



















Multicenter evaluation of the color vision screener test

BENJAMIN E. W. EVANS,^{1,*}  MARISA RODRIGUEZ-CARMONA,¹ 
FRANZISKA G. RAUSCHER,^{2,3}  EMSAL LLAPASHTICA,¹  VILHELM F. KOEFOED,⁴ 
FOCKE ZIEMSEN,⁵  RUDOLPH NITSCHKE,^{2,5}  ALESSANDRO FARINI,⁶ 
ELISABETTA BALDANZI,⁶  LUIS GÓMEZ-ROBLEDO,⁷  AMANDA DOUGLASS,^{8,9} 
MADELINE BAKER,⁸  ROLAND QUAST,¹⁰  SABINE ROELCKE,¹⁰  STEVEN C. C. HO,¹¹  AND
JOHN L. BARBUR¹ 

¹The Henry Wellcome Laboratories for Vision Science, Centre for Applied Vision Research, School of Health and Medical Sciences, City St George's, University of London, London, UK

²Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Härtelstrasse 16-18, 04107 Leipzig, Germany

³Medical Informatics Center-Department of Medical Data Science, Leipzig University Medical Center, Härtelstrasse 16-18, 04107 Leipzig, Germany

⁴University of Bergen, Bergen, Norway

⁵Department of Ophthalmology, Leipzig University Medical Center, Liebigstrasse 10-14, 04103 Leipzig, Germany

⁶ViOLa Visual Optics Lab-National Institute of Optics CNR, 50125 Firenze, Italy

⁷University of Granada, Granada, Spain

⁸School of Medicine-Optometry, Faculty of Health, Deakin University, Waurn Ponds, Australia

⁹Department of Optometry and Vision Sciences, The University of Melbourne, Parkville, Australia

¹⁰Medizinisches Zentrum, Stuttgart Airport, Stuttgart, Germany

¹¹Sunsmile Aeromedical, Hong Kong SAR, China

*benjamin.evans.2@city.ac.uk

Received 16 October 2024; revised 26 December 2024; accepted 28 December 2024; posted 7 January 2025; published 6 February 2025

An international multicenter study was designed and carried out to evaluate the color vision screener (CVS) test for normal trichromats and congenital color deficient. Over 400 participants from nine international Colour Assessment and Diagnosis (CAD) testing centers completed the CVS and the CAD test on calibrated visual displays. The CVS had a sensitivity and specificity [95% confidence intervals] of 1.00 [0.98–1.00] and 0.99 [0.97–1.00] with a positive and negative predictive index of 0.94 and 1.00 for an assumed prevalence of 8%. The CVS is quick, efficient, and easy to use, and its sensitivity is equivalent to the optimal published Ishihara protocol. © 2025 Optica Publishing Group. All rights, including for text and data mining (TDM), Artificial Intelligence (AI) training, and similar technologies, are reserved.

<https://doi.org/10.1364/JOSAA.544985>

1. INTRODUCTION

The ability to establish an individual's type of color deficiency and to quantify the severity of color vision loss is of considerable value, both clinically and within occupational settings [1–4]. The ability to detect efficiently and to identify color signals can enhance the visual performance [5–9], while the reduced chromatic sensitivity can, particularly where no other redundant information is included, result in major accidents and, in the worse instance, loss of life [10]. The majority of individuals have normal trichromatic color vision—approximately 8% of men and ~0.5% of women in Caucasian populations are reported to have congenital deutan and protan deficiencies, while congenital tritan deficiencies, following an autosomal-dominant inheritance pattern, are less common [11–14]—resulting

in the need for an efficient screening test to rapidly detect those who require more time-consuming diagnostic assessment [15,16]. Currently, the Ishihara pseudoisochromatic plate test, a test designed to screen only for congenital red–green (RG) deficiencies, is often used to fulfill this screening requirement [17].

As with other screening tests, the value of a color vision screener is determined primarily by its sensitivity and specificity, the probability that the test will correctly identify individuals with CVD or normal trichromatic color vision, respectively [18]. Diagnostic color assessment tests have a higher resource cost but aim to provide more information, such as the classification of color vision, the presence of combined acquired and congenital deficiency, and a measure of RG and YB loss [19–21]. Tests are combined and employed using different sets

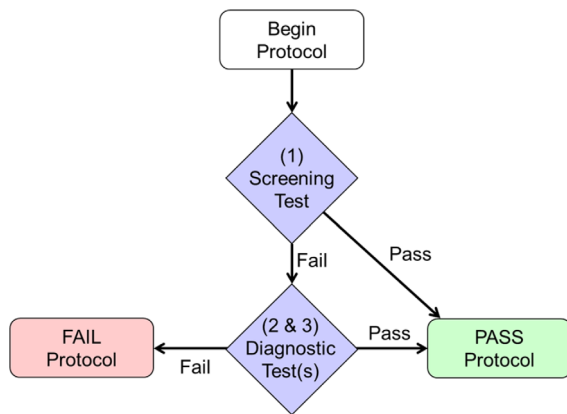


Fig. 1. General stages in all color assessment protocols. The number and type of tests employed at each stage vary across protocols. Note that this procedure was not followed in this research as all participants were assessed with both the screening test and the gold standard reference measure.

of rules or protocols to give clinical significance to the results of the test and/or to achieve a specific outcome [4,22–24]. The most widely used implementation, not specific to color vision, is screening followed by further diagnostic tests if the screening test is failed (Fig. 1).

It is of value to outline the requirements for an “ideal” method for “full” color vision assessment. Ideal color vision assessment would fully isolate color signals, ensuring one can make use of only color cues, allow for the selective simulation of RG and YB chromatic mechanisms to classify accurately the class of any individual’s color vision, quantify the severity of any RG or YB loss, and have a low resource cost (be inexpensive, quick to carry out, and easy to administer). Given the range of physiological properties present in the “normal” population (including variations in the L:M cone ratio, peak wavelength responsivity of cone photoreceptors, differences in photoreceptor pigment optical densities, and the effects of normal aging and pre-receptor filtering of light [25–27]), an “ideal” test also needs to establish a “normal” age-matched range to account for the variation observed in the “normal” population [28].

Diagnostic color threshold tests, such as the Colour Assessment and Diagnosis (CAD) [29] and Cambridge Colour Test (CCT) [30], meet several requirements for an “ideal” color vision assessment. However, both the CAD and CCT have relatively high resource costs and take between 12 and 15 mins to complete in the case of CAD. A potential solution, proposed previously in 2021, is a two-step color assessment protocol which utilizes the CAD test in combination with a quick and efficient color vision screener (CVS) test [31].

