Cambridge Colour Test: reproducibility in normal trichromats

T. P. Fernandes,[1,*] N. A. Santos,[1] and G. V. Paramei[2]

[1]Psychology Department, Federal University of Paraiba, Cidade Universitaria S/N, 58051-900, Joao Pessoa, Brazil
[2]Department of Psychology, Liverpool Hope University, Hope Park, L16 9JD Liverpool, UK
*Corresponding author: paivatm@gmail.com

Received 11 October 2019; revised 17 December 2019; accepted 23 December 2019; posted 3 January 2020 (Doc. ID 380306); published 24 February 2020

This study evaluated reproducibility of the Trivector subtest of the Cambridge Colour Test. Data for normal trichromats were obtained in Brazil (*N* = 111) at T0, six months (T1), and 12 months later (T2), and in the United Kingdom (*N* = 79), with the test directly followed by a retest. Coefficients of repeatability—Bland–Altman indices—for Protan, Deutan, and Tritan vectors were similar for both datasets. Intraclass correlation coefficients (ICCs)—measures of reliability—were low or moderate for these relatively homogeneous datasets; for a heterogeneous dataset, comprising color-normal and abnormal observers, ICCs were 0.80–0.98, indicating the high discriminative accuracy of the Trivector subtest.

https://doi.org/10.1364/JOSAA.380306

1. INTRODUCTION

Normal chromatic discrimination is essential for color-vision demanding occupations, in which color-vision assessment can be required for employment, to avoid risk of professional underperformance associated with certain forms of congenital color-vision deficiencies [1]. Screening of color vision is also part of testing batteries in clinical practice; it serves both diagnostic and monitoring purposes, since various malignant conditions are manifested by chromatic discrimination loss, such as ocular and retinal dystrophies, systemic diseases (e.g., diabetes [2]), neurodevelopmental conditions (e.g., Parkinson's disease [3]), and exposure to [4] or consumption of neurotoxic substances [5,6]. To ensure that the diagnosis is correct, a color-vision test requires an evaluation of its reproducibility—*repeatability*, or test–retest agreement, and *reliability*, or sensitivity to differences between observers [7], characteristics outlined in detail below.

A. The Cambridge Colour Test

In the present study, we undertook to evaluate the reproducibility of the Trivector subtest of the Cambridge Colour Test (CCT), a computerized color-vision diagnostic tool developed by Regan, Reffin, and Mollon [8]. The CCT was commercially released by Cambridge Research Systems Ltd. (Rochester, UK) in 2000 [9], and since then has been broadly used in basic research and studies of eye pathology of various etiologies and systemic diseases manifested by acquired color-vision abnormalities. In clinical practice, as an auxiliary tool, the CCT measures are used as indicators of the onset and development of pathologies, reinforcing the diagnosis and monitoring the disease progression or its reversal following therapeutic intervention. We could identify about 50 studies using the CCT since its development [8] as we show in [10], listing basic and applied investigations with different research foci and reflecting the geographic spread of the studies.

The CCT implements a design that employs the advantages of pseudoisochromatic plates combining the principles of Chibret and Stilling [11]. In particular, the Chibret principle is realized by variation of the chromatic difference between the figure, Landolt "C", and the background, while the Stilling principle is implemented as non-hue noise achieved by variation in luminance and size of composing elements. The opening of the Landolt "C" can have four orientations defined by a chromatic contrast (defined in the CIE 1976 *u'v'* units) superimposed on a textured gray background [Fig. 1(A)]. As a computer-controlled test, the CCT allows precise control of variation of chromaticity parameters of the figure and background and multiple randomized presentations of the figure–background chromaticity differences. The observer's task is to identify the orientation of the gap, where the response is based on chromatic cues only. By means of implemented staircase procedure, the chromaticity of the "C" and, hence, the figure–background chromatic contrast varies adaptively depending on the observer's responses [Fig. 1(B)]. The Trivector subtest estimates discrimination thresholds along the Protan, Deutan, and Tritan confusion lines, resulting in three parameters that reflect sensitivity of the long (L), middle (M), and short (S) wavelength-sensitive cones, respectively [Fig. 1(C)] [3,8,9]. The method allows a rapid estimation of chromatic discrimination thresholds and provides a quantitative outcome that is sensitive to individual differences among normal trichromats (NTs) [9,12,13]. Crucially, it enables identifying observers with congenital and acquired color-vision abnormalities, and it discriminates between the types and severity of dyschromatopsias [2,3,6,8,9,14].

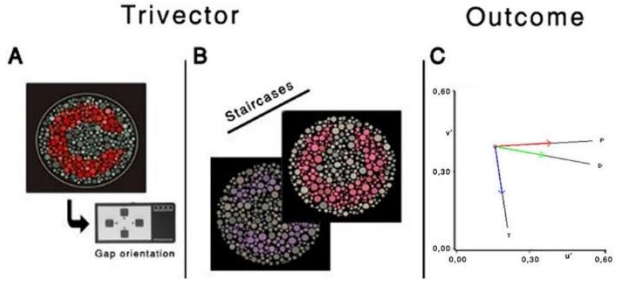

Fig. 1. Illustration of the chromatic targets (A, B), Landolt "C", embedded in the luminance noise background (image source: Cambridge Colour Test Handbook [9], p. 4). Permission has been obtained from Prof. John D. Mollon, who holds the copyright of the CCT. (C) Protan (P), Deutan (D), and Tritan (T) vectors in the CIE 1976 *u'v'* chromaticity diagram.

B. Repeatability of Test Measurements

Repeatability is test–retest agreement measured as variation of repeated measurements made by the same instrument and the same observer, under identical conditions, obtained over a short enough period of time that the underlying value can be considered constant. The test repeatability is used for evaluative purposes to reflect the precision of the measurement obtained on well-separated occasions [15] (sometimes also referred to as "reliability" [7,15,16,17]). Variability of the repeated measurements is attributed to within-participant variation (e.g., diurnal variation, fatigue, etc.) as well as to instrumental variation (errors inherent in the measurement method) and variation in the experimenter [16].

Test repeatability is assessed using Bland–Altman analysis [18]. This allows test–retest "limits of agreement" to be estimated, a particularly valuable approach when tolerance values are unknown for a new test. The *limits of agreement* (*LoA*s) are visualized by a Bland–Altman plot, where the *x* axis shows the means of test (t) and retest (r) measurements [$(t + r)/2$] for each individual subject and the y axis represents the difference between the two measurements for the same individual ($t - r$).

The following estimates quantify test repeatability [7,15,17,19]:

(i) mean difference ($\bar{X}_D$) and standard deviation (SD_D) of pairwise differences across subjects;
(ii) upper and lower *LoA*s [18,19]:

$$LoA = \bar{X}_D \pm 1.96\, SD_D \quad \textbf{(1)}$$

(iii) 95% confidence intervals (CIs) of $\bar{X}_D$, the precision (or statistical uncertainty) of the measure ([20], p. 2217):

$$95\%\ \text{CIs} = \bar{X}_D \pm t_{(95\%,n-1)} \frac{SD_D}{\sqrt{n}} \quad \textbf{(2)}$$

(iv) 95% confidence intervals (CIs) of upper *LoA* and lower *LoA* using mixed design ([20], p. 2217):

$$95\%\ \text{CIs} = \text{LoA} \pm 1.71\, t_{(95\%,n-1)} \frac{SD_D}{\sqrt{n}} \quad \textbf{(3)}$$

Theoretically, if both samples of measurements were identical and free from error, the mean difference would be zero. In practice, an intra-individual variation is observed between test and retest. The further away $\bar{X}_D$ is from zero, the larger is the test-retest bias.

It is expected that, if the test–retest differences are normally distributed, 95% of these differences will lie between ($\bar{X}_D - SD_D$) and ($\bar{X}_D + SD_D$), i.e., within the upper and lower *LoA*s.

