On the need of accurate device calibration in color vision deficiency computer assessment

Luca Armellin[1], Alice Plutino[1], Alessandro Rizzi[1]
[1]MIPS Lab, Department of Computer Science, Università degli Studi di Milano, Milano
Contact: Luca Armellin, luca@armellinluca.com

Abstract

Among the wide offer of Color Vision Deficiency (CVD) assessment tests, only a few are designed to be administered via digital display monitors and thus require accurate color calibration and profiling. On the one hand, using fully characterized and calibrated devices is desirable to reduce the error introduced by inter-device variability. However, the need for a calibrated device restricts the opportunities for users and institutions to perform pre-screenings, low-cost tests, and online/remote tests. Making tests more accessible and available to users might increase the awareness and understanding of CVDs and should help potential CVD observers, whose status might go unnoticed and undiagnosed for years. In this paper, we analyze data gathered using a simple online game (http://qolour.it) to determine whether it is possible to avoid controlled environments and protocols for CVD pre-screening purposes. In particular, we analyze the use of mobiles, tablets, and personal computers to perform CVD pre-screenings, comparing the results and performances obtained by a control group with the rest of the users.

Introduction

Qolour (*www.qolour.it*) is a web-based serious development aimed both at helping people obtain a pre-screening assessment of a possible color vision deficiency (CVD) and collecting big amounts of data inherent to color vision and color perception (Armellin, et al. 2022). It is currently under development, and as such many aspects of the game needs to be analyzed to prove its overall effectiveness, ranging from data collection to data analysis and to user experience. While, at its current version, it proved reliable in correctly identifying the color vision deficiencies of subjects in a control group, we also analyzed the data gathered since the first time Qolour wash published for free usage using statistical metrics that don't require a laboratory setting.

When the game starts, the player is presented with seven differently colored shapes arranged in a circular fashion, with another slightly bigger one in the center surrounded by an animated timer; the purpose of the game is to press, before the timer expires, on the outer shape with the same color as the central one. The central color, which will be referred to as the *target color*, is generated randomly inside the HSL space, and subsequently converted to RGB, with the Hue being completely random, Saturation bounded between 0.6 and 0.8 and Lightness fixed to 0.5; one of the outer colors is the same as the target color, while the remaining six are chosen as to lie on confusion lines corresponding to the three types of dichromats; the background color of each level changes and can randomly be achromatic (having the same Lightness of the target color) or colored (a pseudorandom color computed such that it has the opposite Hue of the target color plus or minus 10 degrees but the same Lightness and Saturation).

The game has been administered to a control group consisting of 8 color-deficient observers (CDOs) and 8 normal color observers (NCOs), each of which had been given a link and required to freely interact with the web interface and play as much times as they wanted, regardless of their physical location, time of day and device; the only constraints given were that they should not wear any glasses, with the exception of prescription ones, nor use any colored filters (both physical or digital), like the yellowish ones implemented by many smartphone OSs to prevent "blue light fatigue". Each of the control subjects has been categorized as a NCO or a CDO based on the results

 ISBN 978-88-99513-22-1

scored on combination of various tests administered in a supervised fashion and a controlled environment, the tests being the 38-plates Ishihara PIP Test (version printed in 2021), the digital version of the Farnsworth D-15 test, and the anomaloscope (model OT-II manufactured by Neitz Instruments)

Qolour proved effective in classifying each of the control subjects as either CDO or NCO, as well as correctly distinguishing the three protan/protanomalous observers from the five deutan/deuteranomalous. It's worth noting that none of the subjects had been screened with tritanopia, being a rare condition often requiring specialized equipment to be correctly diagnosed such as, for example, anomaloscopes using Engelking-Trendelenburg or Pickford-Lakowski match rather than standard Rayleigh match (Pokorny, J; Collins, B; Howett, G; 1981).

Results

At the time of writing, Qolour has been played by 4 012 users from 94 different countries, totaling 6 452 games played and 117 692 total levels. Most of the devices used to access the website have been tablets and smartphones, the number of estimated unique devices being comprised between 800 and 1 900, which is just a rough estimate based on the collected user-agents.