The CAD and CVS tests have been described previously [31,32]. Both tests display moving, color-defined stimuli on a background of dynamic luminance contrast noise to mask any residual perceived luminance contrast signals. A short learning mode is built into the procedure of the CAD and CVS and must be passed before each test is undertaken. The CAD test uses 16 interleaved color directions specified in the CIE 1931 (x, y) color space, with white point chromaticity coordinates of (0.305, 0.323), combined with a four-alternative forced choice (AFC) procedure. The sampling of the hues is arranged to match as closely as possible the expected directions of deutan, protan,

and tritan color confusion bands. The CAD test outputs results in CAD units, where 1 CAD unit is based upon mean color thresholds measured in 330 healthy, young, and normal trichromats [33]. These measured chromatic thresholds are directly proportional to the cone contrasts generated [34]. The CVS test uses a 2AFC, and the chromaticity of the stimuli rotates through the CIE (x, y) color space during each presentation (restricted to R/G, Y/B, and the suprathreshold regions of the corresponding color threshold ellipse) for the observer’s age. The hue directions sampled in the CVS approximately match those employed in the CAD test. Suprathreshold stimuli, generated by adding an additional 150% chromatic contrast and a 45% luminance contrast component to YB CVS stimuli, are used to determine an individual’s response reliability. Measurements where <86% of suprathreshold stimuli are correctly identified are classified as “unusable.” It is important to note that the CVS presents stimuli with different chromatic contrasts for different observers by utilizing the upper normal threshold limits established for normal aging when using the CAD test. The method establishes whether the subject’s chromatic sensitivity falls within the “normal” limits for the corresponding age. The CVS test currently exists in two parallel forms: one version built into the CAD test and designed to run on calibrated visual displays, and a second freely downloadable form as a standalone file designed for use on computers running the Windows operating system connected to uncalibrated visual displays that support the sRGB color mode. It should be acknowledged that colors are coded, and subsequently rendered, to be reproduced accurately on an ideal sRGB display, however, as with any production of color on a visual display, if the screen is uncalibrated one cannot know which color will actually be presented. This paper reports upon results obtained using fully calibrated visual displays, expanding upon preliminary results for the calibrated version of the test [31].

A widely reported inaccuracy surrounding the CAD test can be attributed to a paper by Seshadri *et al.* [35]. The “web-based version of the CAD test” reported by Seshadri *et al.* is an ~90 s video showing the stimuli employed in the CAD test. Unfortunately, the fact that the web video is simply representative of the stimuli used in the full CAD test is not reported in a number of publications which quote Seshadri *et al.*, along with inaccurate statements surrounding the CAD tests’ ability to detect YB loss [36,37].

CAD systems, comprising calibrated hardware and specific software, can be found worldwide at clinical and occupational centers. The resulting international network of CAD centers provides a valuable resource that can facilitate examining a large number of participants in several separate locations using consistent testing conditions and identical hardware and software. This international consortium has been facilitated, in no small part, by the adoption of the CAD test across occupational environments [6,38]. The international network of CAD centers also enables a multicenter study methodology to be used to collect and analyze CAD (and CVS) data. A multicenter study, in which research is conducted in multiple centers following the same protocol, can confer several advantages over single-center studies, including a larger sample size, a more diverse population, and increased generalizability [39–41].

This study aimed to evaluate the recently developed CVS test in an international collaborative multicenter study, carrying out

the test with different examiners in different population groups. The aim was to establish the CVS's outcome and determine its suitability through comparison to the most popular color vision screening test for RG CVD, the Ishihara pseudoisochromatic plate test [37,42]. This builds upon work introduced in 2021 [31] and is the first of two papers evaluating the outcome of the CVS test on calibrated and uncalibrated visual displays.

2. METHODS

In 2019, the international consortium of CAD testing centers was formed, and all centers were invited to participate in the validation of the CVS test. This invitation was re-extended in 2021 following the disruption to research internationally caused by the COVID-19 pandemic. Over the course of the study, data collection was actively paused and resumed to comply with local, national, and international guidelines. Center participation was voluntary, there were no recruitment requirements for inclusion, and every center that contributed results was accepted and included in the study. The participating centers and researchers are shown in Table 1.

Each CAD center within the international consortium is equipped with standardized Advanced and Optometric Test (AVOT) equipment, including a 10-bit visual display, a photometer, the full CAD test, and programs for automatic calibration of the 10-bit visual display employed to generate the visual stimuli. These items were used across all centers to ensure consistency. Participating centers were provided with the CVS (v2.6.1 or v2.8) as an embedded option within the CAD test. The CVS stimuli and psychophysical procedure, including the on-screen instructions given to participants, were the same across centers.

Local research and ethical approval were obtained at each center prior to any participant recruitment, and data collection and participant recruitment were independently managed at each center. Participants completed the CAD and CVS tests

binocularly at a viewing distance of 1.4 m and wore any refractive correction they habitually used for tasks at the tests' working distance. All examiners were familiar with carrying out the CAD test in routine clinical practice. The on-screen CVS instructions, always shown in English, ensured that each participant received standardized instructions, and participants were required to correctly identify all stimuli in the CAD and CVS test learning modes to ensure they understood the test procedure prior to taking the test. Informed consent was obtained from all participants, across all centers. Participants could withdraw from the study at any point and the study followed the tenets of the Declaration of Helsinki.

Data collected at centers were de-identified and securely transferred to the team at City St George's, University of London, in line with international data protection legislation. Records for 488 participants collected at nine CAD centers located in Europe, America, Asia, and Australia were received, and exclusion criteria were applied. Exclusion criteria included duplicates, records with only CVS or CAD test data, participants above the age of 75 or below the age of 16, and participants with acquired color deficiency, as diagnosed by the CAD test.

CAD data for each participant were used as a reference measure to determine the "true status" of participants' color vision (i.e., "normal trichromat," "deutan," "protan," etc.). Participant's CAD and CVS data were analyzed to determine the sensitivity, specificity, test accuracy (or efficiency), and positive and negative predictive value (PPV and NPV, respectively) of the CVS. The CVS accuracy, PPV, and NPV were calculated for an assumed prevalence of 8%, which is in line with estimates for the prevalence of congenital RG CVD in Caucasian male populations [13]. Ninety-five percent confidence intervals were calculated for the sensitivity and specificity using the Wilson method [43], and the outcome of the CVS was compared to the severity of loss, as determined by the CAD test.

