Color Vision and Performance on Color-Coded Cockpit Displays

James P. Gaska; Steven T. Wright; Marc D. Winterbottom; Steven C. Hadley

INTRODUCTION: Although there are numerous studies that demonstrate that color vision deficient (CVD) individuals perform less well than color vision normal (CVN) individuals in tasks that require discrimination or identification of colored stimuli, there remains a need to quantify the relationship between the type and severity of CVD and performance on operationally relevant tasks.

METHODS: Participants were classified as CVN ($N = 45$) or CVD ($N = 49$) using the Rabin cone contrast test, which is the standard color vision screening test used by the United States Air Force. In the color condition, test images that were representative of the size, shape, and color of symbols and lines used on fifth-generation fighter aircraft displays were used to measure operational performance. In the achromatic condition, all symbols and lines had the same chromaticity but differed in luminance. Subjects were asked to locate and discriminate between friend vs. foe symbols (red vs. green, or brighter vs. dimmer) while speed and accuracy were recorded.

RESULTS: Increasing color deficiency was associated with decreasing speed and accuracy for the color condition ($R^2 > 0.2$), but not for the achromatic condition. Mean differences between CVN and CVD individuals showed the same pattern.

DISCUSSION: Although lower CCT scores are clearly associated with lower performance in color related tasks, the magnitude of the performance loss was relatively small and there were multiple examples of high-performing CVD individuals who had higher operational scores than low-performing CVN individuals.

KEYWORDS: color vision, operational performance, color displays.

Gaska JP, Wright ST, Winterbottom MD, Hadley SC. *Color vision and performance on color-coded cockpit displays.* Aerosp Med Hum Perform. 2016; 87(11):921–927.

In aviation, there is an expectation of rigorous safety standards because of the increased potential for injury or fatalities due to mishaps as well as their high economic costs. Given the 5–8% incidence of color vision deficiency in males[6,9] and the overwhelming evidence that color vision deficient (CVD) observers perform less well than color vision normal (CVN) observers in color discrimination tasks,[3,4,15] it is readily apparent that selection and retention standards for aircrew related to color vision are necessary.

The most straightforward way to determine if a color vision deficit will reduce an individual's ability to perform a task is to measure the individual's performance in the operational task. This "task equals standards" method has a long history. Lantern tests that simulate signal lights have been used for the selection and retention of seafarers and railway employees since the 1890s.[5] While operation-based standards tests have the benefit of simplicity, they have two major problems when used to generate color vision standards. The first is that while signal lights are still used to increase aviation safety, they represent only a small proportion of color signals in the modern aviation environment, and a set of standards tests that simulated all of these conditions would be unmanageable and inefficient. The second is that operation-based standards tests such as lantern tests can result in inconsistent metrics for CVD observers.[14]

The Operational Based Vision Assessment (OBVA) laboratory takes an alternative approach. We start with standard tests designed to characterize the magnitude and type of color deficiency using procedures designed to maximize the validity, reliability, and efficiency of the test. We then measure operational performance in high fidelity simulations to quantify

From the U.S. Air Force School of Aerospace Medicine, Aeromedical Research Department, Wright-Patterson AFB, OH.

This manuscript was received for review in April 2016. It was accepted for publication in July 2016.

Address correspondence to: James P. Gaska, USAFSAM/FECO, 2215 First Street, Wright-Patterson AFB, OH 45433-7913; james.gaska.1@us.af.mil.

DOI: 10.3357/AMHP.4630.2016

the relationship between operational performance metrics and the standards test scores.[12]

One of the biggest drivers of OBVA research is technological change. Many of the selection and retention standards currently used in the U.S. Air Force were established in a time when electronic color displays were not part of the operational environment. In this first in a series of studies, we examine the relationship between Rabin Cone Contrast Test (CCT) scores (the current test used by the U.S. Air Force for color vision screening) and the speed and accuracy of target localization on a simulated color-coded situational awareness (SA) screen representative of those used on fifth-generation fighter aircraft. This color coded SA display task was selected by a panel of experts as a critical, operationally-relevant aviation task involving color. The panel consisted of experienced pilots from several different USAF aircraft, aeromedical personnel representing the U.S. Air Force, U.S. Army, and U.S. Navy, and vision scientists from the Air Force, NASA Ames Research Center, and visiting scientists from Canada and the Netherlands.[12]

METHODS

Subjects

There were 94 subjects who participated in this study (75 men and 19 women). The study protocol was approved in advance by the Air Force Research Laboratory Institutional Review Board. Each subject provided written informed consent before participating.

