Hue4True - AR Colorblind Testing-Correction with Natural Preservation

M11 Design Project (2025-1A)

Authors:

Semen Checherin	s2692155	s.checherin@student.utwente.nl
Hieu Chu	s2948923	chuminhhieu@student.utwente.nl
Hoa Dinh	s2841894	dinhthingochoa@student.utwente.nl
Duc Cuong Bui	s2966174	buiduccuong@student.utwente.nl
Justas Gvažiauskas	s2849461	j.gvaziauskas@student.utwente.nl
Mohamed Mohamedin	s2241382	m.mohamedin@student.utwente.nl

Supervisor:

Gwen Qin	gwen.qin@utwente.nl

UNIVERSITY OF TWENTE.

Table of contents

Introduction

Color vision deficiency (CVD) affects daily color perception and is difficult to correct without distorting the natural appearance. Some existing methods like Daltonization enhances red-green separability but produces an unnatural view of the colors. Alternatively, Natural Color Preservation (NCP) aims to improve distinguishability while preserving the naturality of the view.

In this project, we built an AR system on the Meta Quest 3 that combines a FM100 CVD test capable of estimating CVD type and severity with real time personalized natural-preserving color correction based on the test outcomes. The correction for this project is provided with Daltonization and our custom natural preserving LUT method. In user testing, our LUT-based approach improved FM100 performance for most of the Protan users and was preferred in natural scenes. Our natural-preserving recoloring was preferred by the entirety of the Protan group over Daltonization, despite yielding smaller improvements in quantitative testing. While Daltonization had stronger distinguishability for Ishihara plates, it was perceived as completely unnatural when applied to the real world.

1. Background

Colorblindness, or color vision deficiency is a condition that reduces the ability to distinguish certain colors. This disability affects approximately 1 in 12 men and 1 in 200 women globally, which is about 8% of the population. The condition is caused by the anomalies in one or more of the 3 types of photoreceptors in the retinas of the eyes, also known as cones. Depending on which cone is damaged, a person can have different types of colorblindness, including protanomaly (red), deuteranomaly (green) and tritanomaly (blue-yellow), with each varying in severity from anomalous trichromacy (degraded cone sensitivity) or dichromacy (complete absence). Currently, there are no surgical or medical treatments for color-blindness.

Individuals with colorblindness encounter various challenges in their daily lives, which can have an impact on social interactions, professions, emotional well-being and overall life experience. Difficulty in distinguishing colors can lead to misunderstanding of color-coded information. Sports where team colors are of importance, selecting ripe or spoiled fruit, cooking are some of the notable examples of activities that can be problematic and challenging, compared to a person with normal vision. This condition also limits a person’s career choice in fields such as piloting, driving, electrical engineering, graphics design, data visualization and certain military roles. Studies have shown that these issues in daily tasks and interactions can cause frustration and diminish in quality of life.

1.1 CVD Assessment

CVD is often estimated with either quick screening tools or measured precisely with specialized tools i.e. anomaloscope. The Ishihara plates are the most familiar for screening: they are fast to use and effective for spotting red–green deficiencies. However, they are limited, since they only give a pass/fail result, cannot detect blue–yellow (tritan) variations, and are difficult to quantify the severity of the condition. (Simunovic, 2010; Plutino et al., 2023; Zarazaga et al., 2019). To gain a fuller picture, we employ more complex testing participants to arrange colored caps in order of hue. While it does not provide a direct diagnostic measurement like an anomaloscope, the FM100 offers a practical way to estimate the type and severity of CVD based on the pattern of errors.

In our project, we use these two methods together. The Ishihara plates serve as a simple, fast baseline, while the FM100 test offers a more detailed picture of each participant's color vision, capturing both type and severity.

1.1.1 Ishihara Test

The Ishihara test is a fast and simple color vision test for the detection of red-green color deficiencies. The test consists of 38 plates known as Ishihara plates, featuring circles filled with coloured dots. In each circle, dots of different sizes and colors create a number or shape. While these numbers or patterns are clear to the people with normal colour vision, those with colour blindness can not see or have difficulties seeing it. Vice versa, there are plates that are specifically designed in a way such that those with CVD can recognize a number, while to normal vision are color noises.

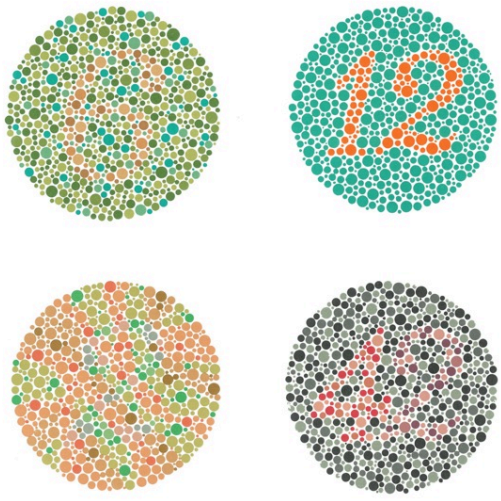


Fig 1.1: Some of the Ishihara plates showing numbers in order: 6, 12, 2, 42

In our test, the Ishihara test serves as a baseline to identify the deficiency type of the participant. Moreover, we also measured the tester's performance in accuracy under different filter types and severity.

1.1.2 FM100 in Desktop

The Farnsworth–Munsell 100 Hue Test (FM100) is a well-known method for assessing color discrimination ability. Online versions, such as the Colorlite Hue Test (Colorlite, n.d.), provide a simplified visual output showing total error scores and confusion regions. Colorlite uses 40 colors instead of the original's 85, influencing its accuracy. These tools, including the official FM100 evaluation software, only indicate where color discrimination is weak and do not clearly or automatically output the user's specific CVD type (protan, deutan, or tritan).

To address this, we upgraded the implementation of FM100 Hue Test in a way that automatically determines and outputs the type of a specific CVD (if there is one) based on which color discrimination regions are dominant. We determined the relevant regions and widened them up so all the colours of the circle would be used for determining the severity and type of CVD.

In addition, the improved system displays results in a more user-friendly way, indicating the specific errors in the test, showcasing the estimated condition and severity and a graph representation of error distribution for both the user and researchers. The integration of quantitative scoring with desktop and AR-based testing provides us with a more reliable and informative assessment process.

1.1.3 FM100 in AR

To showcase how different filters affect the performance in FM100 test, we also implemented the FM100 test in our AR environment. This test is based on our already existing research on color correction in AR (Qin et al., 2025). Instead of arranging physical caps on a desk, participants interact with virtual color caps using the Meta Quest 3 controllers. In this study, we use a simplified implementation of the FM100 test, with the detailed configuration described later in the design/methodology section. This adaptation allows us to maintain diagnostic value while keeping the test practical within the constraints of AR interaction.

1.2 CVD Correction

Many researchers have explored ways to improve color perception for people with CVD by using XR or AR systems. Most of these strategies use Daltonization filters, methods that aggressively remap colors to reduce confusion between colors.

Some of the earliest research on augmented reality (AR) and wearable systems for color vision deficiency used LMS-based color models. A well-known example is the Chroma system

developed by Tanuwidjaja et al. (2014). Chroma was implemented on a wearable device (Google Glass) that captured real-world scenes through its camera and applied real-time Daltonization to adjust colors for users with color vision deficiencies. The system performed LMS-space transformations to modify RGB outputs based on cone-response characteristics, demonstrating that personalized correction was feasible within a wearable setup. However, this method was not adaptive. It relied on fixed transformation matrices that did not change for each user.

Later works brought this concept into mobile and wearable settings. For example, the Wearable Improved Vision System for Color Vision Deficiency Correction (Melillo et al., 2017) built an AR system with a head-mounted display and camera, processed color mapping in real time, and overlaid the corrected image for the user. They evaluated it with individuals with CVD and found improvements in standard tests, such as the Ishihara test. Another recent example is Hue4U (Qin et al., 2025), which advances AR-based color correction by offering a real-time personalized filter that requires no prior clinical diagnosis and continuously adapts to the user's visual behavior.

Daltonization is notorious for its extreme remapping, such as mapping red to white to assist with distinguishing red and green in complete loss of red perception (protanopia). In case of deuteranomaly and protanomaly, however, such correction is too extreme due to the total remapping of colors, making it impractical for color-sensitive tasks in medicine and electronics. Despite the recent gains, existing AR recoloring methods only used and evaluated static Daltonization and did not incorporate any form of Natural Color Preservation (NCP).

Our project builds directly based on this direction. Inspired by the earlier Hue4U framework, we focus on developing and evaluating a NCP algorithm that balances perceptual improvement with realistic color appearance. Our goal is to improve color discriminability while preserving a natural look. To achieve this, we follow a three-stage pipeline: (1) simulate the observer's CVD, (2) transform colors to reduce confusion, and (3) render the corrected scene in real time. We drew inspiration from the NCP recoloring method described by Zhou et al. (2024), amongst earlier works it has been built on top of, adapting the algorithm's principles to design our own color correction. To ensure that the system is compatible with the target device, Meta quest 3, we were constricted to designing a lightweight Look-up Table (LUT) correction for passthrough video and virtual objects. Therefore, our implementation is fundamentally constricted to a one to one map, mapping each color to whatever value we define. This setup provides the correction but keeps the AR experience responsive, however may introduce complications and limitations down the line.

1.3 AR as a Platform

AR provides the setting where all of this comes together. Unlike clinic-based tests or static desktop filters, AR places both assessment and correction directly in the participant's own field

of view. This immediacy makes it possible to measure not only whether correction improves performance on FM100, but also whether it feels usable and comfortable in everyday perception. By using the Meta Quest 3, a widely available consumer headset, we show how it is possible to deliver a full workflow: detailed assessment with FM100, and personalized correction in real time. Prior work has shown that wearable AR systems can yield immediate improvement in Ishihara scores when correction is applied within the device (Melillo et al., 2017), and that mixed-reality adaptations of color tests on headsets can provide engaging, interactive user experiences (Jost, 2024). This points toward AR as a practical accessibility tool for people with CVD, bridging the gap between research methods and real-world support.

2. Planning

In this section, we describe how we organized and managed the development of the project. Our planning approach included regular supervisor meetings, a detailed specification of requirements, and thorough communication between the team members.

2.1 Supervisor Meetings

At the beginning of the project, we interviewed our client and supervisor, to ensure that our project stayed grounded in real needs. Our work builds on her earlier publication, Hue4U: Real-time personalized color correction in augmented reality (Qin et al., 2025), which showed how AR can be used to test colorblindness and deliver personalized color correction for the real world. Our current project extends this line of research by focusing specifically on evaluating novel, natural-preserving correction algorithms. From the beginning, our supervisor's guidance and advice shaped key requirements and design choices. Our meetings were held weekly to ensure consistency with the supervisor's expectations.