Finally, the *coefficient of repeatability* (*COR*), adopted by the British Standards Institution [21], is the modulus (non-negative value) of upper or lower *LoA* and was defined as:

$$COR = 1.96 \times SD_D \quad \textbf{(4)}$$

The *COR* quantifies measurement error in the same units as the assessment tool (here the length of Protan, Deutan, and Tritan vectors) and reflects the smallest within-subject change in the repeated measurement that could be considered meaningful.

C. Test Reliability Measure

Reliability is the ability of the test to distinguish groups of individual observers from each other due to the inherent variability between them, i.e., reliability parameters are used for discriminative purposes. Conventionally reliability is assessed by the *intraclass correlation coefficient* (*ICC*), the measure introduced by Fleiss and Shrout [22,23]. The *ICC* accounts for both intra-subject consistency of performance from test to retest (the test repeatability) as well as for differences in performance between subjects as a group (the test sensitivity). Formally, the *ICC* is defined in terms of measurement variances. Here we use the two-way random model, or *ICC* (A, 1), according to McGraw and Wong ([24], p. 35):

$$ICC = \frac{SD^2{}_b - SD^2{}_w}{SD^2{}_b + (k-1)SD^2{}_w} \quad \textbf{(5)}$$

where $SD^2{}_b$ is the variance of measurements (here vector length) between subjects, $SD^2{}_w$ is the error variance within subjects (due to test–retest variability), and k is the number of (re)tests, i.e., k = 3 for T0, T1, T2 testing sessions for the Brazil sample; k = 2 for each of the Brazil pairwise comparisons, and for the UK sample. The *ICC* parameter is dimensionless and varies between 1 and –1. Values close to 1 or –1 indicate high intraclass correlation, i.e., imply that the test measurement error is small in relation to the variability between the tested subjects. Conversely, *ICC* values close to 0 indicate that the test measurement error is comparable to variability in the tested sample [7,15,25].

There are various suggestions in the literature on what reliability coefficients can be regarded as "low", "moderate" or "high". Cicchetti [26] recommended that *ICC* values <0.4 are "poor", 0.40–0.59 "fair", 0.60–0.74 "good", and 0.75–1.00 "excellent". The cut-off values are, however, to be considered in relation to study purposes: *ICC* values of 0.60, 0.70, and 0.80 are often used as minimum standards for test reliability at group-level comparisons or for research purposes, but "[i]f individual and important

decisions are made on the basis of reliability estimates, values should be at least 0.90 . . . or 0.95" ([27], p. 667).

2. METHOD

A. Participants

1. Brazil

Healthy NTs (N = 111, 59 males), with no abnormalities revealed in fundoscopic or optical coherence tomographic examination, normal or corrected-to-normal vision (with visual acuity of at least 20/20, assessed binocularly by the Snellen chart), and no self-reported ocular or systemic diseases, participated in the study. Their age ranged between 20–49 years (34.5 ± 8.6 years) with 35 observers in the 20–29 y.o. band, 40 observers in the 30–39 y.o. band, and 36 observers in the 40–49 y.o. band. All participants were screened for color blindness using the 24-plate edition of the Ishihara test [28] and the Lanthony D-15d test (Richmond Products, USA). The tests were presented in a room under a compact daylight fluorescent lamp suspended 60 cm above the test. At the test surface, illuminance was 898 lux (measured by a CR-400 Colorimeter; Konica, Minolta).

2. UK

Healthy NTs (N = 79; 40 males), with self-reported normal or corrected-to-normal vision and no ocular or systemic diseases, were part of a large-scale study of chromatic discrimination across eight life decades [13]. Test–retest data were available for the present subsample. Participants were aged 10–69 years (38.7 ± 23.8 years); the subsample included predominantly adolescents (10–19 y.o.; N = 39) and mature observers (60–69 y.o.; N = 30), with few participants in the intermediate life decades. All participants were screened for color blindness using the 24-plate edition of the Ishihara test [28], the Farnsworth D-15, and the Lanthony D-15d tests (Richmond Products, USA). The tests were presented in a viewing booth under D65-metameric illumination (Just Normlicht Mini 5000; Fa. Color Confidence) suspended 40 cm above the test. At the test surface, luminance was 220 cd/m^2 (measured by a PR-650 SpectaScan Colorimeter; Photo Research, Inc.), corresponding to illuminance of 1387 lux.

Outcomes of the screening tests for both samples [the Ishihara test readings and Color Confusion Indices (CCIs) for the D-15 and D-15d tests] can be seen in the Dataset file [29]. One inclusion criterion was correct performance on the Ishihara test; one to three atypical errors were tolerated [30]. Concerning the D-15 and/or D-15d tests, we excluded data of underperforming observers who revealed multiple transpositions with transposition values of 3 or greater (J. Birch, J. Hovis, personal communication). In one case of a 14-year-old participant (UK sample), with perfect performance on the Ishihara test, we tolerated, however, greater transposition values, since arrangement tests, as well as measuring color discrimination, are affected by general nonverbal intelligence [31].

The exclusion criteria were congenital red–green abnormality estimated during the pre-testing screening, history of ophthalmological pathology (ocular or retinal diseases), diabetes, and neurological diseases, i.e., conditions known to reveal elevated CCT thresholds [2,3]. Also excluded were data of observers who self-reported developing cataract or cataract operation and of those participants who had functionally monocular vision or wore tinted glasses/lenses during testing. Finally, we excluded data of (a few) participants who self-reported smoking more than 30 cigarettes daily or consuming weekly a rather high amount of alcohol, i.e., conditions indicative of excessive substance abuse resulting in elevated CCT thresholds [5,6].

Foreshadowing an additional test reliability analysis, we report characteristics of an extended UK dataset (N = 123) with increased sample heterogeneity. Along with the initial 79 observers, it included data of elderly participants (aged 70–88 y.o.; N = 30), males with congenital red–green abnormality (aged 13–83 y.o.; N = 8), two males with acquired dyschromasias caused by excessive use of alcohol (aged 62 and 77), one male with tobacco abuse (aged 64), and three female carriers of color abnormality, who underperformed on the color-vision tests.

At both locations, the study followed the ethical principles of the Declaration of Helsinki, and written informed consent was obtained from all of the participants.

B. Apparatus

1. Brazil

The Cambridge Colour Test v2.0 was used [Cambridge Research Systems Ltd. (CRS), Rochester, UK]. Implementation and calibration procedures were performed with software and hardware provided by the CRS (OptiCAL; VSG 2/5 display card), which was run on a Precision T3500. Stimuli were presented on a gamma-corrected 19" CRT monitor (LG Electronics Inc., South Korea) with 1024 × 768 pixel resolution and frame rate 100 Hz.

2. UK

The Cambridge Colour Test v1.5 was used (CRS). Implementation and calibration procedures were performed with software and hardware provided by the CRS (OptiCAL; VSG interface version 8.12; graphics card VSG 71.02.01E9). Stimuli were presented on a gamma-corrected 21" CRT monitor (Mitsubishi Diamond Pro 2070SB, Japan) with 1024 × 768 pixel resolution and frame rate of 100 Hz.

C. Stimuli

The CCT stimulus is a pattern composed of distributed small circles randomly varying in size (between 2.8° arcmin and 5.7° arcmin in diameter), and luminance (8, 10, 12, 14, 16, and 18 cd/m2). The achromatic background is specified by u' = 0.1977, v' = 0.4689 (CIE 1976 chromaticity diagram). The target, a Landolt "C", is defined by a superimposed chromatic contrast [Fig. 1(A)]; the "C" opening has one of four orientations: top, bottom, left, right. In the Brazil setting, the "C" had an opening of 1.25° of visual angle at a 3 meter viewing distance; in the UK setting, the "C" gap subtended 1° of visual angle at a 4 meter viewing distance.

In randomized presentations, the "C" chromaticity varies, differing from the background by a minimum excursion of 0.002 *u'v'* units and maximum excursion of 0.011 *u'v'* units [9]. To infer the embedded shape and identify the gap position, the participant cannot use spatial or luminance cues and is hence forced to use solely chromatic cues.