The commercially available Rabin CCT[10,11] (Innova Systems, Inc., Moorestown, NJ) was used to determine the type and magnitude of color deficiency. The CCT measures contrast threshold for three stimuli designed to selectively stimulate the long-wavelength sensitive (L), middle-wavelength sensitive (M), and short-wavelength sensitive (S) cones of an average color normal observer. For each cone type, the CCT transforms the contrast thresholds into a score that that ranges from 100 (normal) to 0 (severely color deficient).

The test was administered to each eye separately and averaged. If the average score for any cone type was less than 75, the observer was classified as color deficient. L-cone or M-cone color deficient observers were classified as protans if the L-cone score was less than the M-cone score or deutans if the M-cone score was less than the L-cone score. Using the above procedure, 45 of the participants were classified as color normal and 49 were classified as color deficient (18 protans and 31 deutans). Five of the female observers were color deficient. None of the observers had S-cone scores below 75 (tritan). The proportion of CVD individuals in our sample (0.52) was clearly much higher than that found in the general population (0.05–0.08). As stated above, the purpose of this research is to examine the impact of color deficiency on complex, operationally-relevant tasks, and not to examine the incidence of color deficiency in the population, which has been thoroughly documented in other research.[6,9,11] Thus, we sought the participation of color deficient individuals to more efficiently examine how color deficiency might affect performance.

Equipment

Test images for the simulation task were displayed on a liquid crystal display (LCD) monitor (EIZO FlexScan SX2761W) that was calibrated using a spot colorimeter (Minolta CS-200). Image placement and timing were controlled and keyboard responses collected by custom programs written using the Matlab programing language supplemented by the Psychtoolbox.[2,8]

This experiment simulated the situation awareness (SA) display of fifth-generation fighter aircraft. The SA display uses symbology to depict the location, aspect, and motion of entities in the pilot's local airspace. In the aircraft, this information is presented on an LCD monitor, or multifunction display (MFD). The size, shape, chromaticity, and luminance of symbols used in this simulation were representative of those used on an SA display. We developed a Matlab application that allowed a user to select a symbol and position and rotate it on the monitor. A pilot subject matter expert (F-16 pilot with over 4000 h) who was familiar with fifth-generation fighter aircraft SA display symbology used this application to create a set of operationally representative test images. There were 30 test images generated in this manner. We then rotated the test images over 4 headings (north, south, east, and west) to generate 120 different test images.

Each test image covered 762 $\times$ 762 pixels on the display screen and subtended 14.2 $\times$ 14.2° at the 91-cm (36-in) viewing distance used in this experiment. Aircraft symbols had a width of about 19 pixels and subtended approximately 0.36°. The 36-in viewing distance used here is larger than the viewing distance in fighter aircraft (approximately 25 in). This larger distance was used because the pixel pitch of the test monitor was slightly larger than that of the aircraft MFD.

In the SA display, the shape of a symbol is related to aircraft type (fighter, bomber, etc.), whereas the color is used to represent friend, foe, or unknown. It is important to note that, because color and shape cues are independent, friend and foe aircraft of the same type will differ only in color. In this case, green is used to represent friendly aircraft, while red represents foe. The straight lines in the figure are used to demarcate different areas of the battle space, for example, a national boundary line or a hostile boundary line.

Procedure

The observers were shown a test image accompanied by a trial-start tone and asked to push the up-arrow on the keyboard for a "yes" answer and the down-arrow on the keyboard for a "no" answer. Participants were provided practice trials until they understood the task and were proficient. The test stimulus remained on until the observer responded, and the next trial was initiated 0.5 s after the observer's response. Correct/incorrect feedback was provided by two different tones. There were 3 blocks consisting of 50 trials per block collected for each experimental condition. The observer initiated a block of trials by pressing the space bar on the keyboard. Observers were encouraged to relax and rest their eyes before initiating a new block of trials. The test images used in each 50-trial block were pseudo-randomly selected from the 120 images in the library.

Two conditions were investigated. In the color condition (**Fig. 1**, top row), test images representative of the size, shape, and chromaticity of symbols and lines used on a fifth-generation fighter aircraft SA display were used to measure operational performance. Participants were asked to determine if a red (foe) symbol was located between red and magenta lines. In Fig. 1 (top left) the correct answer is yes and in Fig. 1 (top right) the correct answer is no (i.e., is an enemy aircraft in the "weapons engagement zone").