A review of the literature was carried out to extract data for the Ishihara screening tests in populations of normal trichromats

Table 1. Nine CAD Testing Centers Who Participated in the International Multicenter Study^a

Center Name	Location	Researcher(s)
City St George's, University of London	London, United Kingdom	Professor John Barbur, Dr. Benjamin Evans, Dr. Emsal Llapashtica, Dr. Marisa Rodriguez-Carmona
Deakin University	Victoria, Australia	Ms. Madeline Baker Miss Kate Coffey Dr. Amanda Douglass
Leipzig University, Germany	Leipzig, Germany	Mr. Rudolph Nitsche, Dr. Franziska Rauscher, Professor Dr. med. Focke Ziemssen
Medizinisches Zentrum, Stuttgart Airport	Stuttgart, Germany	Professor Roland Quast, Dr. Sabine Roelcke
Naval Refractive Surgery Center, San Diego, USA: NRSCSD	San Diego, USA	Dr. Vilhelm F Koefoed
Sunsmile Aeromedical	Hong Kong	Dr. Steven C. C. Ho
University of Bergen, Norway: UiB	Bergen, Norway	Dr. Vilhelm F Koefoed
University of Granada	Granada, Spain	Professor Luis Gómez-Robledo
ViOLA Visual Optics Lab-National Institute of Optics CNR	Florence, Italy	Dr. Elisabetta Baldanzi, Professor Alessandro Farini

^aPrincipal researchers at each center are listed in alphabetic order.

and individuals with congenital RG CVD. The search was conducted using Google Scholar and PubMed with the keywords (color vision assessment) AND (Ishihara) AND (anomaloscope). A preselection of papers was performed by screening titles and abstracts for relevance to the topic. Full-text articles meeting the inclusion criteria were then reviewed in detail. The inclusion criteria required studies with sample sizes of at least 140 participants, while exclusion criteria eliminated studies where a reference measure other than CAD or anomaloscopy was used, or where only participants who failed the Ishihara test completed the reference test to confirm the presence of any color vision deficiency. No artificial intelligence (AI) methodologies were used in the search or analysis process. Calculations for sensitivity, specificity, accuracy, PPV, and NPV were carried out using previously published data and compared to the CVS using the current cohort.

3. RESULTS

Data from 180 participants with normal trichromatic color vision, 181 with deutan deficiency, and 69 with protan deficiency were analyzed following the application of the exclusion criteria. Participants ranged from 16 to 71 years of age

with a median [and interquartile range] of 30 [23–40] years. Biological sex data were available for ~82% of the cohort, with the remaining data unavailable due to international data sharing limitations. The distribution of age for male and female normal trichromatic, deutan, and protan participants for this subset of the cohort is shown in Fig. 2. All included participants had YB CAD thresholds within the normal limits for their age and RG CAD thresholds within the expected range for participants' class of color vision deficiency, as shown in Fig. 3. No participant had an “unusable” CVS response reliability as determined by the identification of suprathreshold stimuli throughout the test.

Across all centers and all participants, with assessments carried out by a range of examiners, clinicians, and practitioners, one normal trichromat failed the RG component of the CVS test, and one normal trichromat failed the YB component of the CVS test (Fig. 4). All 181 deutan and 69 protans were correctly identified as having RG loss with normal YB chromatic sensitivity by the CVS, passing the YB and failing the RG component of the CVS, corresponding to a multicenter sensitivity and specificity [and 95% confidence intervals] of 1.00 [0.98–1.00] and 0.99 [0.97–1.00], respectively. For an assumed prevalence of 8%, the CVS has a PPV of 0.94 and an NPV of 1.00. At each

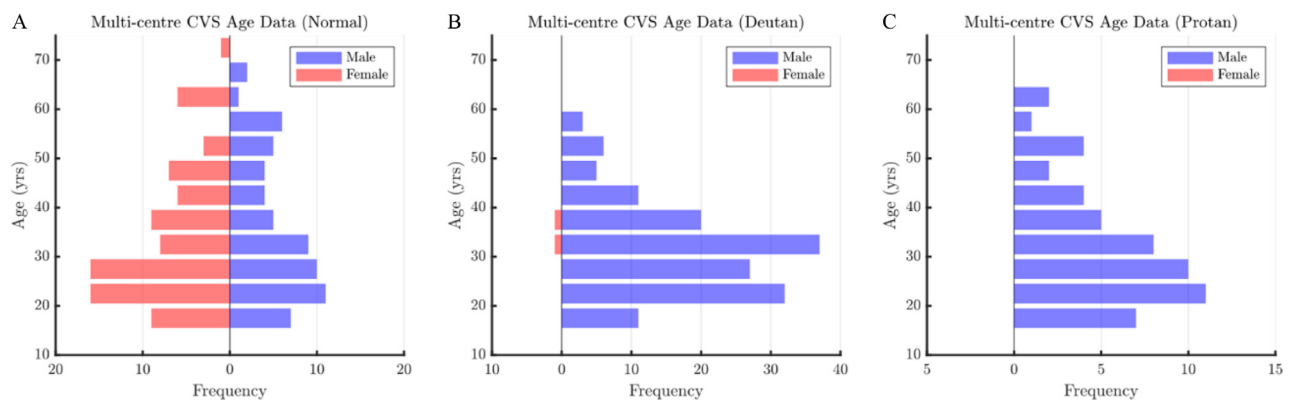


Fig. 2. Age and biological sex distribution for the data collected across all centers. Biological sex data were available for approximately 82% of the cohort, and the distributions are split into (A) participants with normal color vision, (B) participants with a deutan deficiency, and (C) participants with a protan deficiency.