In the Trivector subtest, that estimates discrimination thresholds for L, M, and S cones, chromatic contrast varies along three confusion lines: the Protan (copunctal point u' = 0.678, v' = 0.501), Deutan (copunctal point u' = –1.217, v' = 0.782), and Tritan (copunctal point u' = 0.257, v' = 0.0), respectively.

D. Procedure

In both the Brazil and UK experimental settings, participants were dark adapted at least 10 min, the temporal window ensuring dark adaptation of cones [32], and were tested binocularly. In both labs, participants were instructed to identify the orientation of the Landolt "C" gap—presented randomized in one of the four positions (four-alternative forced choice; 4-AFC)—and to press the corresponding button of the response box (CT6, CRS). They were also instructed to press any button in the cases in which they were unable to see the Landolt "C" and/or its opening. Accuracy over speed was emphasized in the instruction. The response box was held by the participant with both hands, and the thumbs were used for button pressing. The time allowed for observers to respond was 6 s (Brazil) and 8 s (UK). The Trivector subtest was completed by a participant within 3–5 min.

Chromatic contrast of the Landolt "C" was varied relative to that of the background using an adaptive staircase procedure [Fig. 1(B)]. For each of the three confusion lines, the CCT algorithm implements two interleaved staircases presented in a random order using a weighted one up/one down staircase rule, with a ratio of 1/3 to converge on the 75% threshold. Each staircase begins with a target of high saturation (maximum vector length within the monitor color gamut) and then varies between maximum and minimum saturation (0.11 and 0.01, respectively, in CIE 1976 *u'v'* units). Accordingly, the chromaticity of the target is varied to reduce the contrast with the background. In each staircase, the step size and direction of the variation in chromatic contrast is contingent upon the observer's response, specifically: chromatic contrast is halved (24%) after a correct response and doubled (48%) following an incorrect response or no response (within the allocated response time), until the first reversal, and 8% for the remaining reversals. Periodically, a control target at maximum saturation is presented; such catch trials constitute approximately 10% of the stimuli. The test stops after six staircase reversals for each vector; the chromatic discrimination threshold (in *u'v'* units) is computed as the average of the chromaticities corresponding to the six reversals [9].

The Brazil sample was tested at T0, six months later (T1), and 12 months later (T2); in the UK sample, the test was directly followed by retest.

E. Statistical Analysis

For each dataset, distribution of data was assessed, including measures of central tendency: mean (*M*), median (*Med*), measures of dispersion [standard deviation (*SD*), variance (*Var*), semi-interquartile range (*sIQR*)], and measures of skewness and kurtosis. In addition, we calculated coefficient of variation (*CV*), a measure of relative dispersion, as $CV = SD/M$ [33]. Statistical analysis was performed using SPSS 25.0, MedCalc 15.8 (medcalc.be), and MATLAB R2018b (https://www.mathworks.com).

To explore a test–retest bias, related-samples tests of difference were performed. In particular, the Friedman test was used for the Brazil dataset of the three measurements with Bonferroni correction; the Wilcoxon signed-rank test was used for test–retest measurements of the UK dataset.

To assess the agreement between the repeated measures for T0-T1, T0-T2, and T1-T2 (Brazil) and test–retest (UK) datasets, Bland–Altman analysis was conducted. In particular, for each test–retest dataset and the vector (*V*) in question, the following parameters were calculated for each pair of measurements: test–retest mean for individual participants, grand test–retest mean ($\bar{X}_V$) and its standard deviation (SD_V); individual test–retest differences (D); mean $\bar{X}_D$ and SD_D; upper and lower limits of agreement [*LoA*s; see Eq. (1)], and 95% CIs of $\bar{X}_D$ [see Eq. (2)], upper *LoA* and lower *LoA* [see Eq. (3)]. These outcomes were also presented graphically as Bland–Altman plots. In addition, the *COR* was calculated for each vector and each pair of measurements, for the Brazil and UK datasets, according to Eq. (4). In reporting the outcomes of the Bland–Altman analysis, we followed the guiding principles of Abu-Arafeh *et al.* [34].

Finally, according to Eq. (5) (cf. [24]; Table 7, p. 42), the *ICC*, a measure of test reliability that relates the variance of inter-participant difference in a sample to variance of intra-participant differences (or measurement error), was estimated for each test–retest pairing and, for the Brazil dataset, across the three measurements T0, T1, and T2. We opted to use the absolute agreement *ICC* formula (appropriate for the repeated-measures design), and in it, the mean of "*k*" raters (since bias can be observed when a "single-rater" type is used in the repeated-measures design).

3. RESULTS

A. Descriptive Statistics

Figure 2 shows chromatic discrimination thresholds (10^{-4} *u'v'* units) for individual NTs along the Protan (top), Deutan (middle), and Tritan (bottom) confusion lines, at each testing time T0, T1, and T2 (Brazil) and at test and retest (UK) (Fig. 2, left and right, respectively). Within each sample, the distribution of data across the testing times appears to be reasonably stable.

Detailed descriptive statistics of the Trivector measures for each dataset is presented in Table 1. Mean values for Protan and Deutan vectors for test and retest(s) are comparable for the two samples, while mean Tritan vector values for the UK sample are higher, as is expected for a sample that included a substantial number of mature observers in their sixties with age-related crystalline lens changes (cf. [13,35,36]). For normative Trivector data for individual life decades, readers are referred to [13] (Table 2, p. A377). In the Brazil data, dispersion measures are

relatively low compared to the UK measures, which also show greater skewness (see Table 1). Greater dispersion in the UK outcomes probably reflects greater age range in this sample. Notably, despite differences in means and SDs, coefficients of variation (*CV*), i.e., relative data dispersion [33], are comparable for the Brazil and UK samples.

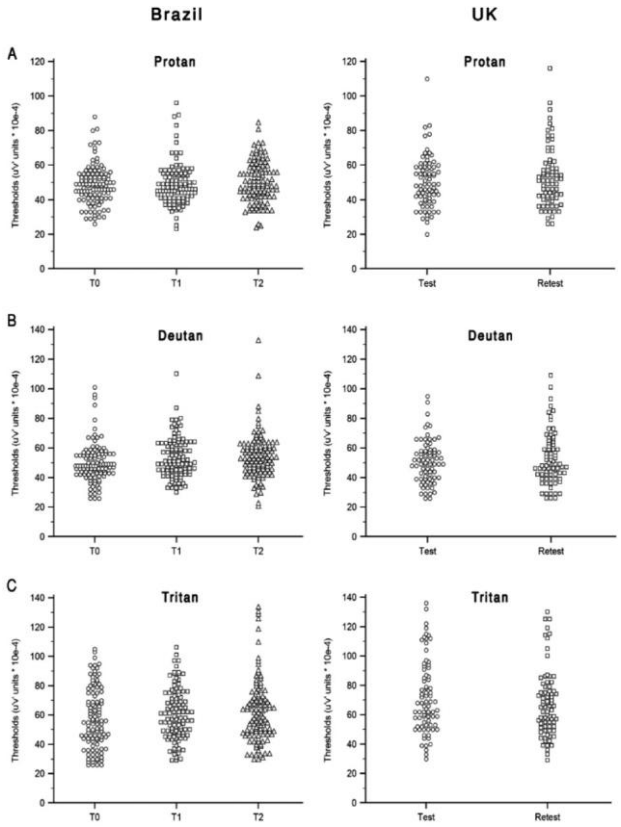

Fig. 2. Chromatic discrimination thresholds (10^{-4} *u'v'* units) of individual NT participants along the (A) Protan, (B) Deutan, and (C) Tritan vectors for T0, T1, and T2 (Brazil, left) and test and retest datasets (UK, right).

To investigate the possibility of a temporal bias, we conducted a Friedman test (T0, T1, and T2; Brazil) and a Wilcoxon test (test and retest; UK). The Friedman test showed no differences across T0, T1, and T2 measurements for either Protan [$\chi^2(2) = 1.47$, $p = 0.23$], Deutan [$\chi^2(2) = 2.46$, $p = 0.09$], or Tritan vectors [$\chi^2(2) = 1.70$, $p = 0.18$]. Also the Wilcoxon test showed no differences between test and retest measurements for Protan ($Z = -1.03$, $p = 0.30$), Deutan ($Z = -0.88$, $p = 0.38$), or Tritan vectors ($Z = 1.47$, $p = 0.14$).

