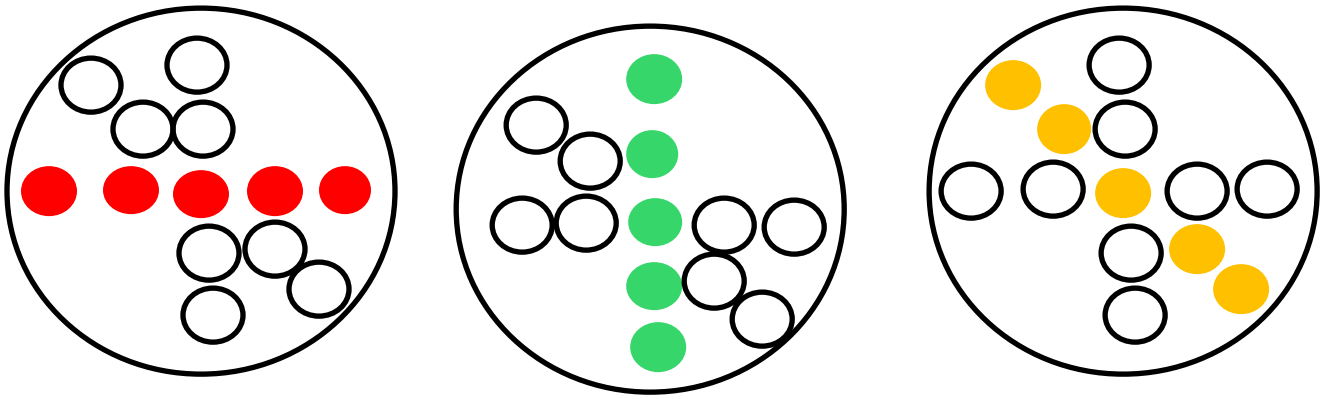
U.S. Department of Transportation

Federal Railroad Administration

Railroad Signal Color and Orientation: Effects of Color Blindness and Criteria for Color Vision Field Tests

Office of Research
and Development
Washington, DC 20590



DOT/FRA/ORD-15/03

Final Report
March 2015

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. Any opinions, findings and conclusions, or recommendations expressed in this material do not necessarily reflect the views or policies of the United States Government, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government. The United States Government assumes no liability for the content or use of the material contained in this document.

NOTICE

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 2015	Technical Report

4. TITLE AND SUBTITLE

Railroad Signal Color and Orientation: Effects of Color Blindness and Criteria for Color Vision Field Tests

5. FUNDING NUMBERS

6. AUTHOR(S)

Thomas G. Raslear [a] and Jordan Multer [b]

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

[a] U.S. Department of Transportation
Federal Railroad Administration
Office of Railroad Policy and Development
Office of Research and Development
Washington, DC 20590

[b] U.S. Department of Transportation
John A. Volpe National Transportation Systems Center
Cambridge, MA 02142-1093

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Department of Transportation
Federal Railroad Administration
Office of Railroad Policy and Development
Office of Research and Development
Washington, DC 20590

10.SPONSORING/MONITORING AGENCY REPORT NO.

DOT/FRA/ORD-15/03

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

This document is available to the public through the FRA Web site at http://www.fra.dot.gov.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This report concerns two issues: 1) whether color vision is necessary for locomotive crews who work on railroads where the signal system is either completely redundant with regard to signal color and signal orientation or the signal system only uses signal orientation; 2) what criteria should the railroad industry use for a valid, reliable, and fair field test of color vision. These two sets of issues are discussed together because they both relate to FRA's Medical Standards Guidelines for locomotive engineers (49 CFR 240, Appendix F) and conductors (49 CFR 242, Appendix D) and to NTSB recommendations (NTSB 2013-18 and 2013-19) that FRA establish a field test for color vision for railroad employees who fail standard tests of color vision such as pseudoisochromatic plate tests. In the event that FRA considers reviewing its regulations regarding color vision, this information will be relevant and useful.

14. SUBJECT TERMS

Railroad Signals, Signal Color, Signal Orientation, Medical Standards for Color Vision for Locomotive Engineers and Conductors, Color Vision, Color Vision Field Tests, Pseeudoisochromatic Plate Tests, Human Error, Signal Detection Theory.

15. NUMBER OF PAGES

34

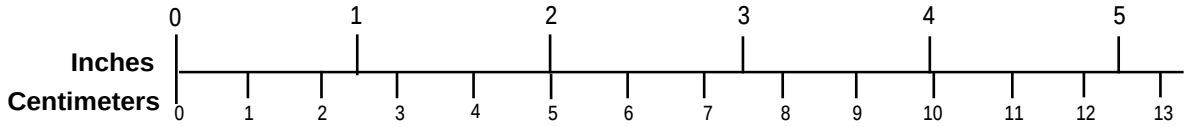
16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

