.11	6.79	6.53

APPENDIX N. (Continued).

Pair	610	620	630	640	650	660	670	680	690	700
Standard X	19.02	20.52	23.11	26.58	30.65	35.83	41.64	48.63	54.15	60.91
47	19.51	20.93	23.45	26.84	30.83	35.98	41.78	48.79	54.34	61.12
48	18.38	19.99	22.69	26.27	30.46	35.71	41.62	48.78	54.48	61.48
49	18.60	19.70	21.79	24.72	28.34	33.12	38.76	45.81	51.67	59.13
50	19.54	21.80	25.44	30.10	35.30	41.36	47.58	54.44	59.31	64.91
51	17.67	19.11	21.58	24.91	28.85	33.87	39.63	46.64	52.34	59.33
52	20.51	22.12	24.90	28.57	32.84	38.19	44.13	51.18	56.69	63.33
Standard XI	18.13	16.4	14.99	13.97	13.27	12.58	11.93	11.7	11.77	13.17
53	18.64	16.95	15.53	14.49	13.78	13.07	12.4	12.18	12.25	13.69
54	16.66	14.96	13.6	12.63	11.99	11.34	10.71	10.53	10.6	11.93
55	16.28	14.58	13.23	12.27	11.63	11	10.36	10.17	10.24	11.52
56	18	17.06	16.39	16.04	15.79	15.42	14.92	14.78	14.88	16.5
57	16.32	14.72	13.4	12.43	11.79	11.16	10.53	10.34	10.41	11.72
58	18.76	17.03	15.6	14.55	13.88	13.19	12.51	12.28	12.35	13.76
Standard XII	4.59	3.95	3.86	4.22	6.12	11.03	19.67	32.24	43.43	55.79
59	4.43	3.87	3.74	4.09	5.90	10.61	19.01	31.28	42.38	54.75
60	4.75	4.09	4.00	4.38	6.32	11.30	19.96	32.50	43.55	55.42
61	4.43	3.88	3.75	4.11	5.94	10.70	19.13	31.45	42.39	54.28
62	4.86	4.19	4.08	4.46	6.44	11.45	20.13	32.69	43.85	56.03
63	3.87	3.42	3.32	3.60	5.09	9.28	17.01	28.64	39.37	51.71
64	5.06	4.35	4.24	4.65	6.78	12.06	21.07	33.94	45.15	57.06
65	6.36	6.56	7.72	9.99	13.70	19.50	27.28	37.30	45.75	55.19
Standard XIII	5.60	5.79	6.83	8.91	12.39	17.82	25.21	34.78	43.05	52.51
66	5.67	5.80	6.77	8.77	12.09	17.38	24.53	33.88	42.06	51.84
67	5.08	5.35	6.20	8.14	11.47	16.83	24.24	34.10	42.65	52.41
68	5.91	6.09	7.15	9.27	12.79	18.33	25.88	35.74	44.19	53.76
69	6.53	6.76	7.93	10.26	14.05	19.85	27.60	37.45	45.56	54.37

Appendix O. SAS Results of Polynomial Procedure Obtain Statistically Reliable Equation for Converting Gray Scale Ratings to Visual Difference.

The GLM Procedure

Dependent Variable: difference

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	56.95154182	14.23788545	653.01	<.0001
Error	5	0.10901818	0.02180364		
Corrected Total	9	57.06056000			

R-Square	Coeff Var	Root MSE	difference Mean
0.998089	3.837332	0.147661	3.848000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	56.85915273	56.85915273	2607.78	<.0001
grade2	1	0.05042727	0.05042727	2.31	0.1888
grade3	1	0.03801441	0.03801441	1.74	0.2439
grade4	1	0.00394741	0.00394741	0.18	0.6882

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	0.25288259	0.25288259	11.60	0.0191
grade2	1	0.00020657	0.00020657	0.01	0.9262
grade3	1	0.00140670	0.00140670	0.06	0.8096
grade4	1	0.00394741	0.00394741	0.18	0.6882

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.0621678322	0.14293838	0.43	0.6817
grade	0.8445221445	0.24797965	3.41	0.0191
grade2	-.0118531469	0.12177570	-0.10	0.9262
grade3	0.0053030303	0.02087796	0.25	0.8096
grade4	-.0004895105	0.00115046	-0.43	0.6882

Appendix O. (Continued)

The GLM Procedure

Dependent Variable: difference

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	56.94759441	18.98253147	1008.23	<.0001
Error	6	0.11296559	0.01882760		
Corrected Total	9	57.06056000			

R-Square	Coeff Var	Root MSE	difference Mean
0.998020	3.565845	0.137214	3.848000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	56.85915273	56.85915273	3019.99	<.0001
grade2	1	0.05042727	0.05042727	2.68	0.1528
grade3	1	0.03801441	0.03801441	2.02	0.2052

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	0.67031175	0.67031175	35.60	0.0010
grade2	1	0.02319989	0.02319989	1.23	0.3095
grade3	1	0.03801441	0.03801441	2.02	0.2052

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.0833146853	0.12453806	0.67	0.5284
grade	0.7564102564	0.12676999	5.97	0.0010
grade2	0.0375874126	0.03386078	1.11	0.3095
grade3	-.0035081585	0.00246889	-1.42	0.2052

Appendix O. (Continued)

The GLM Procedure

Dependent Variable: difference

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	56.90958000	28.45479000	1319.27	<.0001
Error	7	0.15098000	0.02156857		
Corrected Total	9	57.06056000			