2.2 Requirements Specification

Our requirements cover both the technical performance of the system and its purpose as a tool to help people with CVD see colors more clearly and in a way that feels more natural. To keep the project organized, we divided the requirements into several parts: performance benchmarks, improvements to the color pipeline, review of related studies and how our work adds to them, ethical aspects, recruiting participants, defining team roles, and extra steps to confirm and validate our results.

2.2.1 Performance Benchmarks

The following are the Unity profiler graphs, showcasing the CPU usage of our software's main components. The dominant performance impact is seen to come from Rendering (green), followed by Scripts (blue). The impact is measured in frames-per-second (FPS) recorded throughout the duration of the experiment. Each graph records an interval of 1500 frames, roughly corresponding to 25 seconds of running the software.

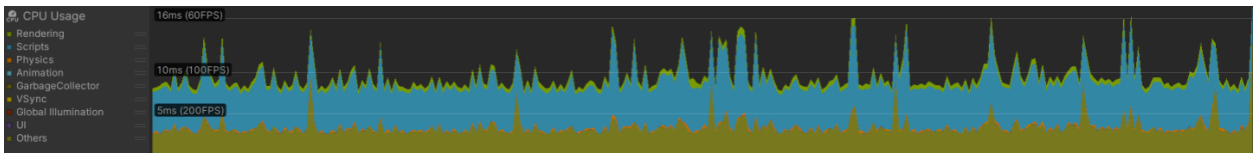


Fig 2.1: Profiler graph with no filter

Without any filter, the rendering impact rarely exceeds 200 FPS, with our main current bottleneck being the lack of optimization in our script logic. Nevertheless, the performance never goes below the comfortable 60 FPS in its worst moments.

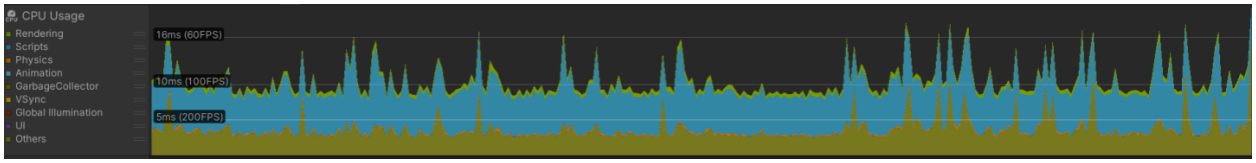


Fig 2.2: Profiler graph with Daltonization LUT applied

Upon applying the Daltonization LUT, we can see that the performance is impacted minimally. Rendering impact now goes slightly below 200FPS at times, and the overall performance dips under 60FPS at most intense moments. This corresponds to a roughly 16ms delay, which is on the border of being acceptable for the real-time AR environment.

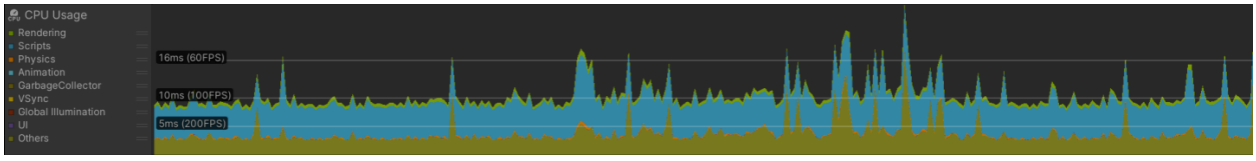


Fig 2.3: Profiler graph with Natural-preserving LUT applied

Applying our customly implemented natural-preserving LUT, we can see that the performance is similar to the previous graph. The overall impact is the same, as both the LUTs are of the same size. The sharper spikes can be explained by other hardware processes interfering with the performance, as we run the software through our laptop.

It shows that the application of LUTs seems to have an effect on the performance that can only be clearly noticed on the graph itself. The game runs below the acceptable real-time delay of 15ms, with rare performance spikes It should be noted that the performance is heavily dependent on the specification of the machine running the Unity project, not on the headset. FPS only measures the rendering performance for virtual objects, while passthrough FPS are dependent on the refresh rate of the camera feed, which for Quest 3 stays at a constant 60Hz, which is the same as FPS in this context. The actual performance is additionally limited by the refresh rate of Quest 3 internal display, which is at 75Hz for both eyes.

2.2.2 Color Pipeline Improvements

With this study, we also aim to improve our color pipeline, specifically with color fidelity and test accuracy. First, we switch from hue to CIELAB space. Instead of hue-based adjustments, we use CIELAB due to its flexibility in working with color and perceptual uniformity. Additionally, we originally planned to increase the number of LUTs, specifically from each 20% down to each 1%, as it gives a finer control over color calibration and correction. To enhance our test accuracy, we also plan to increase the number of caps in the FM-100 test.

2.2.3 Literature Review

As part of the design, we will include an updated review of related AR/XR studies, especially those published after 2023. Due to the long documented history of colorblind testing, we also take inspiration from much earlier fundamental papers alongside the most recent developments in natural-preserving recoloring . We aim to show clearly how our work is different and how it moves beyond earlier approaches, both in technical aspects and in accessibility.

2.2.4 Participants

We will recruit a wider group of people with CVD. To keep this process organized, a private spreadsheet has been created. It will only be used within the project team and will not be shared outside.

2.2.5 Individual Contributions

Responsibilities across the team must be clearly defined. Each member will take on specific tasks in development, testing, analysis, finding participants, and documentation to ensure efficiency and accountability.

Semen Checherin	Team management and direction. Accustomed the team with the topic and idea. CVD simulation software, participated in each part of the project, provided Q&A support on demand. Conducted final experiments.
Hieu Chu	Helped on the development of Python script for LUT generation, Unity project implementation and improvements. Parts of report, diagrams. Conducted final experiments.
Hoa Dinh	Literature review and bibliography, finding participants, conducting experiments for pilot study. Wrote parts of the report.
Duc Cuong Bui	FM100 color moving and circle result visuals, video clip for experiment. Assisted with implementation of FM100 description in the report.
Justas Gvažiauskas	FM100 Hue Test error representation, color generation, CVD type classification and severity estimation. Wrote implementation of FM100 description in the report.
Mohamed Mohamedin	Development and testing of a multitude of LUT-based correction algorithms. Wrote parts of the report

2.2.6 Innovation

We may not be able to fully implement Zhou et al.'s method due to the limitations present with MetaSDK, such as instability and lack of clarity for capturing camera frames. Instead, the key requirement is to show how we use the tools provided by Meta and the innovative aspects of our own implementation, while acknowledging inspiration from prior work and the limitations of our approach.

3. Design/Methodology

In this section, we detail our technical implementation of the FM100 test for both desktop and AR and outline our contributions and improvements over the existing digital versions of the test. We further detail the novel, albeit currently limited, NCP correction, the idea behind our implementation, its strengths and weaknesses over the already implemented Daltonization. Lastly, we discuss the improvements made to the rendering system.

3.1 Implementation and Design of the FM100 Hue test

Compared to the 85 colors tested in the original, physical FM100 test, we have created a 60-cap digital test for both desktop and AR to measure each participant's color discrimination before and during AR use. The reduction in colors was necessary due to the lower fidelity of the AR passthrough and the environment being unfamiliar to the participants. Using full 85 colors would make for a tedious test, with a high risk of normal-vision participants scoring as colorblind due to the poor display or the long time it would take to score perfectly.

The desktop version is given first, and provides us with the baseline for the participant's natural vision. The AR version implements the same test so that any difference in performance can be attributed to passthrough quality and novelty of the environment rather than the participant's vision. Running the desktop test first also ensured that participants were accustomed with the idea of the test itself, reducing confusion when using the unfamiliar technology.

Unlike simpler tests such as the Ishihara plates, which primarily detect red-green deficiencies, the FM100 test measures a participant's ability to perceive and discern hues across the CVD-relevant color spectrum, alongside giving us better insights on the severity of the condition. It provides a quantitative and rather accurate estimate of color vision performance, making it suitable for our research requiring analysis of hue perception accuracy.

Our 60-cap test has been heavily inspired by the Colorlite online FM100 variant, which uses only 40 caps. The increase in tested colors already directly increases the accuracy of our estimation considerably. We retained Colorlite's general interface, but replaced its simplified scoring logic with proper classification informed by the original FM100 publication. This change allowed us to compute more precise color confusion regions for the 3 types rather than simply relying on Colorlite's divisions, and therefore classify the errors to automatically output the estimated type value. Therefore, our system can now automatically output both the user's CVD type and continuous severity estimates, which are not provided by Colorlite, and are not automatically inferred even in the original FM-100 evaluation software provided with the physical test. Finally, we added more clear instructions and error visualization on the test itself for our participants.

3.1.1 Generation of the Colors

First, we defined hard-coded colour values of the anchors (fixed start and end colours of each row of the test) to generate the rows using them. Specifically, we define 4 colours: red (#B2766F), olive (#9D8E48), cyan (#4E9689), and purple (#8575b5). These colours were chosen because they represent hue ranges that people with CVD find difficult to distinguish. Additionally, they closely match the ones used in the online FM-100 Hue Test created by Colorlite.

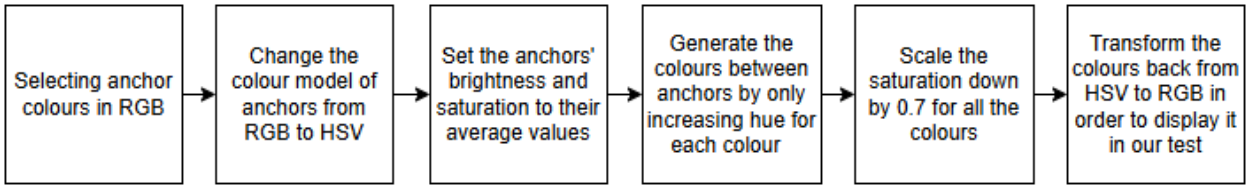


Fig 3.1: diagram explaining how the colours of the FM-100 Hue Test are generated.

After defining the anchor colours, we use them to generate a gradient row per unique tuple of anchors from red to olive, purple, and finally back to red. For this, we use the HSV colour model instead of RGB. The reason for this is that in the RGB model, each color represents a combination of the intensities of three color channels (Red, Green, Blue), while HSV separates the actual color information (Hue) from its saturation and brightness. This separation allows us to normalize brightness and saturation for all the colors, which is necessary for the FM100 Test.

To get fixed values for brightness and intensity for each colour, we calculate the average brightness and intensity of the anchor colours and use these averages to generate the intermediate colours and apply the average to the anchor colors themselves. This ensures that all the colours share the same brightness and saturation, while the hue varies.

However, when we used these values to generate the colours, the colour differences seemed to be too significant which made our test too easy to complete, therefore we scaled down the saturation of each colour by 0.7. We chose this value after trying different saturation levels and finding that it balanced the colors just enough without making the colors look dull. After all this, we transform the colours back to RGB for display on a webpage.