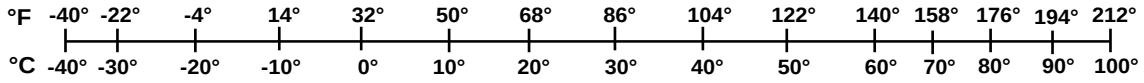
METRIC/ENGLISH CONVERSION FACTORS

ENGLISH TO METRIC	METRIC TO ENGLISH
LENGTH (APPROXIMATE)	**LENGTH (APPROXIMATE)**
1 inch (in) = 2.5 centimeters (cm)	1 millimeter (mm) = 0.04 inch (in)
1 foot (ft) = 30 centimeters (cm)	1 centimeter (cm) = 0.4 inch (in)
1 yard (yd) = 0.9 meter (m)	1 meter (m) = 3.3 feet (ft)
1 mile (mi) = 1.6 kilometers (km)	1 meter (m) = 1.1 yards (yd)
	1 kilometer (km) = 0.6 mile (mi)
AREA (APPROXIMATE)	**AREA (APPROXIMATE)**
1 square inch (sq in, in^2) = 6.5 square centimeters (cm^2)	1 square centimeter (cm^2) = 0.16 square inch (sq in, in^2)
1 square foot (sq ft, ft^2) = 0.09 square meter (m^2)	1 square meter (m^2) = 1.2 square yards (sq yd, yd^2)
1 square yard (sq yd, yd^2) = 0.8 square meter (m^2)	1 square kilometer (km^2) = 0.4 square mile (sq mi, mi^2)
1 square mile (sq mi, mi^2) = 2.6 square kilometers (km^2)	10,000 square meters (m^2) = 1 hectare (ha) = 2.5 acres
1 acre = 0.4 hectare (he) = 4,000 square meters (m^2)	
MASS - WEIGHT (APPROXIMATE)	**MASS - WEIGHT (APPROXIMATE)**
1 ounce (oz) = 28 grams (gm)	1 gram (gm) = 0.036 ounce (oz)
1 pound (lb) = 0.45 kilogram (kg)	1 kilogram (kg) = 2.2 pounds (lb)
1 short ton = 2,000 pounds (lb) = 0.9 tonne (t)	1 tonne (t) = 1,000 kilograms (kg) = 1.1 short tons
VOLUME (APPROXIMATE)	**VOLUME (APPROXIMATE)**
1 teaspoon (tsp) = 5 milliliters (ml)	1 milliliter (ml) = 0.03 fluid ounce (fl oz)
1 tablespoon (tbsp) = 15 milliliters (ml)	1 liter (l) = 2.1 pints (pt)
1 fluid ounce (fl oz) = 30 milliliters (ml)	1 liter (l) = 1.06 quarts (qt)
1 cup (c) = 0.24 liter (l)	1 liter (l) = 0.26 gallon (gal)
1 pint (pt) = 0.47 liter (l)	
1 quart (qt) = 0.96 liter (l)	
1 gallon (gal) = 3.8 liters (l)	
1 cubic foot (cu ft, ft^3) = 0.03 cubic meter (m^3)	1 cubic meter (m^3) = 36 cubic feet (cu ft, ft^3)
1 cubic yard (cu yd, yd^3) = 0.76 cubic meter (m^3)	1 cubic meter (m^3) = 1.3 cubic yards (cu yd, yd^3)
TEMPERATURE (EXACT)	**TEMPERATURE (EXACT)**
[(x-32)(5/9)] °F = y °C	[(9/5) y + 32] °C = x °F

QUICK INCH - CENTIMETER LENGTH CONVERSION



QUICK FAHRENHEIT - CELSIUS TEMPERATURE CONVERSION



For more exact and or other conversion factors, see NIST Miscellaneous Publication 286, Units of Weights and Measures. Price $2.50 SD Catalog No. C13 10286

Updated 6/17/98

Contents

Illustrations

Tables

Executive Summary

This report affirms that color vision is necessary for certain railroad employees, even if the wayside signal system is completely redundant with regard to signal color and signal orientation or the signal system only uses signal orientation. Federal Railroad Administration (FRA) regulations codified in 49 CFR 240 and 49 CFR 242 require that locomotive engineers and conductors have the "ability to recognize and distinguish between the colors of railroad signals." When viewing redundant signals, railroad employees with defective color vision have a much higher relative error risk than employees with normal color vision (the relative risk of an error is nearly 8,000,000 times higher for individuals with defective color vision). Moreover, when employees with normal color vision encounter signals based on signal orientation alone, they are at greater risk of misjudging those signal indications relative to redundant signals.

Additionally, this report discusses four criteria which must be considered in designing a field color vision test: statistical power, validity, reliability and fairness. The National Transportation Safety Board (NTSB) has issued a recommendation (NTSB, 2013-18 and 2013-19) that FRA establish a color vision field test for covered railroad employees who fail the standard tests, such as pseudoisochromatic plate tests, as described in 49 CFR 240, Appendix F and 49 CFR 242, Appendix D.

A color vision field test should be statistically capable of distinguishing between individuals with normal color vision from individuals with defective color vision (i.e., statistical power). In order to accomplish this goal, a criterion must be established that specifies how many errors in a number of trials are sufficient to determine that an individual has defective color vision. There is no single successful approach to setting this criterion since there are multiple decision goals that could guide it (e.g., maximize percent correct decisions, maximize expected value of decisions, maximize correct detections for a fixed value of false detections, etc.). A signal detection theory (SDT) framework is a suggested means to rationally set a criterion if a decision goal has been established. This will result in a defensible criterion that meets an explicit decision goal.

A color vision field test should also be valid, reliable and fair. "Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014, p. 11). This report discusses six possible uses of a field test of color vision and examines the evidence that would support the interpretation of test scores for that purpose.

Reliability refers "...to the consistency of scores across replications of a testing procedure..." (Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014, p. 35). The report asks, "How should the reliability of a field test of color vision be established?"; "What is the acceptable level of reliability?"; and "How should reliability be measured?".

Test standardization, imposing strict control of test administration, test conditions, and test scoring, are important aspects of fair testing methodology. Should a field color vision test be standardized for the railroad industry, or should each railroad be allowed to establish its own

standard test? What aspects of a field test need to be standardized? This can include requirements for types of signals to be used (actual signals, prototypes such as Christmas lights, stimuli that are physically distinct with regard to chromaticity and luminance, etc.), test conditions (viewing distance, from locomotive, weather and ambient light conditions, environmental luminance, etc.), scoring criteria (see section 3.1), and training and qualification of personnel responsible for test administration, etc.

