This is a previous version of the article published in Optometry and Vision Science. 2018, 95(11): 1054-1063. doi:10.1097/OPX.0000000000001293

CAN APPLICATIONS DESIGNED TO EVALUATE VISUAL FUNCTION BE USED IN DIFFERENT IPADS?

Dolores de Fez[a] PhD

Mª José Luque[b] PhD

Mª Carmen García-Domene[b,c] PhD

Mª Teresa Caballero[a] PhD

Vicente J. Camps[a] PhD

[a] Departamento de Óptica, Farmacología y Anatomía, Universidad de Alicante, Carretera San Vicente del Raspeig s/n, 03690 San Vicente del Raspeig, Alicante, Spain

[b] Departamento de Óptica, Universidad de Valencia, C/ Doctor Moliner 50, 46100 Burjassot, Valencia, Spain

[c] Departamento de Calidad Visual, FISABIO-Oftalmología Médica, Avda. de Catalunya 21, 46020 Valencia, Spain

Corresponding author: Dolores de Fez, dolores.fez@ua.es

ABSTRACT

Statement of Significance: Apple devices could be suitable for vision tests, provided that the test has been correctly adapted to the device, taking into account the spatial and colorimetric characterization of the screen.

Purpose: The majority of vision apps have not been developed by vision or colorimetry experts and suffer from conceptual and design errors that may lead average users to an erroneous assessment of their visual capabilities. The reliability of vision tests depends on the accurate generation of the necessary visual stimuli in a particular device. Our aim was to ascertain whether a given colour test, designed for a colorimetrically characterized device, might be used in another similar device.

Method: We evaluated colour reproduction errors in three iPad tablets of different models with Retina screens, using their individual colour characterization models and the model derived for another device.

Results: Our results showed, even with this small sample, the high degree of error caused when disregarding the fact that the colorimetric design valid for a given device may not be correct when displayed in another.

Conclusion: The distortion of the chromatic content may lead to subjects with vision defects to pass as normal or vice versa, compromising diagnosis reliability.

KEYWORDS

iPad, vision apps, stimulus reproduction, colour difference, vision tests

1. INTRODUCTION

Many mobile applications for assessing visual function have been developed in the recent years and are available for all kind of users in Google Play or Apple Store (see Appendix). Even if the developers usually enclose disclaimers stating that the results obtained with these tests are not conclusive and that a vision specialist ought to be consulted, the average user tends to trust the results and often accepts as correct the diagnosis provided by the app.

The problem with these apps is that, even assuming that the test is correctly designed for a particular device of known stimulus reproduction capabilities, the user's device may be different. Are the test results reliable under those conditions?

As we shall discuss below, the number of variables that must be taken into account to evaluate the reproducibility of stimuli in a given device depends on whether the vision tests under analysis is oriented to the assessment of spatial resolution, such as measurements of visual acuity, spatial or temporal contrast sensitivity or color vision. In the Appendix, we list the apps for Android and iOs revised for this study, specifying in each case the type of vision tests they carry out.

Regardless of the aims of the vision test that is being implemented, test designers must in the first place fix the visualization conditions during the measurements[1-3]. Vision test results depend on observation distance, screen tilt, ambience illumination level and screen brightness. The developers should specify the conditions under which the test is to be carried out, but this description usually does not appear in the user's manual. As a general rule[2,3], tests should be administered in a darkened room to avoid reflections or the influence of other light sources different from the device's screen, the observation distance should ensure that the angular size or spatial frequencies of the stimuli are correct, the device must be placed perpendicularly to the subject's line of sight to avoid changes in chromaticity or luminance with the angle of inclination and the

device must be plugged into the electrical supply, to avoid brightness fluctuations. Given that brightness is usually controlled with a slider, brightness should be set to the maximum value, which is the only setting that can be reliably reproduced in different measurements. Measurement conditions must be kept constant during the duration of the test, and reproduced in subsequent measurements, if the patient is being monitored over time, or if different population samples are being compared. Test administrators should adhere to any instructions provided by the application/software company regarding maintenance at maximum brightness, deactivation of auto-brightness changes, as well as ensuring the device is plugged in during testing to avoid inadvertent changes in luminance or chromaticity.

For tests involving stimulus size estimation, the device must be spatially characterized. Both the real screen size and the observation distance must be known to ensure that the stimuli have the adequate spatial characteristics. For instance, due to differences in pixel size, the number of pixels necessary to generate visual acuity optotypes in a 6-inch or a 10-inch screen are not the same. The app usually zooms the image to fit it in the screen, but this process does not necessarily respect the proportions of the optotypes, leading to mismeasurement of stimulus size and to an incorrect diagnosis of the subject.

The number of parameters to be considered increases in the case of tests based on luminance measurements, such as contrast sensitivity measurements with letters or sinusoidal gratings. Besides accurate reproduction of spatial characteristics of the stimulus, correct luminance reproduction in each stimulus pixel must be ensured. If, for instance, we have designed a sinusoidal grating of a particular frequency and contrast, an erroneous spatial characterization may alter the spatial frequency the device is actually displaying, due to changes in stimulus size, or introduce aliasing effects due to different sampling rates. An erroneous luminance characterization of the device will

change pixel intensity and modify the displayed contrast. Again, the diagnostic capabilities of the test are compromised.

The number of variables for color vision tests is, once more, increased. To all the former considerations, we must add that color reproduction characteristics of devices from different makes, and even different models of the same brand, may greatly differ.

The Ishihara plates[4] is the most commonly used color vision test, particularly for screening of red-green anomalies. The colors of the plates are chosen along the confusion line of protanopic and deuteranopic subjects[5]. Dichromats perceive colors belonging to a confusion line as having the same hue and colorfulness, but normal subjects see them as different. This allows discrimination between normal and dichromatic subjects, provided brightness clues are avoided, because dichromats, unlike subjects with normal color vision, cannot see the figure represented in the plate. Anomalous trichromats do not have confusion lines in the strict sense, but their discrimination losses are larger along the confusion lines of the dichromats, and may have difficulties to see certain figures and even be unable to see them, depending on the color difference between figure and background. However, color reproduction errors in mobile devices may be larger than average human thresholds[2], and this may cause normal subjects not to see the test or allow a dichromatic or anomalous subject to perceive it, leading again to an incorrect diagnosis.

Vision testing in the lab or the clinic is carried out using colorimetrically characterized devices, to avoid diagnosis errors. The widespread access to the Internet and to mobile devices have encouraged the proliferation of apps for vision testing. The majority of these apps, though, have not been developed by vision or colorimetry experts, and suffer from conceptual and design errors that may lead average users to an erroneous assessment of their visual capabilities.

Even when app developers have correctly designed the test, they should have information about the colorimetric characterization of the user's device to ensure that no errors will be introduced in their original designs. Unfortunately, the differences between the colorimetric profiles of the existing visualization devices may be large[2,3,6].

Our aim was to study whether the images generated for a test in a colorimetrically characterized device can be used in another device from the same manufacturer with comparable screen size, luminance and color characteristics. To minimize variability, we chose three displays of the same size and manufacturer.

2. METHODS

2.1 Devices

We have selected three iPad tablets (Apple, Cupertino, CA, USA) with screens 19.8 x 14.9 cm (2048 x 1536 pixel), differing in the year of manufacture: iPad4 (2013), iPadAir (2014) and iPad 5th generation (2017), iPad for short. The three devices have a Retina screen, which is based on the IPS (in-plane switching) technology, a specific Liquid Cristal Display type with a Light Emitting Diode backlight.