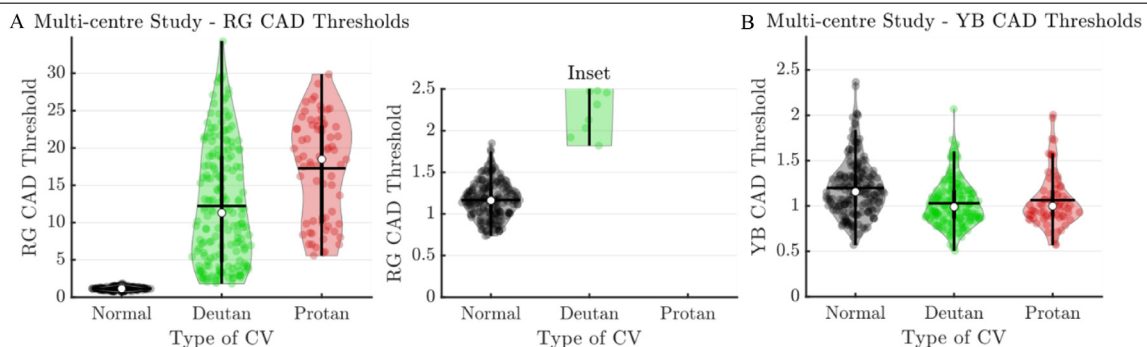


Fig. 3. Violin plots showing (A) RG and (B) YB CAD thresholds for all participants assessed in the multicenter study. Violin plots combine a boxplot, a density trace, and the mean (horizontal black bar) into a single graphic [44]. RG CAD thresholds ranged from 0.74 to 1.85 in normal trichromats, 1.82 to 34.30 in deutans, and 5.56 to 29.87 in protan participants. All normal trichromats, protans, and deutans had YB CAD thresholds within the normal limits established for their age. All deutan participants have CAD thresholds outside the normal limits for their age, and all normal trichromats have thresholds within the normal limits established for their age. The overlap between the RG CAD thresholds for the least affected deuteranomalous subjects and the least sensitive normal trichromats is due to the range of ages in the subject population and the normal age-adjusted limits employed in the CAD test.

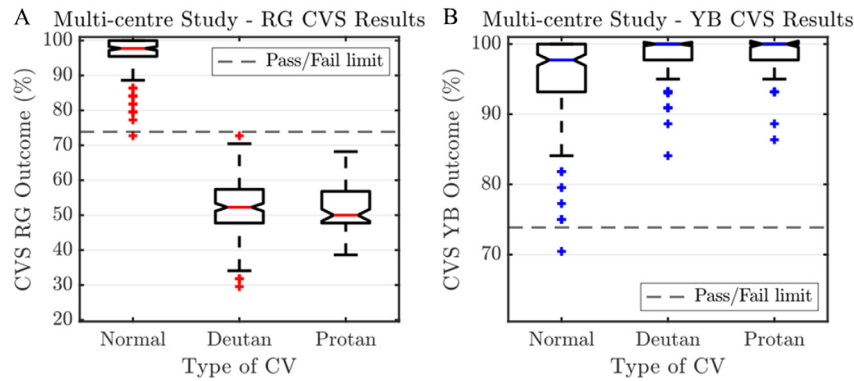


Fig. 4. Notched boxplots showing (A) RG and (B) YB CVS outcomes for participants assessed as part of a multicenter CVS study. Values greater than $q3 + 1.5(q3 - q1)$ or less than $q1 - 1.5(q3 - q1)$ were classified as outliers (plotted as +). All deutan and protans assessed were correctly classified by the CVS test. One normal trichromat was misdiagnosed for RG color vision, and a different normal trichromat was misdiagnosed for YB color vision.

Table 2. Number of Participants Assessed (N), Prevalence of Congenital RG CVD (Prev), Sensitivity (Sens), and Specificity (Spfc) for RG CVS Data^a

Gold Standard	Publication	Test and Pass Protocol	N	Prevalence	Sensitivity (95% CI)	Specificity (95% CI)	Assumed Prevalence: 8%		
							Accuracy	PPV	NPV
CAD	-	CVS RG—All centers	430	0.58	1.00 (0.98–1.00)	0.99 (0.97–1.00)	0.99	0.94	1.00
Nagel and CAD	Rodriguez <i>et al.</i> [23]	Ishihara 38 pl. 0 err pl. 1–25	1827	0.81	0.99 (0.99–1.00)	0.81 (0.76–0.84)	0.82	0.31	1.00
Nagel and CAD	Rodriguez <i>et al.</i> [23]	Ishihara 38 pl. 0 err pl. 1–15	1827	0.81	0.99 (0.99–1.00)	0.89 (0.85–0.92)	0.90	0.44	1.00
Nagel and CAD	Rodriguez <i>et al.</i> [23]	Ishihara 38 pl. ≤ 2 err pl. 1–17	1827	0.81	0.97 (0.96–0.98)	0.99 (0.97–1.00)	0.99	0.88	1.00
Nagel and CAD	Rodriguez <i>et al.</i> [23]	Ishihara 38 pl. ≤ 4 err pl. 1–21	1827	0.81	0.96 (0.95–0.97)	0.99 (0.98–1.00)	0.99	0.94	1.00
Nagel	Birch [42] and Birch and McKeever [20]	Ishihara ≤ 8 errs pl. 2–17	872	0.46	0.81 (0.76–0.84)	1.00 (0.99–1.00)	0.98	1.00	0.98
Nagel	Birch [42] and Birch and McKeever [20]	Ishihara ≤ 6 errs pl. 2–17	872	0.46	0.94 (0.91–0.96)	0.95 (0.93–0.97)	0.95	0.64	0.99
Nagel	Birch [42] and Birch and McKeever [20]	Ishihara ≤ 3 errs pl. 2–17	872	0.46	0.99 (0.97–0.99)	0.94 (0.92–0.96)	0.94	0.59	1.00
Nagel	Birch (2010)	Ishihara 38 pl. ≤ 3 err pl. 2–17	486	1.00	0.98 (0.96–0.99)	-	-	-	-
Nagel	Birch (2010)	Ishihara 38 pl. ≤ 4 err pl. 2–17	486	1.00	0.95 (0.93–0.96)	-	-	-	-
Nagel	Aarnisalo (1979)	Ishihara 38 pl. 0 err pl. 1–25	150	0.33	1.00 (0.93–1.00)	0.67 (0.57–0.75)	0.70	0.21	1.00
Nagel	Aarnisalo (1979)	Ishihara 38 pl. ≤ 1 err pl. 1–25	150	0.33	0.96 (0.87–0.99)	0.95 (0.89–0.98)	0.95	0.63	1.00
Nagel	Aarnisalo (1979)	Ishihara 38 pl. ≤ 4 err pl. 1–25	150	0.33	0.84 (0.71–0.92)	1.00 (0.96–1.00)	0.99	1.00	0.99

^aThe positive and negative predictive values (PPV and NPV, respectively) have been calculated for a prevalence of 0.08, or 8% (the maximum prevalence observed in male populations). Equivalent statistics for published studies that employed the Ishihara pseudoisochromatic plates have been included to allow for a comparison of the CVS test and the Ishihara pseudoisochromatic plates.

individual center, the RG sensitivity was 1.00. The RG specificity was 1.00 at all but two centers, and at both of these centers, one participant with normal RG chromatic sensitivity failed the RG component of the CVS.