R-Square	Coeff Var	Root MSE	difference Mean
0.997354	3.816591	0.146862	3.848000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	56.85915273	56.85915273	2636.20	<.0001
grade2	1	0.05042727	0.05042727	2.34	0.1701

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	5.09256835	5.09256835	236.11	<.0001
grade2	1	0.05042727	0.05042727	2.34	0.1701

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-.0050909091	0.11546990	-0.04	0.9661
grade	0.9181363636	0.05975159	15.37	<.0001
grade2	-.0097727273	0.00639137	-1.53	0.1701

Appendix O. (Continued)

The GLM Procedure

Dependent Variable: difference

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	56.85915273	56.85915273	2258.47	<.0001
Error	8	0.20140727	0.02517591		
Corrected Total	9	57.06056000			

R-Square	**Coeff Var**	**Root MSE**	**difference Mean**
0.996470	4.123419	0.158669	3.848000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	1	56.85915273	56.85915273	2258.47	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	1	56.85915273	56.85915273	2258.47	<.0001

Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	0.1121818182	0.09325842	1.20	0.2634
grade	0.8301818182	0.01746890	47.52	<.0001

Appendix P. Variance Component Analysis: Colombia

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1724	11402.64079	6.61406	6.44	<.0001
Error	3450	3545.54172	1.02769		
Corrected Total	5174	14948.18251			

R-Square	Coeff Var	Root MSE	value Mean
0.762811	44.31687	1.013752	2.287508

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	68	6068.477468	89.242316	86.84	<.0001
observer	24	2299.439165	95.809965	93.23	<.0001
swatch*observer	1632	3034.724162	1.859512	1.81	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	68	6068.477468	89.242316	86.84	<.0001
observer	24	2299.439165	95.809965	93.23	<.0001
swatch*observer	1632	3034.724162	1.859512	1.81	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	68	6068.477468	89.242316	47.99	<.0001
observer	24	2299.439165	95.809965	51.52	<.0001
Error	1632	3034.724162	1.859512		

Error: MS(swatch*observer)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch*observer	1632	3034.724162	1.859512	1.81	<.0001
Error: MS(Error)	3450	3545.541717	1.027693		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	0.4539
swatch*observer	0.2773
Residual	1.0277

Appendix Q. Variance Component Analysis: Czech Republic

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1724	10775.47456	6.25028	6.52	<.0001
Error	3450	3306.70558	0.95847		
Corrected Total	5174	14082.18014			

R-Square	Coeff Var	Root MSE	value Mean
0.765185	37.26129	0.979012	2.627425

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	68	5717.889847	84.086615	87.73	<.0001
observer	24	2447.983261	101.999303	106.42	<.0001
swatch*observer	1632	2609.601449	1.599020	1.67	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	68	5717.889847	84.086615	87.73	<.0001
observer	24	2447.983261	101.999303	106.42	<.0001
swatch*observer	1632	2609.601449	1.599020	1.67	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	68	5717.889847	84.086615	52.59	<.0001
observer	24	2447.983261	101.999303	63.79	<.0001
Error	1632	2609.601449	1.599020		

Error: MS(swatch*observer)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch*observer	1632	2609.601449	1.599020	1.67	<.0001
Error: MS(Error)	3450	3306.705584	0.958465		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	0.4850
swatch*observer	0.2135
Residual	0.9585

Appendix R. Variance Component Analysis: USA

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1724	11203.65564	6.49864	9.75	<.0001
Error	3450	2300.26713	0.66674		
Corrected Total	5174	13503.92277			

R-Square	Coeff Var	Root MSE	value Mean
0.829659	30.94763	0.816544	2.638471

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	68	5799.129808	85.281321	127.91	<.0001
observer	24	3048.687213	127.028634	190.52	<.0001
swatch*observer	1632	2355.838618	1.443529	2.17	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	68	5799.129808	85.281321	127.91	<.0001
observer	24	3048.687213	127.028634	190.52	<.0001
swatch*observer	1632	2355.838618	1.443529	2.17	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	68	5799.129808	85.281321	59.08	<.0001
observer	24	3048.687213	127.028634	88.00	<.0001
Error	1632	2355.838618	1.443529		

Error: MS(swatch*observer)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch*observer	1632	2355.838618	1.443529	2.17	<.0001
Error: MS(Error)	3450	2300.267133	0.666744		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	0.6067
swatch*observer	0.2589
Residual	0.6667

Appendix S. Variance Component Analysis: China

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1724	10154.06277	5.88983	9.74	<.0001
Error	3450	2085.36803	0.60445		
Corrected Total	5174	12239.43080			

R-Square	Coeff Var	Root MSE	value Mean
0.829619	30.51995	0.777467	2.547405

Source	DF	Type I SS	Mean Square	F Value	Pr > F
swatch	68	5397.604459	79.376536	131.32	<.0001
observer	24	2424.396763	101.016532	167.12	<.0001
swatch*observer	1632	2332.061549	1.428959	2.36	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	68	5397.604459	79.376536	131.32	<.0001
observer	24	2424.396763	101.016532	167.12	<.0001
swatch*observer	1632	2332.061549	1.428959	2.36	<.0001

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: value

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch	68	5397.604459	79.376536	55.55	<.0001
observer	24	2424.396763	101.016532	70.69	<.0001
Error	1632	2332.061549	1.428959		
Error: MS(swatch*observer)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
swatch*observer	1632	2332.061549	1.428959	2.36	<.0001
Error: MS(Error)	3450	2085.368030	0.604455		

Covariance Parameter Estimates

Cov Parm	Estimate
observer	0.4811
swatch*observer	0.2748
Residual	0.6045