1. Introduction

This report discusses the following topics:

(1) Whether color vision is necessary for locomotive crews who work with signal systems that are either completely redundant with regard to signal color and signal orientation or only use signal orientation;
(2) What criteria should the railroad industry use for administering a valid, reliable, and fair color vision field test.

These two issues are discussed together because they are related to to FRA's Medical Standards Guidelines for locomotive engineers (49 CFR 240, Appendix F) and conductors (49 CFR 242, Appendix D), as well as an NTSB recommendation (NTSB, 2013-18) that FRA establish a field test for color vision for railroad employees who fail standard tests of color vision such as pseudoisochromatic plate tests. In the event that FRA considers reviewing its regulations regarding color vision, this information will be relevant and useful.

Section 2 discusses how color blindness influences the viewing of railroad signals with a particular emphasis on redundant signal indications, which combine signal color and signal orientation. The ability to discriminate color and orientation is discussed with regard to how people combine and act upon such information, and the estimated error rate for persons with normal color vision and defective color vision when viewing signals that present only color, only orientation, or color and orientation. This directly addresses issue (1) above.

Section 3 discusses the criteria needed to ensure a field test of color vision is valid, reliable and fair. It examines the statistical requirements that are essential for distinguishing between persons with normal color vision and persons with defective color vision, then discusses how modern measurement theory decides if a test is valid, reliable and fair.

2. Signal Color[1] and Orientation: Effects of Color Blindness

In some railroad wayside signal systems, signal color and signal orientation are completely redundant. The question has been raised as to whether color vision is necessary for locomotive crews if the railroad they work on has such completely redundant signals. One passenger railroad asserted that its signaling aspects/indications are generally positional; and engineers and conductors with color vision deficiency can pass a Field Test without use of chromatic/chromagen/tinted lenses (or with use of such lenses) because the signals are positional.

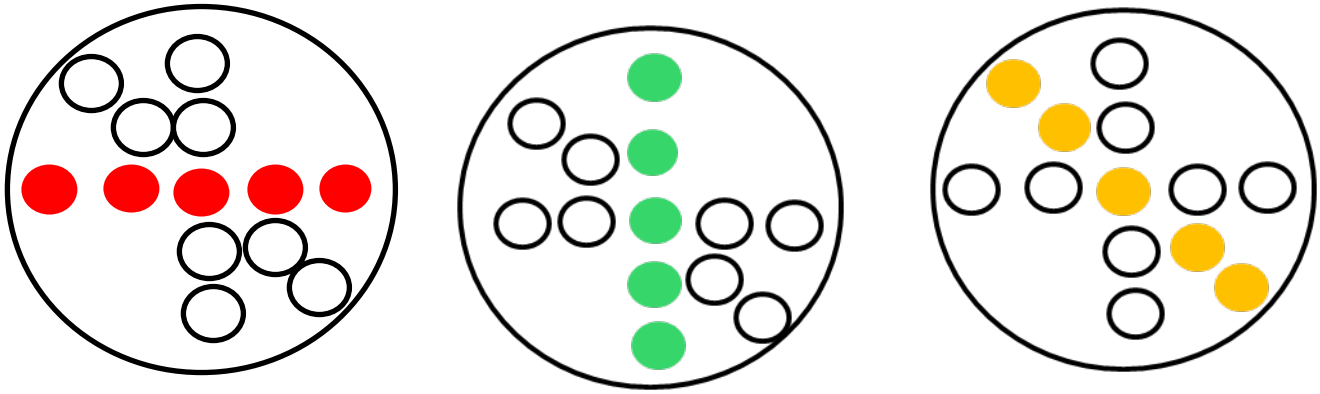


Figure 1. Redundant Color and Orientation Signal System

Figure 1 shows a redundant color and orientation signal system. In such a system, signal orientation alone can indicate the intended meaning of the signal, but there are additional issues to consider.

If signal color and signal orientation aspects are orthogonal (independent of each other), and if the discriminability of color and orientation are approximately equal, then the discriminability of the compound (color and orientation) is 1.41 times either method for accurately conveying the meaning of the signal. Consequently, a color blind individual would have reduced ability to properly discriminate the redundant signal compared to an individual with normal color vision.

This calculation is based on signal detection theory (SDT). In SDT, the ability to detect a stimulus or to discriminate between two stimuli is called sensitivity or d' (see Macmillan and Creelman, 2005).

[1] The correct technical term is wavelength, specified in nanometers (nm). Color names are arbitrary, but useful in that they have a common understanding in certain contexts.

$$d'_{compound} = \sqrt{(d'_{orientation})^2 + (d'_{color})^2}. \quad (1)$$

If $d'_{orientation} = d'_{color}$, then $d'_{compound}$ = √2 x d' = 1.41 d'. Differences in sensitivity between color and orientation would enhance the reduction in sensitivity suffered by the color blind operator. However, if $d'_{orientation} \ll d'_{color}$, the color blind operator's ability to discriminate the different signal aspects would be limited by orientation sensitivity which could be substantially less than color sensitivity.

