https://doi.org/10.2352/ISSN.2169-2629.2017.32.106
Reprinted from
Journal of Imaging Science and Technology® 60(5): 050409-1–050409-9, 2016.

Using a Behavioral Match-to-Sample Method to Evaluate Color Vision Deficiency Simulation Methods

Joschua Thomas Simon-Liedtke and Ivar Farup
The Norwegian Colour and Visual Computing Laboratory, Faculty of Computer Science and Media Technology, NTNU—Norwegian University of Science and Technology, Gjøvik, Norway
E-mail: joschua.simonliedtke@ntnu.no

Abstract. *Color vision deficiency (CVD) simulation methods are used both to simulate color vision of color-deficient people and as input for image enhancement methods for the color-deficient. However, a standardized method to compare simulation methods has not yet been defined. We propose a behavioral methodology to evaluate, compare and rank different simulation methods. By using accuracy and response time data from a match-to-sample experiment, we can assess behavioral performance of simulation methods. We show firstly that the match-to-sample paradigm is well suited to show performance differences between observer groups; secondly, that the simulation methods do indeed simulate the desired CVD to some degree; and thirdly, we show that the proposed methodology can be used to rank different simulation methods. Our results indicate that the simulation method proposed by Brettel et al. depicts deutan CVD more accurately than the simulation method proposed by Kotera.*

INTRODUCTION

Around 4% of the population faces problems in differentiating colors due to color vision deficiencies (CVDs). CVD simulation methods emulate color-deficient vision for normal-sighted people. Simulation methods are also used in image enhancement methods for the color-deficient. However, a standardized method for evaluating simulation methods does not exist. This method would examine, if simulation methods work, and how well they perform. We introduce a behavioral method to evaluate simulation methods by analyzing accuracy and response time data from a match-to-sample experiment. In the proposed method, the original and the adjacent simulated version of the image are presented to the observer. He/she is asked to point out which one is the original or the simulated version, respectively. The response time and accuracy data are recorded, and this data is used to evaluate and rank the simulation methods. In this article, we investigate the questions

(i) if the match-to-sample methodology can be used to evaluate simulation methods;
(ii) if differences between the observer groups show that the simulation methods do simulate CVD as desired; and
(iii) if differences in the accuracy and/or response time data can be used to rank the simulation methods.

Received Apr. 10, 2016; accepted for publication June 8, 2016; published online Aug. 16, 2016. Associate Editor: Susan Farnand.

BACKGROUND

The majority of people with *trichromatic* color vision can differentiate millions of colors under certain viewing conditions. Photosensitive sensors called cones on the retina of the human eye contain pigments that make them sensitive to light of different wavelength: L-, M- and S-cones are sensitive to long, medium and short wavelengths, respectively [1, Chapter 5] [2, Chapter 6]. For the color-deficient, however, the cone sensitivities are either slightly shifted as compared to the normal-sighted (*anomalous trichromacy*) or certain cone types are missing all together (*dichromacy*) [3, Chapter 4] [2, Chapter 6]. This leads to either a reduced ability to differentiate some colors for anomalous trichromats, or the inability in perceiving certain colors at all for dichromats. CVDs can be found in about 8% of the male population, where CVDs along the red–green axis are most common[4] [3, Chapter 3]: CVDs related to the L-cones are called protan anomaly and protanopia [3, Chapter 4] [2, Chapter 6]. The respective terms for CVDs related to M-cones are deutan anomaly and deuteranopia. CVD tests are used to find out if a person is color-deficient, and/or to investigate the type and severeness of a CVD. Popular CVD tests are, for example, the Ishihara test,[5] the Hardy–Rand–Rittler (HRR) test,[6] the Farnsworth D15 test,[7] the Lanthony D15 desaturated test[8] or the anomaloscope.[9] Some digital CVD tests exist as well.[10,11]

Color-deficient people perceive colors differently. Colors along the red–green axis, for example, are mapped toward colors along the yellow–blue axis or toward neutral colors for protan and deutan color-deficient people.[12] Therefore, CVD simulation methods have been developed to emulate color-deficient vision for the normal-sighted. A variety of simulation methods exist.[10–20] Brettel et al.[12] proposed a dichromat simulation using a projection in three-dimensional LMS color space onto a plane that is visible for both trichromats and dichromats. Viénot et al.[20] simplified this method for the usage on CRT screens by computing a color map of the standard palette with 256 colors in order to simulate images for protanopes and deuteranopes. Capilla et al.[14] derived a more general analytical blueprint from the Brettel method for color vision models that fulfill certain requirements. The CVD simulation method

by Meyer and Greenberg[11] projects colors along confusion lines onto an axis that looks the same for dichromats and normal-sighted in the CIEXYZ chromaticity diagram. Kotera's approach[17] converts the original image into the perceptual IPT color space,[21] and reducing the contribution of the P-channel for protan and deutan color-deficient. This last simulation method provides both the simulation of dichromatic and anomalous trichromatic vision. Yang et al.[10] proposed another CVD simulation for anomalous trichromats based on shifting the sensitivities of the traditional LMS color matching functions by a λ wavelength. A unified color vision model to simulate both anomalous trichromacy and dichromacy based on the state model has been discussed by Machado et al.[19] Anomalous trichromacy can be obtained similar to the simulation method by Yang et al. Finally, Flatla and Gutwin[15] presented a personalized CVD simulation based on empirical measurement of color differentiation abilities that can also simulate anomalous trichromacy. Accurate simulation methods are essential for most so-called daltonization methods, which are automatic methods that improve the perception of images for the color-deficient by, for example, supporting the attentional mechanism through increased contrast between confusion colors and/or other strategies. Daltonization methods often require the simulation of the original image as input. Thus, it is important that CVD simulation methods simulate color perception of the color-deficient as accurately as possible for optimal daltonization results. How well CVD simulation methods represent color perception of the color-deficient is thus the key research question of this article.