In the achromatic condition (Fig. 1, bottom row), all symbols and lines had the same chromaticity ($x = 0.37$, $y = 0.36$) but differed in luminance. The formerly red symbols and lines were then set to a luminance value of $171\ \text{cd} \cdot \text{m}^{-2}$, creating the brightest symbol and line in the test image. The formerly magenta line was set to $86\ \text{cd} \cdot \text{m}^{-2}$, creating the second brightest line in the test image. The luminance values of all other symbols and lines were set to $42.8\ \text{cd} \cdot \text{m}^{-2}$, creating the least bright symbology. These levels were chosen to provide highly discriminable perceived lightness levels. The spatial properties of the images were the same as those used in the color condition. Participants were asked to determine if the brightest symbol was located between the brightest and second brightest lines. In Fig. 1 (bottom left), the correct answer is yes and in Fig. 1 (bottom right) the correct answer is no.

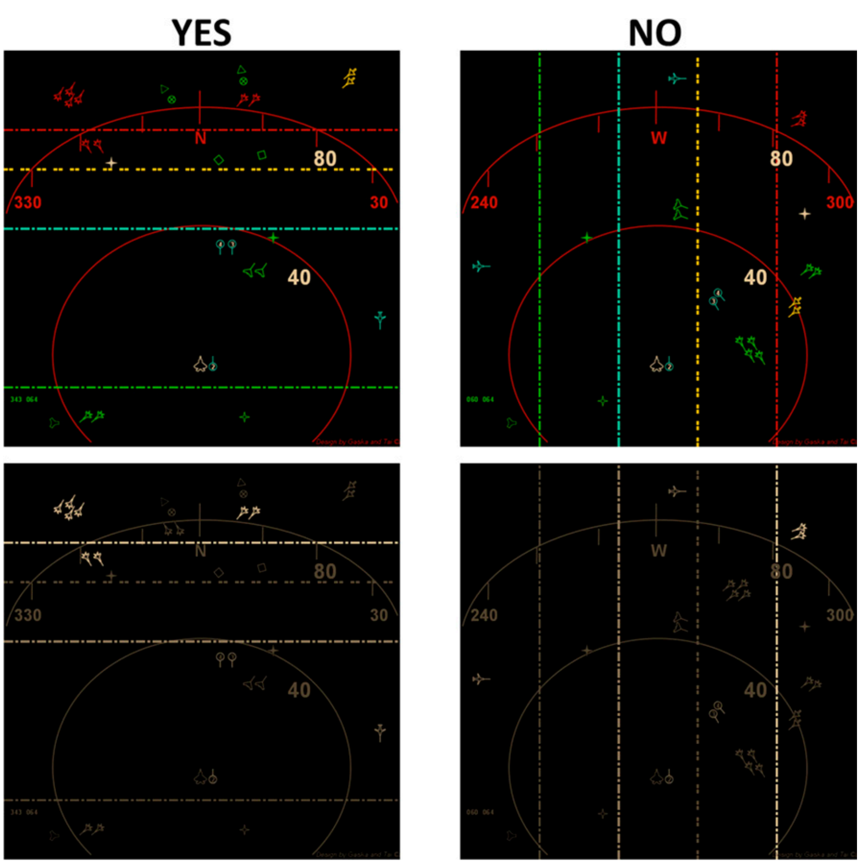

Fig. 1. Examples of test images used in the experiment. Color condition images are shown in the top row and achromatic images are shown in the bottom row. The correct response was "yes" for left-column images and "no" for the right-column images. See text for details.

RESULTS

Each individual scatterplot in **Fig. 2** depicts the CCT scores of color normal and color-deficient observers against a particular operational performance metric. Thus each scatterplot depicts how operational performance is related to color discrimination as measured by the CCT. Focusing on the horizontal axis (CCT scores), it is clear the large gap between the CCT scores in the 75 to 90 range allows us to easily distinguish between CVN and CVD individuals. In this sample, the mean CCT score for the CVN sample was 97.5 and the minimum score was 87.5. This low score is well above the score of 75 that is used to classify CVN and CVD individuals in the CCT test and replicates the previous studies that show that the CCT can reliably classify CVN and CVD individuals.[10,11]

Using this classification, the white circles depict the results from color normal individuals and solid black (top row) and solid grey (bottom row) symbols depict results from protans and deutans, respectively. The plots in the top row show CVN and protan data. Shown in the bottom row are data for CVN and deutan individuals.