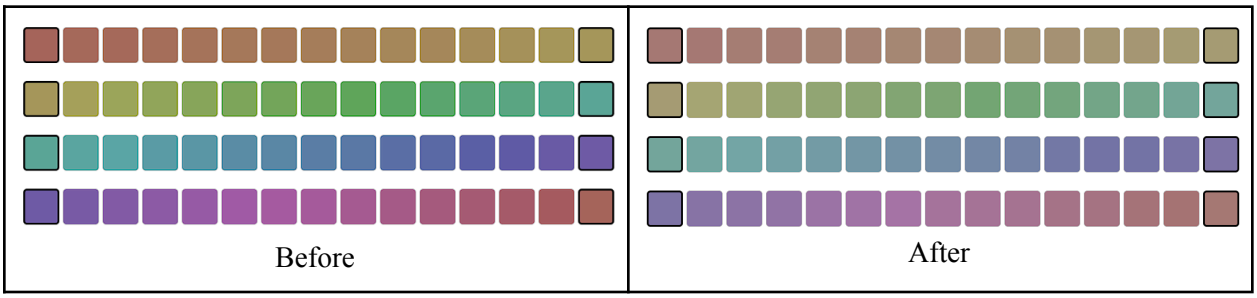


Fig 3.2: The colours before and after scaling their saturation down by 0.7.

After the colors are ready, they are randomly rearranged within the row for the user to take the test. When the user finishes the system generates the results, including:

- Numerical values generated on the squares that indicate the correct position of the colour,
- Graphical representation of errors
- Final result output.

Later chapters explain each of these explicitly.

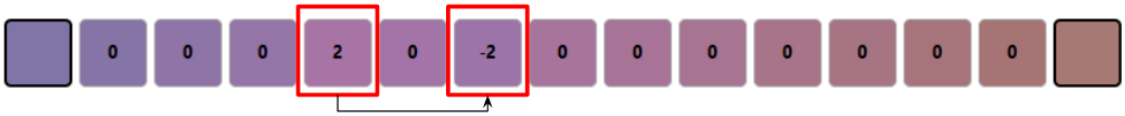
3.1.2 Error representation of colour displacement

After the test is finished, in order to present their actual results to the user we implemented a way to visually represent what the correct positioning of each of the movable squares should be. For this, an integer in the middle of each of them is used. It serves as an additional visualization of errors, since in the graphical plot representation it is not possible to determine the direction of the error. Therefore, making the direction of the error more understandable, the participant who wants to review their error can immediately know whether they have placed a square too far to the left or right of its actual position.

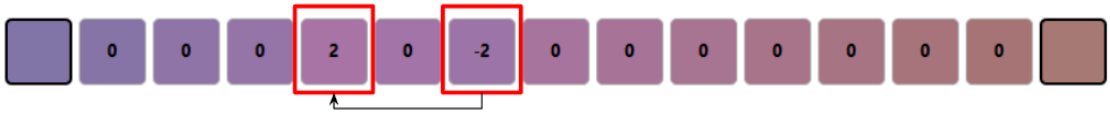
Fig 3.3: One of the rows in the test completed with significant errors, marked with numbers explained on the next page.

There are 3 ways an integer can describe the intended correct position of the coloured square in the row:

Zero: indicates the current position of the square is correct.



Positive number: shows that the square should be moved that many positions to the right. For example, the square in the image that has an integer 2 should be moved 2 squares to the right, to the position of the square with -2.



Negative number: indicates that the square should be moved that many positions to the left. For example, the square in the image that has an integer -2 should be moved 2 squares to the left, to the position of the square with 2.

Fig 3.4: Error representation on the test after completion.

3.1.3 Graphical Representation of Test Results

The circular plot visualises the participant's colour arrangement errors from the FM-100 Hue Test:

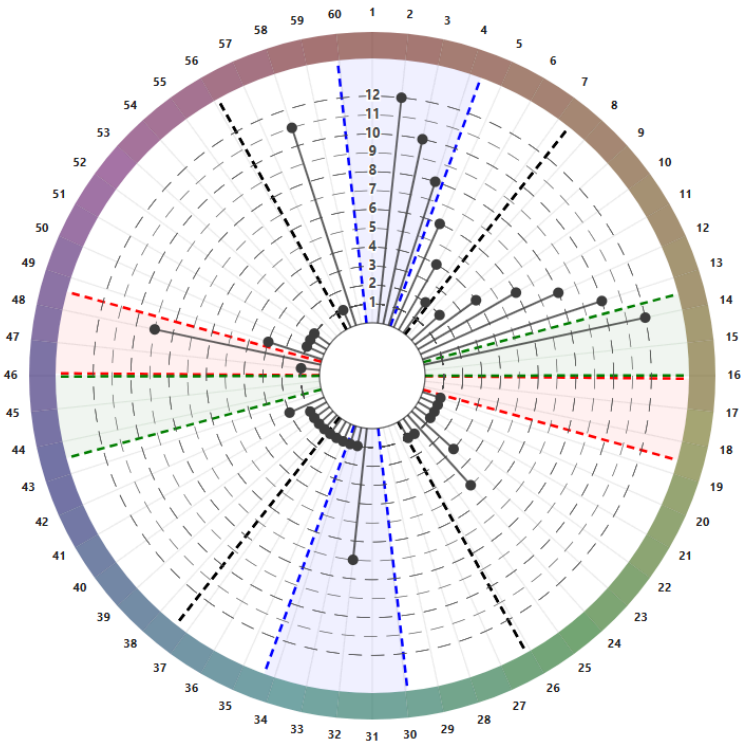


Fig 3.5: Complete circular plot that shows errors made for each colour, represented by black dots. Levels of displacement error are shown as the grey dashed circular lines. The circle contains the regions separating confusion ranges for protanomaly (red), deutanomaly (green) and tritanomaly (blue). The black lines in between define the extensions of these confusion ranges for type estimation.

The chart is divided into 60 equally spaced segments, each representing one hue cap from the test sequence. Together, these hues form a continuous spectrum, showing the natural progression of colors within the relevant range to CVD estimation. The outermost colored ring shows the correct hue order, with smooth transitions between colors.

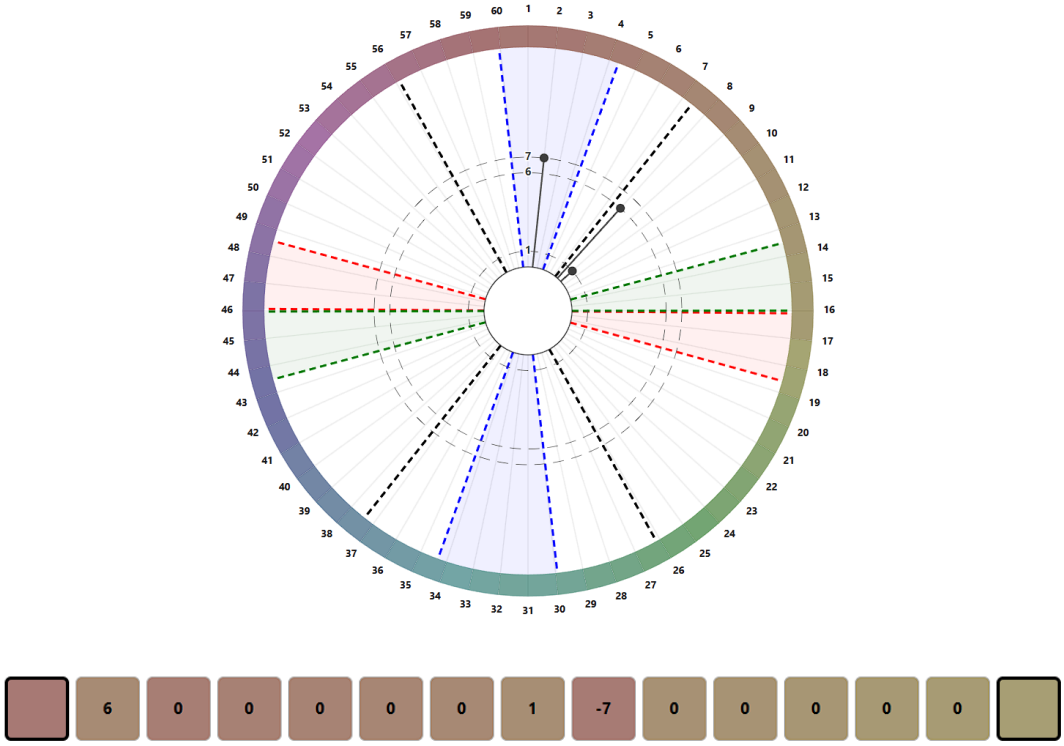


Fig 3.6: Example of the colour displacement error in the test, matched with its graph representation. Color at index 2 is misplaced at index 9, giving the error distance of 7.

Black dots represent the FM-100 Hue Test errors aligned with corresponding hues on the outer colour wheel. Each dot lies on one of the black dashed circular lines, which represent the level of displacement error for the colour made by the participant. The ground truth is the middle of the plot, with an error value of 0. Therefore, if the user has placed the colour correctly, the black dot for that colour is not drawn in order to minimize information.

The plot is further divided into six coloured regions, with each color corresponding to a different type of colorblindness: blue for tritanopia, green for deuteranopia, and red for protanopia. These regions were created based on the original FM-100 Hue Test, presented in the Farnsworth-Munsell 100 Hue Test: Instructions (n.d.). Using this paper, we approximated the CVD ranges of the 85-cap version of the test to our 60-cap variation.

The only significant difference between the original plot and our version of the plot, besides the number of colors used, is the hue circle direction. The original hue circle introduced by Munsell colours begin at hue 1 (red), and go anti-clockwise to 85, which is in fact the opposite of the color sequence in our plot. Therefore, the ranges align only if the original plot is mirrored across the vertical y-axis. As a result of changing the number of tested colors, the color ranges may not

match perfectly. While the current version is aligned with the Ishihara protan/deutan classification, this possible limitation could be improved with further participant testing.

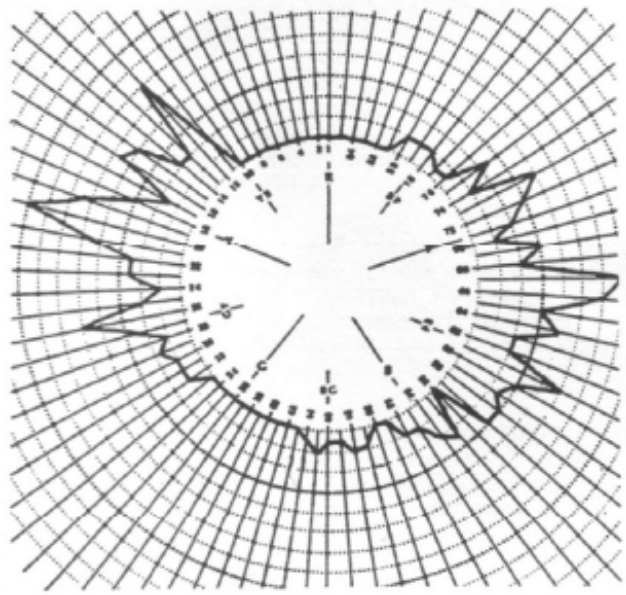


Figure: color defective pattern of Protan.