Color perception is in its core an entirely subjective experience. But internal subjective processes can be assessed somewhat objectively. Behaviorism has the goal of forming psychological theories from observable events. Experiments are used, in which a sensible stimulus causes a measurable reaction in an observer [22, Page 14]. Match-to-sample is a behavioral methodology, in which a *sample stimulus* or *target stimulus* is presented to an observer adjacent to two or more *comparison stimuli* [22, Page 251f]. It is used to measure the performance of the short-term memory. The observer is asked to choose the correct stimulus that matches the sample stimulus. For each observation, it is recorded whether or not the observer answered correctly, and the sum of all correct answers for all observations and observers results in the *accuracy* value of the experiment. In a so-called *delayed match-to-sample* experiment, the comparison stimuli are presented after a short delay and in absence of the target stimuli allowing the measurement of further short-term memory influence on perception [22 Pages 251–252]. In this case, the *response times* from the observers are included in the analysis.[23,24] We showed in previous studies how the behavioral paradigm can benefit the field of image processing: as mentioned before, daltonization methods are meant to improve visual perception of the color-deficient by supporting the visual attention mechanism. Two different visual-search methods have been introduced to evaluate and rank daltonization methods based on measuring the effect of daltonization on the performance of the *attentional mechanism*.[25,26] Simulation methods, on the other hand, emulate the subconscious reflexive mental imaginary of color-deficient observers rather than a subjective conscious perception. An image simulated with an accurate simulation method will make the simulated and the original image look close to identical. This will consequently create a significant challenge on the *short-term memory* when both image versions are presented adjacent to each other. We argue therefore that a behavioral match-to-sample method is well suited to evaluate CVD simulation methods.

METHOD AND IMPLEMENTATION

We presented in a previous article[27] a behavioral experiment using the match-to-sample paradigm in order to evaluate CVD simulation methods, called Sample-to-Match Simulation Evaluation Method (SaMSEM). The method is used to show whether or not observers can perceive the difference between an original image and its simulated version. The SaMSEM methodology can be summarized as following (Figure 1):

(i) A fixation screen is shown for 1250 ms. Then, a target image is randomly shown either in its original version or in one of its simulated versions for 1250 ms.

(ii) After disappearing and showing the fixation screen for another 250 ms, the image reappears in two comparison versions, the original and the simulated version, located next to each other. One of the comparison images is the target image that has just been presented to the observer before, placed randomly either to the left or to the right side of the screen.

(iii) The observer is asked to click on either the left or the right arrow key of the keyboard depending on the position of the correct version on the screen. He/she is asked to answer as quickly and as accurately as possible. The program records the response time and whether or not the observer has answered correctly. The randomization for the target image version and its location is meant to minimize several biases.

Forty-four different images have been chosen according to the color image quality classification attributes by Pedersen[28] with minor adjustments for the color-deficient: predominant color hue/s (red, yellow, green, cyan, blue, magenta and/or multicolored), protan/deutan/normal color contrast, overall lightness and saturation, memory colors like skin color (African, Asian and/or Caucasian), sky blue and/or grass green, large areas of the same color including neutral colors, color transitions, fine details and busyness. There are natural images and graphical images such as drawing, transportation plans and flags. Images have been chosen with vivid and/or neutral colors, and colors that were difficult to differentiate for the color-deficient and/or colors that were easy to distinguish for everybody (Figure 2).

In the present study, we investigated only dichromacy simulation methods because these methods are commonly