What is meant by orientation sensitivity and color sensitivity? This terminology describes how normal humans discriminate between different colors in a group of signal aspects and between different orientations of the signal aspects (see Fig. 1). The smallest detectable difference between any two stimuli on the same continuum (i.e., colors of light) is known as a just noticeable difference (jnd). The human eye is sensitive to wavelengths of light in the range from approximately 400 nm (blue) to 700 nm (red). Across that range of wavelengths there are 128 jnds (Geldard, 1972, p. 63). By contrast, across stimulus orientations that range, as in Figure 1, from 0◦ to 180◦, there are 28 jnds (Leibowitz et al, 1955). This means that humans have more sensitivity to color differences than to orientation differences. The ratio of color jnds to orientation jnds (color jnds/orientation jnds) is 4.54. $d'_{orientation}$ has an estimated value of 1.27 (see Figure 2, Taylor, 1963), which means that judgments of orientation would be correct[2] approximately 81.56% of the time. Color has far higher sensitivity because we perceive many colors as totally distinct (red does not have any similarity to green, for instance). Stevens (1961) has argued that dimensions like color are metathetic continua in which changes in jnds are distinguished qualitatively as opposed to prothetic continua like brightness, in which changes in jnds are quantitative. Consequently, there are no estimates of d'_{color}. However, since it is known that judgments distinguishing colors that are separated by a large number of jnds (such as is the case for red, green and yellow) have very high accuracy, an arbitrarily high percent correct rate of 99.995% is assumed. This corresponds to a d'_{color}= 5.5. The ratio, $d'_{color}/d'_{orientation}$ = 4.33, which is close to the jnd ratio and supports the reasonableness of this value for d'_{color}.

For the compound stimulus, $d'_{compound} = \sqrt{(d'_{orientation})^2 + (d'_{color})^2}$ = $\sqrt{1.27^2 + 5.5^2}$ = 5.64. The percent correct judgments would be 99.99999917%. If an individual was totally color blind, judgments would be based solely on orientation, and the percent correct judgments would be 81.56%. However, complete color blindness is very rare. Dichromatism is more frequent (red-green or yellow-blue color blindness). Dichromats can distinguish colors, but with less sensitivity, so the actual percent correct judgments would lie somewhere between 99.99999917% and 81.56%.

The number of jnds in the visible spectrum could be used to estimate the relative sensitivity of dichromats. The average jnd[3] for a dichromat is 15.82 nm. For color vision normals the average

[2] Equation 7.6 in Macmillan and Creelman (2005) provides a translation between d' and percent correct for a two alternative forced choice discrimination paradigm. If a "yes-no" paradigm was used, the percent correct could be somewhat lower.

[3] Based on Fig. 4-13 from Geldard, 1972, p. 109.

jnd[4] is 2.58 nm. If we assume the range of wavelengths in the visible spectrum is 300 nm (400-700 nm), the number of jnds for normals is 116, which is close to the value of 128 noted above. For dichromats, the number of jnds, estimated in this fashion, is 19. The ratio of dichromat jnds to normal jnds is 0.15. Consequently, the estimated value of $d'_{dichromat}$ is 0.81, which corresponds to 79.23 % correct judgments. For the compound stimulus, $d'_{compound} = \sqrt{(d'_{orientation})^2 + (d'_{dichromat})^2} = \sqrt{1.27^2 + 0.81^2} = 1.51$. This means that the percent correct judgments would be 93.4%. This represents a 6.6% reduction in correct judgments for dichromat individuals if color and orientation redundant signals are used.

2.1 Summary and Conclusions

Table 1 (on the next page) summarizes the above discussion about orientation, color, color and orientation sensitivity, percent correct judgments, percent erroneous judgments and the expected number of signals that would be viewed before an erroneous judgment is made. The last measure is an easy way to understand the relative sensitivities for color and orientation in normals and dichromats. For instance, a 20% error rate means that a judgment error will be made 1 in 5 times when viewing a signal. In statistics, the null hypothesis is rejected if the error rate is 5%, which means that a judgment error will be made 1 in 20 times.

Table 1 shows that in a normal population, a signal with only color is expected to have only 1 erroneous judgment in 20,000 signal viewings. A signal with orientation only is expected to have 1 erroneous judgment in 5 signal viewings. Combining signal color and orientation is clearly superior: a redundant signal is expected to have 1 erroneous judgment in over 120,000,000 viewings.

Table 1 shows that the dichromats are at a clear disadvantage. With a redundant signal a dichromat is expected to experience 1 erroneous judgment in 15 viewings. In the absence of a redundant color, this drops to 1 erroneous judgment in 5 viewings.

2.1.1 Acceptable risk

Ultimately, a decision concerning what constitutes acceptable risk must be made. Table 1 allows an order of magnitude estimate of operator-relative error risk for dichromats relative to normals. Relative error risk (RR) is the percent error for dichromats divided by the percent error risk for normals. If RR = 1, the risk of an error is equal for the two groups. If RR > 1, the risk of an error is greater for the dichromats. If RR < 1, the risk of an error is less for the dichromats. The last column in Table 1 shows that risk is equal for orientation only signals. Error risk is much higher for dichromats for color only and color and orientation signals. It should be noted that for dichromats, adding orientation to color in a signal does reduce risk (RR = 6.56/20.77 =0.32) by about one-third. The question for policy makers is "Is that an acceptable level of risk?"

[4] Based on Fig 3-17 from Geldard, 1972, p. 63.

Table 1. Sensitivity (d'), Percent Correct, Percent Error, and Number of Signals Viewed per Error for Signal Orientation, Signal Color, Signal Color and Orientation for the Normal and the Dichromatic Population.

	Sensitivity				
Modality by Population	d'	Percent Correct	Percent Error	Signals Viewed per Error[a]	Relative Error Risk (Dichromats vs. Normals)
Normal Population					
Orientation only	1.27	81.56	18.44	5	
Color only	5.5	99.995	0.005	20,000	
Color and Orientation	5.64	99.99999917	0.00000083	120,481,927	
Dichromatic Population					
Orientation only	1.27	81.56	18.44	5	1
Color only	0.81	79.23	20.77	5	4,154
Color and Orientation	1.51	93.44	6.56	15	7,903,614

[a] Values are rounded to the nearest whole number

2.1.2 Limitations of the current approach

The estimate of color sensitivity is a rough estimate. It is based on the assumption that the number of jnds in the visible spectrum corresponds to the value of d'. It was assumed that colors could be distinguished correctly 99.995% of the time, which translates into a d' value of 5.5. That value may be too low. On the other hand, the number of color jnds was determined under ideal laboratory conditions. Under less than ideal conditions, people may not be able to distinguish 128 jnds in the visible spectrum. Consequently, d' for railroad conditions for normals might be less than 5.5.