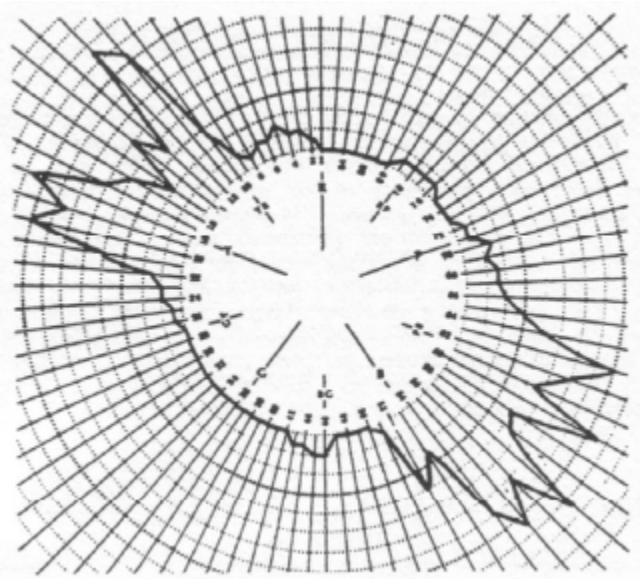


Figure: color defective pattern of Deutan.

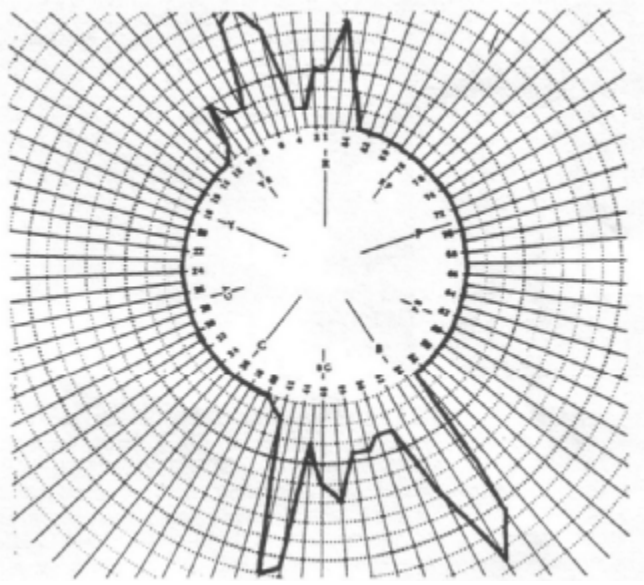


Fig 3.7: color defective pattern of Tritan.

The ranges for each CVD type were extended and represented by the black dashed lines drawn in order to also include the errors that do not fall directly within the coloured regions into the error score calculation. These lines divide the unclassified areas (white color areas), with each divided area assigned to the nearest region for the error score calculation. This ensures errors in the entire colour circle are covered. The regions are then narrowed vertically to reduce the tritan region, ensuring the regions are distributed evenly, as tritan would otherwise border more unclassified areas than protan or deutan. Such errors are not in the main coloured regions, therefore they are weighted as 50% of their original value. For instance, if a colour in a white region is misplaced by 6 positions, the weighted error for that colour is then 6 * 0.5 = 3. This allows us to estimate the user's deficiency type while regarding the full range of the color wheel.

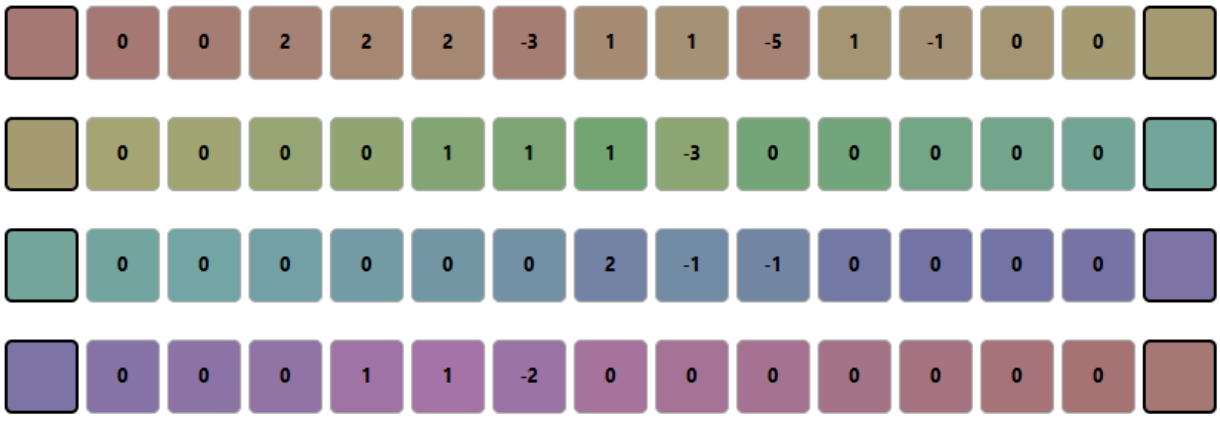
3.1.4 Total Error Score (TES) and Severity

Total Error Score (TES) represents total colour displacement for the whole experiment and it is calculated as the sum of all errors found in 4 rows.

To find the total errors per row, we calculate the sum of absolute values of all integers in that row.

Fig 3.8: Possible colour placement for one row. The total error for this row is: 0 + 0 + 2 + 2 + 2 + |-3| + 1 + 1 + |-5| + 1 + |-1| + 0 + 0 = 18.

Finally, the errors of all 4 rows are added up to obtain TES.



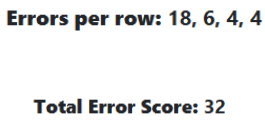


Fig 3.9: errors for each row are calculated and added up together to get the TES.

After calculating TES, we can estimate the severity of the CVD type.

To make our calculated severity value comparable to the original FM-100 Hue Test which uses 85 movable caps, we normalized our TES. The normalized version of TES is calculated as:

normalizedTES = TES * (85 / 60),

where 60 represents the number of colours used in our version of the FM-100 Hue Test.

Then final severity is calculated as the minimum between 1 and the normalizedTES divided by 100. The constant 100 was selected as a threshold defining the upper boundary beyond which higher error values no longer increase the severity:

normalizedSeverity = min(normalizedTES / 100, 1).

As a result, we get a severity represented as a normalized value ranging from 0 to 1, where 0 means that the colors were arranged flawlessly and 1 means that the upper limit of severity was reached, representing complete dichromacy.

3.1.5 Deficiency Type

When the participant completes a test, a separate score is calculated for each of the 3 CVD type regions by summing the errors made within that region. For example, all the errors that occur in the blue (tritan) region and its adjacent white regions are summed up to obtain the total error score for the blue region.

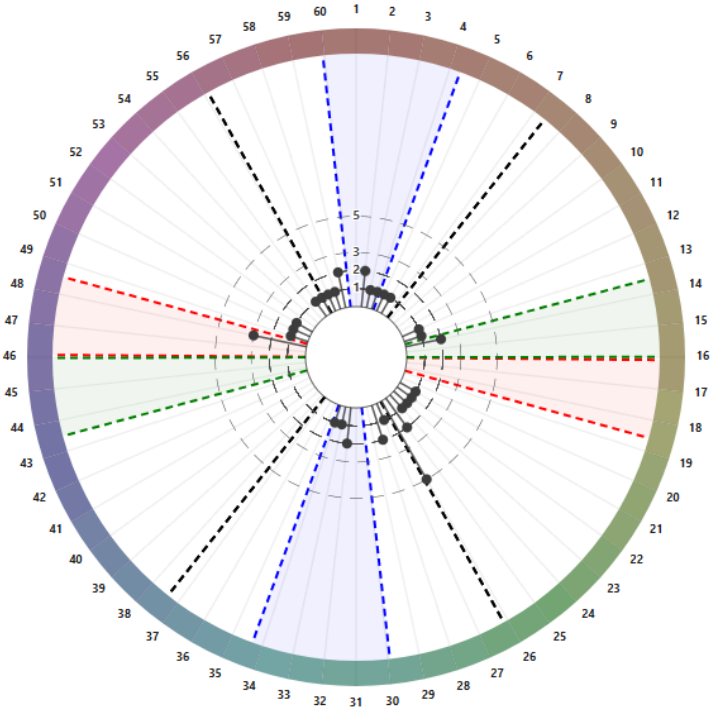


Fig 3.10: Determination of type Tritan. Here, TES is 40, and tritan has the highest error score of 15.5, compared to protan with the score of 11.5 and deutan with the score of 3.5. In this scenario, 15.5 > (10 = 0.25 * 40), therefore the deficiency type of tritan is determined. If this max value was lower or equal to the TES, then no type can be determined because either there were no errors, or the errors are distributed too evenly across the CVD regions.

As mentioned before, if an error falls within one of the coloured regions, 100% of the displacement error value is used, however, if an error falls into the white region, 50% of the displacement error value is used.

Once all three scores are computed, the region with the highest total error is identified as max, and this score is compared with TES.

If max > 25% of TES, the result is significant enough to say that the person potentially has a specific CVD type that is more dominant than others. Then the CVD type which represented the max variable (had the most errors) is determined as the type.

If max ≤ 25% of TES, it shows that the errors are distributed too evenly across regions, which means that no specific CVD type can be determined. In this case, the type is classified as "None".

Selecting 25% makes sense as the analysis of the distribution of errors works by adding up errors in the regions and then comparing them by each other. We have taken into consideration that small errors of size 1 or 2 could appear that do not have much to do with the CVD itself, and more with concentration to the task. Therefore we allow some human error to occur, which would have more impact on the severity calculation, but not the type itself. For this, instead of selecting 100 / 3 = 33.3% of TES as the norm, we lowered it down to 25%.

3.2 CVD Recoloring Method: Natural Preservation

In this section, we introduce our NCP recoloring method for the Meta quest 3. The previous version of the software Hue4U has only provided personalized simple Daltonization, which is far from ideal for trichromats for the reasons discussed above. We must therefore prioritize maintenance of realistic color appearances while increasing distinguishability.

Our methodology gets its inspiration from the NCP recoloring principles described by Zhou et. al., whose work specifically emphasises preserving naturalness with a complex algorithm comparing each pixel per image it is applied to. Unfortunately, a direct implementation of such an algorithm is not feasible due to the hardware constraints of the Meta quest 3 passthrough system.

3.2.1 LUT Constraint

The fundamental challenge of implementing our desired recoloring algorithm in real-time AR is the limitation of Meta's passthrough rendering pipeline. The latest and fastest algorithm currently, implemented by Zhou et. al., relies on image-dependent processing. It analyzes relationships between the neighboring pixels, and only works on a single image. As reported by Zhou, a standard 1920x1080 image may take 6 seconds to be processed. The algorithm by itself simply cannot work in real-time.