2.2 Measurements of the Tristimulus Values

Colorimetric data were obtained with two telespectroradiometers. The iPad4 was measured with a SpectraScan PR-650 telespectroradiometer (Photo Research, Chatsworth, CA, USA), while the iPadAir and iPad were measured with a Spectroradiometer CS-2000A (Konica Minolta Inc., Tokyo, Japan). A Bland-Altman analysis showed the two devices to be in agreement (at the 95% confidence level), with an average CIEDE2000 difference of 0.3, below the minimum distinguishable difference. The measurement device was always placed perpendicularly to the screen,

focused on its central point, and was controlled by means of the specific manufacturer's software (Spectrawin and CS-S10w respectively). The images were true-color Matlab-generated TIFF files, 2048x1536 pixels in size, and were displayed at the maximum size, using the default image visualization application of the device. The telespectroradiometer provided the spectral radiance in W/sr/m^2 in the 380 nm to 780 nm range, at 4-nm steps. With this information, the XYZ tristimulus values for the 2º CIE1931 standard observer were computed by the software.

Measurements were carried out in a dark room with the screen brightness set to the maximum value (no auto-brightness) and with the devices plugged to the electrical supply, to avoid possible automatic adjustments in brightness as a function of battery charge level. The temporal stability check did not reveal relevant luminance or chromaticity changes with time, but to ensure temporal stability, measurements began 15 minutes after the tablets were turned on.

2.3 Colorimetric Characterization Using 3D Lookup Tables[7-10]

The procedure consists in measuring the tristimulus values of a large number of luminous stimuli generated in an array covering the domain of digital levels, including cases where one, two, or three (R, G, B) channels are active. The image generation criteria, as well as the process to obtain the 3D tristimulus value matrix and the color reproduction errors, has been described in detail in de Fez et al.[2]. We generated 1000 images, corresponding to a 10x10x10 uniform sampling of the space defined by the RGB digital values[7]. Once the tristimulus values of our 1000-color sample were measured, the tristimulus values of any desired color were calculated by interpolation. To reduce computation time for all the image pixels in a given test design, we obtained by cubic interpolation the tristimulus values of the $(2^8)^3$ colors that can be generated by the device and saved them in table format.

The processes of stimulus definition and generation, interpolation of the 3D Lookup Tables and computation of tristimulus values from digital levels and vice versa, have been carried out using our own library of functions for Matlab[11,12].

3. RESULTS

The primary intents of our study were to evaluate the chromaticity constancy of the primaries, luminance additivity and color reproduction errors for each of the three tablets used. In our study of color reproduction errors, we will address both the case of colors generated in a calibrated device (reproduction errors) and colors generated in a device using color characterization data of another standard device (cross-reproduction errors).

3.1 Chromaticity constancy of primaries

From the data obtained in 3D Lookup Tables, we represent in Figure 1a the CIE1931 chromaticity coordinates and luminance of the red, green and blue primaries in isolation, as a function of the digital levels. As digital level increases, the chromaticity of the primaries changes (triangles, squares and circles) along a straight line (dashed line) containing the screen's achromatic stimulus. Therefore, for each digital level, the resulting color is the additive mixture of a low-luminance constant color and the color of the primary at the maximum digital level. The triangle defined by the chromaticities of the primaries at the maximum digital level (continuous line) defines the limits of the device's color gamut.

3.2 Luminance Additivity

153 The results of the luminance additivity check for each device can be seen in Figure 1b.
154 We represent the luminance of achromatic stimuli, obtained by the mixture of the three
155 primaries with equal digital levels between 0 and 255 (continuous line) and the sum of
156 the luminances of each isolated primary (dashed line), as a function of digital level. The
157 average differences between these two measurements are [1.51, 0.66, 0.83] and the
158 differences for the maximum digital level [0.41, 0.15, 0.43] for ipad4, iPadAir, and iPad,
159 respectively. Although the behaviour of the three devices slightly differ, all are additive.
160 For digital level 255, additivity deviations are practically the same in all cases. The
161 average deviation does show an improvement in the newer devices, with values below
162 1%, almost half the value of iPad4.

163

164 3.3 Color reproduction errors in a color characterized device

165 Color reproduction errors were computed with the same set of 100 randomly generated
166 colors for all devices. Experimental tristimulus values were measured as in the
167 characterization process and theoretical tristimulus values were derived from the
168 device's colorimetric profile. Tristimulus values were transformed to CIELab, using as
169 reference white the achromatic stimulus of each device, at maximum luminance, and to
170 the lightness (L*), hue (H*) and chroma (C*) CIELab perceptual descriptors. Color
171 differences, ΔE, between theoretical and experimental L*H*C* values were obtained
172 with the CIEDE2000 formula[13].

173 The color reproduction errors are summarized by the mean values shown in Table 1
174 and the color difference histograms in Figure 2. The contributions of lightness, chroma
175 and hue angle differences (ΔL, ΔC and ΔH, respectively) to the total color difference,
176 ΔE, are represented in Figure 3, and the average values of these magnitudes are
177 shown in Table 1. The ranges of ΔL and ΔC are similar and smaller than those of ΔH in
178 the three devices. There is a single outlier in iPadAir in the hue angle difference

179 distribution (ΔH=-1.1). In iPad, a single color gives ΔH and ΔC very different from the
180 rest of the sample (-2.2 and -1.7 respectively), and this sample also gives the greatest
181 ΔL value (0.6).

182 The Kolmogorov-Smirnov tests indicates that only the ΔL distributions follow the normal
183 distribution (P<.05). Therefore, devices were compared by means of one-way ANOVA
184 with Tamhane's post-hoc –the hypothesis of equal variances was not assumed- for ΔL
185 and the Kruskal-Wallis test for the remaining variables, employing the Mann-Whitney's
186 U test as a post-hoc. For each variable Δx, the p-values cited in the text follow the
187 notation $P_{\Delta x}$ when referring to the result of the tests for 3 samples, and $P_{\Delta x(A,B)}$ for the
188 comparisons between samples A and B. Correlation between devices was checked
189 using Pearson's correlation coefficient in the normal samples, and Spearman's rho for
190 the rest.

191 The ΔE, ΔL and ΔC color differences indicate that, for equal digital values, the colors
192 displayed by the iPad4 are different from those displayed by both the iPadAir and the
193 iPad ($P_{\Delta E(iPadAir,iPad)}$<.001), and that the characterization model of iPad4 tends to
194 overestimate lightness and chroma. The differences between the colors reproduced by
195 iPadAir and iPad are not statistically significant ($P_{\Delta E(iPad,iPadAir)}$=.69, $P_{\Delta L(iPad,iPadAir)}$=.15 and
196 $P_{\Delta C(iPad,iPadAir)}$=.11). Although lightness and chroma may be overestimated in some
197 samples and underestimated in others, the magnitude of these errors are smaller than
198 those of iPad4 (P<.001 in all cases). In all devices, the reproduction errors with the
199 largest variation ranges are those of the hue angle, which do not follow any definite
200 trend.