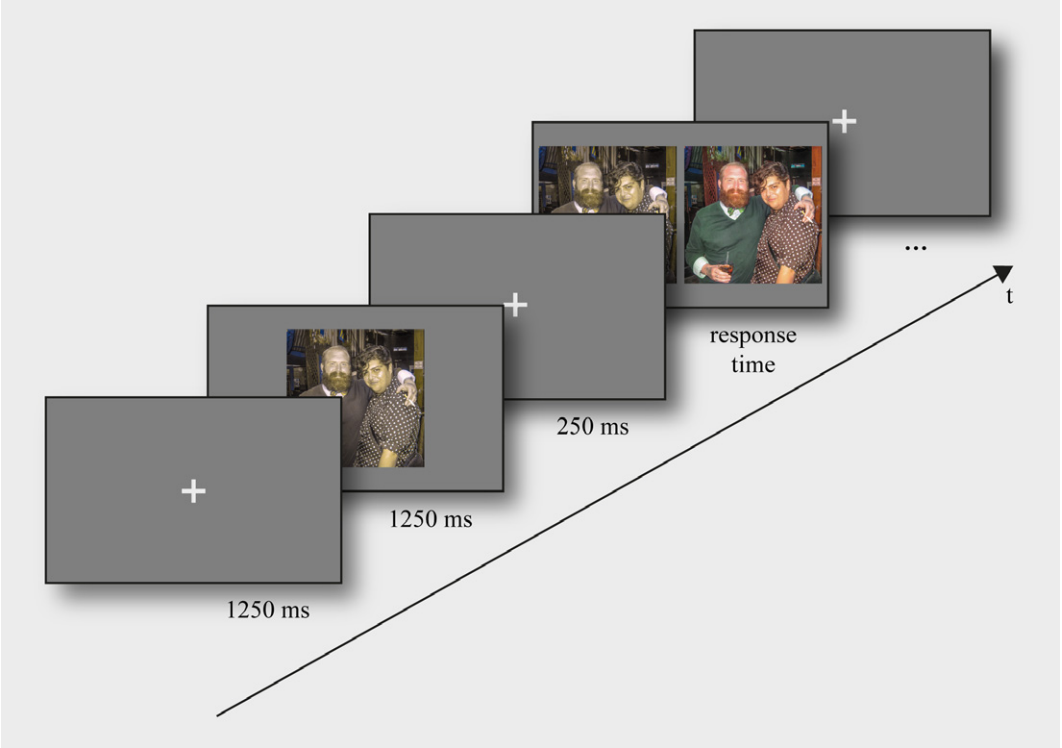


Figure 1. The workflow of the proposed method: at first, the target image is presented to the observer. In the above example the simulation is shown. Secondly, the image is presented in its original and one simulated version, and the observer has to decide whether the target image that he has just seen before is now located on the left or on the right side of the screen. In the above example, the observer is correct, if he/she clicks on the left arrow key. The program records response time and correctness, and moves on to the next target image.

Figure 2. Selected examples of images used in the experimentation. Images in 2(i)–2(iii), 2(vi) and 2(viii) by Joschua Simon-Liedtke, image in 2(iv) by Flickr@nhoulihan (CC BY 2.0), image in 2(v) by Ivar Farup, image in 2(vii) reproduced by courtesy of Carl Pilon (Flickr@Pilou), image in 2(ix) scanned by the authors from the Hardy–Rand–Rittler (HRR) test,[6] and image in 2(x) reproduced by courtesy of Truls Lange.

used in multimedia devices like browsers, smart phones, and so forth. We decided to concentrate on three main CVD simulations representing three major color spaces and gamuts used in color science and the industry: the Brettel method is based on the LMS color space, the Kotera method is based on the IPT color space, and the Vienot method

is based on a typical CRT screen gamut. We additionally introduced some adjustments to the Vienot method. Since its publication in 1998 when CRT screens were the standard, LCD screens that follow the sRGB standard have been taking over. In the original article, Viénot et al. convert the original DAC values to linear RGB by using a noncomplex gamma transform. This step has been replaced with the standard companding procedure defined by the sRGB specification.[29] From the same specification, the sRGB2XYZ matrix has been used, in addition to the XYZ2LMS matrix by Smith and Pokorny [1, Page 615]. The adjusted Vienot method represents therefore the widely used sRGB color space. Lastly, we included a control dummy method that reduces the chroma of the image to zero. The control dummy method makes the difference between original and "simulated" version obvious to both normal-sighted and color-deficient observers.

We expect several observations:

(i) We expect differences in both accuracy and response time data between normal-sighted and color-deficient observers. Namely, we predict lower accuracies and higher response times from color-deficient observers. Since original and simulated version look close to identical for the color-deficient, they are expected to make more mistakes in spotting the differences and/or need more time to find the correct target image among the comparison images. Normal-sighted observers are expected to see differences between original and simulated versions quickly, and to make less mistakes.

(ii) For the ranking of different simulation methods, we expect low accuracies and high response times from color-deficient observers for a good simulation method. We expect an accuracy of around 0.5 for a perfect simulation method, which would represent a random choice of two indistinguishable representations.

(iii) The dummy method is expected to have the highest accuracy and the lowest response times. This method is mainly used to prove the validity of our method.

The accuracy is given by the formula [30, Chapter 6]:[31]

$$\mathrm{ACC} = \hat{p} = \frac{n_{\mathrm{corr}}}{n_{\mathrm{total}}} \tag{1}$$

n_{total} is the total number of observations, and n_{corr} is the number of correct observations. The confidence interval has been computed with the Wilson interval score:[31]

$$\frac{1}{1+\frac{z^2}{n_{\mathrm{total}}}}\left[\hat{p}+\frac{z^2}{2n_{\mathrm{total}}} \pm z\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n_{\mathrm{total}}}+\frac{z^2}{4n^2}}\right]. \tag{2}$$

Furthermore, we implemented the χ^2 test and the paired Student-t tests (The latter did not give any further insights and has therefore been excluded in the following analysis.) [30, Chapter 8] for the statistical analysis of the accuracy data, and the Mood's median test [32, Pages 394–399] for the analysis of the response time data. A test level of 5% for statistical significance has been chosen for all tests. Furthermore, we plotted Q–Q plots in order to investigate the (log-)normality of the response time distributions.[33] The statistical data was analyzed using various Python libraries, namely the NumPy, the Matplotlib, the SciPy (especially the stats package for the statistical tests), and the Pandas libraries.[34–37]