Data for the previously reported Ishihara pseudoisochromatic plate test are shown in Table 2. As previously reported across multiple studies, maximizing sensitivity occurs at a cost to specificity and vice versa. The highest reported sensitivity of the

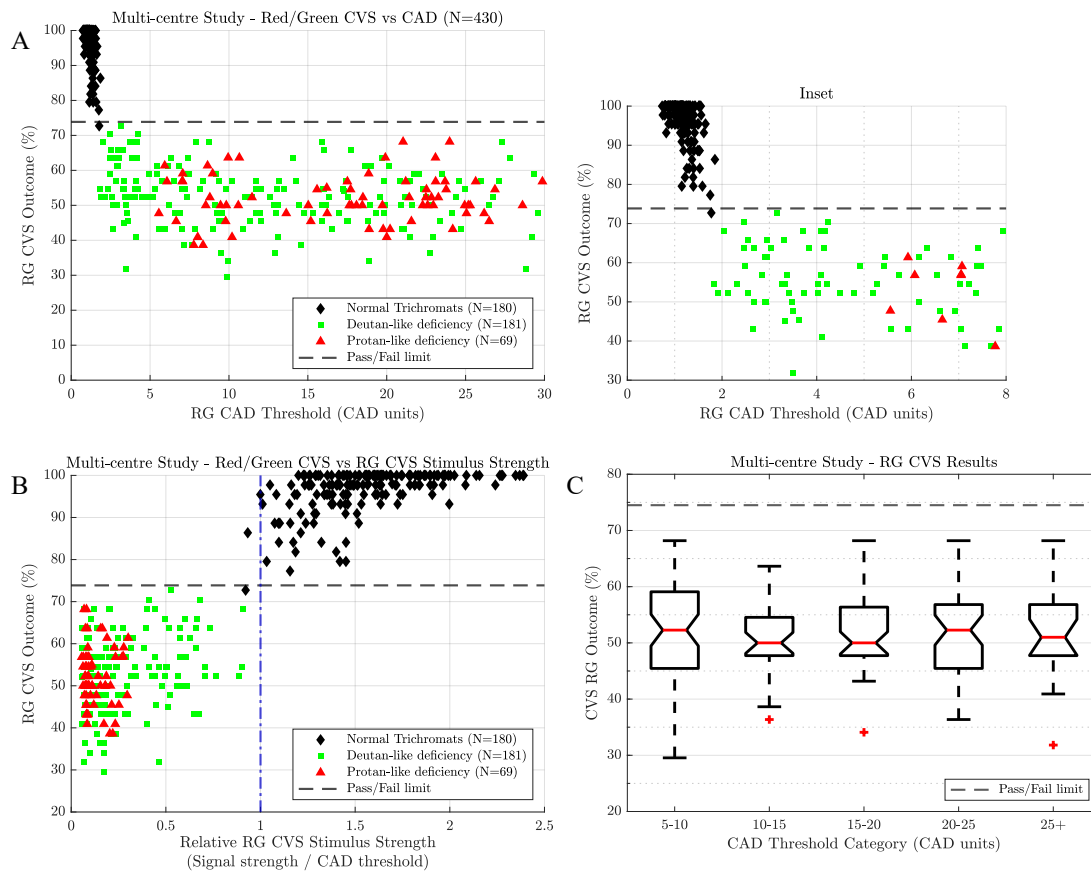


Fig. 5. (A) Outcome of 430 participants on the CVS test compared to their outcome on the reference CAD test for RG stimuli. The CAD results are plotted in terms of CAD units; units based upon a standard young observer who has a threshold of 1 CAD unit. These units are the standard output of the CAD test. An inset shows a resized version of the same data. The one normal trichromat who failed had a RG CAD threshold of 1.77, and the normal CAD upper limit for their age (22 years) is 1.79. (B) The same data shown in (A) are plotted using the relative RG CVS stimulus strength employed. The relative RG CVS stimulus strength is calculated by dividing the signal strength used for each participant (which varies based on their age) by their measured RG CAD threshold. For example, an observer with a RG CAD threshold two times larger than the stimulus strength employed in the CVS test would have a relative RG CVS stimulus strength of 2, whereas an observer with a CAD threshold that is half the signal strength employed would have a relative RG CVS stimulus strength of 0.5. The least affected, or most sensitive, deutan observers will be shown stimulus strengths $\sim 0.8 \times$ their threshold, whereas the least affected, or more sensitive, protan observers are shown stimulus strengths $\sim 0.25 \times$ their threshold. (C) A selection of the data shown in (A) in notched boxplots, grouped in CAD threshold categories, showcasing the equitability for the outcome of the RG CVS in individuals with CAD thresholds over 5 CAD units.

Ishihara is statistically equivalent to the outcome of the CVS test when carried out on calibrated visual displays.

The normal trichromat who failed the RG components of the CVS had a RG CAD threshold of 1.77 with an upper normal age-adjusted limit of 1.79. The relationship between the severity of RG color deficiency and the outcome of the CVS screener is shown for all participants in Fig. 5. Figure 5B reveals the experimental agreement with the predicted test outcome, the probability of passing the CVS is proportional to an individual's sensitivity to color, as quantified by the CAD test. The least affected deutan participants, with the smallest CAD thresholds, have the highest probability within the group of congenital color deficient of passing the RG component of the CVS. The RG CVS outcome for individuals with RG CAD thresholds over 5 RG CAD units was statistically equivalent and reflects the expected spread around the change probability of a correct response (Fig. 5C).

4. DISCUSSION

The CVS is a rapid screening test, typically taking 2–3 mins to complete, and is easy for participants to understand and testers to administer. The present evaluation of this test in 430 participants across eight countries reveals that the test simultaneously achieves high sensitivity (1.00) and specificity (0.99). Table 2 demonstrates the trade-off between sensitivity and specificity present in the most commonly used color screener, the Ishihara pseudoisochromatic plate test. This compromise is further elucidated as the study sample size increases. With a singular pass criterion, the CVS achieves simultaneously a statistically equivalent sensitivity and specificity to the highest sensitivity and specificity obtained in separate protocols for the most common screening test in current use, the Ishihara pseudoisochromatic plates.

Unlike the Ishihara test, the CVS can also screen for YB loss with high specificity (0.99 [0.97–1.00]). While the specificity of the YB CVS was high (only one normal trichromat failed the

YB component of the CVS), the sensitivity of the YB CVS has yet to be experimentally established, not being determined here due to the exclusion criteria for this study. A potential challenge with doing so is the low prevalence of acquired color vision loss, particularly in younger participants, and confounding factors that may drive such acquired loss.