Using the currently implemented metric (which slightly changed since the first release in Q2 of 2022), out of the 4 012 players, 3 660 have been classified as NCO whereas the remaining 352 as CDO, leading to a percentage of around 8.8% of CDOs among the whole population of players; Tab. 1 shows the actual numbers of players with their classification, players categorized as "deutan/protan" are the ones that make similar errors in both deutan and protan directions for which a clear distinction cannot be made having an high uncertainty, while the differentiation of anomalous trichromats from dichromats is not being carried out in the current version of Qolour.

Deficiency	Total players	% (total)	
Normal	3 660	91.23 %	91.23 %
Protan	146	3.64 %	8.77 % (all deficiencies)
Deutan/Protan	27	0.67 %	
Deutan	174	4.34 %	
Tritan	5	0.12 %	

Tab. 1 – Estimate of the distribution of deficiencies among the players population of Qolour.

To simplify visualization, in the following plots target colors have been grouped in larger classes derived from the sampling of all the target colors shown to the players from 8 bit to 4 bit per channel (reducing the possible colors to 4096 representative classes); these classes have been used to compute both the percentage of correct answers for each color class (Fig. 1) as well as the variance of the answers (Fig. 3).

Fig. 1 shows the percentage (along the radius, ranging from 0% to 100%) of correct answers given by all the non-deficient observers for each color class hue (in degrees along the circumference). It can be noted that all the target colors having a hue ranging from red to green were more easily correctly identified by the players (with a success rate above 90% for most of the hues), whereas a higher error rate occurred when challenged in correctly identifying light blue to magenta hues, with an error rate getting as high as 40%.

This tendency needs to be further analyzed, since it appeared with almost every individual labeled as NCO, leading to higher error rates along the tritan confusion lines with respect to CDOs. Given

this anomaly and the unavailability of tritan control subjects, as of now a user is flagged as tritan only if all the errors committed fell onto tritan confusion lines in the first 20 levels, the easier ones where NCOs doesn't seem to commit as much errors as in further levels.

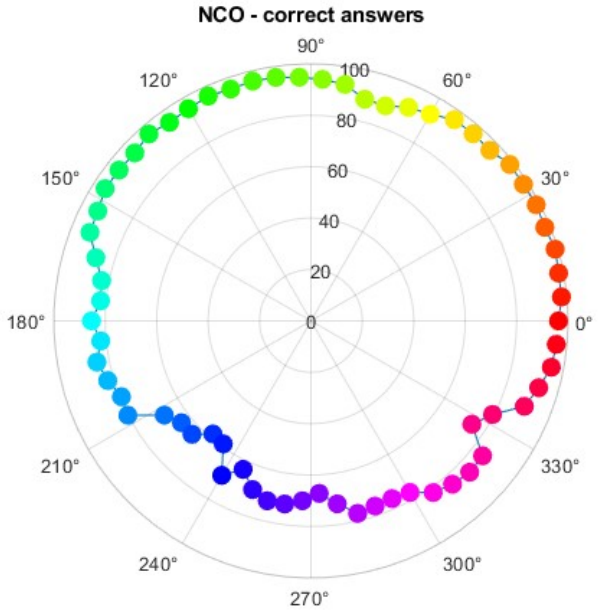

Fig. 1 – Percentage of correct answers by hue given by normal color observers between level 10 and 39.

Later the variance of the answers given by both the CDOs and the NCOs has been evaluated; a choice has been made to evaluate only the variance among the hues for two main reasons:

- Each level of the game presents the user with colors with roughly the same brightness and saturation, mainly the hue changes
- Visualization of variances in 3D spaces (such as RGB or xyY) is not trivial

The variance among the hues has been computed using the following algorithm:

```
for each level l {
      for each target color tc {
            for each background bg {
                  compute σ²h as the variance of the n hj
                  hues of the colors chosen by the players.
            }
      compute σ²tc as the average of all the m σ²h,
      weighted by the numerosity qh of the samples.
      }
}
```