Even worse, Meta's development environment is restricted, with the only ways to access raw camera frames only being implemented in summer of 2025, therefore being unstable and still restricted, requiring deployment on device and explicit permissions. The only stable available method we have to edit the video feed, therefore, are the structures explicitly suggested by Meta: the Look-up Tables, or LUTs. They are static mapping functions f(c)→c′, where the original color c is always mapped to a pre-defined color c'. While this allows for minimal performance overhead, it stops us from performing image-dependent processing. As an attempt to work against this restriction, we had to compress an estimate of the original algorithm's logic into a static color map in a creative way which seems to not really have been done before.

3.2.2 Color Transforms and LUT Format

We developed a custom Python script, utilizing the color-science API to pre-generate all the desired LUTs for Deutanomaly and Protanomaly. We decided to go with Meta's suggested resolution of 32x32x32, meaning we sample 32768 colors to apply our transformation logic on. After the LUTs are generated, we simply upload them to the Unity project to be stored and loaded later on by the controller after the FM100 test is submitted.

Our transformation logic is defined in the CIEL*a*b* color space, which has been chosen over LMS or other colorspaces due to its flexibility in modifying colors and the particular relevance to our task. In this color space, L refers to lightness or brightness, which can be adjusted independently of chromacity or color itself. a* stands for a Green-Red component, which directly mirrors the protan/deutan confusion line we are interested in. b* then represents the blue-yellow component, which we will use for subtle color shifting, as it is necessary for improving color discernment due to the LUT limitations.

The goal of our transformation is to widen the perceptual gap between red and green hues without remapping the colors aggressively. In simple Daltonization, the remapping is done by, for example, simply mapping the red color to white. To solve this problem, we have considered the specific wavelengths to which the colorblind types are sensitive to.

The general principle for our transformation is to first convert the input RGB color to CIEL*a*b*, which is easily done with methods provided in colour-science API. While our original approach relied on simply stretching the red to make oranges and pinks have more of it, we quickly realized this simplistic approach results in unnaturalness, as showcased further in the exploration of our pilot study. Therefore, we had to introduce a technique that we borrowed directly from Zhou's paper: luminance preservation, and added more selective color-dependent saturation boosting. For each color, we store the L* value and calculate its chroma (saturation):

$$C = sqrt(a^{*2} + b^{*2}).$$

After we can manipulate the color as we like, our approach branches between Protanomaly and Deuteranopy correction. An important distinction must be noted in the a* parameter. a* > 0 corresponds to Red hues, while a* < corresponds to Green hues.

In Protanomaly, the red color is problematic, and is typically dampened, making reds appear darker. Therefore, for protanomaly, for any a* > 0 we boost C and L* in a way that can be represented with the formula:

$$C^*_{new} = \begin{cases} C^* \cdot (1 + k_s \cdot \alpha), & \text{if } a^* > 0 \text{ (Red hues)} \\ \beta^* \cdot (1 - k_s \cdot \beta), & \text{if } a^* < 0 \text{ (Green hues)} \end{cases}$$

$$L^*_{new} = \begin{cases} L^* \cdot (1 + k_l \cdot c), & \text{if } a^* > 0 \\ L^* \cdot (1 - k_l \cdot \gamma), & \text{if } a^* < 0 \end{cases}$$

Fig 3.11: The contrast boosting equation for protanomaly. Ks and Kl correspond to parameters based on input severity, while α,β,c,γ correspond to pre-defined constants to prevent color clipping (out of bounds assignment)

For Deutanomaly, we hypothesized the reverse of our Protan logic would yield similar benefits, thus we applied the reverse of the formula above. To further assist with color discretion, we decided to add a small blue component to green hues to push them away from the red spectrum in perceptual space, as blue is not affected by red-green colorblindness. The resulting color correction provides the following result for protan of severity 1 (total dichromacy, maximum possible correction):

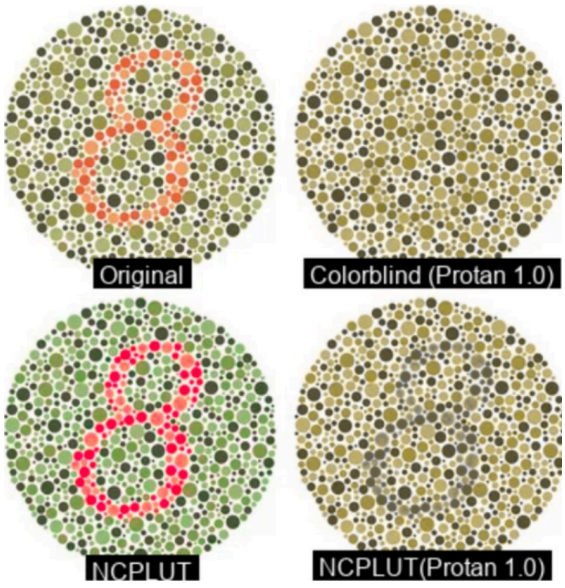


Fig 3.12: top left - original image. Top right - protanope view simulation. Bottom left - Our correction. Bottom right - protanope view of corrected image. Note the enhanced visibility of the 8, while other colors are relatively stable.

3.2.3 CVD Levels in LUT Generation

For personalized color correction, we have generated distinct LUTs for varying severity levels. For protanomaly and deutanomaly each, just like in Hue4U, we have generated ten LUTs for each type in increments of 10%. The choice of 10% intervals was made for two practical reasons. First, through testing, we observed that visible differences in color correction only start to become noticeable roughly at every 10% change in severity. Smaller increments produce barely perceptible changes. Second, generating each LUT takes computation time as we process 32768 colors for each level of type and severity. Therefore, using 10% steps provides a good balance between visual accuracy and processing time.

3.3 AR Rendering

The generated LUTs are then all stored into the AR environment to be loaded on demand, and are sorted by each deficiency and severity. We use a controller class as a centralized control over the rendering logic. It is responsible for handling FM100 test results, and deciding on which LUT to apply according to the deficiency type and severity. Meta Quest's rendering necessitates separate pipelines for our LUT application. Virtual objects are rendered through Unity's URP (Universal Render Pipeline), where we apply a shader which performs one-to-one color transformation from the LUT color values in real-time. Meanwhile, the passthrough view requires a separate method from Meta Quest's official SDK, which directly applies the LUT image onto the camera feed. Both virtual objects and camera feed are applied with color correction, together making the complete view.

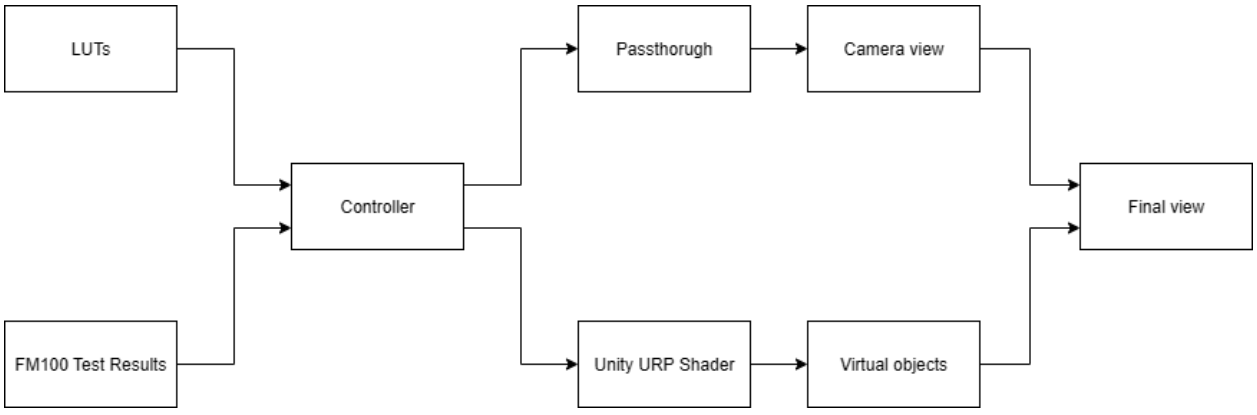


Fig 3.13: An abstract view of the rendering pipeline.

Our current approach differs considerably from the previous version of the software. Instead of correcting the pass-through with LUTs, and implementing a custom HLSL shader separately for correcting the AR elements, we now simply apply the LUT to both pass-through and AR elements. This was done due to the new complexity introduced by the natural-preserving intention of our color correction, which would be wildly inefficient to be implemented within a shader. Additionally, it became clear that a change in color correction with the old system would require modifying both the LUT generation python script and the HLSL shader separately, which is inconvenient for further software development and maintenance.

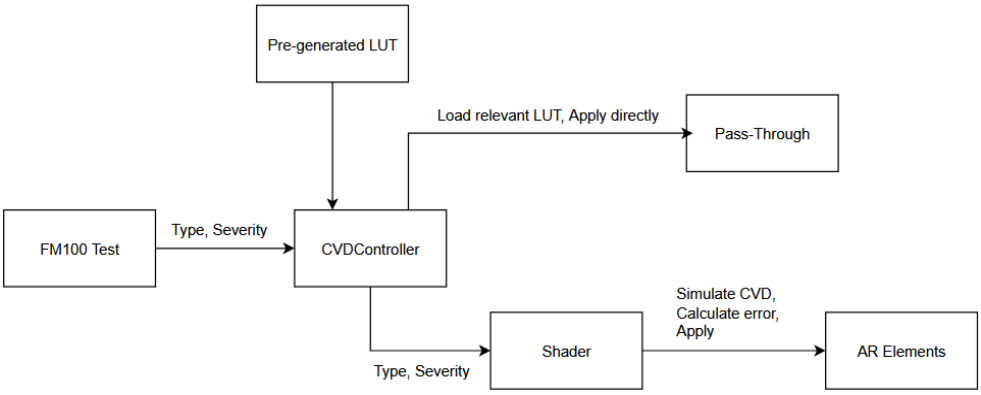


Fig 3.14: The old system architecture, using the shader separately for AR elements

4. Experiment Procedure

4.1 Ethical Review

All research procedures in this study were approved by the Computer and Information Sciences (CIS) Ethics Committee at the University of Twente. This approval confirmed that the study met the university's ethical standards for research involving human participants, including procedures for recruitment, consent, and data protection.

Before taking part in the study, all participants were fully informed about the purpose of the research, what their participation would involve, and how their data would be used. Each participant then provided written informed consent before the experiment began. This process ensured that participation was voluntary, transparent, and based on informed understanding.

All data was collected and handled in accordance with GDPR guidelines to guarantee participant privacy and anonymity. Participants were signed with numbers linked to research results, and their names were not recorded for anonymity.

This ethical approval and consent process ensured that the study was conducted responsibly, respecting the safety and well-being of our participants.

4.2 Participants

Twelve men with different types and levels of CVD participated in our experience. They were all between 18 and 25 years old and living in the Netherlands. Among them, we have eight people directly involved in experiencing the product to give the evaluation results, and four men volunteered to participate in the pilot study during the project implementation.