In addition to classification, the CCT provides a score that estimates the degree of color vision deficiency and we wanted to examine whether decreased color discrimination capabilities (lower CCT scores) are linearly related to reduced operational performance. To quantify this relationship we computed bivariate correlation statistics and report the coefficient of determination (R^2) as well as the probability that this coefficient could result from random variation when the true correlation is zero. The bivariate-correlation statistics were computed using both the CVN and CVD individuals and are shown at the top of each scatter plot. The statistics were computed separately for the CVN-plus-protan and CVN-plus-deutan datasets. In addition, **Table I** summarizes the results from a set of *t*-tests that estimate the statistical significance of the mean differences between CVN and CVD subjects.

Speed and accuracy of operational decisions are important factors for mission success. In this section we examine how operational task accuracy (proportion correct), speed (1/reaction time), and their product are related to the degree of color deficiency. The product of speed and accuracy, called throughput, was used because previous research has demonstrated that throughput is a more sensitive measure of performance than speed or accuracy alone.[16]

Fig. 2 shows the results from the color condition in which the three performance metrics, accuracy, speed, and throughput, are grouped by column. It can be seen that the coefficient of determination (R^2) values

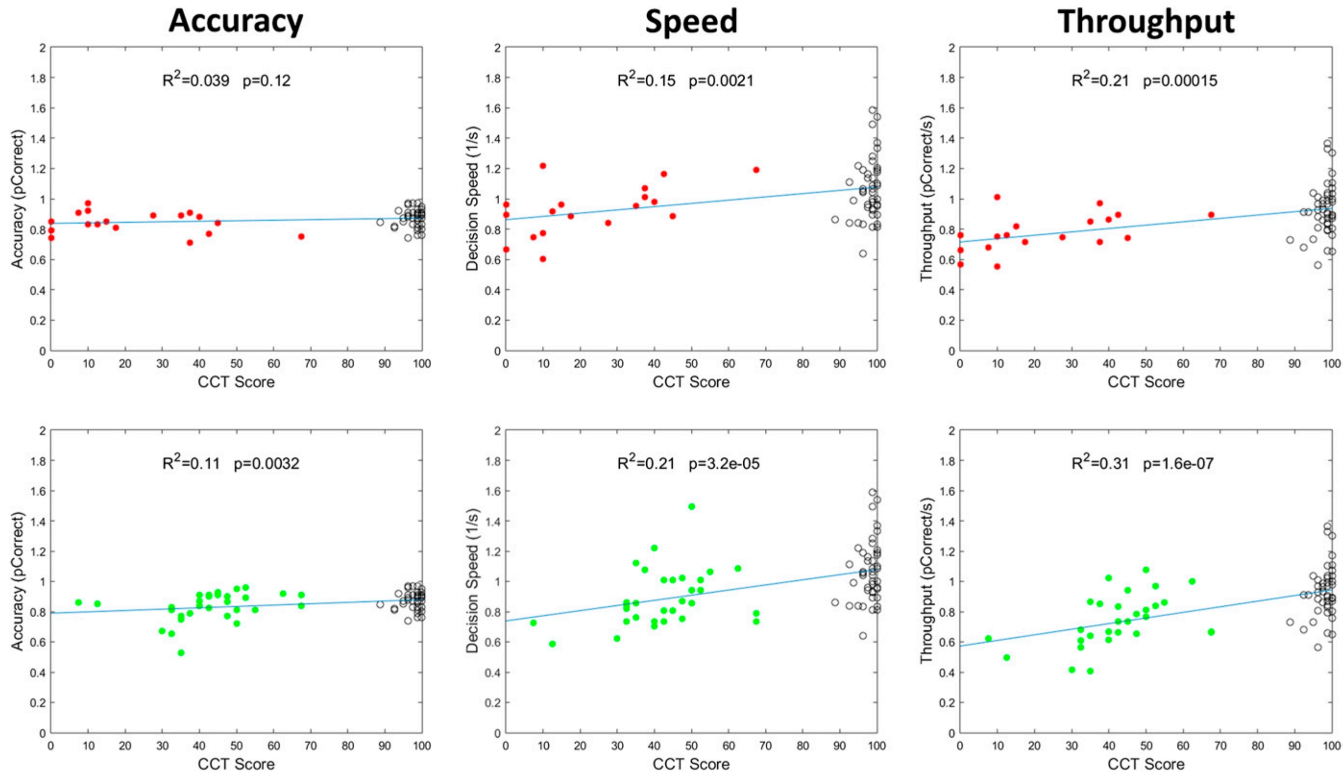

Fig. 2. Accuracy, speed, and throughput as a function of cone contrast test (CCT) scores for the color condition. White circles depict the results from color normal individuals. Solid black and solid grey symbols (red and green, respectively, in the online article) depict results from protans and deutans, respectively. The three metrics, accuracy, speed, and throughput, are grouped by column.