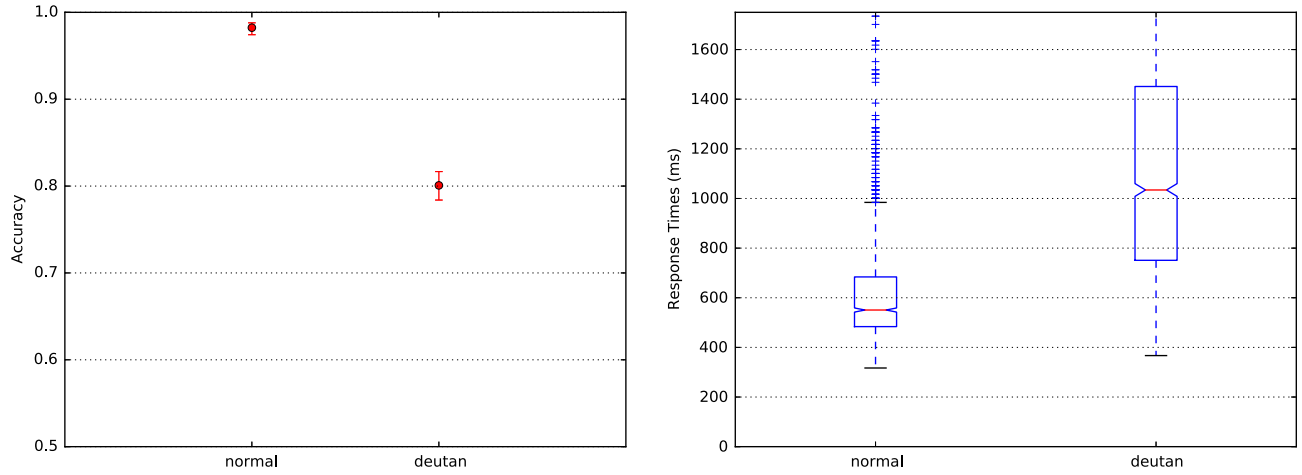
Four CVD tests have been used to investigate type and strength of deficiency: the Ishihara, the HRR, the Farnsworth D15 and the Lanthony desaturated D15 tests. The Ishihara test was used to identify normal-sighted from color-deficient observers. Type and severeness of the CVD were investigated with the HRR test. These results were confirmed with the Farnsworth D15 and the Lanthony D15 tests. The applied CVD tests, however, could not separate clearly between dichromats and anomalous trichromats, which would require an anomaloscope. We decided to focus on deutan color-deficient observers only for this experimentation because this observer group makes up about 75% of all color-deficient people [3, Chapter 3]. We conducted the experiment with 24 observers: 10 with normal color vision, and 14 with some sort of deutan CVD. According to the HRR test, 9 observers showed signs of strong, 3 observers showed signs of medium, and 2 observers showed signs of mild deutan CVD.

The method was implemented with PsychoPy2[38] on two PCs running Windows 7, and calibrated with Eye-One Match Pro to medium white, a gamma of 2.2 and illuminance of 120 lux. The luminance of the surrounding fluorescent D50-like lights was set to ca. 30 lux (CCT of 4230 K) for the experiments, and ca. 200 lux (4230 K) for the CVD tests. The luminance of the table for the CVD tests was about $38 cd/m^2$ measured with a Konica Minolta CS 100 spectroradiometer under a 45°/0° viewing angle.

RESULTS AND ANALYSIS

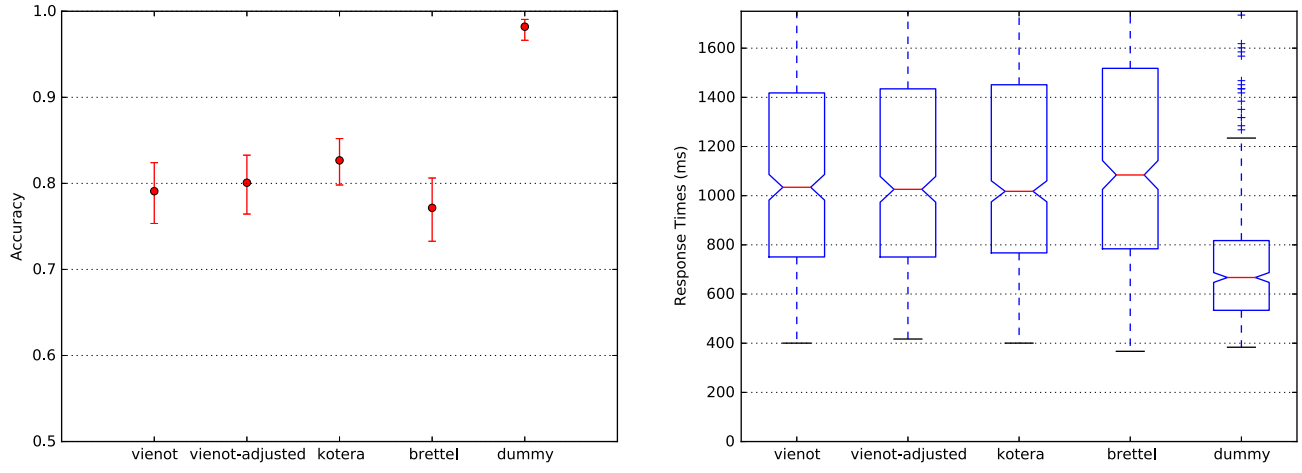
Analysis of the Accuracy Data for Deutan Color-Deficient Observers

All deuteranopia simulation methods excluding the dummy method collapsed result in the following accuracies (Figure 3(i)): normal-sighted observers with 0.98 (from 1471 observations), deutan color-deficient observers with 0.80 (from 2284 observations). The χ^2 test reveals that both values are indeed statistically significantly different: a p-value of 2.3×10^{-59}, which is virtually identical to zero, agrees with the observation from the accuracy graph since the accuracy values are very different from each other and their confidence intervals do not overlap. The observers in the first rounds of experimentation were shown the four simulated versions of all 44 images once. Then, we reduced the number of images gradually to 32 images, in order to reduce experimentation time and to avoid observer fatigue. We removed images that had similar or identical image quality attributes as explained in Method and Implementation. This is the reason why the observations are more than $32 \times 4 \times 10 = 1280$ but less than $44 \times 4 \times 10 = 1760$ for normal-sighted



(i) Accuracy data. (ii) Response time data.