4.3 Software and hardware

We conducted the experiment on the Meta Quest 3 headset. The testing environment was created using the Unity game engine, utilizing Meta SDK and OpenXR libraries for development.

4.4 Pilot Study

In the middle of our project, we decided it was time to evaluate our current implementation. Thankfully, testing the system early on multiple colorblind people, and not just the one team member, has shown glaring downsides and inconsistencies across the board. In this section, we explain the pilot study that gave us immeasurable insights on what to change, fix and how to conduct further research.

4.4.1 Old Experiment Procedure

We started by giving participants a baseline desktop test. Here, they worked through a 60-cap Farnsworth–Munsell 100 Hue Test, which gave us a clear picture of their CVD, both the type and how severe it was.

Next, we moved into the headset for some AR pre-tests with Ishihara plates from the book. At this stage, no filter was applied; participants simply viewed the plates through the AR pass-through. This step was important because it showed us how they performed in the AR environment itself, before we added any correction.

After that is the AR version of the FM100 Hue Test. We implemented it with 60 caps, and participants used the controllers to drag and arrange the colors in order. The errors they made told us not just how strong their CVD was, but also pointed to which type they had. This information then became the backbone for creating a personalized, natural-preserving filter for them.

Once the filter was in place, we asked them to repeat the Ishihara test in AR. Now they were seeing the plates through the corrected pass-through, and we could see right away how much easier (or not) the plates became compared to their first attempt.

To wrap things up, we sat down for a semi-structured interview. Here, participants shared their thoughts: how the Ishihara plates felt before and after the filter, whether the corrected colors still looked natural, how they experienced the AR FM100 test, and what their overall impressions were of the system. These conversations gave us insight into not just the numbers, but also how the system actually felt to use.

4.4.2 Pilot Study Results

The feedback we received was not at all positive, and Some of the reasons for this were technical limitations in the early version of our natural color correction. It did not preserve luminance and was too aggressive in its color mapping. That saying it failed to maintain the original brightness and contrast of each color while adjusting hues. As a result, some areas became overly bright or “glowing,” while others lost detail. Several participants mentioned that reds appeared brighter but sometimes looked too strong or “emissive,” and resulting blooming effects caused loss of contrast. These reactions directly reflected the lack of luminance preservation in the algorithm. In other words, although the filter mathematically improved color separation, it unintentionally changed the perceived brightness of the scene, leading to discomfort and unnatural color appearance.

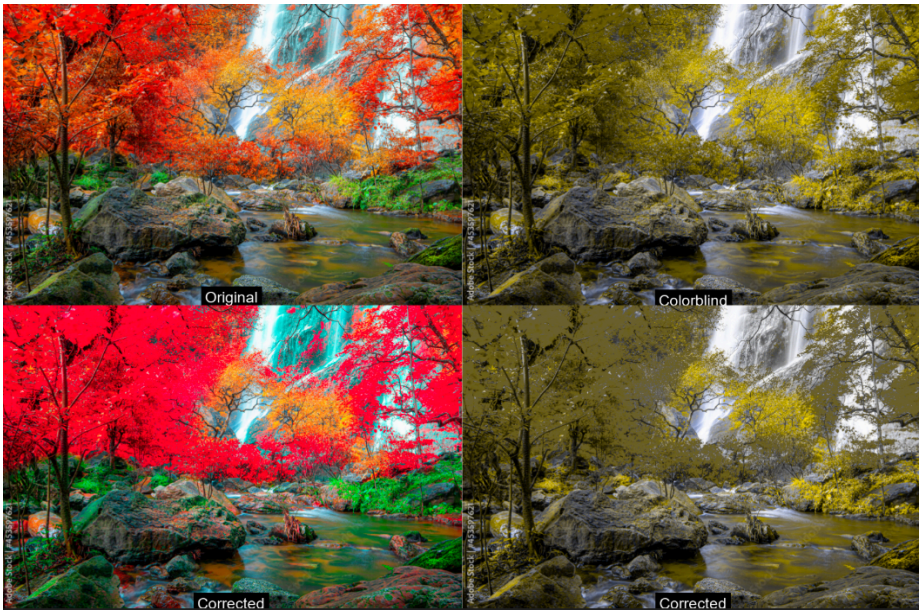


Fig 4.1: A grid showing in order: The original image, the original image under Protan simulation, image after applying our original Protan filter and its simulated counterpart.

During the experiments and after reviewing the numerical data in Table 1 and Table 2, we noticed that the severity values were not reliable at all, and led to a total inconsistency with our results. Some participants showed contradictory or ambiguous results between the desktop and AR sessions. For instance, Participant 1 was classified as “None” in the AR FM100 yet displayed a severity of 62 percent in the AR version, while Participants 2 to 4 were all identified as protan but their severity values varied without a clear pattern. These inconsistencies suggested that the current scoring logic still needed significant refinement.

Another issue was related to the experimental procedure. The Ishihara test was not fully utilized in the workflow, as we did not include a direct comparison after applying our filter. This absence left out an important evaluation step that could have shown how the correction affected color perception. We also forgot to include the Daltonization algorithm as a reference for comparison with our new method.

Based on these findings, we began to improve our system across the board, reviewing the code behind the scoring logic in both the desktop and headset environment to ensure their consistency. We have conducted further literature review into natural-preserving color correction to get more insight on how to possibly implement it into a LUT structure. By borrowing additional techniques, such as adding luminance preservation and reducing the overall color-mapping strength, resulting in a more balanced and natural correction. We also refined our experimental pipeline significantly by reordering the testing procedure and including the Daltonization filter testing environment for proper comparison between different filter types. Additionally, we added a user interface in which the participants can use to switch between different filters to their

preferences. These adjustments made the experiment more stable and prepared the setup for the main study.

Participant ID	Ishihara (Correct/Total)	AR FM100 Score	Severity	Deficiency Type
ID 1	8 / 24	108	62%	None
ID 2	15 / 24	16	—	Protan
ID 3	12 / 24	28	—	Protan
ID 4	13 / 23	66	47%	Protan

Table 1. Combined Ishihara and AR Test Results

Participant ID	Total Error Score	Deficiency	Severity
ID 1	46	None	—
ID 2	28	Protan	0.05
ID 3	50	Protan	0.09
ID 4	38	Protan	0.07

Table 2. Desktop FM100 Test Results

4.5 Final Experimental Procedure

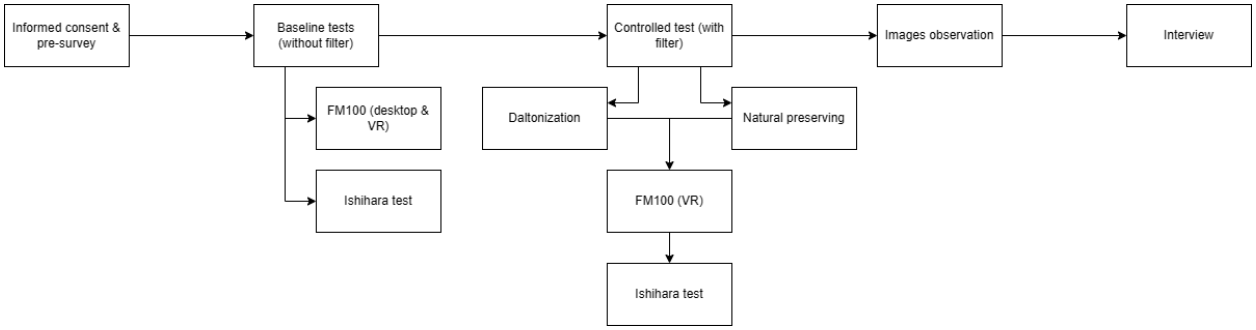


Fig 4.2: Overview of the Experimental Procedure: Flowchart illustrating the full sequence of tasks performed by participants during the study. The procedure began with informed consent and a short pre-survey, followed by baseline color vision tests, including the FM100 Hue Test on a desktop computer and the Ishihara Plate Test without filters. Participants then completed the AR version of the FM100 test (no filter) using the Meta Quest 3 headset. Next, personalized color correction filters were applied: the NCP (natural preserving) filter, followed by the Daltonization filter. For each condition, participants repeated both the FM100 AR test and the Ishihara test to compare performance across filters. The session concluded with an exploration phase, where participants viewed a short video, adjusted filter intensity manually, and shared their impressions through a recorded interview.

Based on what we learned from the pilot study, several important changes were made to improve both the structure of the experiment and the participant experience.

At the beginning of each session, we asked participants to sign the informed consent form, their experience with digital devices, any known color vision issues and the availability of proper CVD diagnosis in the past. This step helped us understand their background and ensured ethical compliance before the visual testing began.

We began each session with our FM100 Hue Test on a desktop computer, which served as the baseline measurement of color discrimination ability. We asked participants to arrange a series of colored caps in what felt to them like the correct order of hue. This step provided a recorded estimate of the baseline for each participant's color perception, and helped determine both the type (protan, deutan, or tritan) and severity of their CVD.

Following the FM100, participants completed the Ishihara test without any filters. Although the Ishihara test cannot measure the severity of deficiency or detect tritan-type anomalies, it remains an effective and fast screening tool for distinguishing between protan and deutan color vision types. Conducting this test early in the sequence gave us a quick metric, upon which we could evaluate the effectiveness of the provided color correction.

Participants then performed the FM100 test again, but this time in AR using the Meta Quest headset. In the AR version, the color caps appeared as floating virtual objects that could be picked up and rearranged using the controllers. This allowed us to see how their performance changed when colors were shown through the AR pass-through view instead of on a flat screen.

Once participants completed the unfiltered AR test, we applied the pre-generated personalized NCP filter first based on their individual FM100 results. The purpose of this filter was not to transform colors drastically but to gently enhance hue differences in regions where participants tended to make mistakes, while maintaining an overall natural appearance. Participants then were given the chance to review their AR FM100 test with the NCP filter applied and also repeated the Ishihara plates through the same filter. This phase helped us assess whether the filter improved color distinguishability without introducing distortions or unnatural tones.

In addition to the NCP filter, we introduced a Daltonization filter as a comparison condition. Daltonization represents a more conventional approach to color correction, where colors are remapped to maximize separability, often by shifting hue values based on cone response models. Participants reviewed AR FM100 and repeated the Ishihara test under the Daltonization condition.

After completing the structured testing, participants moved on to a video evaluation phase. In this part of the study, they watched short video clips of natural scenes, including forests, flowers, lakes, and skies. These scenes were chosen deliberately because they contain a wide range of colors that are often challenging for individuals with red-green or blue-yellow color vision deficiencies. For example, forest scenes show many shades of green leaves, along with red and yellow tones that appear during autumn. This step was designed to help us better understand how each color correction filter changed the way participants saw colors in scenes that looked more like everyday life.

However, during this phase, we noticed that the video evaluation was not as effective as expected. The clips were short and visually intense, and several participants found the task unfamiliar or unnatural. One possible reason is that the Meta Quest 3 is primarily designed for interactive, spatial activity rather than passive video watching, which can make static or pre-recorded scenes feel distant and less natural.