B. Repeatability of the Trivector Subtest

Repeatability analysis was conducted using Bland–Altman outcomes. Individual participants' Trivector outcomes at T0, T1, and T2 (Brazil) and test and retest (UK) can be found in the Trivector Reproducibility file [29]. Plots for Protan (Fig. 3), Deutan (Fig. 4), and Tritan (Fig. 5) measurements are presented for all datasets. Numeric outcomes of the Bland–Altman analysis are presented in Table 2. The results show that mean test–retest differences ($\bar{X}_D$) deviated only slightly from zero; specifically, no deterioration or systematic improvement due to the learning effect is apparent and thus good repeatability of outcomes is indicated not only for the immediate test–retest (UK) but also six months or one year later (Brazil). For different pairwise comparisons, $\bar{X}_D$ for the Protan vector varies between −0.20 and −1.90 (Brazil) and is −2.38 (UK); for the Deutan vector, it varies between −1.18 and −4.64 (Brazil) and is −1.86 (UK); and for the Tritan vector, $\bar{X}_D$ varies between −2.09 and −4.48 (Brazil) and is 4.38 (UK).

Table 1. Descriptive Statistics of Trivector Measures (10^{-4} *u'v'* Units) for the Brazil and UK Samples of Normal Trichromats (NTs) and the UK Extended Sample Including Color-Vision Deficient Observers (CVDs)[a]

	Brazil (N=111) NTs			**UK** (N=79) NTs		**UK** (N=123) NTs & CVDs	
Test sess.	**T0**	**T1**	**T2**	**Test**	**Retest**	**Test**	**Retest**
Protan							
M	48.6	48.8	50.5	50.3	52.7	80.1	77.3
SD	11.4	12.0	120.	14.8	17.2	131.9	120.1
Var	130	144	141	219	296	17396	14420
CV	0.24	0.25	0.23	0.29	0.33	1.64	1.55
Med	48	46	49	49	51	54	53
sIQR	7	7	8.5	9.5	9.5	11.5	13.5
Skw	0.71	1.39	0.27	0.93	1.10	5.84	6.53
Krt	1.23	3.42	0.60	2.40	1.62	35.86	48.17
Deutan							
M	50	53.5	54.7	50.8	52.7	75.7	70.4
SD	13.3	13.5	15.1	14.3	17.6	103.4	68.3
Var	178	180	228	204	310	10703	4662
CV	0.27	0.25	0.28	0.28	0.33	1.36	0.97
Med	48	51	53	50	49	53	54
sIQR	6.5	9.5	8.5	9.5	10.5	12	13.5
Skw	1.30	0.92	1.59	0.64	0.92	5.13	3.97
Krt	3.29	1.74	6.92	0.67	0.84	27.06	16.46
Tritan							
M	58.1	60.2	62.6	72	67.7	94.3	86.1
SD	20.5	17.3	21.9	24.6	22.4	54.7	46.5
Var	420	299	479	607	502	2992	2169
CV	0.35	0.29	0.35	0.34	0.33	0.61	0.54
Med	56	58	60	66	65	77	73
sIQR	16	12	12	17	12	28	24
Skw	0.32	0.39	1.18	0.68	0.95	2.39	1.94
Krt	0.84	0.35	1.80	0.21	0.76	8.68	5.41

[a]*Mean (M), standard deviation (SD), variance (Var), coefficient of variation (CV), median (Med), semi-interquartile range (sIQR), skewness (Skw), and kurtosis (Krt).*

The upper and lower *LoA*s are very similar across all test–retest pairs and both observer samples (Table 2). The precision measure (95% CIs) of the *LoA*s is slightly larger for the UK sample. In accordance with similar *LoA* values in both samples, also comparable are the *COR*s highlighted in Table 2; specifically, for the Protan vector, *COR* varies between 29.40–32.42 for the three test–retest pairings (Brazil) and is 35.30 (UK); for the Deutan vector, 34.00–39.73 (Brazil) and 33.62 (UK); for the Tritan vector, 49.12–55.55 (Brazil) and 46.69 (UK).

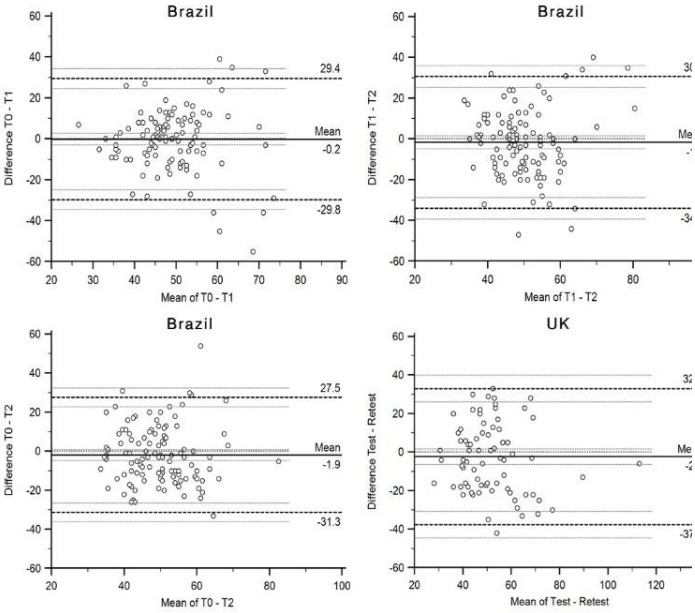

Fig. 3. Bland–Altman graphs for Protan vector measurements (10^{-4} *u'v'* units): *x* axis—mean of test–retest ($\bar{X}_V$); *y* axis—the corresponding difference of test–retest (*D*) for each participant. Open circles indicate data for individual observers; solid lines show mean test–retest difference ($\bar{X}_D$); dashed lines indicate upper and lower *LoA*s; and gray lines indicate 95% CIs for $\bar{X}_D$ and *LoA*s. (A) Data for T0–T1, (B) T1–T2; (C) T0–T2 (Brazil), and (D) for test–retest (UK).

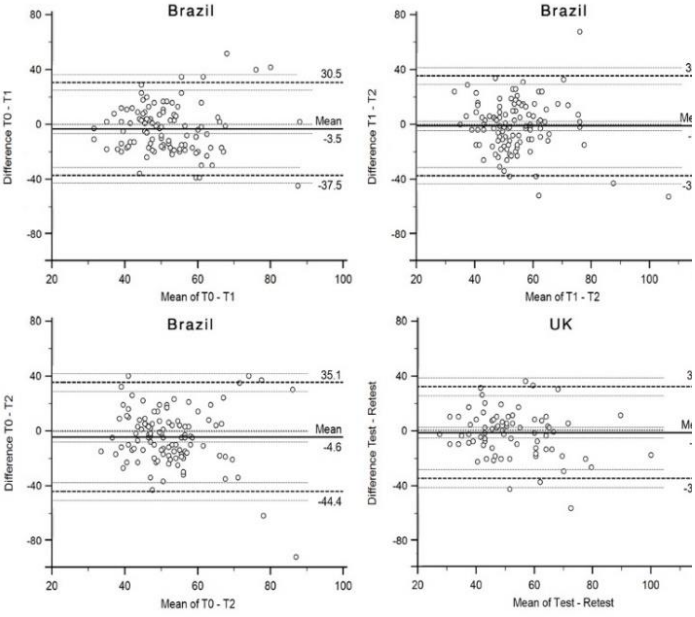

Fig. 4. Bland–Altman graphs for Deutan vector measurements (10^{-4} *u'v'* units): *x* axis—mean of test-retest ($\bar{X}_V$); *y* axis—the corresponding difference of test–retest (D) for each participant. Open circles indicate data for individual observers; solid lines show mean test–retest difference ($\bar{X}_D$); dashed lines indicate upper and lower *LoA*s; and gray lines indicate 95% CIs for $\bar{X}_D$ and *LoA*s. (A) Data for T0–T1, (B) T1–T2; (C) T0–T2 (Brazil), and (D) for test–retest (UK).