201 Correlation coefficients between devices are always lower than 0.52, indicating that the
202 relationships between the color differences obtained in the three devices are not
203 strong, though not due to chance (P<.05). The correlation between the ΔE values of
204 iPad4 and iPad is significant but weak (Spearman's rho=0.291, P<.01) whereas those

of iPadAir are moderately but significantly correlated with those of iPad (Spearman's rho=0.419, *P*<.01). The ΔL values of iPadAir are moderately correlated with those of iPad (Pearson's coefficient=0.517, *P*<.01) and more weakly correlated with those of iPad4 (Pearson's coefficient=0.387, *P*<.01). ΔC values show moderate but significant correlation (Spearman's rho=0.472, p<0.01) between iPad and iPadAir. The correlation between iPad and iPad4 is negative and weak (Spearman's rho=-0.258, *P*=.01). In spite of the large variability of ΔH, the correlation between iPad and iPadAir is moderate and significant (Spearman's rho=0.480, *P*<.01).

3.4 Color-reproduction errors when using the characterization model of a similar device

In the previous section, we have used with each device the best model that we could derive from the color characterization data of each individual Retina screen. One of our aims is to test whether a color vision test specifically created using color characterization data from a particular device can be used in another. That is, if we have computed the RGB values that yield the desired color for a given test in a particular device, can these values be used in another device with negligible error? To this end, we analyse the cross-reproduction errors, that is, the color differences (ΔE, ΔL, ΔC, ΔH) between the colors actually measured in a given device (M) and the theoretical predictions based on the 3D Lookup Table characterization of another device (T). Figure 4 shows the errors induced in each device when the characterizations of the other two are used. This makes a total of 6 cross-reproduction error samples.

We have obtained the tristimulus values corresponding to the digital levels values of the testing colors, using the 3D Lookup Tables derived for iPadAir and iPad, and computed the color differences with the triestimulus values of the colors generated in iPad4. As can be seen in Figure 4, the largest contributions to the total color difference

231 ΔE are in both cases due to the cross-reproduction errors in lightness and hue, where
232 both positive and negative values can be found. The prediction derived from iPad data
233 tends to underestimate lightness. In general, the value of chroma is underestimated.

234 If we repeat the process with the colors measured on the iPad and the predictions
235 derived from iPad4 and iPadAir characterization data, again we find that the largest
236 contributions to ΔE are due to lightness and hue. The two theoretical predictions tend
237 to greatly overestimate lightness, while for hue both larger and smaller values than the
238 experimental one are found. Both predictions tend to overestimate chroma.

239 Finally, comparing the experimental measurements with iPadAir and the theoretical
240 predictions made from iPad4 and iPad data, again the largest contributions to ΔE are
241 those of lightness and hue, which show both positive and negative values. iPad tends
242 to underestimate in a greater degree the value of lightness. Chroma is overestimated
243 by iPad4 in the majority of samples, whereas iPad tends to underestimate it.

244 The range width for cross-reproduction errors, ΔE, is around 4-6 units. Our results
245 imply that the color reproduction errors incurred by estimating the tristimulus values
246 generated in a device using a model derived from data of another device are much
247 greater than when the colorimetric characterization model for that device is used.

248 Not all the distribution of color differences are normal, so differences between the
249 cross-reproduction errors obtained with all the possible M-T pairs were assessed with
250 the Kruskal-Wallis test. We obtained a p-value under 0.05 in the four (ΔE, ΔL, ΔC, ΔH)
251 parameters, indicating statistically significant differences between the six M-T
252 combinations tested. To determine between which pairs of M-T combinations the
253 differences are significant, we compared them using the Mann-Whitney U test.

254 The global cross-reproduction differences ΔE, significantly differ between M-T pairs
255 (P<.05). Therefore, the errors associated to using stimuli derived from the colorimetric
256 characterization of a given device in another they were not designed for, change

257 depending on the pair of devices involved. Only the differences between the iPad4 (M)-
258 iPad (T) and iPad (M)-iPadAir (T) pairs (*P*=.22) and between the iPadAir (M)-iPad(T)
259 and iPad (M)-iPadAir (T) (*P*=.23), are not significant.

260 All the ΔL pairs compared showed significant differences (*P*<.05). For ΔC, only 2:iPad4
261 (M)-iPadAir (T) is not significantly different from 6:iPadAir (M)-iPad17 (T) (*P*=.23).

262 The ΔH cross-reproduction errors are the most irregular of all, due to the high
263 dispersion of the results. In more than half the comparisons of M-T pairs, there are no
264 significant differences between samples (*P*>.05). In all these cases, the iPad4 device
265 was involved, either as M or as T. Let us remember that, in the previous section, this
266 device was shown to be significantly worse than the other two.

267 The analysis of correlation between samples, carried out with the procedure described
268 in the previous section, shows that correlations between the results obtained with the
269 six M-T combinations are weak, though not due to chance.

270

271 **4. DISCUSSION**

272 The analysis of the chromaticity and additivity of the primaries shows a similar
273 behaviour of the three devices chosen. The three are reasonably additive, have a
274 similar color gamut (Figure 1a) and the greater differences seem to arise from the
275 luminances vs. digital level curves (Figure 1b) although in Figure 1a the differences in
276 the rate of change in the chromaticity of the primaries with digital level are also evident.
277 Designers work to improve the device's brightness, and this can clearly be seen in the
278 higher luminance values yielded by iPad. As greater brightness is achieved, the
279 chromaticity of the point with zero digital levels and the rate of chromaticity change with
280 digital level change as shown in Figure 1a.

281 The best characterization procedure for the three devices are 3D Lookup Tables,
282 although the color reproduction errors associated with this model significantly differ
283 between devices. Average differences are always smaller in iPadAir and iPad, and
284 therefore these devices reproduce better the sample's lightness, chroma and hue. This
285 is confirmed by the statistical analysis, and therefore we may conclude that the two
286 more recent devices exhibit significantly smaller reproduction errors, about 0.3 units in
287 average, below the color discrimination capabilities of the human visual system. In
288 iPad4, color reproduction errors are close to the minimum distinguishable difference of
289 one CIEDE2000 unit[2,14].

290 All the differences found between the three devices are consistent with the
291 considerable change in screen technology during the four year gap separating our
292 oldest and newest devices. The two telespectroradiometers used in the experimental
293 setup can only justify an error in color reproduction (0.3) far below the color differences
294 found (1 unit between oldest and newest devices). .

295 In spite of the similarity between devices, the analysis of cross-reproduction errors
296 show that the colorimetric characterization data of a device cannot be used to predict
297 color in another with an acceptable error. The majority of the ΔE values are greater
298 than the minimum perceptible difference for a human being[2,14] and are basically due to
299 errors in lightness and hue predictions. Given the different luminance ranges of the
300 three devices shown in Figure 1b, the large contribution of lightness to the total
301 reproduction errors are not surprising. The smallest differences were obtained with the
302 combination iPad4-iPadAir and the largest with iPad4-iPad, ranging from ΔE_{min}=0.6 and
303 ΔE_{max}=8.1. Given that the iPadAir and iPad were the two devices whose models
304 yielded the smallest color reproduction errors, we expected that the cross-reproduction
305 errors involving any of these two devices would not be the worst among the six
306 combinations tested. But this is not what has happened, confirming our initial idea that
307 using a single colorimetric design for all devices is not feasible.

No significant trends were found for ΔL, ΔC and ΔH, since both positive and negative values are present in the sample. We can infer from the results that the major color differences may be due to reproductions with both higher and lower luminance and chroma. Due to the large variability range of ΔH, no significant trends have been identified.