The PPV and NPV are the proportion of individuals who fail a test who are correctly diagnosed as having a CVD and the proportion of individuals who pass a test who are correctly diagnosed as not having a CVD, respectively. The PPV and NPV were calculated for a fixed prevalence of 8% to allow for a comparison between studies and provide a more accurate representation of the tests' expected performance in a male population. As with sensitivity and specificity, a good screening test maximizes both PPV and NPV. A high NPV ensures one is confident that individuals who pass have normal trichromatic color vision, which is of particular importance in occupational settings where one wishes to detect all CVDs at the screening stage. The NPV and the sensitivity are maximized when the pass protocol becomes more stringent. In practical terms, doing so maximizes the "safety" of an occupational protocol at an increased resource cost, as more applicants need to complete further diagnostic testing to confirm the presence of any CVD. The alternative is to maximize specificity, ensuring all, or almost all, normal trichromats pass screening at the cost of allowing some CVDs to pass, potentially with moderate-to-severe CVD [23].

While a multicenter approach confers several advantages, it should be noted that the results of a large multicenter study are not necessarily applicable to less heterogeneous populations [45]. Within the context of multicenter color vision research, the age of participants is a key consideration (an older cohort would have a worse mean chromatic discrimination), along with the varying prevalence of CVD in different populations.

Participants assessed as part of the multicenter study were primarily young and of working age with a median [and IQR] age of 30 [23–40] years. The age-adjusted nature of the classification made by the CAD test ensures that any systematic intercenter differences in the median age of participants are taken into account by the CAD test when the class of CVD is determined. The incorporation of the normal age-matched limits into the CVS also means that while two participants of the same age at different centers will have been tested using the same stimuli, two participants of different ages within the same center will be shown different stimuli, and participants are screened based upon whether their RG and YB chromatic sensitivities are within the normal limits *for their age*, not an arbitrary fixed standard.

A larger proportion of male CVD participants is expected and consistent with the established prevalence of deutan and protan deficiencies in male and female populations. Specific to this multicenter study, CAD testing centers, by virtue of their position, are established, advertised, and used to assess those with known or suspected CVD and to determine whether individuals have the level of chromatic discrimination required to work in a specific occupational setting. Hence, the large prevalence of congenital RG CVD (58%) is not unexpected, and the study design minimizes the potential impact of varying CVD prevalence across centers located in different continents. The sensitivity and specificity can be calculated independent of

prevalence, and, as previously, the PPV, NPV, and test accuracy (or test efficiency) was calculated for a fixed prevalence level of 8%.

The potential impact of the screening protocol employed upon the observed prevalence of CVD has been highlighted in a study by Arnegard *et al.* [46] in which 193 young Norwegian males were screened for congenital RG CVD. Screening was carried out with the 24-plate edition of the Ishihara test (≥ 3 errors) and genetic testing using the Agena MassARRAY system. Genetic screening revealed a 10.4% prevalence of congenital RG CVD, yet the results of the Ishihara test only indicated a prevalence of 5.2% in the same sample. A similar discrepancy between the two screening methods was reported for a female sample. It should, however, be noted that several other studies have used the Ishihara test to identify the prevalence of RG CVD using larger samples (≥ 5000 male participants) in European Caucasian populations and found prevalences $\sim 8\%$, including a study carried out in Norway with 9049 male participants [13,47].

One of the primary sources of error in multicenter studies can be a lack of protocol adherence across centers [39,48,49]. This limitation was minimized through the clear and concise user instructions displayed as part of the CAD and CVS software every time CAD or CVS testing was carried out. Both sets of instructions describe the test procedure centers should follow, and the CVS also provides on-screen instructions to all participants prior to starting the test, at the end of the learning mode, and at the end of the final CVS test, ensuring that the instructions provided to participants remained constant across all centers. The protocol adherence was not externally validated, but the participating centers were involved in color assessment in aviation and were inspected and had to comply with the requirements of their Civil Aviation Authorities. The CAD test also records the date of every display calibration check made by users for compliance with specified requirements.

The severity of loss measured by the CAD test for the cohort suggests the inclusion of both anomalous trichromats and dichromats, as shown by the "double peak" within violin plots shown in Fig. 3. The secondary peak is likely attributable to dichromats with worse chromatic discrimination, and the primary peak attributable to anomalous trichromats. The maximum severity of loss capable of being measured and quantified by the CAD test is determined by the gamut of the visual display and the amplitude of the dynamic luminance contrast noise employed in the test. For standard CAD test parameters, the maximum severity of loss is the same for deutan and protans. Only one participant, a 49-year-old deutan, had a CAD threshold of >30 RG CAD units. The least affected deutan had a threshold of 1.82 RG CAD units while the least affected protan had a CAD threshold of 5.56. This discrepancy has previously been reported and attributed, at least in part, to the difference in $\delta\lambda_{\max}$ between each group [26,27,32].

The observed variation within presumed dichromats is of interest. This variation is likely due to factors that also drive the observed variation in the chromatic sensitivity of normal trichromats such as the L:M cone ratio, small shifts in peak wavelength responsivity, differences in pigment optical density, and variations in prereceptoral filtering such as the lens and macular pigment. Such changes contribute to the observed

intersubject variability in normal color vision. A potential limitation of the multicenter study procedure is the lack of testing via anomaloscopy, instead relying solely upon the CAD test for a reference measure. The agreement between CAD and the anomaloscope has previously been found to be high [50], and while the anomaloscope is renowned for its accuracy in distinguishing between protan and deutan observers—frequently employed as a gold standard for this purpose—the parameters of the match have poor agreement with an individual's chromatic discrimination thresholds, the latter being more relevant for occupational environments [51,52].

The CVS test results align with the predicted binomial outcome of the test [31]. The probability of correctly identifying CVS stimuli is determined by two factors, an individual's chromatic sensitivity and the stimuli shown during the test. As the stimuli's chromatic signal strength is based solely on an individual's age, the probability of a correct response is based upon how close an individual's chromatic sensitivity is to the limit for their age. Individuals with high chromatic sensitivity for their age—the majority of normal trichromats tested—have the highest *relative* CVS stimulus strength and, hence, the highest probability of correctly identifying the stimuli (i.e., the stimulus strength used is twice as large as their threshold). Individuals with lower chromatic sensitivity for their age—those with color deficiency—have the lowest *relative* CVS stimulus strength and subsequently the lowest probability of correctly identifying the stimuli. As shown in Fig. 5B, the normal trichromat who failed the RG component of the CVS had the lowest *relative* CVS RG stimulus strength of all normal trichromats and hence had the smallest probability of a correct response for the stimulus strength employed in the CVS test. This is acceptable for occupational settings where failing the CVS indicates the need for further diagnostic testing.