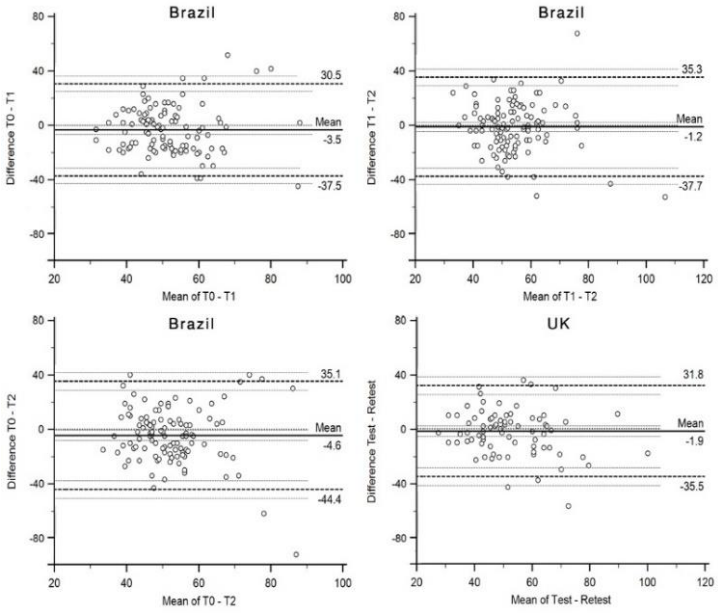

Fig. 5. Bland–Altman graphs for Tritan vector measurements (10^{-4} *u'v'* units): *x* axis—mean of test-retest ($\bar{X}_V$); *y* axis—the corresponding difference of test-retest (*D*) for each participant. Open circles indicate data for individual observers; solid lines show mean test–retest difference ($\bar{X}_D$); dashed lines indicate upper and lower *LoA*s; and gray lines indicate 95% CIs for $\bar{X}_D$ and *LoA*s. (A) Data for T0–T1, (B) T1–T2; (C) T0–T2 (Brazil), and (D) for test–retest (UK).

Table 2. Bland–Altman Parameters of the Trivector Test–Retest Measures (10^{-4} *u'v'* Units) for Samples of NTs in Brazil and the UK[a]

		Brazil (N=111)			**UK** (N=79)
Vector	**Statistic**	**T0-T1**	**T1-T2**	**T0-T2**	**Test-Retest**
Protan	$\bar{X}_V$ (SD_V)	48.71 (8.96)	49.66 (8.63)	49.56 (8.90)	51.53 (13.27)
	$\bar{X}_D$ (SD_D)	−0.20 (15.10)	−1.70 (16.54)	−1.90 (15.00)	−2.38 (18.01)
	95% CI $\bar{X}_D$	(−3.04)–(2.64)	(−4.82)–(1.40)	(−4.72)–(0.92)	(−6.41)–(1.65)
	Upper *LoA*	29.40	30.72	27.50	32.92
	95% CI Upper *LoA*	23.57–35.24	24.33–37.10	21.71–33.30	25.93–39.90
	Lower *LoA*	−29.80	−34.12	−31.31	−37.68
	95% CI Lower *LoA*	(−35.63)–(−23.97)	(−40.51)–(−27.73)	(−37.10)–(−25.51)	(−44.66)–(−30.69)
	COR	**29.60**	**32.42**	**29.40**	**35.30**
Deutan	$\bar{X}_V$ (SD_V)	51.76 (10.23)	54.08 (10.87)	52.33 (10.01)	51.72 (13.57)
	$\bar{X}_D$ (SD_D)	−3.46 (17.35)	−1.18 (18.64)	−4.64 (20.27)	−1.86 (17.15)
	95% CI $\bar{X}_D$	(−6.72)–(−0.19)	(−4.68)–(2.32)	(−8.45)–(−0.82)	(−5.70)–(1.98)
	Upper *LoA*	30.54	35.35	35.09	31.76
	95% CI Upper *LoA*	23.84–37.24	28.15–42.54	27.26–42.92	25.11–38.41
	Lower *LoA*	−37.46	−37.71	−44.37	−35.48
	95% CI Lower *LoA*	(−44.16)–(−30.76)	(−44.90)–(−30.51)	(−52.20)–(−36.54)	(−42.13)–(−28.83)
	COR	**34.00**	**36.53**	**39.73**	**33.62**
Tritan	$\bar{X}_V$ (SD_V)	59.15 (14.25)	61.39 (14.43)	60.34 (15.78)	69.86 (20.33)
	$\bar{X}_D$ (SD_D)	−2.09 (25.06)	−2.39 (26.94)	−4.48 (28.34)	4.38 (23.82)
	95% CI $\bar{X}_D$	(−6.81)–(2.62)	(−7.45)–(2.67)	(−9.80)–(0.85)	(−0.95)–(9.72)
	Upper *LoA*	47.03	50.41	51.06	51.07
	95% CI Upper *LoA*	37.35–56.71	40.01–60.82	40.12–62.01	41.83–60.31
	Lower *LoA*	−51.21	−55.19	−60.02	−42.31
	95% CI Lower *LoA*	(−60.89)–(−41.53)	(−65.59)–(−44.78)	(−70.96)–(−49.08)	(−51.55)–(−33.07)
	COR	**49.12**	**52.80**	**55.55**	**46.69**

[a]Note: $\bar{X}_V$, mean of the vector (V) test−retest; SD_V, standard deviation of $\bar{X}_V$; $\bar{X}_D$, mean of test−retest difference; SD_D, standard deviation of *D*; *LoA*, limit of agreement; CI, confidence interval; *COR*, coefficient of repeatability (highlighted in bold).

Table 3. Trivector Test–Retest Intraclass Correlation Coefficients (*ICCs*) and Their 95% Confidence Intervals (in Square Brackets) for the Brazil and UK Samples of Normal Trichromats (NTs) and for the UK Extended Sample Including Color-Vision Deficient Observers (CVDs)

	Brazil (N=111) NTs			**UK** (N=79) NTs	**UK** (N=123) NTs & CVDs
Vector	**T0-T1**	**T1-T2**	**T0-T2**	**Test-Retest**	**Test-Retest**
Protan	0.29 [(−0.03)–(0.51)]	0.08 [(−0.33)–(0.36)]	0.29 [(−0.03)–(0.51)]	**0.54** [(**0.28**)–(**0.70**)]	**0.98** [(**0.97**)–(**0.99**)]
	(T0-T1, T0-T2, T1-T2) = **0.30** [(**0.04**)–(**0.50**)]				
Deutan	0.28 [(−0.04)–(0.50)]	0.26 [(−0.07)–(0.49)]	−0.02 [(−0.47)–(0.28)]	**0.60** [(**0.38**)–(**0.74**)]	**0.90** [(**0.87**)–(**0.93**)]
	(T0-T1, T0-T2, T1-T2) = **0.24** [(**−0.03**)–(**0.45**)]				
Tritan	0.23 [(−0.12)–(0.46)]	0.13 [(−0.27)–(0.40)]	0.19 [(−0.17)–(0.44)]	**0.65** [(**0.46**)–(**0.78**)]	**0.80** [(**0.71**)–(**0.86**)]
	(T0-T1, T0-T2, T1-T2) = **0.25** [(**−0.02**)–(**0.46**)]				

C. Reliability of the Trivector Subtest

For the Brazil sample, results of the Trivector reliability analysis (Table 3) for NTs revealed relatively low *ICC* magnitudes: across the three vectors, *ICC* varied between (–0.02) and 0.29 for individual test–retest pairs and between 0.24–0.30, when all three datasets (T0–T1, T0–T2 and T1–T2) were included. In comparison, for the UK sample, the *ICC* magnitudes were moderate, varying between 0.54–0.65. These results can be explained by taking into account that *ICC* is estimated by relating within- and between-participants measurement variances {see Eq. (5), [24]}, whereby the lack of variability among the sample observers is related in low *ICC* [23].