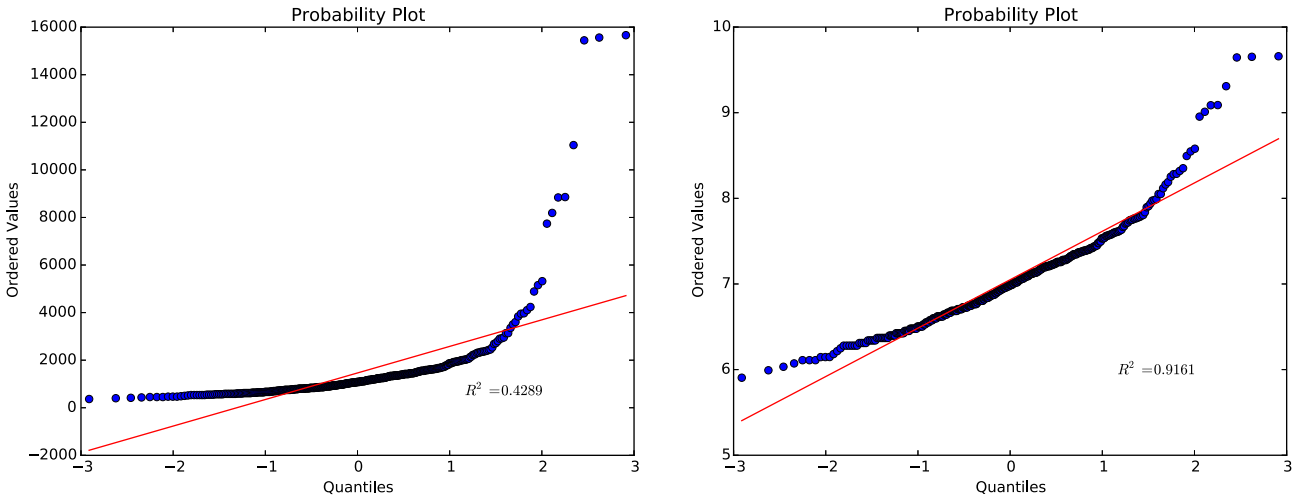
Figure 3. Accuracy (left) and response time data (right) of the deuteranopia simulation methods for normal-sighted and deutan color-deficient observers. The versions of all simulation methods excluding the dummy method are collapsed.



(i) Accuracy data. (ii) Response time data.

Figure 4. Accuracy (left) and response time data (right) of the individual deuteranopia simulation methods for deutan color-deficient observers.

observers, and more than $32 \times 4 \times 14 = 1792$ but less than $44 \times 4 \times 14 = 2464$ for deutan color-deficient observers. Since no paired data is used in this analysis, we included all observations in the following discussion. The high accuracy value for normal-sighted observers shows that this observer group can clearly perceive the difference between the original and the simulated versions. The accuracy of below 1.0 can be explained by target images containing only a few confusion colors. Thus, these images were little affected by the simulations, and the simulated versions looked close to identical to the originals for the normal-sighted as well. The lower accuracy for deutan color-deficient observers indicates that they have difficulties in seeing the differences between the originals and their simulated versions. This supports the hypothesis that the simulation methods do simulate deutan CVD to some degree. The fact that the accuracy value is statistically significantly higher than 0.5, as expected for a perfect simulation, suggests that none of the simulation methods is completely accurate. Color perception and CVDs are very personal and subjective experiences as mentioned in Background. There are different types and severities of CVDs, and many of our observers were anomalous dichromats for example. Dichromacy simulation methods can therefore represent individual color deficiencies only to a certain degree. At the same time, the investigated deuteranopia simulation methods have been chosen because they represent the lowest common denominator for a given observer group since the methods were designed to emulate an average deutan color-deficient observers. The individual deuteranopia simulation methods ranked by accuracies result in the following accuracy values (Figure 4(i)): the dummy method with 0.98 (from 502 observations), the Kotera method with 0.83 (from 756 observations), the Vienot-adjusted method with 0.80 (from 522 observations), the Vienot method with 0.79 (from 507 observations), the Brettel method with 0.77 (from 499 observations). The data



(i) The Q–Q normality plot shows a curved line in contrast to the best fit solution in red.

(ii) Not even the Q–Q log normality plot of the distribution results in a straight line.

Figure 5. Normality plots for the response time data of deutan color-deficient observers for the Brettel deuteranopia simulation.

Table I. Statistical analysis of the accuracy data (top) and response time data (bottom) between the individual simulation methods for deutan color-deficient observers. Statistically significant values are emphasized.

	Vienot-adjusted	Kotera	Brettel	Dummy
	(a) p-values of the χ^2 test			
Vienot	0.70	0.11	0.46	$\mathbf{1.3 \times 10^{-21}}$
Vienot-adjusted	x	0.24	0.25	$\mathbf{2.1 \times 10^{-20}}$
Kotera	x	x	**0.02**	$\mathbf{9.5 \times 10^{-18}}$
Brettel	x	x	x	$\mathbf{3.5 \times 10^{-24}}$
	(b) p-values of the Mood's median test			
Vienot	0.92	0.95	0.10	$\mathbf{4.4 \times 10^{-34}}$
Vienot-adjusted	x	0.98	0.22	$\mathbf{2.7 \times 10^{-35}}$
Kotera	x	x	0.14	$\mathbf{7.5 \times 10^{-51}}$
Brettel	x	x	x	$\mathbf{3.8 \times 10^{-40}}$

set is not complete for all observers due to practical reasons, and thus the observation numbers are different. The accuracy of the dummy method is statistically significantly higher than the accuracies of any other method (Table Ia). This supports our hypothesis that the accuracy data can indeed be used for the evaluation of CVD simulation methods as expected. Moreover, the Brettel method has a statistically significant lower accuracy than the Kotera method (Table Ia).