In this way, the variances are computed only among the users who have been shown the same target (as to avoid comparing different stimuli between them), with the same background color (since different backgrounds may introduce simultaneous contrast) and at the same level (since different levels comes with different difficulties derived from a varying distance between the target color and all the possible colors shown to the player). The var_h variances have been computed as shown in Eq. 1, while the average variance for each target color has been computed using Eq. 2; from Eq. 1 can be seen that the resulting variance is always normalized between 0 and 1, where 1 represents the distance between two colors having their hues 180° apart.

$$\bar{h} = \arctan2\left(\frac{1}{n}\sum_{j=1}^{n}\sin\left(h_j\right),\ \frac{1}{n}\sum_{j=1}^{n}\cos\left(h_j\right)\right)$$
$$\sigma_h^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{\pi}\arctan2\left(\sin\left(\bar{h}-h_j\right),\ \cos\left(\bar{h}-h_j\right)\right)\right)^2$$

Eq. 1

$$\sigma_{tc}^2 = \frac{1}{\sum_{h=1}^{m} q_h}\sum_{h=1}^{m}\sigma_h^2 \cdot q_h$$

Eq. 2

The variances have been evaluated only between level 10 and level 39, the lower limit being chosen because levels prior to 10 are purposefully easy to show the user how to interact with the game and are not meant for actual data collection nor are taken into consideration for evaluating the player's color perception, the upper limit being chosen because level 39 coincide with the 3rd quartile of all the observations, as can be seen in Fig. 2.

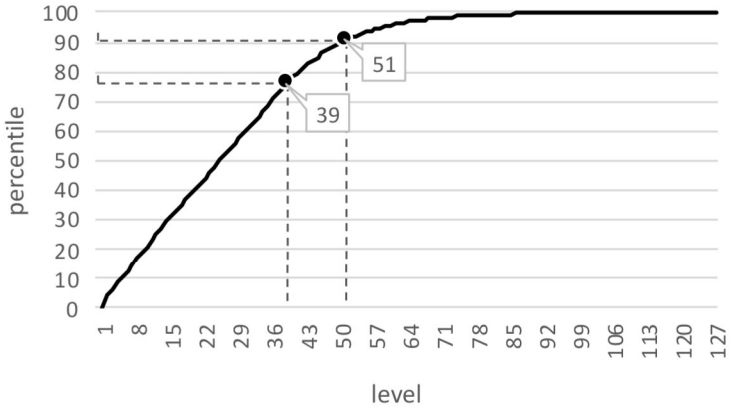

Fig. 2 – 75% of all the observations lies between level 1 and 39, with just 25% of all the observations being made in higher levels. This indicates how most of the players struggles to get past the first 40 levels, with levels above 51 comprising only 10% of all the observations.

The plots in Fig. 3 show the normalized variances for each target color class (computed by sampling all the possible target colors with 4 bits per channel), the hues are shown on the circumference and the variances on the radius axis. For the sake of readability only some representative levels and variances up to 99th percentile are shown in the following plots, instead a synthetic report for all the levels is shown in Tab. 2.

	NCOs			CDOs		
Level	**Mean σ²**	**Mean hue variability**	**Numerosity**	**Mean σ²**	**Mean hue variability**	**Numerosity**
10-14	$2.71\pi^2 \cdot 10^{-5}$	± 0.94 °	9600	$1.05\pi^2 \cdot 10^{-3}$	± 5.83 °	2763
15-19	$5.50\pi^2 \cdot 10^{-5}$	± 1.33 °	9431	$7.42\pi^2 \cdot 10^{-4}$	± 4.90 °	1917
20-24	$8.99\pi^2 \cdot 10^{-5}$	± 1.71 °	9205	$5.78\pi^2 \cdot 10^{-4}$	± 4.33 °	1584
25-29	$8.94\pi^2 \cdot 10^{-5}$	± 1.70 °	9771	$5.36\pi^2 \cdot 10^{-4}$	± 4.17 °	1326
30-34	$1.46\pi^2 \cdot 10^{-4}$	± 2.17 °	9384	$2.47\pi^2 \cdot 10^{-4}$	± 2.83 °	1121
35-39	$3.63\pi^2 \cdot 10^{-4}$	± 3.43 °	10928	$3.32\pi^2 \cdot 10^{-4}$	± 3.28 °	717

Tab. 2 – Average variances and hue variability in each group of levels, divided between NCOs and CDOs, with the numerosity referring to the number of times the corresponding levels have been played.