Indeed, in a (homogeneous) sample of NTs, the test–retest measurement error for individual observers is expected to be comparable with measurement variability between participants. Low *ICC* values for the Brazil sample indicate that the data of young and middle-aged participants (20–49 y.o.) were fairly homogeneous compared to the UK sample as is apparent from inspection of Figure 2 and Table 1. Greater heterogeneity of the UK data is due to significant variation of age of the sample participants and, in particular, a substantial number of adolescent and older participants, age groups whose chromatic sensitivity is lower and more variable compared to that of observers of middle life decades [12,35,36].

To assess Trivector reliability, the measure of the test's discriminative ability, we explored ICC for an extended UK sample that, along with the data for the NTs analyzed above, also included data for participants with mild and severe color vision abnormalities as well as elderly observers in their seventies and eighties, i.e., life decades with accelerated decline of chromatic sensitivity [13,35].

As presented in Table 1 (rightmost column), the between-participants variability in this sample was substantial, particularly in the Protan and Deutan vector measurements, reflecting significantly increased thresholds of observers with congenital red–green abnormality. The *ICC* measures obtained for this heterogeneous participant sample were 0.98 for the Protan, 0.90 for the Deutan and 0.80 for the Tritan vectors (see Table 3, rightmost column), i.e., in the range expected for a highly discriminative test (cf. [26]).

4. DISCUSSION

The purpose of this study was to evaluate the reproducibility, i.e., repeatability (*evaluative measure*) and reliability (*discriminative measure*), of the Trivector subtest of the CCT in healthy NTs comparing test–retest measures over (relatively) short periods of time and, also, between participant samples in two distant geographic locations. In particular, we aimed to estimate Trivector *COR*, the measure of test precision, and Trivector reliability, measured by *ICC*, the discriminative measure of the test. In addition, we investigated the reliability of the Trivector subtest for an extended sample comprising NTs and observers with congenital and acquired color vision abnormalities.

A. Trivector Repeatability

The Bland–Altman analysis showed that the estimates of Trivector mean test–retest differences ($\bar{X}_D$) only slightly deviate from zero (Table 2), indicating no systematic learning effect and good test repeatability. The Bland–Altman plots (Figs. 3–5) also show that across all test–retest pairs, 88%–98.3% (Brazil) and 89%–94% (UK) of differences lie between the upper and lower *LoA*s, i.e., values close to 95% adopted by the British Standards Institution [21].

Remarkably, the upper and lower *LoA*s are very similar across both observer samples (Table 2). Compared to the Brazil estimates, slightly lower precision measures (95% CIs) of *LoA*s in the UK outcomes result from greater SD_D values in this more age-heterogeneous sample and its smaller size [cf. 19]. The UK sample size was also lower than the at least 100 subjects recommended for agreement studies [17].

Trivector *COR* values, characteristic of the measurement precision used for evaluative purposes, are very similar for the two participant samples, in spite of these populations' varying heterogeneity and slight variation in the equipment specification and viewing angle.

COR is also referred to as the Smallest Detectable Change (SDC) [7,37] or the Smallest Real Difference (SRD) [19] and is the useful index that quantifies test–retest measurement error. Values above *COR*, i.e., test–retest difference after a relatively short lapse of time, indicate a "real" change in individual participant's performance and can be used to guide decision-making with individual observers.

Leaning upon *COR* parameters reported in Table 2 for young and middle-aged NTs (Brazil), i.e., life decades of best or next-to-best chromatic discrimination [12,35,36], we suggest the following *COR* values of Trivector measurements: for the Protan vector, *COR* = 29–32, for the Deutan vector, *COR* = 34–40, and for the Tritan vector, *COR* = 49–56.

For Tritan measurement, the greater *COR* values obtained for both Brazil and UK samples can be explained by observations in other studies employing Bland–Altman analysis, namely, that when the value being measured is larger the variability of test–retest differences is larger too [15]. Larger Tritan values are indeed the case, regardless of a NT's age, compared to Protan and Deutan measurements (see Table 1), in line with previously reported Trivector data for NTs [9,12,13,38].

We endeavor to compare repeatability of the Trivector measures for NTs recently reported in two studies that, too, employed Bland–Altman analysis [39,40].

In particular, for a sample of NTs (N = 93) aged 18–56 years, Bodduluri *et al.* [39] present averaged data for Protan and Deutan vectors (to enable comparison with a red–green discrimination index of a tablet computer-based application developed by the authors) and for Tritan vector. The absence of Trivector differences in test–retest over a short time in [39] is in accord with the present findings.

Medians of the Protan and Deutan vectors, varying between 46 and 53 in the present study across the Brazil and UK samples (Table 1), are higher than median of the averaged Protan and Deutan vectors (*Med* = 35.6) in [39]. In addition, medians of Tritan vectors in the present study (T0: *Med* = 56; T1: *Med* = 58; T2: *Med* = 60, Brazil; *Med* = 66, UK) are slightly greater than the corresponding value in [39], *Med* = 54. The lower estimates (better discrimination) reported in [39] might have resulted from a more "friendly" conditions of stimulus presentation: a shorter distance (3 m) to a monitor ([39], p. 676), as well as the monitor employed in [39], HP CRT p1230 (i.e., from a 22''-series, as we see at https://support.hp.com/us-en/document/c00355202) probably rendered a slightly greater visual angle of the "C" target and hence of the gap. In addition, according to the HP p1230 monitor specification, the employed monitor had a higher resolution than that used for the present data collection.

As is in the present study, test–retest mean difference $\bar{X}_D$ in [39] [Figs. 6(a) and 7(a)] reveals small deviation from zero. However, upper and lower *LoA*s in [39], with $COR \sim 19$ for the red–green parameter [read from their Fig. 6(a)] and $COR \sim 37$ for the Tritan vector [read from their Fig. 7(a)] are much lower than those reported here. We cannot exclude that lower variability of test–retest differences in [39] might have originated from several statistical aspects (cf. [34]) such as data structure (distribution) and/or the computing method and software used that differed from those in the present study.

Repeatability of the Trivector measures—$\bar{X}_D$, upper and lower *LoA*s, and CIs of *LoA*s—was also estimated by Hasrod and Rubin [40] for a small sample (N = 20) of young NTs (aged 19–24 years). Participants were tested monocularly, twice, either of the same day or on two consecutive days. The authors report mean test–retest difference ($\bar{X}_d$; in our notation $\bar{X}_D$), standard deviation of differences (in our notation SD_D), and upper and lower *LoA*s ([40], Table 4). Based on these parameters and applying the Eq. (1) (introduced above [18,19]), we obtained the following *COR* values for their data: *COR* = 38.41 for the Protan, *COR* = 35.67 for the Deutan, and *COR* = 100.94 for the Tritan vectors. The values for the Protan and Deutan measurements are very similar to those reported here; however, their Tritan value is noticeably higher due to a significant outlier in retest as the authors remark and as is apparent in their Figs. 4 and 6.

The overestimation of *COR* for the Tritan vector in [40], along with slightly higher repeatability parameters for more heterogeneous UK sample in the present study are instructive: they remind of Terwee *et al.*'s [37] caution of a selective bias or of an extremely heterogeneous study population that result in indeterminate evaluation of test measurement properties. (Invoking de Vet *et al.*'s [7] graphic comparison, one cannot use the SDC obtained for adult body weight to monitor babies' weight, since the two scales are very different.) If Trivector repeatability studies were extended in the future, this caution prompts stratifying tested samples with regards to certain populations of observers, whose type and/or degree of color-vision abnormality is known. The caution also instigates taking into account the life decade to improve precision of the age-tailored SDC measure in order to monitor genuine changes in the observer's (patient's) color vision over time or evaluate the effects of interventions (improvement or deterioration).