are smallest for the accuracy metric, intermediate for the speed metric, and highest for the throughput metric. Note that the regression *P*-values for protan-normal using the accuracy metric were not statistically significant at the 0.05 alpha level, although the rest of the results were highly significant ($P < 0.005$). Finally, the R^2 values are larger and *P*-values are smaller for the deutan-normal analysis compared to the protan-normal analysis.

These results demonstrate that lower CCT scores (decreased color discrimination capabilities) are clearly associated with lower performance on an SA task that simulates fifth-generation fighter cockpit symbology. One exception is that protan-normal vs. accuracy correlations were not statistically significant on this operational color discrimination task.

Fig. 3 depicts throughput as a function of CCT scores for the achromatic condition, color condition, and their ratio (i.e., the color throughput divided by the achromatic throughput). As in Fig. 2, white circles depict the results from color normal individuals and solid black and solid grey symbols depict results from protans and deutans, respectively. It can be seen that that the R^2 values are small and not statistically significant for the achromatic condition (left column). To aid comparison, the throughput results for the color condition shown in Fig. 2 are carried over to this figure (center column).

Throughput values for the achromatic condition are, in general, higher than those in the color condition, particularly for color deficient individuals. For color normal individuals, the mean achromatic throughput is 1.21, the mean color throughput is 0.92, and the mean throughput ratio is 0.77. This suggests that the large luminance differences in the achromatic condition provided a more salient cue than the color and luminance differences of the red and green symbols used in the color condition.

The spread of the color normal sample values is greatly reduced for the throughput ratio metric (shown in Fig. 3, right column). For color normal individuals, the standard deviation for the achromatic, color, and ratio conditions are 0.21, 0.17, and 0.09, respectively. The reduced spread of the throughput ratio condition results in the highest correlation values, and were highly significant. For the deutan-normal data, a linear fit to CCT scores accounts for 54% of the throughput ratio

Table I. Results of *t*-tests for Mean Differences Between Color Normal (CVN) and Color Deficient (CVD) Observer Samples.*

		PERFORMANCE METRIC		
		ACCURACY	SPEED	THROUGHPUT
TEST CONDITION	CVN-CVD TYPE	DIFFERENCE, *P*	DIFFERENCE, *P*	DIFFERENCE, *P*
Color	Normal - Protan	0.033, 0.057	0.137, 0.013	0.154, $< 10^{-3}$
	Normal - Deutan	0.048, 0.018	0.175, $< 10^{-3}$	0.190, $< 10^{-3}$
Achromatic	Normal - Protan	0.013, 0.23	0.043, 0.24	−0.031, 0.54
	Normal - Deutan	−0.008, 0.26	0.014, 0.65	−0.026, 0.56

* For the normal -protan test, df = 61, and for the normal-deutan test, df = 74.