 ISBN 978-88-99513-22-1

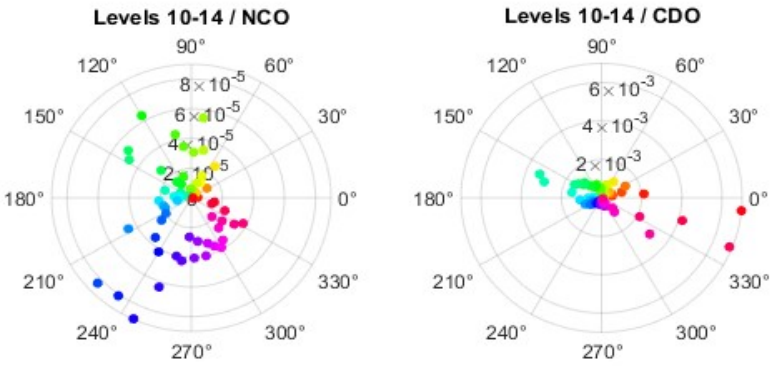

Fig. 3-a – Average variances among the levels from 10 to 14.

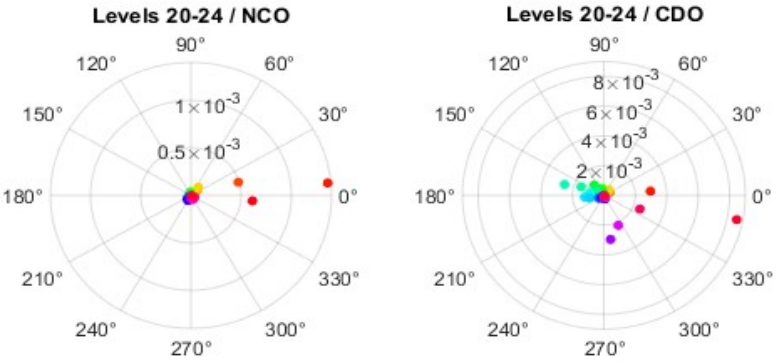

Fig. 3-b – Average variances among the levels from 20 to 24.

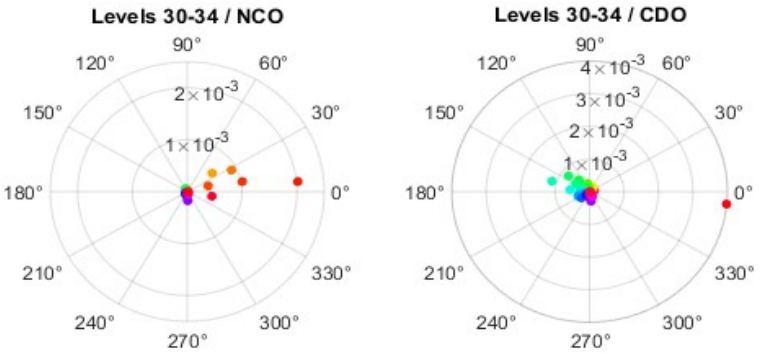

Fig. 3-c – Average variances among the levels from 30 to 34.

Discussion

Even though results obtained among the control group seems to show a certain degree of reliability in guessing the deficiencies of the subjects in the control group, the numerosity of the group is relatively low (being composed of only 16 subjects), it is thereby mandatory to include further analysis on the gathered data, which we included in the "Results" section.

The incidence of CDOs in the players population shown in Tab. 1 is close to the global incidence, which is of around 8.8% in males and 0.4% in females (Birch, 2012; Hunt and Carvalho, 2016). With respect to the specific deficiencies, the protan/protanomalous subjects appears in a higher incidence than in the actual global population (Padgham and Saunders ,1975); this can anyway be a result of the non-optimal segmentation carried out by Qolour, which has been tuned on an overall small sample of control subjects, as discussed above, showing thus room for improvement once a control set with a greater number of subjects will be tested.