For dichromats, the number of jnds was estimated on the basis of the average jnd. The relationship between the size of jnds and wavelength is a non-linear function, so this estimate is certainly inaccurate.

Similar arguments can be made concerning signal orientation. Estimates of sensitivity for orientation were made under controlled laboratory conditions and a different, lower value of d' might be obtained under railroad conditions. Sensitivity to oblique orientations (45 or 135 degrees) is lower than sensitivity to vertical and horizontal stimuli (Appelle, 1972). The value of d', estimated from Taylor (1963), is an average across orientations, so it overestimates sensitivity at oblique orientations.

The equation for compound stimuli, shown earlier, assumes that orientation and color are totally independent. While this may be the case in a purely physical sense, it is not known whether orientation affects color psychologically. To the uninitiated, this may sound nonsensical but it is

known, for instance, that the perceived hue of a stimulus changes with stimulus intensity (the Bezold-Brucke effect). So it is possible that orientation and color are correlated (perhaps hue changes between oblique and vertical/horizontal orientations). If this is the case, the compound stimulus sensitivity is overestimated.

The translation between d' and percent correct is based on the assumption that an unbiased method was used to determine thresholds (e.g., the two alternative forced choice procedure). The extent to which these methods were used in the older psychophysical literature cited here is not known. The use of biased measures introduces additional uncertainty in the true values that are being estimated.

2.1.3 Comparison with Pseudoisochromatic Plate Test Criteria

49 CFR 240, Appendix F and 49 CFR 242, Appendix D – Medical Standards Guidelines, have failure criteria for eight different pseudoisochromatic plate tests of color vision. These are shown in Table 2. These tests are used to establish whether individuals do not have normal color vision (i.e., are dichromats). For instance, if you make 5 errors over 15 trials (plates) in the American Optical Company test, you would be considered a dichromat, a person who does not have normal color vision.

Table 2 shows the number of errors in a specified number of trials that constitute a failure on each of the eight tests. The last column indicates the percent error on each test that constitutes a failure for acceptable color vision. The mean percent error for these various tests is 23.19%. The percent error estimate for dichromats in Table 1 is 20.77%.

Figure 2 compares the percent error estimate for dichromats from Table 1 with the mean percent error in Table 2, which results from applying the threshold values for making the determination that a person has a color vision deficit. The error bars in Figure 2 represent the 95% confidence interval for the pseudoisochromatic plate test mean. It is clear that the estimate of Table 1 for dichromats is not statistically different from the error rate established by the thresholds in 49 CFR 240 and 49 CFR 242. Considering the limitations noted in the previous section of this report, it is comforting to see that our estimate is consistent with known standards.

Table 2. Failure Criteria for Pseudoisochromatic Plate Tests per 49 CFR 240, Appendix F and 49 CFR 242, Appendix D. Percent Error for Each Test, Mean Percent Error, Standard Deviation (SD), and Standard Error of the Mean (SEM) Are Also Shown.

	Failure Criteria		
Pseudoisochromatic Test	Errors	Trials	% error
American Optical Company	5	15	33.33333
Hardy-Rand-Ritter (AOC)	1	6	16.66667
Dvorine	3	15	20
Ishihara 14 plate	2	11	18.18182
Ishihara 16 plate	2	8	25
Ishihara 24 plate	3	15	20
Ishihara 28 plate	4	21	19.04762
Richmond Plates	5	15	33.33333
		Mean	23.19535
		SD	6.699654
		SEM	2.532231

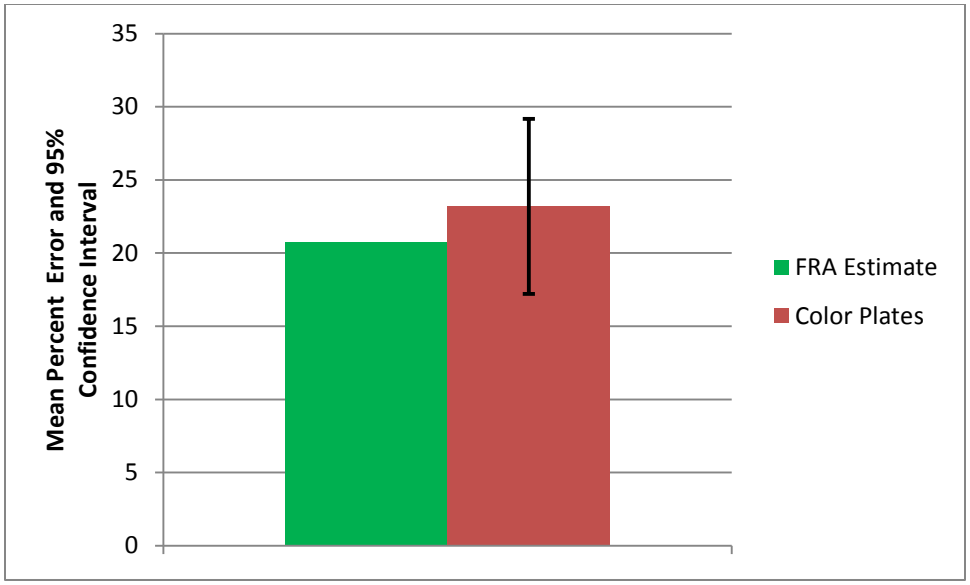


Figure 2. Estimates Of Error Rates For Color Discrimination From This Report And Pseudoisochromatic Plates Tests Per 49 CFR 240, Appendix F And 49 CFR 242, Appendix D.