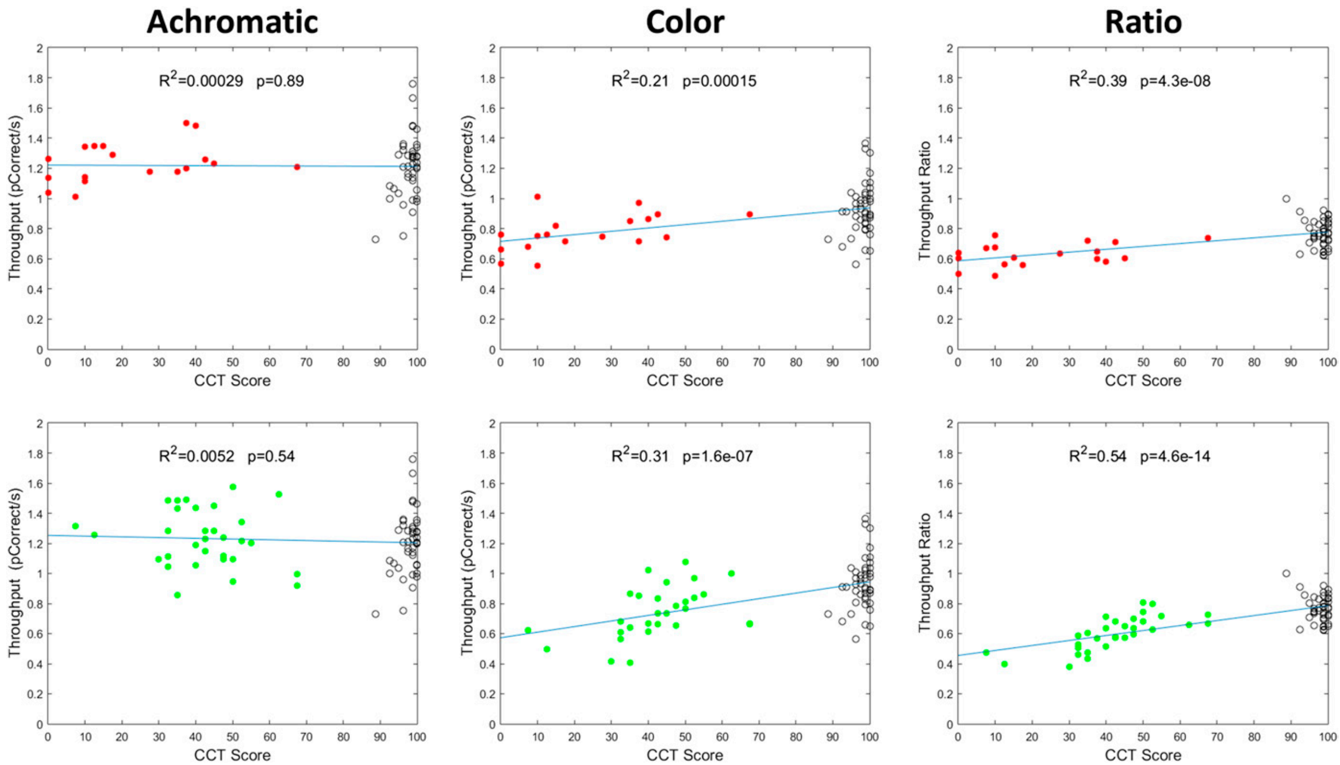

Fig. 3. Throughput as a function of cone contrast test (CCT) scores for the achromatic condition, color condition, and their ratio (color/achromatic). White circles depict the results from color normal individuals. Solid black and solid grey symbols (red and green, respectively, in the online article) depict results from protans and deutans, respectively.

variance. A potential explanation of this finding is provided in the discussion section.

Table I provides the results from *t*-tests of mean differences between CVN and CVD individuals. The results replicate the correlation analysis. For the color condition, statistical significance is marginal for the accuracy metric, significant for the speed metric, and highly significant for the throughput metric. The differences were not significant for any metric in the achromatic condition.

DISCUSSION

As stated in the introduction, the purpose of the OBVA research laboratory is to generate data that can be used to establish data-driven selection and retention standards. The data from the color condition demonstrate that lower CCT scores are associated with lower performance on a symbol localization and discrimination task that simulates the color-coded SA display of fifth-generation fighter aircraft. However, association strength was lowest when accuracy alone was used, and the accuracy performance metric was not statistically significant for the protan-normal analysis. Thus, accuracy metrics alone may underestimate the performance loss in CVD individuals. The strongest associations between color vision capabilities and operational performance were demonstrated when the speed-accuracy product, or throughput, was used as the performance metric. This is consistent with previous work by Thorne, who demonstrated that throughput is a more sensitive measure of performance than speed or accuracy alone.[16]

Whereas achromatic throughput was not associated with CCT scores, the color/achromatic throughput ratio metric revealed a very strong association. The most likely explanation of this finding is that the achromatic throughput metric measures individual differences in task competency unrelated to color discrimination capabilities, such as luminance discrimination, overall processing speed, or sensory-motor transport times. By normalizing the color throughput by the achromatic throughput, the contribution of these individual differences is reduced in the resulting ratio metric. This can be readily appreciated by comparing the range of scores for normal observers in the color, achromatic, and ratio conditions shown in Fig. 3. This reduction of individual differences in noncolor related capabilities generates a metric that better isolates an individual's ability to use chromaticity information in task performance. The tight correlations also suggest that CCT scores may be highly predictive of tasks that rely primarily on chromaticity discrimination, even in a relatively complex, operationally-relevant SA display task such as the one used here, involving not just simple color discrimination, but visual search within a complex display with many other elements.