To sum up, color reproduction errors due to displaying in a color reproduction device the colorimetric design optimized for a different display are considerably bigger than the minimum perceptible difference[2,14] of the average human observer. They are also considerable larger than the color reproduction errors associated to the optimal model of each device. For instance, Lookup Tables will yield average reproduction errors of about 2 units for Cathode Ray Tube[9,15], Liquid Cristal Display[9,15] and In-Plane Switching displays[2]. For these reasons, the use of a generic test design for all devices is unadvisable. Each individual device should be characterized, and the necessary stimuli for any vision test implemented in this device should be designed using this characterization, for measurements of visual function to be reliable.

One of the main limitations of our study is that we have chromatically characterized the centre of the screen only, although other error sources have been avoided[3], such as time stabilization, directionality and connection to a power source. We had formerly proposed a procedure for a position-dependent screen characterization, using a customizable measurement grid[11,12] but, due to the long time required for data acquisition, we decided to work in a single position. Bodduluri et al.[3], using mini iPads, similar to the ones employed in this study, obtain that the chromaticities in the centre and the periphery are different. This would affect the design of a test covering the whole screen, if only the characterization of the central point is used. Color reproduction errors would be, in such a case, even larger than the ones we have reported. Dain et al.[6] in their comparison between smartphones, do measure color at different screen locations, but their results do not agree with ours, because, even when

recommending Lookup Tables for these devices, they conclude that individual characterizations for each smartphone are not necessary. This is what is usually assumed to happen with Cathode Ray Tube screens, with a power function linking digital levels to luminance, whose parameters where similar even for different manufacturers. Our previous experience[2], corroborated for the present study, is that the same characterization cannot be used, not just for devices developed by different manufacturers, but even for different models of the same device.

Even if the color reproduction devices used are different, our study has similitudes with the work by Lee & Honson[16] on the comparison of the colorimetric design of different versions of printed Ishihara plates. They found chromaticity differences between test editions, which could justify the diagnosis disagreements reported by different authors. Lee & Honson suggest a periodical quality analysis of the plates, which should be discarded when the aging of the materials should have introduced excessive colorimetric changes. However, it must be taken into account other reported sources of diagnosis error[17], linked with the introduction of spurious signals, detectable by visual mechanisms that should have been silenced.

The errors induced when displaying in a device a color stimulus generated with the colorimetric characterization derived for a different device may alter the results of a vision test, resulting in the inability to distinguish between different types of defect (protans from deutans), in false negatives or in false positives. In Figure 5 we show an example of false positives when a plate designed to detect severe protan defects is displayed in a different device. The chromaticity of background and letter have been chosen on the same protanopic confusion line, and achromatic noise has been introduced to silence the patient's achromatic mechanism. We show the appearance of the test when generated in the device it has been designed for (IPad in Figure 5A) and another device (IPadAir, Figure 5C). The column on the right shows the appearance of both stimuli for a protanopic subject, simulated using the corresponding pair

algorithm[18]. The protanope gives the expected answer when the plate is displayed in the device it has been designed for (Figure 5B, where the letter has vanished), but would pass as normal in the other device (Figure 5D).

Regarding colorimetric studies based on visualization devices comparable with our own, we find that Dain[19] evaluates five iOs apps (only two of which remain active), based on the Ishihara plates, by comparing the colors of the dots on the display and the printed version. The results show that none of the apps was used a specific colorimetric design and that the developers did just scan the plates. For this reason, their applicability does not go beyond curiosity driven self-testing.

Publications experimentally testing the agreement between conventional and electronical versions of a vision test are scarce. Campbell et al.[20] compare the performance of 70 eyes affected by optic neuritis in the printed version of the Ishihara test and an app for iPad (Color vision test HD, no longer available). The app's developers do not provide information either about the plate's colorimetric design or about visualization conditions in the devices of different size and make of their potential users. Even if Campbell and co-workers conclude that the agreement between devices is good, this is true only for the subjects that do not fail the test, but is debatable for the five patients that do present abnormal color vision.

The paper by Kosikowski et al.[21] analyses their own app for Apple devices (iPad, iPod and iPhone), consisting of Ishihara plates and a contrast discrimination test. Even if the authors affirm to have previously characterized several devices of the same type, they simply mention that these devices have reproducible and similar parameters. Their design considers only the necessary size changes to make the plates fit the different screens used in their study. Their validation data, however, come from whatever visualization device app's users around the world would employ. Although they report a large mass of testing results, their data have not been obtained under uniform and

388 controlled experimental conditions and the reliability of their tests has not been studied
389 by comparing their diagnosis with a trustworthy reference test.

390 Studies on vision test design for PCs are more numerous. The viability of the use of
391 Cathode Ray Tube, Liquid Cristal Display or Thin Film Transistor screens have been
392 studied[22,23]. Pardo and co-workers[22], after colorimetrically characterizing Thin Film
393 Transistor-Liquid Cristal Display screens, concluded they were valid for color vision
394 research and diagnosis, and Marey et al.[23], using a scanned version of the Ishihara test
395 in monitors calibrated by a procedure of their own design, finds that the specificity and
396 sensitivity of the electronic and printed versions of the test are similar.

397 Algorithms simulating dichromatic color perception have also been developed, to
398 illustrate the kind of color-coded information losses and the difficulties for object
399 detection suffered by dichromatic subjects[18,24-27]. Different more or less successful
400 algorithms have also been proposed to modify the color palette of an image in order to
401 allow dichromatic subjects to perceive them with minimum information loss[28,29]. These
402 algorithms are intended for use on a computer, for ease in programming, and the
403 resulting images may be visualized in a mobile device.

404 To ensure a reliable result when exporting to a mobile device these design and
405 reproduction processes, the colorimetric characterization of the final visualization
406 device must be known. Ideally, the colorimetric characterization should be carried out
407 for each pixel in the screen, since screens are not spatially homogeneous. Our results
408 show that cross-reproduction errors, induced when forgetting that the digital levels
409 values that yield the desired image in a given device will not produce the same result in
410 a different device, may be large.

411

412 4. CONCLUSIONS

The new generation iPad devices incorporate a Retina screen with good color reproduction characteristics. This would make them apt for vision testing, provided the test has been correctly adapted to the device, taking into account both the spatial and colorimetric characterization of the screen.

Any app for visual testing should inform the user in clear terms about whether it is intended as an entertainment or professional medical app/tool. The vision specialist is the only professional who is qualified to value whether the device and the test meet the requirements necessary to yield a reliable result.

Even if we start from a correctly designed vision test, our results show that the stimulus digital levels levels computed for a given device will yield different chromaticities when reproduced in another. The distortion of the chromatic content may lead to subjects with vision defects to pass as normal or vice versa, compromising diagnosis reliability.

APPENDIX

Applications consulted based on vision tests (Google Play and Apple Store). In each one it has been indicated if they have special tests for color vision and spatial vision.

5. REFERENCES

1- Dain SJ. Clinical Colour Vision Tests. Clin Exp Optom 2004;87:276-93.

2- de Fez MD, Luque MJ, García-Domene MC, et al. Colorimetric Characterization of Mobile Devices for Vision Applications. Optometry Vision Sci 2016;93:85-93.

3- Bodduluri L, Boon MY, Dain SJ. Evaluation of Tablet Computers for Visual Function Assessment. Behav Res 2017;49:548–58.

4- Birch J. Colour vision tests: general classification. In: Foster DH. ed. Vision and Visual Dysfunction, vol.7. Inherited and Acquired Colour Vision Deficiencies: Fundamental Aspects and Clinical Studies. Boca Raton, FL: CRC Press; 1991:215-31.