2.1.4 Conclusions

Given all these limitations and concerns, we recommend that the estimates provided in this report should be order of magnitude estimates. Some differences between normals and dichromats are so large that the error risk is much higher for the dichromats, even with redundant signals. On the other hand, some differences within dichromats are probably statistically significant (e.g., dichromat color vs. dichromat color and orientation), but not overwhelmingly large. It is a matter of judgment and risk aversion as to whether the redundant signal adds sufficient safety. This analysis suggests that dichromats are not the only employees at risk of misjudging signals when position is the primary cue for determining the meaning of the signal. Individuals who rely on position alone are at greater risk of misjudging a signal than individuals who can make use of color alone or color and position. This means that when operators with normal vision encounter signals based on position alone, they are at greater risk of misjudging those signals. The National Transportation Safety Board (NTSB) noted this in their report on the Chase, MD accident in 1987 (NTSB, 1988, p. 9):

> However, the Safety Board believes the use of the same color in all aspects is a weakness in the position-light signal used on the corridor. At great distances, it is difficult to distinguish one aspect from another. The amber lights can be seen best at night and in overcast daylight; bright sunlight illuminates the black background and reduces the definition between the backgrounds and the lights. This was evident in the Safety Board's postaccident sight distance tests. Overhead catenary wires often prevent a full view of signal aspects in curves, somewhat diminishing the value of the position indication. This problem is aggravated by all the aspect lights being the same color.
>
> The color red is universally recognized as a warning of danger. When locomotive crewmembers watch for the amber aspects of a signal on the NEC, they must first detect this display and then decide, based on the position of the display, what action the aspect requires. However, if the "stop" aspect lenses were red, the engineer would know that on detection of the color red, he would be required to stop. This would save the time otherwise required to perceive the position of the aspect lights. It may be necessary to use a bulb of greater intensity for the red aspect to enable the engineer to detect it from the same or greater distance needed to detect the amber, but this should not present a problem.

3. Criteria for Color Vision Field Tests

This section discusses criteria for color vision field tests that must be considered when devising a test that is valid, reliable and fair. There are many ways that a field test for color vision could be devised to meet these criteria, and this report does not specify how to meet the criteria. Rather, this report provides guidance for determining that a field test will perform adequately with regard to the criteria.

In addition to validity, reliability and fairness, the design of a field test must have a rational basis for the number of trials included in the test and the number of errors that constitute a test failure. However, there are conflicting goals inherent when implementing this principle in any test design. A field test takes time and resources, so it is desirable to restrict the number of trials to the minimum. However, it is also desirable to have a sufficient number of trials to reliably detect defective color vision. Furthermore, since any test is imperfect there will be individuals with defective color vision who are not detected by the test (missed detections) and individuals with normal color vision who fail the test (false detections). The decision about the number of trials and errors affects missed detections and false detections. Consequently, a decision must be made concerning the number (or probability) of missed detections and false detections that are acceptable. These issues are discussed in detail in the next section.

3.1 Number of Trials Required for Color Vision Field Tests

NTSB has recommended that FRA establish a color vision field test for railroad employees who fail standard tests such as pseudoisochromatic plate tests (NTSB, 2013-18). FRA will need to carefully specify certain aspects of the color vision field test for it to yield valid and consistent results. Under consideration here are the number of errors in a fixed set of trials that constitute a failure of the field test (i.e., evidence that the person does not have sufficient color vision to be considered operationally safe). Appendix F to 49 CFR 240, for instance, has failure criteria for eight pseudoisochromatic plate tests (see Table 2). The purpose of this section is to provide statistical support to the definition of such failure criteria for a field test.

3.1.1 Statistical Background

The pseudoisochromatic tests and field tests for color vision can be viewed as Bernoulli-trials processes. In a Bernoulli-trials process there are only two outcomes on a given trial: yes and no (or 0 and 1, heads and tails, success and failure). A sequence of independent Bernoulli-trials results in a binomial distribution, which is the sum of the outcomes of *n* Bernoulli-trials. The number of trials required to get the rth error results in the negative binomial distribution (see Egan, 1975, p. 140). This is the type of criterion for failure that is specified in 49 CFR 240 Appendix F and 49 CFR 242 Appendix D, which implies the use of the negative binomial distribution to specify a failure criterion for a field test. The negative binomial probability density function (pdf) is

$$\binom{r+n-1}{n-1} p^r q^n. \tag{2}$$

The failure criterion is *r* errors in *n* trials. p is the probability of an error, and q = (1- p).

3.1.2 Signals with Redundant Color and Orientation

Table 1 shows that people with normal color vision (normals) would require approximately 120,000,000 trials on a color vision test with redundant color and orientation to make one error. By contrast dichromats would only require 15 trials to make one error. This suggests that as few as 1 error in 15 trials in a field test might be sufficient to determine that a person was a dichromat. However, the probability of a correct detection of dichromaticy, a missed detection of dichromaticy, a false detection of dichromaticy, and a correct rejection of dichromaticy are not known from the information in Table 1. Table 3 summarizes the information that would be necessary to rationally determine a failure criterion for a field test. The four cells of the table show the 'Yes' and 'No' decision outcomes from a field test for dichromats and normals. If a dichromat meets the test criterion for being dichromatic, a 'Yes' results in a Correct Detection. If a normal meets the test criterion for being dichromatic, a 'Yes' results in a False Detection. If a dichromat does not meet the test criterion for being dichromatic, a 'No' results in a Missed Detection. If a normal does not meet the test criterion for being dichromatic, a 'No' results in a Correct Rejection.

Table 3. Decision Outcome Matrix for a Color Vision Field Test.