In Method and Implementation, we correlated lower accuracies to more accurate simulation performance. This means that the Brettel method simulates deutan CVD more accurately than the Kotera method because deutan color-deficient observers confuse the originals more often with the Brettel versions than with the Kotera versions. This makes sense because the Brettel method has a unique implementation for deutan CVD, whereas the Kotera method represents a simulation strategy for both protan and deutan CVDs combined. In the Kotera method, the P-channel of the IPT color space is omitted. The P-channel represents the information that is predominantly lost by *protan* color-deficient observers.[21]

Finally, the Vienot and the Vienot-adjusted methods have no statistically significant difference to the Kotera and the Brettel methods. The similarity to the Brettel method is not surprising because both methods are simplified versions of the same idea. However, we would have expected more obvious differences since different color spaces are used in their respective simulation models.

Analysis of the Response Time Data for Deutan Color-Deficient Observers

In the analysis of the response time (RT) data, we included only response times of correct observations. This is common for the analysis of this data type.[24] The response time data is not distributed normally, neither is the log distribution (Figure 5). Consequently, we analyzed the median values of the distributions rather than the mean values. The box plots in Figs. 3(ii) and 4(ii) show the median as a red line; the notches represent the width of the confidence interval of the median, computed by a Gaussian-based asymptotic approximation;[39] and the whiskers indicate the interquartile range of the distribution: it is computed as 1.5 times the third minus the first quartile following statistical standards.[39] When the min or max of the distribution lies within the whiskers, the min respectively the max are plotted as whiskers instead.

The response time data of all deuteranopia simulation methods collapsed excluding the dummy method are shown in Fig. 3(ii). The median RT of deutan observers is statistically significantly higher than the median RT of normal-sighted observers. This is confirmed by the Mood's median test: a p-value of 1.4×10^{-243} that is virtually identical to zero, shows that the median RTs of

deutan color-deficient observers is statistically significantly higher than the median RT of normal-sighted observers. Deutan color-deficient observers need more time than normal-sighted observers to spot differences as expected. This supports the hypothesis that response time data could be used to measure behavioral performances of simulation methods.

The median RTs of the individual deuteranopia simulations for deutan color-deficient observers are more homogeneous than the accuracies (Fig. 4(ii)). The dummy method has a statistically significantly lower median RT than any of the remaining methods as expected (Fig. 4(ii) and Table I(b)). This and the higher accuracy supports the validity of our setup: deutan color-deficient observers see the originals more similar to the simulated versions than to the grayscale versions. The statistical analysis of the response time data shows no further differences between any of the remaining simulation methods (Fig. 4(ii) and Table I(b)). The median RTs fail to be usable for the ranking of the investigated simulation methods. The results suggest that the response time data is much less sensible in detecting differences between simulation methods. One reason might be that at least the Vienot, Vienot-adjusted and the Brettel methods are too similar.

DISCUSSION

The analysis revealed two major points: Firstly, there are measurable differences in both accuracy and response time data between deutan color-deficient and normal-sighted observers. This proves the general validity of using the behavioral paradigm for the evaluation. It also shows that the investigated simulation methods emulate color-deficient vision adequately. Secondly, the accuracy data shows differences in at least one pair of the investigated simulation methods. The response time data in contrast is not sensible enough to detect any differences. Since only one statistical test reveals measurable differences, a complete ranking of the investigated methods is not possible. Our assumption a priori that simulation methods based on different color spaces would result in more measurable differences could not be confirmed. However, the results from the different observer groups indicate that our proposed method *can* be used for ranking in future studies under certain circumstances. In order to obtain statistically significant differences, the widths of the accuracy confidence intervals have to be reduced. This can be done by increasing the number of observers. Thus, the presented results serve mainly as proof of concept for using the behavioral paradigm in the evaluation of CVD simulation methods.

There is some more interesting feedback from the results. The chosen simulation methods give similar visual results to begin with, and the accuracy data echo this similarity. The Vienot and Vienot-adjusted methods for example are simplified variations of the Brettel method. The theoretical correctness of these simulations has been reduced to obtain a simpler implementation. Thus, the three methods could be used in different settings depending on the requirements for correctness and performance. In cases where a more accurate solution is necessary, the Brettel method could be used. Likewise, the Vienot or Vienot-adjusted methods could be used in cases, where a simpler and more resource preserving solution is needed.

As mentioned in Method and Implementation, the images in the image database have been chosen according to different image quality attributes. Most observers reported during discussions after the experimentation that images with strong red–green contrast were easier to distinguish than images with huge areas of neutral colors or predominantly yellow–blue contrast. However, the data obtained for this experimentation did not give enough data points in order to support these observations statistically.

A general advantage can be seen in contrast to psychometric scaling experiments based on Thurstone's law of comparative judgment[40] that are most commonly used in color imaging, like for example pairwise comparison [41, Chapter 8]. These experiments map visual stimuli varying along more than one measurable physical dimension to subjective perceptions like overall quality, naturalness, and so forth [42, Page 10], in order to assess subjective perception of certain image quality attributes. However, simulation methods do not emulate singular image attributes only, but the overall visual mental imagery of the color-deficient. Behavioral experiments in contrast assess the more or less subconscious, reflexive but objective reactions of the HVS rather than a subjective perception.[23,24] Especially match-to-sample experiments are known for measuring these behavioral responses to a given stimuli [22, Page 251]. Thus, our proposed match-to-sample method show more accurately how similar or different simulation methods perform because it is based on the behavioral response of the HVS rather than on subjective judgments.