5- Wyszecki G, Stiles WS. Color Science: Concepts and Methods, Quantitative Data and Formulae, 2nd ed. New York: John Wiley & Sons; 2000.

6- Dain SJ, Kwan B, Wong L. Consistency of Color Representation in Smart Phones. J Opt Soc Am (A) 2016;33:A300-5.

7- Thomas JB, Hardeberg JY, Foucherot I, Gouton P. The PLVC Display Color Characterization Model Revisited. Color Res Appl 2008;33:449–60.

8- Jennings E, Holland RD, Lee CC. Error analysis of look-up-table implementations in device-independent color imaging systems. Proceedings SPIE 2170, Device-Independent Color Imaging 98. 1994. Available at: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=956135. Accessed October 25, 2017.

9- Bastani B, Cressman B, Funt B. Calibrated Color Mapping Between Liquid Cristal Display and Cathode Ray Tube Displays: A Case Study. Color Res App 2005;30:438–47.

10- Balasubramanian R. Reducing the Cost of Lookup Table Based Color Transformations. J Imaging Sci Technol 2000;4:321-7.

11- Garcia-Domene MC. Design and Testing of a Perimetric Device of Incremental Threshold for Projection [Doctoral dissertation] University of Alicante, Spain; 2013. Available at: http://rua.ua.es/dspace/handle/10045/35676. Accessed October 25, 2017.

12- García-Domene MC, Luque MJ, de Fez MD. Software: Design of a Perimetric Device of Incremental Threshold for Projection. Property registration 09 / 2015 / 447; University of Alicante, Spain; 2015.

13- Luo MR, Cui G, Rigg B. The Development of the CIE 2000 Colour-Difference Formula: CIEDE2000. Color Res Appl 2001;26:340–50.

14- Day EA, Taplin L, Berns RS. Colorimetric Characterization of a Computer-controlled Liquid Crystal Display. Color Res Appl 2004;29:365Y73.

15- Gibson JE, Fairchild MD. Colorimetric Characterization of Three Computer Displays (LCD and CRT). Munsell Color Science Laboratory Technical Report, January, 2000. Available at: http://www.sgidepot.co.uk/vw/PDFs/Colorimetric_Characterization.pdf. Accessed October 25, 2017.

16- Lee DY, Honson M. Chromatic Variation of Ishihara Diagnostic Plates. Color Res Appl 2003;28:267-76.

17- Miyahara E. Chromaticity Co-ordinates of Ishihara Plates Reveal That Hidden Digit Plates Can Be Read by S-Cones. Clin Exp Optom 2009;92:434–39.

18- Capilla P, Díez-Ajenjo MA, Luque MJ, Malo J. The Corresponding Pair Procedure: a New Approach to Simulation of Dichromatic Colour Perception. J Opt Soc Am (A) 2004;21:176-86.

19- Dain SJ. Colorimetric Evaluation of iPhone Apps for Colour Vision Tests Based on the Ishihara Test. Clin Exp Optom 2016;99:264-73.

20- Campbell TG, Lehn A, Blum S, et al. iPad Colour Vision Apps for Dyscrhomatopsia Screening. J Clin Neurosci 2016;29:92-4.

21- Kosikowski R, Kosikowski L, Odya P, Czyzewski A. Senses-what u see? Vision screening system dedicated for IOS-based devices development and screening results. Proceedings of the International Conference on Signal and Multimedia Applications (SIGMAP); 2011. Available at: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6731297. Accessed October 25, 2017.

22- Pardo PJ, Pérez AL, Suero MI. Validity of TFT-Liquid Cristal Display Displays for Colour Vision Deficiency Research and Diagnosis. Displays 2004;25:159-63.

23- Marey HM, Semary NA, Mandour SS. Ishihara Electronic Color Blindness Test: an Evaluation Study. Ophthalmol Res 2015;3:67-75.

24- Capilla P, Luque MJ, Díez-Ajenjo MA. Looking for the Dichromatic Version of a Colour Vision Model. J Opt A: Pure Appl Opt 2004;6:906-19.

25- Brettel H, Vienot F, Mollon JD. Computerized Simulation of Color Appearance for Dichromats. J Opt Soc Am (A) 1997;14:2647–55.

26- Luque MJ, de Fez MD, Acevedo P. Software for Simulating Dichromatic Perception of Video Streams. Color Res Appl 2014;39:486-91.

27- de Fez MD, Acevedo P, Garcia C. A Library for Matlab to Simulate Video as Perceived by a Dichromatic Subject (Colorblind). Property registration 09 / 2012 / 2110, Universidad de Alicante, Spain; 2012. Available at: http://hdl.handle.net/10045/23471. Accessed October 25, 2017.

28- Jefferson L, Harvey R. An Interface to support color blind computer users, Proceedings Conference on Human Factors in Computing Systems, CHI 2007, California, USA, 2007. Available at: https://dl.acm.org/citation.cfm?id=1240855. October 25, 2017.

29- Dougherty B, Wade A. Daltonize: Color blind image correction (Vischeck). Available at: http://www.vischeck.com/daltonize/. Accessed October 25, 2017.

FIGURE LEGENDS

Figure 1: a) Variations of the chromaticities of the device's primaries as a function of DAC values (Digital to Analog Converter) plotted in the CIE1931 chromaticity diagram. The continuous line represents the limits of the color gamut of each device. Circles (iPad4), triangles (iPadAir) and squares (iPad) represent the chromaticities of the primaries at each digital level. b) Luminance (cd/m^2) vs. DAC for achromatic stimuli (White: R=G=B, continuous line) and sum of the luminances of the isolate primaries (R+G+B, dashed line). Magenta: iPad4, Green: iPadAir, Blue: iPad.

Figure 2: Color reproduction errors (ΔE) histograms for a test set of 100 randomly generated colors generated by the three displays, computed by the CIEDE2000 color difference formula. Black bars: iPad4, grey bars: iPadAir, white bars: iPad.

Figure 3: Distributions of lightness (ΔL, black bars), chroma (ΔC, grey bars) and hue angle (ΔH, white bars) differences for the set of 100 randomly generated colors in each device. A: iPad4, B: iPadAir, C: iPad

Figure 4: Histograms for color, lightness, chroma and hue differences (ΔE, ΔL, ΔC, ΔH) between the experimental measurements (M) of colors generated in one device and the theoretical predictions (T) obtained with the 3DLUTs (three dimensional Lookup Table) for the other two devices. A: iPad4(M)-iPad(T). B: iPad4(M)-iPadAir(T). C: iPad(M)-iPad4(T). D: iPad(M)-iPadAir(T). E: iPadAir(M)-iPad4(T). F:iPadAir(M)-iPad(T). Black bars: ΔE, dark gray bars: ΔL, light gray bars: ΔC, white bars: ΔH.

Figure 5. Effect of cross-reproduction errors in a test for detection of severe protan defects. A simple vanishing plate for detection of severe protan defects has been designed using the colorimetric profile of our iPad (A), and then the same image has been presented in iPadAir (C). A color vision model has been used to simulate how a protanopic subject would perceive the test in iPad (B) and iPadAir (D).