	YES, Dichromatic	**NO, Normal**
DICHROMATIC	Correct Detection	Missed Detection
NORMAL	False Detection	Correct Rejection of Dichromaticy

A separate value of p (probability of an error in equation 2) is known from Table 1 for dichromats and normals. The negative binomial distribution[5] for dichromats is used to determine the probability of Correct Detections (P_{CD}). Missed Detections (P_{MD}) come from the same distribution as P_{CD}, so $P_{MD} = 1- P_{CD}$. Similarly, the negative binomial distribution for normals is used to determine the probability of False Detections (P_{FD}), Correct Rejections (P_{CR}) come from the same distribution as P_{FD}, so $P_{CR} = 1 - P_{FD}$. For normals p = 0.0000000083 and for dichromats p = 0.0656. For any value of *r*, the four probabilities associated with the decision outcomes in Table 3 can be determined for various values of *n*. For example, Fig. 3 shows the pdf of three errors for normals and for dichromats as a function of number of trials (*n*). The pdf for normals appears as a flat line at a probability density of nearly zero because the probability of an error is so small.

[5] The distribution function is the sum or integral of the pdf.

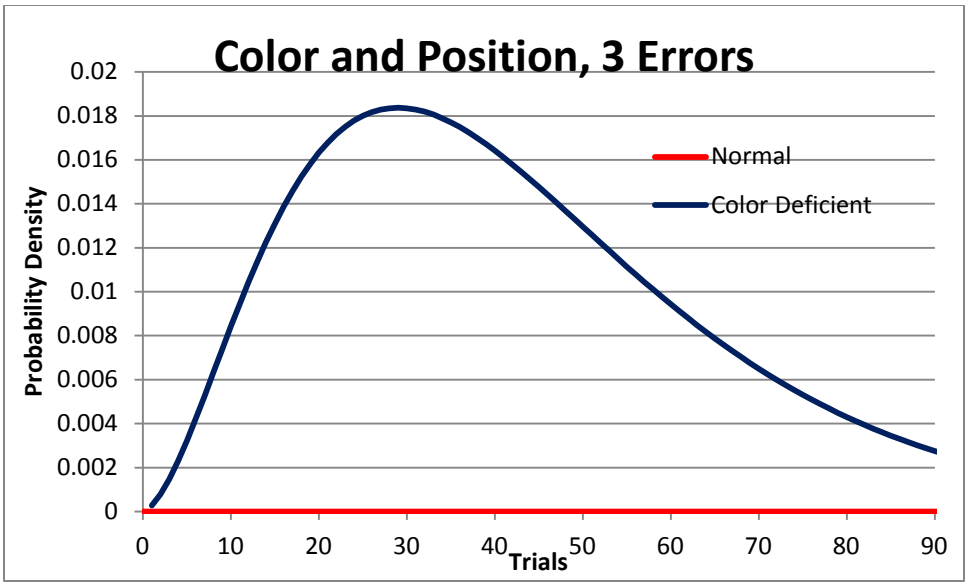


Figure 3. Probability Density Of 3 Errors As A Function Of Number Of Trials For Dichromats And Normals With Color And Orientation Redundant.

The mean number of trials to make three errors for dichromats is 42.7,while for normals the mean number of trials is 361,445,780. Obviously, the focus should be on the dichromats because the normals are unlikely to fail this test.

If one uses the mean number of trials in the criterion for failure, what are the probabilities associated with Table 3? Table 4 shows these probabilities[6]. It is clear from this table that 43 trials allows more than 41% of dichromats to miss detection or 2,778 individuals (the approximate number of locomotive engineers and conductors is indicated by N in each cell of the table). If we wanted to make correct detections 95% of the time, 91 trials would be required (Table 5).

Table 4. Probabilities for 3 Errors in 43 trials, Color and Orientation. [a]

	YES	NO
DICHROMATIC	Correct Detection $P_{CD} = 0.588$ N = 3,965	Missed Detection $P_{MD} = 0.412$ N = 2,778
NORMAL	False Detection $P_{FD} = 8.68E\text{-}21$ N = 0	Correct Rejection of Dichromaticy $P_{CR} \approx 1$ N = 79,120

[a] N is the estimated number of locomotive engineers and conductors corresponding to the probabilities in each cell. Eight percent of the male population is dichromatic, and there are 86,000 locomotive engineers and conductors of whom 98% are male (Gertler and DiFiore, 2009).

[6] The cumulative probability summed from 1 to 43 trials in the Color Deficient pdf in Fig. 3 is P_{CD}. Similarly, the cumulative probability summed from 1 to 43 trials in the Normal pdf in Fig. 3 is P_{FD}.

Table 5. Probabilities for 3 Errors in 91 Trials, Color and Orientation.[a]

	YES	NO
DICHROMATIC	Correct Detection P_{CD} = 0.95 N = 6,405	Missed Detection P_{MD} = 0.05 N = 337
NORMAL	False Detection P_{FD} = 7.66E-20 N = 0	Correct Rejection of Dichromaticy $P_{CR} \approx 1$ N = 79,120

[a] N is the estimated number of locomotive engineers and conductors corresponding to the probabilities in each cell. Eight percent of the male population is dichromatic, and there are 86,000 locomotive engineers and conductors of whom 98% are male (Gertler and DiFiore, 2009).

There is an obvious trade-off in a practical test between statistical power (ability to correctly detect a dichromat) and the amount of time required to conduct the test. If we reduce the error criterion we would also reduce the number of trials required. Figure 4 shows the pdfs for dichromats and normals for 1 error in *n* trials.