 ISBN 978-88-99513-22-1

The percentage of correct answers given by NCOs for each hue shown in Fig. 1 shows two different behaviors based on the considered target colors' hue, for the blue-to-magenta being the most difficult to correctly identify, and the green-to-red being the easiest. The rationale behind the analysis of the correct answers given by NCOs, and NCOs only, is such that correctly identified NCOs should show a high percentage of correct answers; this approach still relies on the segmentation of the data between estimated player's deficiencies which, as stated, is not optimal.

The analysis of the variance among the chosen hues shows an overall low variance for both NCOs and CDOs, indicating a high level of consistency between the gathered data even using uncalibrated devices in uncontrolled environments. As the level increases, the colors shown to the player get less and less different between one another, effectively reducing the intra-level variance of displayed colors, while increasing the difficulty; the fact that variances from the lower to the higher levels remains roughly the same and of the same magnitude (with few exceptions), as can be seen in Tab. 2 seems to indicate that, regardless of the specific difficulty of the levels, the majority of the players are prone to give the same answers when presented with the same target and background colors. This trend is also confirmed when compared with the plot in Fig. 2 showing correct answers against hues for the NCOs, here we can see from the plots in Fig. 3 that variance remains low even for the blue-to-magenta hues, where NCOs seems to commit multiple errors; the same applies for CDOs observers that shows a low variance even though their error rates is significantly higher than that of NCOs, especially in levels above 20. The maximum variance is seen in the first 5 levels played by the CDOs, showing ±5.83 degrees in hue, which is definitely noticeable even if small, while staying under ±3 degrees for the NCOs with the exception of the last 5 levels considered.

Conclusions

The results of the analysis on the data gathered in the past several months show a certain consistency, while also outlining some issues. The low overall variance in hue among the answers given by the players presented with similar visual stimuli suggests that on average users respond in similar ways regardless of the device they are using or the environment in which they are playing, which pose the basis for a visual test that can be rendered accessible to the whole population and enables data collection for research purposes on a potentially worldwide scale.

In testing against the control group, Qolour proved effective in correctly discriminating subjects with color deficiency from those without; as the number of players increased over the months, the overall incidence of deficient subjects settled to a value close to that of the estimated global average.

Some issues emerged and remain open, for example the difficulty in discriminating tritan observers from normal observers, given the high error rates both in the direction of the tritan confusion lines as well as with blue to magenta target colors, analyzing data collected from thousands of session enabled the discovery of such critical issues which could've been not so evident in a controlled laboratory setting with a test population of orders of magnitude smaller.

It is clear that an app cannot be an alternative to a professional diagnosis carried out in a controlled setting, but the development of freely accessible apps and games might help in rapidly carrying out pre-screenings that could motivate individuals to seek for a professional opinion on a condition they might've not even known, as well as raising awareness in the general public on the existence of color vision defects and their meaning. The tradeoff between accuracy and accessibility should not be overlooked, especially since most of the most common and studied tests need particular and expensive equipment.

 ISBN 978-88-99513-22-1

References

Armellin L., Plutino A., Rizzi A., (2022), "Online games for colour deficiency data collection", *Colour and Colorimetry. Multidisciplinary Contributions. Vol. XVII A* (Open Access), pp. 79-86, RCASB, ISBN: 978-88-99513-18-4, DOI: 10.23738/RCASB.006

Birch, J. (2012), 'Worldwide prevalence of red-green color deficiency', JOSA A, 29(3), pp. 313-320.

Hunt, D. M. and Carvalho, L. S. (2016), 'The genetics of color vision and congenital colordeficiencies', Human Color Vision, Springer, pp. 1 - 32.

Padgham, C. A. and Saunders, J. E. (1975). 'The perception of light and colour'. London, UK: Bell.

Pokorny, J; Collins, B; Howett, G (1981), 'Procedures for Testing Color Vision: Report of Working Group 41.', National Research Council (US) Committee on Vision, National Academies Press (US).