TABLES

Table 1: Mean values and range of the color, lightness, chroma and hue differences (ΔE, ΔL, ΔC and ΔH, respectively) for the set of 100 randomly generated colors.

	iPad4	**iPadAir**	**iPad**
ΔE **[min, max]**	1.3 [0.3, 2.8]	0.3 [0.0, 1.1]	0.3 [0.0, 2.6]
ΔL **[min, max]**	0.6 [-0.2, 1.8]	0.0 [-0.6, 0.5]	-0.0 [-0.4, 0.7]
ΔC **[min, max]**	0.5 [-0.8, 1.9]	-0.0 [-0.4, 0.5]	-0.1 [-1.7, 0.4]
ΔH **[min, max]**	0.0 [-2.1, 2.7]	-0.0 [-1.1, 0.5]	-0.0 [-2.2, 0.6]

Figure 1

Click here to download Figure Figure 1.tif

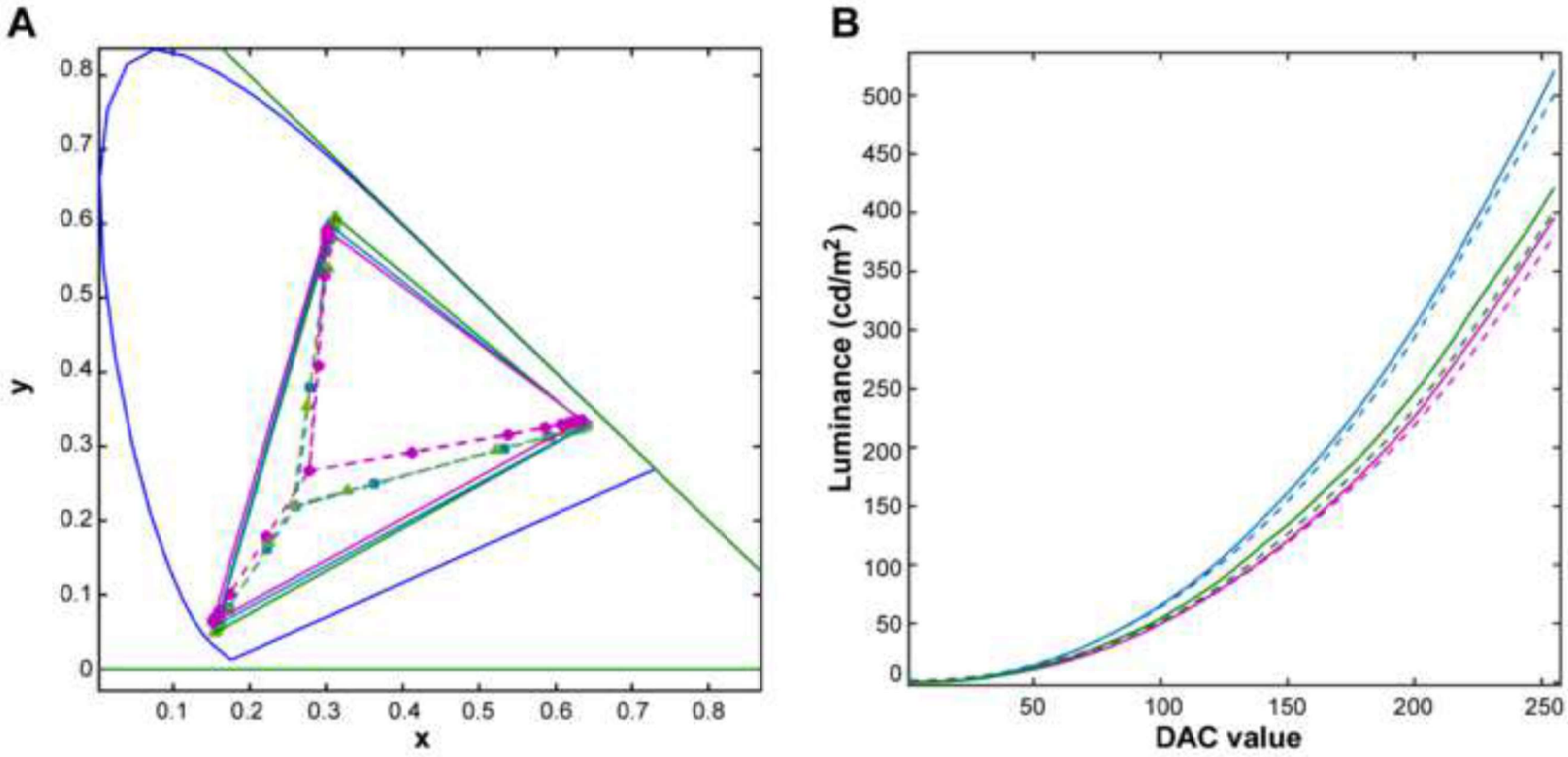


Figure 2

Click here to download Figure Figure 22.tif

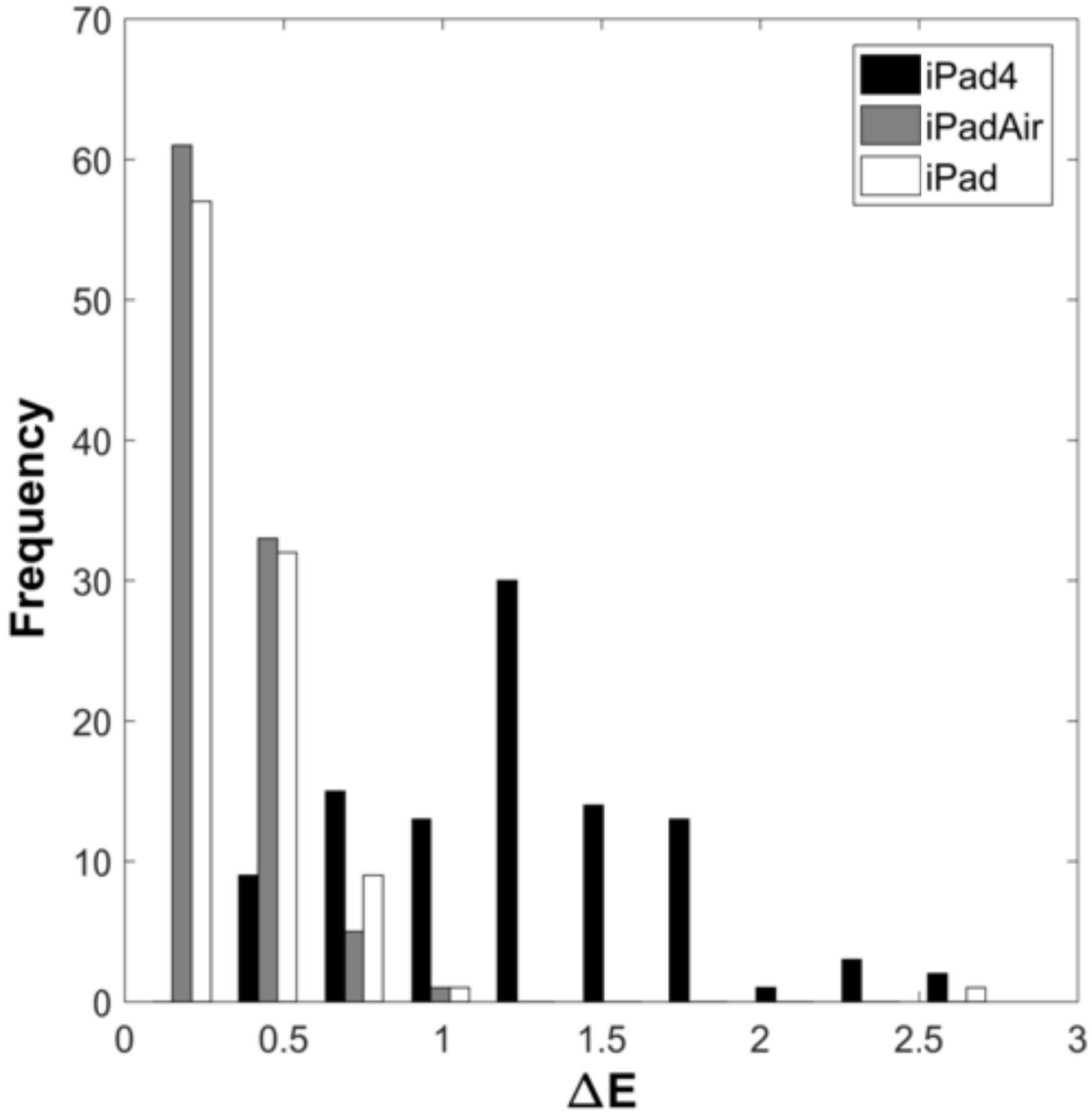


Figure 3

Click here to download Figure Figure 3.tif

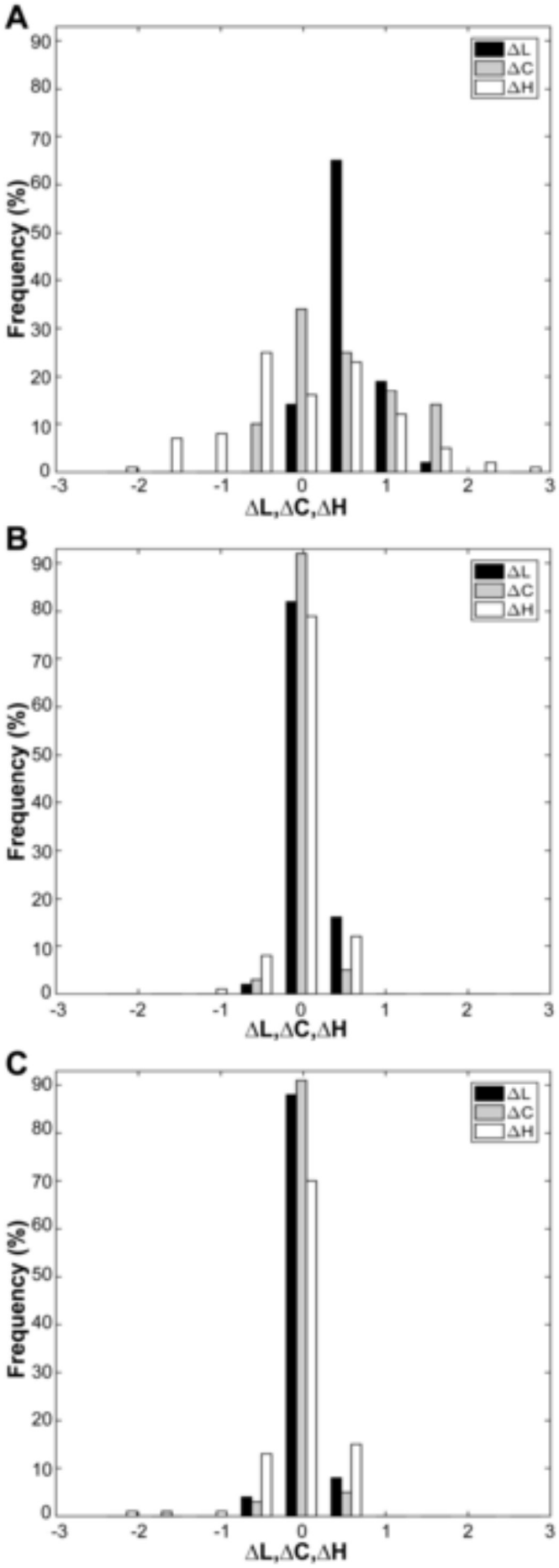


Figure 4

Click here to download Figure Figure 4.tif

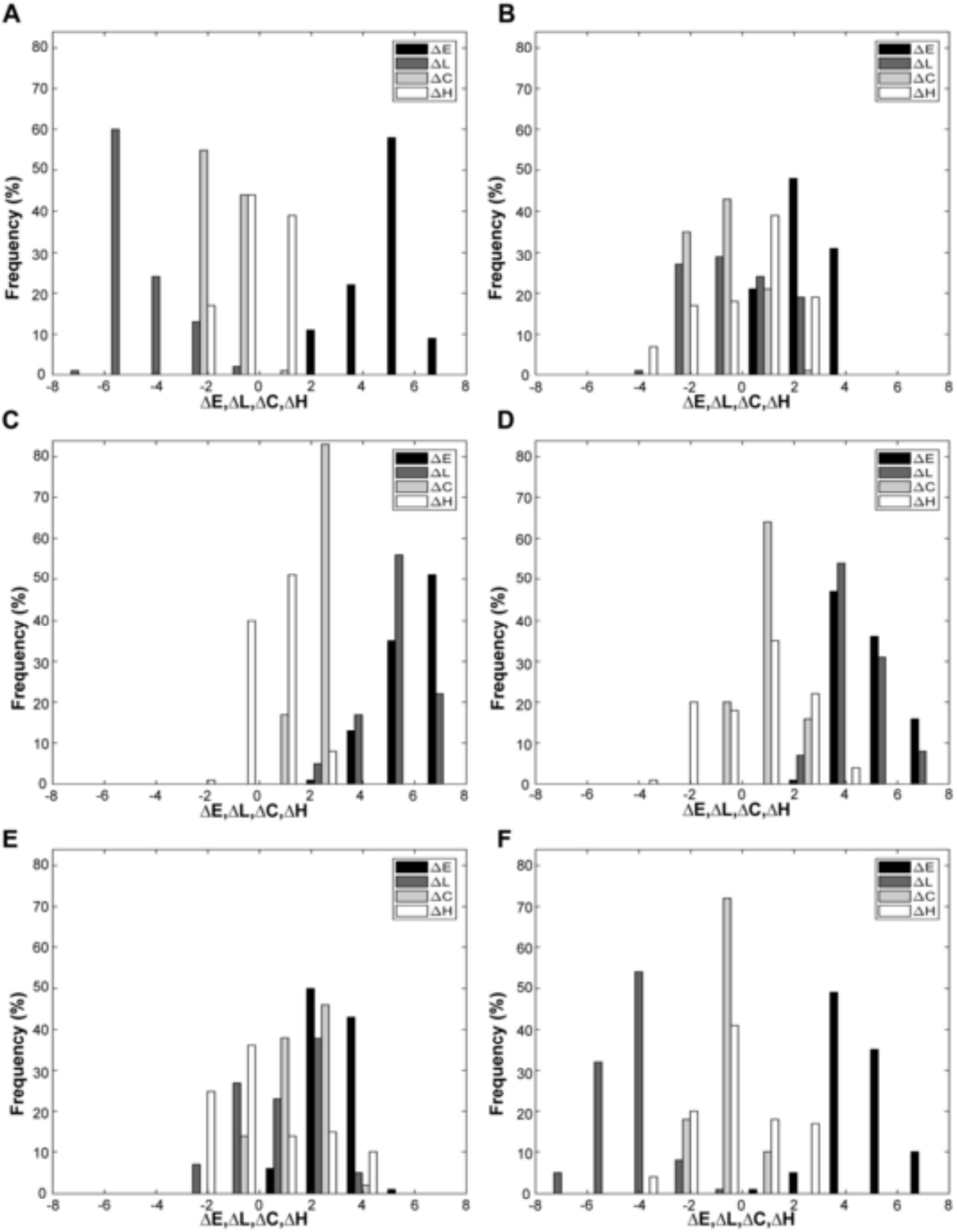
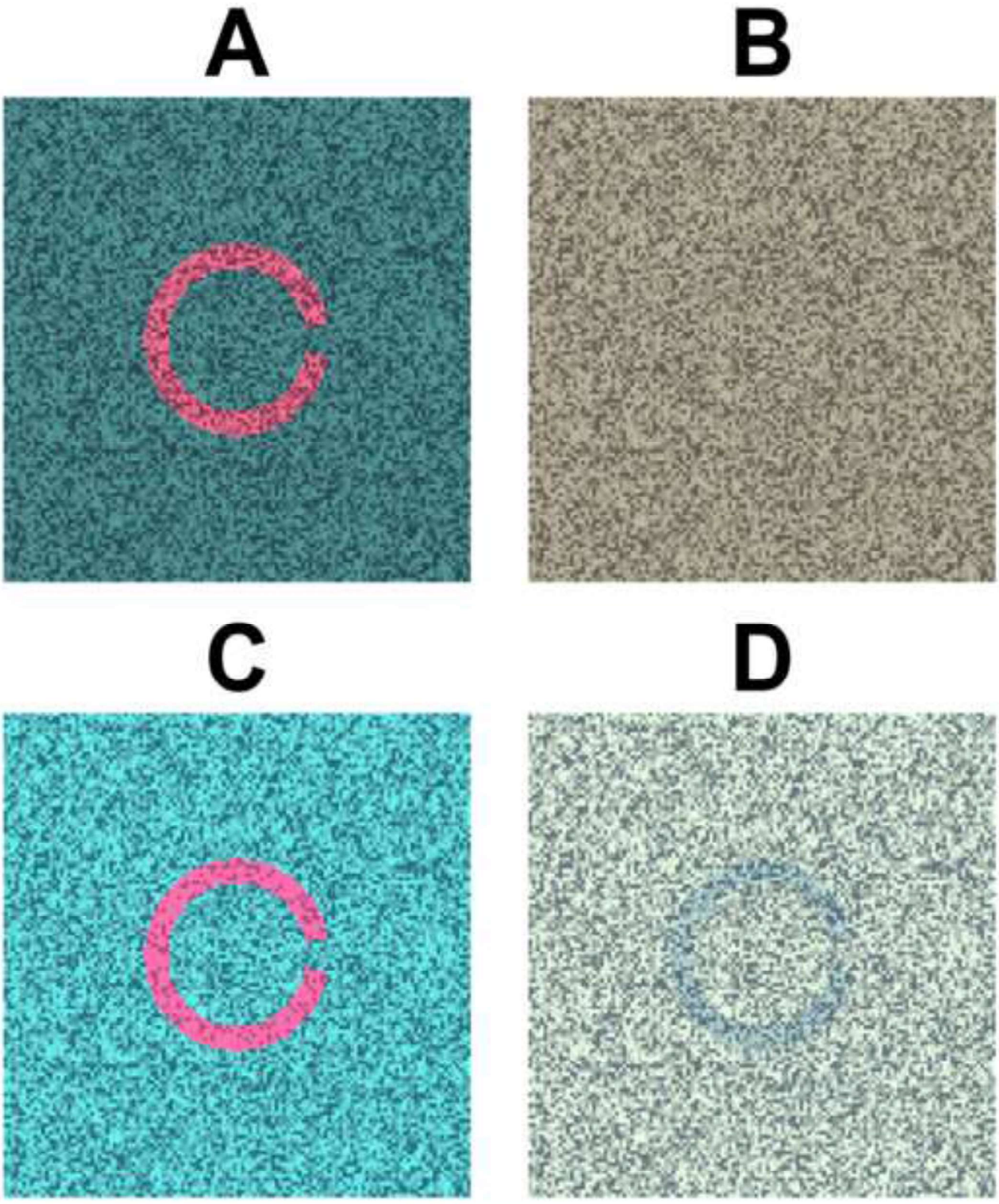


Figure 5

Click here to download Figure Figure 5 new.tif



APPENDIX

Android App	Evaluation	Developer
Central Vision Test	General	healthcare4mobile
Colour blind Tester	Color vision	chachacode
Colour Test	Color vision	App2U