A unique strength of the CVS test is that the signal strength displayed is proportional to the normal upper CAD limit for each participant's age. This property of the CVS allows the test to work with subjects of any age (from ~10 years of age). Additionally, when viewed from this relative frame of reference, the separation between the least affected deutan and protans is strikingly evident. No individual out of the 250 individuals with color deficiency assessed across all centers passed the CVS. Color deficient with the highest probability of passing the RG CVS are the least affected deutan, a small subsection of the deutan population. Such an outcome is advantageous and has benefits within occupational health domains.

Most participants with congenital CVD, ~78% of deutan and ~99% of protans, have RG CAD thresholds greater than or equal to 5 CAD units [32]. The outcome of the CVS was statistically equivalent across all participants within this threshold range. For relative stimulus strength, the participant's probability of correctly identifying the CVS stimuli is chance (50% for the 2AFC procedure) and, as with all CVS outcomes, is governed by a binomial distribution determined entirely by the probability of a correct response.

The 2AFC nature of the test carries advantages and limitations. The primary limitation is that there is a non-zero chance that an individual with CVD can pass the RG and YB components of the CVS. However, as one can observe from Table 2, such a limitation is present in current screening tests,

including the Ishihara pseudoisochromatic plates. This limitation is offset by the CVS test result being directly proportional to an individual's level of chromatic sensitivity. The probability of an individual with CVD passing the CVS is less than 0.1%. Individuals with the highest probability of passing are the least affected and have the highest chromatic sensitivity within the CVD group. The randomized nature of the CVS test presentation also presents an advantage by eliminating the possibility for an individual to memorize the test components.

5. CONCLUSION

A multicenter study was carried out to evaluate the outcome of the new CVS test in a large clinically relevant population containing normal trichromats and individuals with congenital CVD. Testing carried out by multiple examiners and across multiple locations revealed that the CVS has high sensitivity and specificity when carried out on spectrally calibrated visual displays. The outcome of the test achieves at least as good an outcome as the most stringent protocol for RG color vision screening using the Ishihara pseudoisochromatic plate test. The new CVS test is, however, much more efficient since almost all normal trichromats pass. When the CVS is employed in a two-step color assessment protocol, the number of applicants needing a full color assessment is much reduced. The CVS test offers the potential to rapidly and efficiently screen for congenital CVD in clinical and occupational settings and integrates cleanly into a “two-step” protocol for color vision assessment.

Funding. Colt Foundation; UK Civil Aviation Authority; Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación, European Regional Development Fund, European Union (Project PID2022-139056NB-I00).

Acknowledgment. We wish to thank and acknowledge the Colt Foundation and the UK Civil Aviation Authority for their financial support. We also wish to thank Stuart Mitchell, who has supported us from the UK CAA. Some of this work was carried out under Project PID2022-139056NB-I00, supported by MICIU/AEI/10.13039/501100011033 and ERDF, EU.

Disclosures. The authors declare no conflicts of interest. Professor John Barbur designed all the Advanced Vision and Optometric Tests (AVOT), including the CAD test, but no longer has any commercial interests in these tests or current conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

REFERENCES

1. M. O'Neill-Biba, S. Sivaprasad, M. Rodriguez-Carmona, *et al.*, “Loss of chromatic sensitivity in AMD and diabetes: a comparative study,” *Ophthalmol. Physiol. Opt.* **30**, 705–716 (2010).
2. C. Feitosa-Santana, G. V. Paramei, M. Nishi, *et al.*, “Color vision impairment in type 2 diabetes assessed by the D-15d test and the Cambridge colour test,” *Ophthalmol. Physiol. Opt.* **30**, 717–723 (2010).
3. M. P. Simunovic, “Acquired color vision deficiency,” *Surv. Ophthalmol.* **61**, 132–155 (2016).
4. A. Stockman, “Report on colour vision standards for STU officers,” Rep. 6.2 (College of Policing, 2018), <https://assets.college.police.uk/s3fs-public/2022-02/FOIA-2021-153.pdf>.
5. H. C. Walkey, A. Hurden, I. R. Moorhead, *et al.*, “Effective contrast of colored stimuli in the mesopic range: a metric for perceived contrast