Based on these observations, we replaced the video task with a more natural alternative: allowing participants to freely look around their surroundings or out the window while using the filters. This approach aligned better with how the headset is meant to be used, taking advantage of its real-world passthrough capability. Participants could explore familiar spaces, compare colors under both the Daltonization and NCP filters, and judge the differences directly in context. This made the evaluation more intuitive, realistic, and comfortable for them.

Finally, the session concluded with a semi-structured interview, which was audio-recorded for later analysis. During this discussion, participants reflected on their experiences with both filters, describing which colors appeared clearer, which seemed exaggerated or unnatural, and how the corrected view compared to their normal perception. They also shared their thoughts on the comfort and usability of the AR FM100 task and their overall impressions of the system interface.

By combining these qualitative personal reflections with the quantitative performance data from the FM100 tests, we could see what improved or got worse, alongside with how our participants experienced the changes. The data variety allowed us to evaluate whether the NCP recoloring provided a positive effect on the quantitative tests, and compare its real-world comfort against the classic Daltonization recoloring.

5. Results and Data

In this section, we discuss the qualitative and quantitative results we have received from the final study.

5.1 CVD Assessment and Correction Results

The table in Appendix 1 provides the quantitative record of the 8 participants going through our experimental procedure. From it, it is evident that the FM100 test successfully provided a consistent baseline for assessing CVD type and severity on the desktop, with all participants' classified type matching their performance in type-dependent Ishihara plates perfectly. Repeat testing in AR has then reliably shown a lower severity score, which, as noted in the original FM100 paper, is to be expected for repeat testing due to the participants becoming more familiar with the test. This is contrary to the expected worsening in performance due to the different environment, screen resolution and color gamut of Meta quest 3.

Looking at the Desktop FM60 and No filter Ishihara plates reveals an inconsistency in their results, such as a 0.62 severity (rather strong colorblindness) guessing 15 plates, comparable with a 0.09 severity individual. There are a multitude of possible reasons causing this discrepancy. Both FM100 and Ishihara are estimators and not diagnosis tools, as such, they will inherently provide slightly ranging results and are affected by several different factors. Such reasons are explored further in the discussion of our results later on.

Comparing the performance on FM100 between no filter, NCP and Daltonization reveals a positive trend. Our natural preserving color correction, although limited, has increased the participant's performance once applied. While this could be attributed to repeat testing, the participants were simply asked to fix any newly visible mistakes from their previous results. Daltonization, once applied afterwards, has yielded mixed results, as due to the aggressive color remapping and presence of colorblindness, many participants simply got confused and began rearranging rows that were mapped to one color.

Ishihara numbers reveal a positive trend as well, however the NCP improvement here is less significant than with FM100. While most participants were able to see slightly more numbers with the natural filter, it could not compete with Daltonization, which completely remaps the colors to different wavelengths. It is worth noting that the Daltonization filter provided overall worse performance on the Ishihara plates when compared with the previous study. This is most likely due to the fact that the test is now evaluated in pass-through instead of AR, and the correction is applied through an LUT, and not a custom shader. As such, it is even more influenced by camera resolution, not just the screen, alongside the lighting conditions.
The final important distinction to be made is the profound differences in the metrics between Deutan and Protan participants. This is entirely due to the implementation of our NCP correction for deutanomaly, which was not strong enough and had the aforementioned flaws. As such, participants with protanomaly observed overall better results for both FM100 and Ishihara with NCP than those with deutanomaly. This particular distinction has been reflected profoundly in the semi-structured interviews.

Lastly, some outliers deserve to be mentioned. For example, participant 7 has showcased an excellent performance in all of our FM100 tests, however was only able to recognize less than half the plates, which appears to be the highest inconsistency between the two metrics in the table. Participant 6 has not observed a meaningful improvement in Ishihara with Daltonization, indicating that the filter was not applied, however due to time constraints we were unable to re-test them.

5.2 Interview Results

5.2.1 Interview Structure

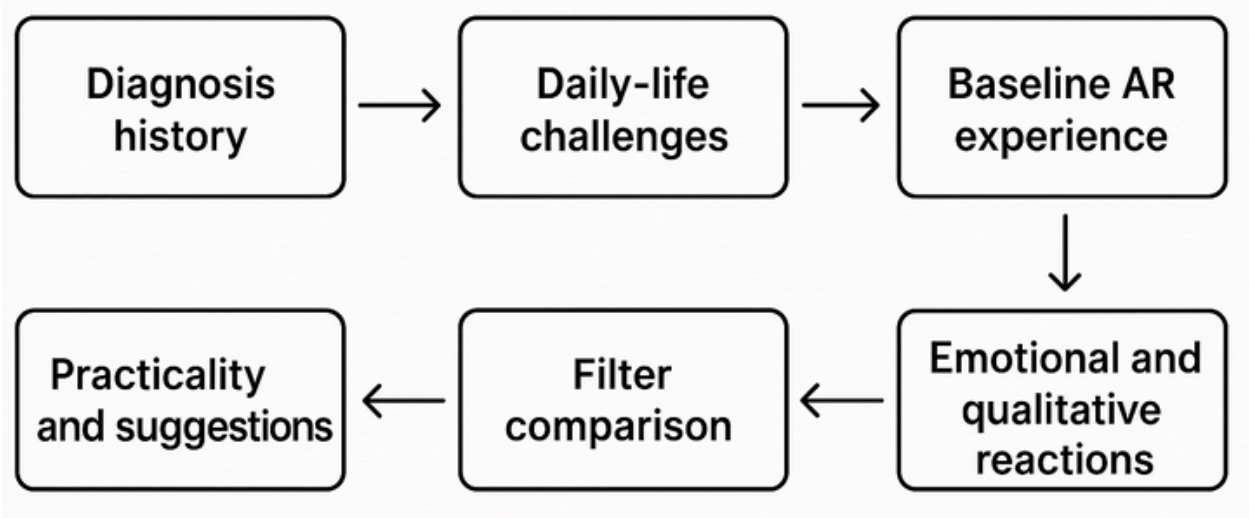


Fig 5.1: Flow diagram of the semi-structured interview

For each participant, our semi-structured interview followed the same progression. We began with an introduction, touching on prior diagnosis, how the participant found out, and whether their type and severity is known. We then moved on to a discussion of daily challenges encountered as a result of their condition, such as cooking meat in cases of participants 3, 5 and many others.

We then began the evaluation of our software with questions about the testing experience, comparing the desktop or AR FM100 test in convenience and difficulty. We also noted any issues with screen resolution, freezing, color difference or unstable AR objects (as specifically reported by participant 7).

We moved on to the evaluation of both filters, asking participants about their newfound confidence with UI elements using either and their preference of one over the other for practical and visual reasons.

Finally, we proceeded to the emotional impact of each correction, as the feelings of suddenly observing previously unseen gradients, numbers and colors is an important aspect of our study. We have also asked about any discomfort or motion sickness, and the participants' opinions on the realism of the applied correction. We finished by querying the participants on the potential of using something like this in real life, and any future improvements they would find necessary to do so.

5.2.2 Practical Use

Participants have overall criticized the system as impractical for daily use, with by far the most significant reason, as mentioned by all participants, being the bulky size of the headset. It was impossible to imagine wearing such a system daily in its current form for the participants. Participants like 2 and 8 stated that the added value was too small to currently justify wearing such apparatus on the head, and were often concerned with wiring or battery limitations. Participant 5 was unhappy with the environment and stated that AR experience was too imprecise for productive use, and eye-strain made it impossible for long-term use.
Ignoring the discomfort caused by the size and weight of the Meta quest 3, however, leads to a much different conversation. When asked to consider the current software in a lightweight form, almost all participants made a strong distinction between original Daltonization compared to our NCP recoloring.

Daltonization was found technically impressive when applied to Ishihara plates, however, observing the surroundings, only participants 2 and 7 explicitly stated this to be a practical correction. Most participants found it outright unusable outside of the Ishihara test, such as participant 8, who reported the real-world colors to be washed out, “uglier” and “less alive”, and noted an uncomfortable change in people’s skin tones. Participants 3,4,5 and 8 explicitly rejected Daltonization over NCP filtering despite the improvements in Ishihara scores, as the real-world colors were drastically altered such that objects lost their expected appearance.

The NCP correction was received much differently by most participants. Participant 8, with one of the strongest improvements in the table, described the outdoor scenes to be more vibrant while noting the newfound distinguishability of certain colors. Specifically, 8 was now able to distinguish certain fall leaves that beforehand blended, and was more confident in naming their colors. Participant 3, among others, stated they would particularly consider NCP on lightweight glasses. Participants 5 and 6 did not observe a practical preference of one filter over the other. Participants 1, 2 and 7 did not observe much of a difference from the normal vision when using the NCP filter, and as such preferred Daltonization for practical applications.

The results we received from the interview aligns with the fact that our Deutanomaly NCP has come out faulty, had very little effect, and thus Daltonization was preferred by the three participants. They do, however, show a general preference for practical use of our implementation in participants with Protanomaly.

5.2.3 Emotional Impact

Same as with practical use, there is an important separation between Daltonization and NCP, and also between Protanomalous and Deutanomalous participants. In terms of emotional impact, however, the divide is even larger for both distinctions.

Daltonization has been received negatively almost across the board for the Protanomaly group, while it was surprising to see more numbers in the test, it was described as unnatural, unpleasant and dead by most participants when looking at videos, around the room or outside the window. Participant 6 was particularly concerned with a concrete example of a table in the room whitening, calling it misleading. Daltonization showed itself as significantly more unique to the Deutanomaly group due to the aforementioned lack of NCP effect, thus they gave an emotional preference in favor of Daltonization for this study.

NCP for Protanomaly is a different story. Participant 8 has once again noted on the vibrancy of colours, described the world as “more alive”, noted that FM100 gradient was seen “much better”. This was the general sentiment observed with other participants. Participants also felt slightly more confident with identifying the ishihara plates, which everyone found surprising. The same group found it easier to correctly guess the Ishihara plates targeted specifically to distinguish protan versus deutan, making it impossible to note the resulting deficiency from the test with the NCP filter applied. Overall, participants found NCP to boost their confidence, participants like 3 expressed surprise when working on FM100, comparing it to a puzzle “clicking”, despite only a mild improvement in performance. NCP was once again not noted as anything special by the Deutanomaly group.

6. Discussion

In this section, we expand more on the results we have observed in the previous section.

6.1 Natural Color Preservation vs Daltonization

The main objective of our study was to evaluate if our NCP implementation could compete with Daltonization on the quantitative metrics and evaluate their differences in emotional effects on the colorblind. Despite our extremely limited implementation, constricted to a one-to-one look-up table structure, mostly implemented by borrowing and simulating modern color correction techniques for the colorblind, and trying to implement a per-image algorithm that would otherwise compare neighboring pixels, our results being positive at least for one colorblind group is a great showcase on the potential of this technology.