Colour vision test	Color vision	Colour Vision
Contrast Sensitivity Test	Contrast Vision	healthcare4mobile
Examen visual	General	andrew.brusentsov
Eye Care Plus	General	healthcare4mobile
Eye Test	General	Nguyen Phuc Khanh
Eye Test - Eye Exam	General	healthcare4mobile
Eye Test - Ishihara	Color vision	NTIMobile
Eye Test Charts	General	App Park
Eye Test: colour vision	Color vision	Shpand
Icare Vision Test	General	Icare Eye Hospital
Prueba del ojo	General	designveloper
Prueba de daltonismo	Color vision	iGreen Software
Prueba de ojos-cuidado de ojos	General	Icare Fit Studio
Pruebas Visual Completas	General	gfsm
REST Rapid Eye Screening Test	General	Epic Egg Studio
Smart Optometry	General	Smart Optometry
Test Visual	General	Barraquer
Test your vision	Color vision	Y_Novic
Vision Test 2.0	General	Rocktime Ltd
Visual Acuity Test	General	healthcare4mobile

iOs App	Evaluation	Developer
aidColours	Color vision	Tilenus Consultores, SL
Chromatic Glass	Color vision	Kazunori Asada
Chromatic Vision Simulator	Color vision	Kazunori Asada
Clinic CSF	Contrast vision	Manuel Rodriguez Vallejo
Colour Blind	Color vision	Zoom Inc