- based on achromatic luminance contrast," *J. Opt. Soc. Am. A* **22**, 17–28 (2005).
6. J. Barbur, M. Rodríguez-Carmona, S. Evans, *et al.*, "Minimum Colour Vision Requirements for Professional Flight Crew," CAA 80 (2009).
 7. J. L. Barbur, M. Rodríguez-Carmona, J. Hickey, *et al.*, "Analysis of European colour vision certification requirements for air traffic control officers," Civil Aviation Authority CAP 1429 (2016).
 8. P. Sumner and J. D. Mollon, "Chromaticity as a signal of ripeness in fruits taken by primates," *J. Exp. Biol.* **203**, 1987–2000 (2000).
 9. B. C. Regan, C. Juliot, B. Simmen, *et al.*, "Fruits, foliage and the evolution of primate colour vision," *Philos. Trans. R. Soc. London B* **356**, 229–283 (2001).
 10. J. D. Mollon and L. R. Cavonius, "The Lagerlunda collision and the introduction of color vision testing," *Surv. Ophthalmol.* **57**, 178–194 (2012).
 11. M. Neitz and J. Neitz, "Molecular genetics of color vision and color vision defects," *Arch. Ophthalmol.* **118**, 691–700 (2000).
 12. J. Neitz and M. Neitz, "The genetics of normal and defective color vision," *Vision Res.* **51**, 633–651 (2011).
 13. J. Birch, "Worldwide prevalence of red-green color deficiency," *J. Opt. Soc. Am. A* **29**, 313–320 (2012).
 14. L. N. Went and N. Pronk, "The genetics of tritan disturbances," *Hum. Genet.* **69**, 255–262 (1985).
 15. D. M. Goodman and E. H. Livingston, "Screening tests," *J. Am. Med. Assoc.* **309**, 1185 (2013).
 16. P. A. Davison and G. Scanlon, "Comparative analysis of four color vision screening tests benchmarked by anomaloscopy for detection and investigation of protanomaly and deuteranomaly," *Color Res. Appl.* **49**, 474–485 (2024).
 17. S. J. Dain, "Clinical colour vision tests," *Clin. Exp. Optom.* **87**, 276–293 (2004).
 18. D. G. Altman, "Practical statistics for medical research," *Stat. Med.* **10**, 1635–1636 (1991).
 19. J. Birch and M. Rodríguez-Carmona, "Occupational color vision standards: new prospects," *J. Opt. Soc. Am. A* **31**, A55–A59 (2014).
 20. J. Birch, "Classification of anomalous trichromatism with the Nagel anomaloscope," in *Colour Vision Deficiencies XI*, B. Drum, ed. (Kluwer Academic, 1993), Vol. **56**, pp. 19–24.
 21. S. J. Dain, A. Casolin, J. Long, *et al.*, "Color vision and the railways: part 1. The Railway LED lantern test," *Optom. Vis. Sci.* **92**, 138–146 (2015).
 22. T. H. Margrain, J. Birch, and C. G. Owen, "Colour vision requirements of firefighters," *Occup. Med.* **46**, 114–124 (1996).
 23. M. Rodríguez-Carmona, B. E. W. Evans, and J. L. Barbur, "Color vision assessment-2: color assessment outcomes using single and multi-test protocols," *Color Res. Appl.* **46**, 21–32 (2021).
 24. V. F. Koefoed, T. Miles, J. B. Cason, *et al.*, "Colour vision classification—comparing CAD and CIE 143:2001 International recommendations for colour vision requirements in transport," *Acta Ophthalmol.* **98**, 726–735 (2020).
 25. M. Rodríguez-Carmona and J. L. Barbur, "Variability in normal and defective colour vision," in *Colour Design*, J. Best, ed., 2nd ed. (Woodhead, 2017), pp. 43–97.
 26. P. B. M. Thomas and J. D. Mollon, "Modelling the Rayleigh match," *Vis. Neurosci.* **21**, 477–482 (2004).
 27. J. L. Barbur, M. Rodríguez-Carmona, J. A. Harlow, *et al.*, "A study of unusual Rayleigh matches in deutan deficiency," *Vis. Neurosci.* **25**, 507–516 (2008).
 28. J. Birch, *Diagnosis of Defective Colour Vision* (Butterworth-Heinemann, 2001).
 29. M. Rodríguez-Carmona, J. Harlow, G. Walker, *et al.*, "The variability of normal trichromatic vision and the establishment of the 'Normal' range," in *Proceedings of 10th Congress of the International Colour Association* (International Colour Association, 2005), pp. 979–982.
 30. J. D. Mollon and J. P. Reffin, "A computer-controlled colour vision test that combines the principles of Chibret and of Stilling," *Proc. Physiol. Soc.* **414**, 20 (1989).
 31. J. L. Barbur, M. Rodríguez-Carmona, and B. E. W. Evans, "Color vision assessment-3. An efficient, two-step, color assessment protocol," *Color Res. Appl.* **46**, 33–45 (2021).
 32. J. L. Barbur and M. Rodríguez-Carmona, "Colour vision requirements in visually demanding occupations," *Br. Med. Bull.* **122**, 51–77 (2017).
 33. M. Rodríguez-Carmona, M. O'Neill-Biba, and J. L. Barbur, "Assessing the severity of color vision loss with implications for aviation and other occupational environments," *Aviat. Space Environ. Med.* **83**, 19–29 (2012).
 34. M. Rodríguez-Carmona and J. L. Barbur, "Variability in normal and defective colour vision," in *Colour Design*, J. Best, ed. (Woodhead, 2012), pp. 43–97.
 35. J. Seshadri, J. Christensen, V. Lakshminarayanan, *et al.*, "Evaluation of the new web-based 'Colour Assessment and Diagnosis' test," *Optom. Vis. Sci.* **82**, 882–885 (2005).
 36. A. French, K. Rose, E. Cornell, *et al.*, "The evolution of colour vision testing," *Aust. Orthopt. J.* **40**, 7–15 (2008).
 37. A. Fanlo Zarazaga, J. Gutiérrez Vásquez, and V. Pueyo Royo, "Review of the main colour vision clinical assessment tests," *Arch. Soc. Esp. Oftalmol.* **94**, 25–32 (2019).
 38. T. Carter and J. Barbur, "Colour vision assessment for maritime navigational lookout: review for UK Maritime and Coastguard Agency (MCA)" (Maritime and Coastguard Agency, 2015), <https://www.gov.uk/government/publications/colour-vision-assessment-for-maritime-navigation-lookout>.
 39. C. L. Meinert and S. Tonascia, "Single-center versus multicenter trials," in *Clinical Trials: Design, Conduct and Analysis* (1986), pp. 23–29.
 40. L. J. Appel, "A primer on the design, conduct, and interpretation of clinical trials," *Clin. J. Am. Soc. Nephrol.* **1**, 1360–1367 (2006).
 41. R. M. Lucas, A. L. Ponsonby, A. J. McMichael, *et al.*, "Observational analytic studies in multiple sclerosis: controlling bias through study design and conduct. The Australian multicentre study of environment and immune function," *Mult. Scler.* **13**, 827–839 (2007).
 42. J. Birch, "Efficiency of the Ishihara test for identifying red-green colour deficiency," *Ophthalm. Physiol. Opt.* **17**, 403–408 (1997).
 43. L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Stat. Sci.* **16**, 101–133 (2001).
 44. J. L. Hintze and R. D. Nelson, "Violin plots: a box plot-density trace synergism," *Am. Stat.* **52**, 181–184 (1998).
 45. S. Seifirad and L. Alquran, "The bigger, the better? When multicenter clinical trials and meta-analyses do not work," *Curr. Med. Res. Opin.* **37**, 321–326 (2021).
 46. S. Arnegard, R. C. Baraas, J. Neitz, *et al.*, "Limitation of standard pseudoisochromatic plates in identifying colour vision deficiencies when compared with genetic testing," *Acta Ophthalmol.* **100**, 805–812 (2022).
 47. G. H. M. Waaler, "Über die Erblichkeitsverhältnisse der verschiedenen Arten von angeborener Rotgrünblindheit," *Z. Indukt. Abstamm. Vererbungslehre.* **45**, 279–333 (1927).
 48. A. Cheng, D. Kessler, R. Mackinnon, *et al.*, "Conducting multicenter research in healthcare simulation: Lessons learned from the INSPIRE network," *Adv. Simul.* **2**, 6 (2017).
 49. M. K. Das, "Multicenter studies: relevance, design and implementation," *Indian Pediatr.* **59**, 571–579 (2022).
 50. B. E. W. Evans, M. Rodríguez-Carmona, and J. L. Barbur, "Color vision assessment-1: visual signals that affect the results of the Farnsworth D-15 test," *Color Res. Appl.* **46**, 7–20 (2021).
 51. W. D. Wright, *Researches on Normal & Defective Colour Vision* (Kimpton, 1946).
 52. L. M. Hurvich, "Color vision deficiencies," in *Visual Psychophysics* (Springer-Verlag, 1972), Vol. **704**, pp. 582–624.