The question arises if the given observers are representative for deutan color-deficient in general. The number of observers has been chosen according to an ISO guideline recommending at least 10 observers in psychophysical experimentations.[43] A survey of relevant literature shows that between 10 and 20 observers are commonly used for experimentations in image processing and behavioral science.[23,44–46] The MacAdam ellipses, for example, are based on the observations from one observer only. We argue that the chosen number is high enough for a proof of concept of our proposed method. Another question is related to multiple inference: due to multiple statistical comparisons between simulation methods, it could be necessary to adjust the test level of 5% according to the Bonferroni correction, and so forth. However, it is a common praxis in image processing to use an individual test level of 5% for each comparison. This can be seen, for example, for the comparison of gamut mapping algorithms.[47] Also, a CIE guideline recommends to solely account for the number of observers in the analysis, and not to account for the number of algorithms [48, Section 15.2.3].

In future work, we will use our proposed method to evaluate more CVD simulation methods with a higher

number of observers. We will implement a web-based application of the proposed method in order to reach a broader spectrum of relevant observers. A higher number of observers will lead to smaller confidence intervals and more statistically significant differences between the simulation methods as mentioned before. Using more observers will also give us the opportunity to group images from the database by different image attributes like for example predominantly protan, deutan or tritan contrast. Moreover, we will analyze how observers with different CVD severities perform for a given dichromacy simulation method. Using more homogeneous observer groups will also result in less variances for the results and narrower confidence intervals. In future work, we will also analyze performance for anomalous trichromacy and other personalized CVD simulation methods.[10,15,19] In that way, we will investigate if personalized CVD simulation methods perform better than standardized dichromacy simulation methods.

CONCLUSION

We presented a proof of concept for using the behavioral paradigm for the evaluation of CVD simulation methods. Our proposed method analyzes the response time and accuracy data from a match-to-sample task in order to evaluate, compare and rank performance of simulation methods. The results from the experiment of normal-sighted and deutan color-deficient observers show a clear difference in both response times and accuracies for deuteranopia simulation methods. This supports firstly the hypothesis that the behavioral paradigm is indeed suited for the evaluation. Secondly, it shows that the simulation methods do actually simulate color-deficient vision. A consistent ranking was not possible in the present study, but the accuracy data revealed that the Brettel method seems to perform better than the Kotera method. It could be seen that the accuracy data is much more sensible in detecting behavioral differences than the response time data. In future work, we will compare more simulation methods including personalized methods for anomalous trichromacy, with more observers including protan color-deficient observers.

ACKNOWLEDGMENTS

The authors thank Peter Nussbaum from NTNU for his feedback on writing the article, and Bruno Laeng from the University of Oslo for his help during the setup of the behavioral experiment and the data analysis. This research has been funded by the Research Council of Norway through project no. 221073 "HyPerCept – Colour and quality in higher dimensions".

REFERENCES

[1] G. Wyszecki and W. Stiles, *Color Science*, 2nd ed. (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2000).

[2] A. Valberg, *Lys Syn Farge*, 1st ed. (Tapir Forlag, Chichester, England, United Kingdom, 1998).

[3] E. Hansen, *Fargeblindhet*, 1st ed. (Gyldendal Norsk Forlag AS, Oslo, Norway, 2010).

[4] C. Rigden, "The eye of the beholder'—designing for colour-blind users," Br. Telecommun. Eng. **17**, 2–6 (1999).

[5] S. Ishihara, *Tests for colour-blindness—24 plates*, 24 Plates ed. (Kanehara Shuppan Co., Ltd., Tokyo, Japan, 1972).

[6] B. L. Cole, K.-Y. Lian, and C. Lakkis, "The new Richmond HRR pseudoisochromatic test for colour vision is better than the Ishihara test," Clin. Exp. Optom. **89**, 73–80 (2006).

[7] D. Farnsworth, *The Farnsworth Dichotomous Test for Color Blindness: Panel D-15* (Psychological Corporation, San Antonio, TX, US, 1947).

[8] P. Lanthony, "The desaturated panel D-15," Doc. Ophthalmol. **46**, 185–189 (1978).

[9] W. Nagel, "Zwei apparate für die augenärztliche funktionsprüfung," Z. für Augenheilkunde **17**, 201–222 (1907).

[10] S. Yang, Y. M. Ro, E. K. Wong, and J.-H. Lee, "Quantification and standardized description of color vision deficiency caused by anomalous trichromats—Part I: simulation and measurement," EURASIP J. Image Video Process. 9 (2008).

[11] G. W. Meyer and D. P. Greenberg, "Color-defective vision and computer graphics displays," IEEE Comput. Graph. Appl. **8**, 28–40 (1988).

[12] H. Brettel, F. Viénot, and J. D. Mollon, "Computerized simulation of color appearance for dichromats," J. Opt. Soc. Am. A **14**, 2647–2655 (1997).

[13] C.-N. Anagnostopoulos, G. Tsekouras, I. Anagnostopoulos, and C. Kalloniatis, "Intelligent modification for the daltonization process of digitized paintings," *5th Int'l. Conf. on Computer Vision Systems* (Universität Bielefeld, Bielefeld, 2007).

[14] P. Capilla, M. A. Díez-Ajenjo, M. J. Luque, and J. Malo, "Corresponding-pair procedure: a new approach to simulation of dichromatic color perception," J. Opt. Soc. Am. A **21**, 176–186 (2004).

[15] D. Flatla and C. Gutwin, "'So That's What You See!' building understanding with personalized simulations of colour vision deficiency," *ASSETS '12: Proc. 14th Int'l. ACM SIGACESS Conf. on Computers and Accessibility* (Association for Computing Machinery (ACM), Boulder, Colorado, 2012), pp. 167–174.