Further, the discrepancies in Trivector *COR* measures addressed above in relation to the outcomes in [39] prompt (i) examination of the role of specific monitor characteristics and of the target visual angle as well as (ii) scrutiny of the data structure and a full report of the computing method and software used (cf. [34], Table 1, p. 571). If these technical and procedural (cf. [16]) and/or statistical factors affect Trivector repeatability measures, they should be taken into consideration when outcomes at other research or clinical locations are related to the *COR* estimates reported here or in other studies.

In this relation, we consider it important to add that in our analysis we used measurements that represent exact parametric *LoA*s (considering asymmetrical), rather than approximations, of parameters such as *LoA*s (and 95% CIs for *LoA*s), since we compare each upper and lower *LoA* separately using a paired *t*-test (an adaptation of the MOVER formula proposed by [41]). In addition, as argued by Shieh [42], one needs to compare each *LoA* individually, while the two endpoints of the two-sided CIs generally do not meet the assumption of equal-tailed error rates, as proposed by MOVER and other formulae that are more appropriate for small sample size.

B. Trivector Reliability

While undertaking the study of Trivector reliability, which has not been assessed previously, in order to gain an insight into reliability estimates, we explored *ICC* values for computerized tests of achromatic contrast sensitivity as a proxy of the CCT. Specifically, we considered tests that use (near) threshold variation of contrast of gratings and employ adaptive forced-choice staircase procedures. For a 2AFC test that required discrimination of grating orientation, *ICC* values varied between 0.45–0.74 for individuals with normal contrast sensitivity and 0.76–0.96 for patients [43]. For the 2AFC Metropsis test, requiring detection of a vertical grating at the right or left side of a screen, *ICC* varied between 0.63–0.80 for a sample of healthy participants [44]. For recently developed contrast sensitivity tests with a 4AFC procedure (i.e., comparable with that of the CCT), the reported *ICC* values are higher. In particular, for the Spaeth/Richman contrast sensitivity test [45], *ICC* varied between 0.90–0.98 for glaucoma patients, 0.36–0.95 for glaucoma suspects, and 0.90–0.97 for controls (the latter had a wide range of visual acuity and optic nerve damage). For the TuebingenCSTest [46], *ICC* varied between 0.88–0.96 for a small sample of young participants.

The *ICC* values for Protan, Deutan and Tritan vectors for NTs obtained here, between 0.19–0.30 (encompassing T0-T1-T2; Brazil) and 0.54–0.65 (UK) (Table 3), imply "poor" and "fair/"good" reliability, respectively (cf. [26]). We considered the possibility that low *ICC* values might reflect the relatively small size of samples tested by us. This

concern was, however, eased by the recent finding (in a simulation study) that an increase in sample size beyond n = 80 does not have a major impact on *ICC* [47].

Note that "poor" reliability of the Trivector subtest for a NT sample was also reported by Hasrod and Rubin [40], namely, *ICC* = 0.27 for the Protan, *ICC* = 0.32 for the Deutan, and *ICC* = 0.13 for the Tritan vectors.

Among other color vision tests, to our knowledge, the *ICC* was assessed only for the D-15d test, for NTs, and revealed moderate reliability, *ICC* = 0.56; as commented by the authors [48], this is less than recommended in clinical testing or research.

We observe that higher ICC values for the contrast sensitivity tests are reported for heterogeneous patient groups compared to relatively homogeneous control groups as pointed out by, e.g., Rubin [43]. Indeed, *ICC* estimates considerably depend on the variance in the tested population [24]: all else being equal, the more similar to each other are the measurements of participants as a group (i.e., more homogeneous), the smaller the *ICC* magnitude [7,19,25]. Apparently, the NT populations tested in the present study and by Hasrod and Rubin [40] were quite homogeneous.

To further explore Trivector reliability performance, we undertook an additional analysis of data from an extended UK sample that was highly heterogeneous (Table 1) and, along with NTs (N = 79), included observers (N = 44) with substantial variation in either congenital (red–green) or acquired (predominantly Tritan) abnormality. The *ICC* values reported in Table 3 (rightmost column), 0.98, 0.90, and 0.80, for Protan, Deutan, and Tritan measurements, respectively, are very promising, all falling within the "excellent" range of test reliability parameter [26].

In the design of future studies of reliability of Trivector (and Ellipses) subtests of the CCT, in tested subsamples, a uniform distribution of observers is desirable to achieve more precise *ICC* values as demonstrated by Mehta *et al.* [47].

C. Conclusions

In summary, our study reports reproducibility measures, repeatability and reliability, for Trivector subtest—Protan, Deutan, and Tritan vectors—of the CCT for healthy NTs.

COR, or smallest test–retest difference in the vector length after a relatively short lapse of time, can serve as a Trivector *evaluative characteristic* useful for guiding clinicians in decision-making about chromatic discrimination change of an individual patient.

ICC, *discriminative characteristic*, assessed for a heterogeneous sample comprising color-normal and color-abnormal observers, shows that the Trivector subtest has high discriminative accuracy and is a highly reliable tool for measuring chromatic sensitivity.

Our findings buttress the current practice of using the CCT Trivector subtest in clinical settings for assessing chromatic sensitivity in patients with various retinal and optic nerve dystrophies, systemic diseases, or neurodevelopmental conditions (cf. [10]).

Funding. Liverpool Hope University (REF1011/20, RES01400); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes; 001); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; 309778/2014-0).

Acknowledgment. T.P.F. and N.A.S. express their gratitude to their colleagues from the *Perception, Neuroscience and Behaviour Laboratory*, for technical and theoretical assistance. Natalia Almeida is kindly acknowledged for data collection. G.V.P. thanks John Mollon and Caterina Ripamonti for valuable advice and Robert Hewertson for technical assistance, as well as psychology students Beata Oakley, Krishni Kunasingham, and Katie Robinson for help with data collection. The authors are grateful to all participants for their time and goodwill. David Bimler is gratefully acknowledged for comments on and proofreading of an earlier version of the manuscript. We also thank three anonymous reviewers and Marina Danilova, Topical Editor, for constructive comments that greatly helped to improve an initially submitted manuscript version. Results of this study were partly presented at the 25th Symposium of the International Colour Vision Society, July 5–9, 2019, Riga (Latvia).

REFERENCES

1. J. L. Barbur and M. Rodriguez-Carmona, "Colour vision requirements in visually demanding occupations," Brit. Med. Bull. **122**, 51–77 (2017).
2. C. Huchzermeyer, J. Kremers, and J. Barbur, "Color vision in clinical practice," in *Human Color Vision*, J. Kremers, R. C. Baraas, and N. J. Marshall, eds. (Springer, 2016), pp. 269–315.
3. M. F. Silva, P. Faria, F. S. Regateiro, V. Forjaz, C. Januário, A. Freire, and M. Castelo-Branco, "Independent patterns of damage within magno-, parvo- and koniocellular pathways in Parkinson's disease," Brain **128**, 2260–2271 (2005).
4. G. V. Paramei, M. Meyer-Baron, and A. Seeber, "Impairments of colour vision induced by organic solvents: a meta-analysis study," NeuroToxicology **25**, 803–816 (2004).
5. A. J. O. Castro, A. R. Rodrigues, M. I. T. Côrtes, and L. C. L. Silveira, "Impairment of color spatial vision in chronic alcoholism measured by psychophysical methods," Psychol. Neurosci. **2**, 179–187 (2009).
6. T. P. Fernandes, S. M. Silverstein, N. L. Almeida, and N. A. Santos, "Visual impairments in tobacco use disorder," Psychiatry Res. **271**, 60–67 (2019).
7. H. C. W. de Vet, C. B. Terwee, D. L. Knol, and L. M. Bouter, "When to use agreement versus reliability measures," J. Clin. Epidemiol. **59**, 1033–1039 (2006).
8. B. C. Regan, J. P. Reffin, and J. D. Mollon, "Luminance noise and the rapid determination of discrimination ellipses in colour deficiency," Vision Res. **34**, 1279–1299 (1994).
9. J. D. Mollon and B. C. Regan, *Cambridge Colour Test Handbook* (Cambridge Research Systems Ltd., 2000), http://www.crsltd.com/tools-for-vision-science/measuring-visual-functions/cambridge-colour-test/.
10. G. V. Paramei, Overview of the studies using CCT, figshare (2019), https://doi.org/10.6084/m9.figshare.11440791.v3.
11. J. D. Mollon and J. P. Reffin, "A computer-controlled colour vision test that combines the principles of Chibret and Stilling," J. Physiol. **414**, 5P (1989).
12. G. V. Paramei, "Color discrimination across four life decades assessed by the Cambridge Colour Test," J. Opt. Soc. Am. A **29**, A290–A297 (2012).
13. G. V. Paramei and B. Oakley, "Variation of color discrimination across the life span," J. Opt. Soc. Am. A **31**, A375–A384 (2014).
14. B. C. Regan, N. Freudenthaler, R. Kolle, J. D. Mollon, and W. Paulus, "Colour discrimination thresholds in Parkinson's disease: results obtained with a rapid computer-controlled colour vision test," Vision Res. **38**, 3427–3431 (1998).
15. J. W. Bartlett and C. Frost, "Reliability, repeatability and repro-