Colour Blind Pal	Color vision	Vincent Fiorentini
Colour Inspector	Color vision	Aaron L'Heureux
Colour Perception Test	Color vision	David Liu
Colour Vision (for Colour Blindness)	Color vision	Rasmus Barringer
Colour Vision Test Lite	Color vision	Rila Software
Colour Vision Test Pro	Color vision	Linton Intergroup Inc.
Colourblind Avenger	Color vision	Brian Wardle
ColourDeBling	Color vision	Elektron software
ColourDetect	Color vision	Sunset Software Ltd Liab. Co
ColouredEye	Color vision	Sanhita Choudhury

Colour Blind Test	Color vision	Lee Kah Seng
Eye Handbook	General	Cloud Nine Development
EyeChart-Vision Screening App	Spatial vision	Dok LLC
EyeXam	General	Global EyeVentures
How well do you see colour?	Color vision	Sergey Skosyrev
HueVue: Colourblind Tools	Color vision	AppFoundry
Kolorami	Color vision	Comparatel
Kuku Kube- Colour Test	Color vision	Hien Nguyen
Odd Colour	Color vision	VM Mobile Team
PseudoChromatic Colour Test	Color vision	Cassiopeia Information Technologies
Rinnegan	Color vision	Mario Vega
Say Colour	Color vision	HotPaw Productions
Vision Scan Lite	General	Cygnet Infotech LLC
Vision Test	General	Rocktime LTD