[16] J.-B. Huang, C.-S. Chen, T.-C. Jen, and S.-J. Wang, "Image recolorization for the colorblind," *IEEE Int'l. Conf. on Acoustics, Speech and Signal Processing, 2009* (ICASSP, IEEE, Piscataway, NJ, 2009), pp. 1161–1164.

[17] H. Kotera, "Optimal daltonization by spectral shift for dichromatic vision," *Proc. IS&T/SID 20th Color and Imaging Conf.* (IS&T, Springfield, VA, 2012), pp. 302–308.

[18] G. R. Kuhn, M. M. Oliveira, and L. A. Fernandes, "An efficient naturalness-preserving image-recoloring method for dichromats," *IEEE Transactions on Visualization and Computer Graphics* (IEEE, Piscataway, NJ, 2008), Vol. 14, pp. 1747–1754.

[19] G. M. Machado, M. Oliveira, and L. A. F. Fernandes, "A physiologically-based model for simulation of color vision deficiency," *IEEE Transactions on Visualization and Computer Graphics* (IEEE, Piscataway, NJ, 2009), Vol. 15, pp. 1291–1298.

[20] F. Viénot, H. Brettel, and J. D. Mollon, "Digital video colourmaps for checking the legibility of displays by dichromats," Color Res. Appl. **24**, 243–252 (1999).

[21] F. Ebner and M. D. Fairchild, "Development and testing of a color space (IPT) with improved hue uniformity," *Proc. IS&T/SID Sixth Color Imaging Conf.* (IS&T, Springfield, VA, 1998), pp. 8–13.

[22] J. E. Mazur, *Learning and Behavior*, 6th ed. (Pearson Education Inc., Upper Saddle River, NJ, USA, 2005).

[23] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," Cogn. Psychol. **12**, 97–136 (1980).

[24] I. Bramão, L. Faísca, K. M. Petersson, and A. Reis, "The contribution of color to object recognition," *Advances in Object Recognition Systems* (InTech, Rijeka, Croatia, 2012), pp. 73–88.

[25] J. Simon-Liedtke and J. Y. Hardeberg, "Task-based accessibility measurement of daltonization algorithms for information graphics," *12th Congress of the Int'l. Colour Association (AIC 2013)* (International Colour Association (AIC), Newcastle, UK, 2013), pp. 108–111.

[26] J. T. Simon-Liedtke and I. Farup, "Evaluating color vision deficiency daltonization methods using a behavioral visual-search method," J. Vis. Commun. Image Represent. **35**, 236–247 (2016).

[27] J. T. Simon-Liedtke, I. Farup, and B. Laeng, "Evaluating color deficiency simulation and daltonization methods through visual search and sample-to-match: SaMSEM and ViSDEM," Proc. SPIE **9395**, 939513 (2015).

[28] M. Pedersen, N. Bonnier, J. Y. Hardeberg, and F. Albregtsen, "Attributes of image quality for color prints," J. Electron. Imag. **19**, 01101601–01101613 (2010).
[29] "International Electrotechnical Commission (IEC)" Technical Report IEC 61966-2-1:1999 International Electrotechnical Commission (IEC, 1999).
[30] G. G. Løvås, *Statistikk for universiteter og høgskoler*, 2nd ed. (Universitetsforlaget, Oslo, Norway, 2008).
[31] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," J. Am. Stat. Assoc. **22**, 209–212 (1927).
[32] A. M. Mood, *Introduction to the Theory of Statistics* (McGraw-Hill Book Company, Inc., New York City, NY, USA, 1950).
[33] M. B. Wilk and R. Gnanadesikan, "Probability plotting methods for the analysis of data," Biometrika **55**, 1–17 (1968).
[34] NumPy Developers, NumPy documentation, http://www.numpy.org/, Last checked: 04/23/2015. (2013).
[35] SciPy Developers, SciPy documentation, http://www.scipy.org/, Last checked: 04/23/2015. (2013).
[36] Matplotlib Development Team, Maplotlib documentation, http://matplotlib.org/, Last checked: 04/23/2015 (2014).
[37] PyData Development Team, Pandas documentation, http://pandas.pydata.org/, Last checked: 04/23/2015. (2012).
[38] J. Peirce, PsychoPy documentation, http://www.psychopy.org/index.html, Last checked: 04/23/2015. (2014).
[39] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," Am. Stat. **32**, 12–16 (1978).
[40] L. L. Thurstone, "A law of comparative judgment," Psychol. Rev. **34**, 273–286 (1927).
[41] P. G. Engeldrum, *Psychometric Scaling: a Toolkit for Imaging Systems Development* (Imcotek Press, Winchester, MA, USA, 2000).
[42] B. Keelan, *Handbook of Image Quality: Characterization and Prediction* (CRC Press, Boca Raton, FL, USA, 2002).
[43] B. W. Keelan and H. Urabe, "Iso 20462: a psychophysical image quality measurement standard," Image Qual. Syst. Perform. **5294**, 181–189 (2003).
[44] J. Guild, "The colorimetric properties of the spectrum," Phil. Trans. R. Soc. A **230**, 149–187 (1932).
[45] D. L. MacAdam, "Visual sensitivities to color differences in daylight," J. Opt. Soc. Am. **32**, 247–273 (1942).
[46] W. D. Wright, "A re-determination of the trichromatic coefficients of the spectral colours," Trans. Opt. Soc. **30**, 141 (1929).
[47] J. Morovič, "*To Develop a Universal Gamut Mapping Algorithm*," Ph.D. thesis, (University of Derby, October 1998).
[48] J. Morovič, TC 8-03—"Guidelines for the evaluation of gamut mapping algorithms", Technical Report, Commission Internationale de l'clairage (CIE) (2003).