ducibility: analysis of measurement errors in continuous variable," Ultrasound Obstet. Gynecol. **31**, 466–475 (2008).
16. J. D. Mollon, J. M. Bosten, D. H. Peterzell, and M. A. Webster, "Individual differences in visual science: What can be learned and what is good experimental practice?" Vision Res. **141**, 4–15 (2017).
17. C. McAlinden, J. Khadka, and K. Pesudovs, "Statistical methods conducting agreement (comparison of clinical tests) and precision (repeatability and reproducibility) studies in optometry and ophthalmology," Ophthalmic Physiolog. Opt. **31**, 330–338 (2011).
18. J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet **327**, 307–310 (1986).
19. S. Vaz, T. Falkmer, A. E. Passmore, R. Parsons, and P. Andreou, "The case for using the repeatability coefficient when calculating test–retest reliability," PLoS ONE **8**, e73990 (2013).
20. D. Stöckl, D. Rodríguez Cabaleiro, K. Van Uytfanghe, and L. M. Thienpont, "Interpreting method comparison studies by use of the Bland–Altman plot: reflecting the importance of sample size by incorporating confidence limits and predefined error limits in the graphic," Clin. Chem. **50**, 2216–2218 (2004).
21. British Standards Institution, *Precision of Test Methods I. Guide for the Determination and Reproducibility for a Standard Test Method* (BS 5497, part 1) (BSI, 1987).
22. J. L. Fleiss and P. E. Shrout, "Approximate interval estimation for a certain intraclass correlation coefficient," Psychometrika **43**, 259–262 (1978).
23. P. E. Shrout and J. L. Fleiss, "Intraclass correlation: uses in assessing rater reliability," Psychol. Bull. **86**, 420–428 (1979).
24. K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," Psychol. Meth. **1**, 30–46 (1996).
25. R. Müller and P. Büttner, "A critical discussion of intraclass correlation coefficients," Stat. Med. **13**, 2465–2476 (1994).
26. D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," Psychol. Assess. **6**, 284–290 (1994).
27. J. Kottner, L. Audige, S. Brorson, A. Donner, B. J. Gajewski, A. Hróbjartsson, C. Roberts, M. Shoukri, and D. L. Streiner, "Guidelines for reporting reliability and agreement studies (GRRAS) were proposed," Int. J. Nurs. Stud. **48**, 661–671 (2011).
28. S. Ishihara, *Test for Colour-Blindness, 24 Plates Edition* (Kanehara Shupan Co., Ltd., 1972).
29. T. P. Fernandes and G. V. Paramei, Dataset_Reproducibility. Dataset, figshare (2019), https://doi.org/10.6084/m9.figshare.11374449.v2.
30. G. V. Miyahara, "Errors reading the Ishihara pseudoisochromatic plates made by observers with normal colour vision," Clin. Exp. Optom. **91**, 161–165 (2008).
31. M. B. Cranwell, B. Pearce, C. Loveridge, and A. Hurlbert, "Performance on the Farnsworth-Munsell 100-Hue test is signifi- cantly related to non-verbal IQ," Invest. Ophthalmol. Vis. Sci. **56**, 3171–3178 (2015).
32. M. H. Pirenne, "Dark adaptation and night vision," in *The Eye*, H. Davson, ed. (Academic, 1962), Vol. **2**, pp. 93–122.
33. H. Abdi, "Coefficient of variation," in *Encyclopedia of Research Design*, N. Salkind, D. M. Dougherty, and B. Frey, eds. (Sage, 2010), pp. 169–171.
34. A. Abu-Arafeh, H. Jordan, and G. Drummond, "Reporting of method comparison studies: a review of advice, an assessment of current practice, and specific suggestions for future reports," Br. J. Anaesth. **117**, 569–575 (2016).
35. J. L. Barbur and M. Rodriguez-Carmona, "Color vision changes in normal ageing," in *Handbook of Color Psychology*, A. J. Elliot, M. Fairchild, and A. Franklin, eds. (Cambridge University, 2015), pp. 180–196.
36. K. Knoblauch, F. Vital-Durand, and J. L. Barbur, "Variation of chromatic sensitivity across the life span," Vision Res. **41**, 23–36 (2001).
37. C. B. Terwee, S. D. M. Bot, M. R. de Boer, D. A. W. M. van der Windt, D. L. Knol, J. Dekker, L. M. Bouter, and H. C. W. de Vet, "Quality criteria were proposed for measurement properties of health status questionnaires," J. Clin. Epidemiol. **60**, 34–41 (2007).
38. D. F. Ventura, L. C. L. Silveira, A. R. Rodrigues, J. M. De Souza, M. Gualtieri, D. Bonci, and M. F. Costa, "Preliminary norms for the Cambridge Colour Test," in *Normal & Defective Colour Vision*, J. D. Mollon, J. Pokorny, and K. Knoblauch, eds. (Oxford University, 2003), pp. 331–339.
39. L. Bodduluri, M. Y. Boon, M. Ryan, and S. J. Dain, "Normative values for a tablet computer-based application to assess chromatic contrast sensitivity," Behav. Res. Meth. **50**, 673–683 (2018).
40. N. Hasrod and A. Rubin, "The Cambridge Colour Test: reliability of discrimination trivectors in colour space," Afr. Vision Eye Health **78**, a451 (2019).
41. G. Y. Zou, "Confidence interval estimation for the Bland–Altman limits of agreement with multiple observations per individual," Stat. Meth. Med. Res. **22**, 630–642 (2013).
42. G. Shieh, "The appropriateness of Bland–Altman's approximate confidence intervals for limits of agreement," BMC Med. Res. Methodol. **18**, 45 (2018).
43. G. Rubin, "Reliability and sensitivity of clinical contrast sensitivity tests," Clin. Vision Sci. **2**, 169–177 (1988).
44. T. P. Fernandes, N. L. Almeida, P. D. Butler, and N. A. Santos, "Spatial contrast sensitivity: effects of reliability, test–retest repeatability and sample size using the Metropsis software," Eye, **33**, 1649–1657 (2019).
45. J. Richman, C. Zangali, L. Lu, S. S. Wizov, E. Spaeth, and G. L. Spaeth, "The Spaeth/Richman contrast sensitivity test (SPARCS): design, reproducibility and ability to identify patients with glaucoma," Br. J. Ophthalmol. **99**, 16–20 (2015).
46. T. Schilling, A. Ohlendorf, A. Leube, and S. Wahl, "TuebingenCSTest– a useful method to assess the contrast sensitivity function," Biomed. Opt. Express **8**, 1477–1487 (2017).
47. S. Mehta, R. F. Bastero-Caballero, Y. Sun, R. Zhu, D. K. Murphy, B. Hardas, and G. Koch, "Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies," Stat. Med. **37**, 2734–2752 (2018).
48. G. W. Good, A. Schepler, and J. J. Nichols, "The reliability of the Lanthony Desaturated D-15 test," Optometry Vision Sci. **82**, 1054–1059 (2005).