The NCP filter generally facilitated better performance in the FM100 test for the protanomalous group than simple Daltonization, as such aggressive remapping made one of the rows appear as one color, damaging the results. By preserving relative luminance and saturation relationships, we allow the users to improve their scores without changing the colors.

On the other hand, just like in the previous study, Daltonization has outperformed NCP on the Ishihara test. The aggressive shift of the technique is suited very well to distinguish the problematic colors observed in Ishihara. This was, however, overall seen as “overcompensating” by the participants.

6.2 Discrepancies in Assessment

The main issue in our data is the inconsistency between the desktop FM100 baseline, AR FM100 baseline, and the ishihara testing scores. As detailed in the results, participants have shown varying performance across the tests. The discrepancies are caused by a multitude of reasons, ranging from lighting setup and problems with experimentation procedure to the fundamentally different nature of the two tests.

First, the slight discrepancy between the desktop and AR FM100 baseline tests is to be expected, and, as mentioned even in the original FM100 manual, first re-test will almost always lead to a better performance due to the user’s familiarity with the test.

The discrepancy between FM100 and ishihara in this study deserves more explanation. The best explanation for this discrepancy is the fundamentally different nature of both tests. While FM100 tests for the specific confusion ranges, the Ishihara test is designed for rapid screening and red/green classification, and as such, their results are generally difficult to correlate consistently with CVD severity. This is due to the different factors influencing the final results, for example, the FM100 test was shown to correlate with non-verbal Iq and pattern recognition (M.B. Cranwell et. al., 2015), while no such factors affect the Ishihara test.

The original work observed a similar, although less dramatic effect. The worsening of the discrepancy between FM100 and Ishihara in our study can be further explained by the fact that Ishihara plates were now being evaluated in real life, and thus significantly more affected by lighting, distance, and even the angle of the book, which were not controlled precisely in this study. Finally, the logging for correctly guessed ishihara plates has been done manually, which could have led to errors or missing numbers. Therefore, automated logging is necessary to be implemented in the further study to mitigate this.

6.3 User Experience

The qualitative feedback we received has shown a significant split in the Protanomalous and Deutanomalous participants due to our implementation of NCP. As one of the project team members is protanomalous, we had an excellent reference on what exactly to do with the algorithm, with deutanomaly being less clear. As result, our implementation was not effective

enough for deutans, and participants with Protanomaly reported a higher degree of satisfaction with the NCP filter and the entire study as a whole. The environment felt richer and more vibrant for them, and their quantitative metrics saw an improvement. Often, the only reason the technology wasn't considered for daily use was simply the bulky form factor of the VR headset we have used.

7. Conclusion

Our study was first set to validate whether personalized, natural-preserving color correction for colorblind trichromats could be applied to the real-time context of AR. We have built on top of the previous work, Hue4U, that has implemented a simplified FM100 test into the environment and provided simple Daltonization based on the result. As a result of our study, we have not only validated the original goal but improved on the previous software, increasing its testing accuracy, the experiment procedure, new UI elements giving more control to the participant
Diving into the totally new area of XR and colorblind accessibility has been a profound challenge for the team and a valuable learning experience both in terms of communication and collaboration, scientific theory behind the condition, and irreplaceable software engineering experience. The many limitations and inconveniences we have encountered with the MetaSDK and the quest itself have slowed us down considerably, yet, considering these limitations, we were able to provide a satisfying and extremely promising result for the protanomalous natural-preserving recoloring. Our currently limited implementation already observes significant improvements in color-sorting tasks such as FM100, and the evaluation has shown that it is a much more user-friendly and practical way of providing accessibility to colorblind trichromats. Another contribution of our work is the technical enhancement of the digital FM100 test, as the existing implementations online mostly observe 40 colors, and no software other than Hue4U attempted to automatically determine and pass the result as CVD type and severity parameters. In this study, we have shown the logic and the basis behind this in more detail, however there are still improvements to be made.

Our evaluation has revealed the limitations with our current implementation clearly. The relatively simplistic logic of determining the CVD type from FM100 struggles at low severities and can be expanded further by properly analyzing the error distribution. Current NCP algorithm for deutanomaly has not come out as effective, meriting further improvement and evaluation due to the promising results with protanomaly.

To conclude, our project confirms that, while Daltonization remains useful for dichromats, NCP offers a more viable and practical direction for trichromats. Implementing such correction in AR still requires lighter and more seamless frames, which is the precise direction where the industry is headed. We hope this serves to be an important step in creating colorlind diagnostic and correction software in this field.

8. Future Plans

In this section, we quickly theorize on what the next steps are for our project.

8.1 FM100 Test improvements

Our implementation of the FM100 Test could be improved in a couple of ways.

First of all, after the experiments ended, we realized that this error representation potentially contributed to getting better results in the AR environment as it exposed the main regions where participants made the mistakes. For this reason, it would be beneficial if the numbers would stay hidden after the participant completed the test. Introducing controls such as buttons or checkboxes to show or hide these results whenever needed would be a valid addition.

Second of all, the color ranges for each CVD type are not as accurate as they could be due to the difference of color direction, as well as the smaller number of colors used in our test, compared to the original version of the test. Therefore, more thorough work could be done here as this would definitely have an impact on the severity of the CVD type.

Finally, it could be beneficial to implement a feature of saving the experiment results automatically. It could be done by saving a PDF file into the system instead of doing it manually or making screenshots of the results.

8.2 Natural-Preserving color correction with Machine learning

The LUT approach is clearly not the path to take when implementing image-dependent algorithm estimations. As machine learning develops, it becomes increasingly more possible to implement per-image, natural-preserving algorithms such as the one by Zhou in real time. We hope that by training a model such as HDRNet, the algorithm can be replicated more accurately and applied to the passthrough video feed with minimal delay. This would allow for proper application of this technique both in the real-time context of XR and any other display.

9. Bibliography

- Plutino, A., Di Scipio, F., Danese, M., & Balestri, M. (2023). Aging variations in Ishihara test plates. *Color Research & Application.* Advance online publication. https://doi.org/10.1002/col.22877
- Simunovic, M. P. (2010). Colour vision deficiency. *Eye (London, England), 24*(5), 747–755. https://doi.org/10.1038/eye.2009.251
- Fanlo Zarazaga, A., Gutiérrez Vásquez, J., & Pueyo Royo, V. (2019). Review of the main colour vision clinical assessment tests. *Archivos de la Sociedad Española de Oftalmología (English Edition), 94*(1), 25–32. https://doi.org/10.1016/j.oftal.2018.08.006
- Amos, J. F., & Piantanida, T. P. (1977). *The Roth 28-hue test. American Journal of Optometry and Physiological Optics, 54*(3), 171–177. https://doi.org/10.1097/00006324-197703000-00008
- Erb, C., Adler, M., Stübiger, N., Wohlrab, M., Zrenner, E., & Thiel, H. J. (1998). Colour vision in normal subjects tested by the colour arrangement test “Roth 28-hue desaturated.” *Vision Research, 38*(21), 3467–3471. https://doi.org/10.1016/S0042-6989(97)00433-1
- Zhou, H., Huang, W., Zhu, Z., Chen, X., Go, K., & Mao, X. (2024). Fast image recoloring for red–green anomalous trichromacy with contrast enhancement and naturalness preservation. *The Visual Computer, 40*, 4647–4660. https://doi.org/10.1007/s00371-024-03454-8
- Jost, P. (2024). *Advancing Colour Perception: Exploring young children’s colour discrimination in mixed reality* (CELDA 2024). https://files.eric.ed.gov/fulltext/ED665429.pdf
- Melillo, P., Riccio, D., Di Perna, L., Sanniti di Baja, G., De Nino, M., Rossi, S., Testa, F., Simonelli, F., & Frucci, M. (2017). Wearable improved vision system for color vision deficiency correction. *IEEE Journal of Translational Engineering in Health and Medicine, 5*(1). https://doi.org/10.1109/JTEHM.2017.2679746
- Qin, J., Checherin, S., Li, Y., van der Zwaag, B.-J., & Durmaz-Incel, Ö. (2025). *Hue4U: Real-time personalized color correction in augmented reality*. arXiv. https://arxiv.org/abs/2509.06776
- Tanuwidjaja, E., Huynh, D., Koa, K., Nguyen, C., Shao, C., Torbett, P., Emmenegger, C., & Weibel, N. (2014). *Chroma: A wearable augmented-reality solution for color blindness*. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp ’14). ACM. https://doi.org/10.1145/2632048.2632091
- Melillo, P., Riccio, D., Di Perna, L., Sanniti di Baja, G., De Nino, M., Rossi, S., Testa, F., Simonelli, F., & Frucci, M. (2017). Wearable improved vision system for color vision deficiency correction. *IEEE Journal of Translational Engineering in Health and Medicine*, *5*, Article 2679746. https://doi.org/10.1109/JTEHM.2017.2679746

- Colorlite. (n.d.). *Farnsworth Munsell 100 Hue Color Blind Test (Limited).* Retrieved October 13, 2025, from https://www.colorlitelens.com/images/huetest/Farnsworth100.html
- Farnsworth Munsell 100 Hue Color Blind Test - Colorlite https://www.colorlitelens.com/images/huetest/Farnsworth100.html
- D. Farnsworth (1945) - 100 HUE TEST https://www.xritephoto.com/documents/literature/gmb/en/gmb_fm100_instructions_en.pdf
- M.B. Cranwell et. al. (2015) - Performance on the Farnsworth-Munsell 100-Hue Test Is Significantly Related to Nonverbal IQ https://pubmed.ncbi.nlm.nih.gov/26024100/

10. Appendix

Participant number	**Farnsworth-Munsell 60** (Deficiency, Severity)				**Ishihara Plates** (Out of 38, screened deficiency)		
-	(Desktop)	(AR)	(Natural color filter)	(Daltonization filter)	(No filter in AR)	(Natural color filter)	(Daltonization filter)
1	Deutan 0.28	Deutan 0.14	- 0.1	- 0.1	14 Deutan	16 Deutan	28 -
2	Deutan 0.2	Deutan 0.16	Deutan 0.14	Deutan 0.3	8 Deutan	8 Deutan	32 -
3	Protan 0.62	Protan 0.47	Protan 0.38	- 0.24	15 Protan	19 -	22 -
4	Protan 0.42	Protan 0.36	- 0.16	- 0.04	14 Protan	19 -	34 -
5	Protan 0.52	Protan 0.45	- 0.19	Protan 0.22	8 Protan	12 Protan	21 -
6	Protan 0.38	Protan 0.21	Protan 0.14	Protan 0.2	6 Protan	8 Protan	8 Protan
7	- 0.09	- 0.12	- 0.06	- 0.04	15 Deutan	18 Deutan	21 -
8	Protan 0.4	Protan 0.37	- 0.04	- 0.08	7 Protan	14 -	26 -

Appendix 1: Final test results

Dashes indicate the inability to determine CVD type as participant approaches results observed in normal vision.