In this study we measured performance in an achromatic and color task and found large individual differences for each task. We have previously shown that an observer's performance in an operationally-relevant air-to-air target identification task is closely related to U.S. Air Force standards tests that measure acuity and contrast sensitivity.[12] Here we studied the relationship between CCT scores and an operationally-relevant color-coded SA task.

While the data presented in this research show that, on average, CVD individuals are slower and less accurate than CVN individuals on this operationally-relevant task, it also demonstrates that the current selection standard (a score of 75 or greater on the CCT test) does not reliably discriminate between high or low operational performance on the color task. In particular, some CVD subjects could perform the SA task acceptably well, and some CVN subjects performed less well. This is an example of a binary classification test, where performance is quantified in a clinical or operational context by using a sensitivity metric (true positive rate) and a specificity metric (true negative rate). In this study, the test is used to identify color-based performance impairment, so sensitivity represents the proportion of CVD individuals classified as low performers on the SA task, and specificity represents the proportion of CVN individuals classified as high performers. In a perfect test, both sensitivity and specificity would be 1.0. For example, when a binary classification test is used to compare CCT and anomaloscope classification of color vision deficiency, both specificity and sensitivity are near 1.0.[10,11]

Fig. 4 illustrates the relatively poor binary classification afforded by CCT scores in predicting SA performance by deutans and CVN individuals in the color condition; note (see Fig. 2) that this was the combination of group and task that yielded the highest operational performance correlation. The correlation analysis and mean differences were highly significant ($P < 10^{-4}$).

As previously discussed, the CCT scores provide a clear and valid classification of CVN and CVD individuals. However, due to overlap of the performance metrics of CVN and CVD individuals, there is no performance criterion that can cleanly separate high and low performers. In Fig. 4, the dashed line marks a performance criterion of 2 SD below the CVN throughput mean. In this case, the criterion yields a reasonably high specificity of 0.94 but a poor sensitivity of 0.089. We can increase sensitivity by increasing the value of the performance criterion,

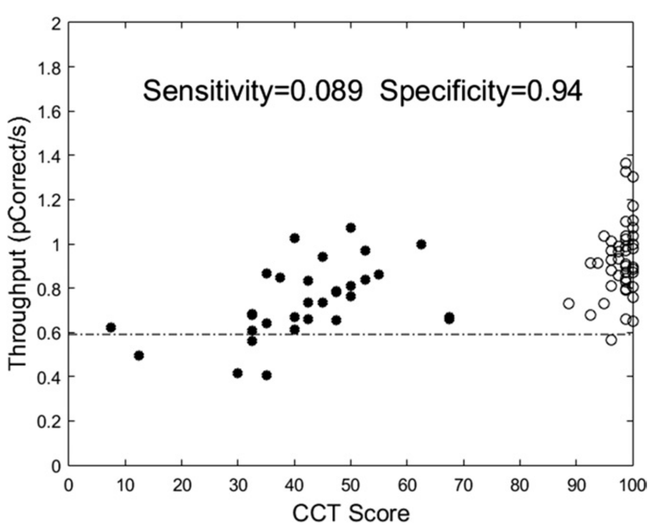

Fig. 4. Sensitivity and specificity for the throughput ratio of deutan (black symbols) and normal (white symbols) observers in the color condition.

but this will reduce specificity. In short, because the variability of the performance metric is large relative to differences between the mean of the CVN and CVD samples, discriminability is low. The d' (d-prime) metric, which is computed as the difference between the means of CVD and CVN divided by their standard deviation, is often used to quantify discriminability. Using the procedures outlined in "Psychophysics: A Practical Introduction,"[7] we computed a d' of 0.25 for this dataset, which is indicative of low discriminability.

It is important to note that different operational tasks may demonstrate higher discriminability and there are several factors that may have contributed to the low discriminability in this dataset. First, this task was chosen because discrimination between friend and foe is important in an operational environment, and is representative of the actual colors used on the real aircraft display (i.e., red vs. green). However, it has been shown[13] that discrimination of yellow-green cockpit symbology results in larger mean differences between CVN and CVD observers than red-green pairs. Second, the green symbols on SA displays (and in this experiment) are more luminous than red symbols. Because luminance differences can be used by color deficient individuals to discriminate between targets, other studies that measured performance using simulated operational stimuli have randomized the luminance of the targets over trials to eliminate this cue.[1] While randomizing the luminance of targets may increase the performance differences between CVN and CVD individuals, this does not happen on the SA display and we believe that our method, which replicates the fixed luminance ratio of symbols, provides a better estimate of operational performance. Finally, data in this experiment were collected in a darkened room. Thus, luminance and chromatic contrast were always high. While this experimental environment may replicate a more controlled operational environment, such as a remotely piloted aircraft control station, contrast on a cockpit display in a fighter aircraft can be greatly reduced by sunlight falling on the screen. We are currently investigating if reduced contrast increases the differences in operational performance between CVN vs. CVD individuals.

In summary, the research presented here demonstrates that lower CCT scores are clearly associated with decreased performance. However, the magnitude of the performance loss was relatively small and there were multiple examples of high-performing CVD individuals who had higher operational scores than low-performing CVN individuals. Under ideal, high-contrast viewing conditions, the performance of CVD individuals on a color discrimination task representative of a fifth-generation fighter cockpit situation awareness display cannot be clearly distinguished from that of color normal individuals.

ACKNOWLEDGMENTS

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the U.S. Air Force, the Department of Defense, or the U.S. Government. The authors would like to thank John White and Chi-Feng Tai for their technical help in designing the operational task, and all the members of the OBVA Integrated Product Team that supported this research, and contributed to the selection of the operational color task described in the research.

Authors and affiliations: James P. Gaska, Ph.D., Steven T. Wright, O.D., Marc D. Winterbottom, Ph.D., and Steven C. Hadley, M.D., U.S. Air Force School of Aerospace Medicine, Aeromedical Research Department, Wright-Patterson AFB, OH.

REFERENCES

1. Barbur J, Rodriguez-Carmona M, Evans S, Milburn N. Minimum color vision requirements for professional flight crew, part III: recommendations for new color vision standards. Washington (DC): Federal Aviation Administration, Office of Aerospace Medicine; 2009. DOT/FAA/AM-09/11.
2. Brainard DH. The psychophysics toolbox. Spat Vis. 1997; 10(4):433–436.
3. Cole BL. Does defective colour vision really matter? In: Drum B, editor. Colour vision deficiencies XI. Documenta Ophthalmologica Proceedings Series. 1993; 56:67–86.
4. Cole BL. The handicap of abnormal color vision. Clin Exp Optom. 2004; 87(4–5):258–275.
5. Cole BL, Vingrys AJ. A survey and evaluation of lantern tests of color vision. Am J Optom Physiol Opt. 1982; 59(4):346–374.
6. Delpero WT, O'Neill H, Casson E, Hovis J. Aviation-relevant epidemiology of color vision deficiency. Aviat Space Environ Med. 2005; 76(2):127–133.
7. Kingdom FA, Prins N. Psychophysics: a practical introduction. London, UK: Academic Press Elsevier; 2010.
8. Kleiner M, Brainard D, Pelli D. What's new in Psychtoolbox-3? [Abstract]. In: Abstract Supplement of 30th European Conference on Visual Perception; 2007 Aug. 27–31; Arezzo, Italy. Perception. 2007; 36(Suppl):14.
9. Neitz J, Neitz M. The genetics of normal and defective color vision. Vision Res. 2011; 51(7):633–651.
10. Rabin J. Cone-specific measures of human color vision. Invest Ophthalmol Vis Sci. 1996; 37(13):2771–2774.
11. Rabin J, Gooch J, Ivan D. Rapid quantification of color vision: the cone contrast test. Invest Ophthalmol Vis Sci. 2011; 52(2):816–820.
12. Rader J. Operational based vision assessment. Wright-Patterson AFB (OH): U.S. Air Force School of Aerospace Medicine; 2014. Technical Report AFRL-SA-WP-TR-2014-0005. [Accessed 1 Mar. 2015]. Available from: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA605857.
13. Reddix M, Kirkendall C, Gao H, O'Donnell K, Williams H, et al. Assessment of color vision screening tests for U.S. Navy special duty occupations [Abstract]. Aviat Space Environ Med. 2014; 85(3):299.
14. Squire TJ, Rodriguez-Carmona M, Evans AD, Barbur JL. Color vision tests for aviation: comparison of the anomaloscope and three lantern types. Aviat Space Environ Med. 2005; 76(5):421–429.
15. Steward JM, Cole BL. What do color vision defectives say about everyday tasks? Optom Vis Sci. 1989; 66(5):288–295.
16. Thorne DR. Throughput: a simple performance index with desirable characteristics. Behav Res Methods. 2006; 38(4):569–573.