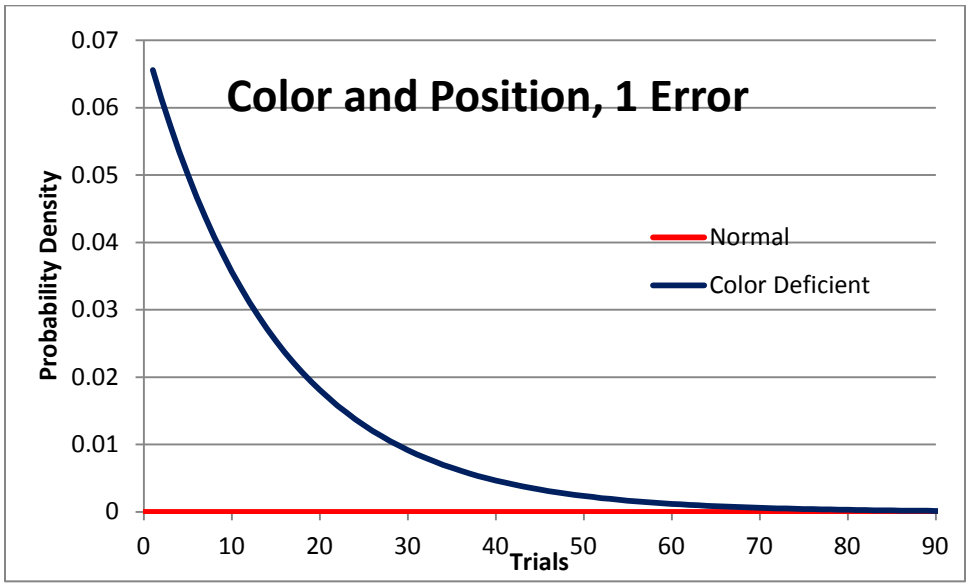


Figure 4. Probability Density Of 1 Error As A Function Of Number Of Trials For Dichromats And Normals With Color And Orientation Redundant.

Since the error probabilities are same as for Fig. 3, the normal pdf again appears as a flat line at a probability density of nearly zero.. The mean number of trials to make one error for dichromats is 14.2, while for normals the mean number of trials is 120,481,926. The number of trials necessary to achieve 95% correct detection of dichromats is 44. Table 6 shows the probabilities for 1 error in 44 trials for dichromats and normals.

Table 6. Probabilities for 1 Error in 44 Trials, Color and Orientation.[a]

	YES	NO
DICHROMATIC	Correct Detection P_{CD} = 0.953 N = 6,425	Missed Detection P_{MD} = 0.047 N = 317
NORMAL	False Detection P_{FD} = 3.73E-07 N = 0	Correct Rejection of Dichromaticy P_{CR} = 0.999999627 N = 79,119

[a] N is the estimated number of locomotive engineers and conductors corresponding to the probabilities in each cell. Eight percent of the male population is dichromatic, and there are 86,000 locomotive engineers and conductors of whom 98% are male (Gertler and DiFiore, 2009).

3.1.3 Signals with Only Color

Table 1 shows that color vision normals would require approximately 20,000 trails on a color vision test with only color to make one error. By contrast dichromats would only require 5 trials to make one error. For normals p = 0.005 and for dichromats p = 0.2077. With these different p values, the same logic can be followed to determine the number of trials required to have 95% correct detection of dichromats. Figure 5 shows the negative binomial pdfs for dichromats and normals for 1 error. Again, the normal pdf appears as a flat line because of the low p value.

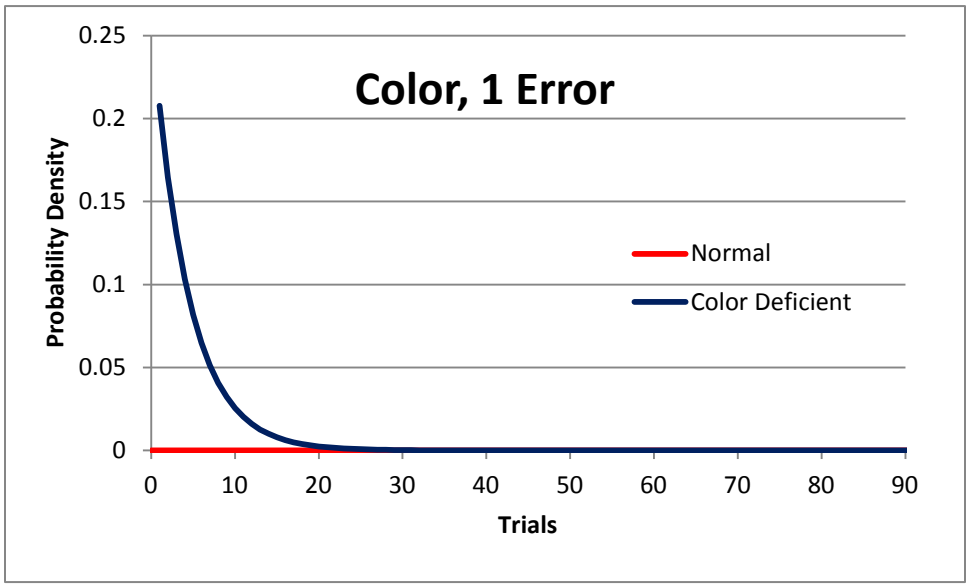


Figure 5. Probability Density Of 1 Error As A Function Of Number Of Trials For Dichromats And Normals For Color Only Test.

The mean number of trials to make one error for dichromats is 3.8, while for normals, the mean number of trials is 19,999. The number of trials necessary to achieve 95% correct detection of dichromats is 12. Table 7 shows the probabilities for 1 error in 12 trials for dichromats and normals. If the purpose of a color vision field test is to detect dichromats, this is more efficiently done with a test that uses color only.

Table 7. Probabilities for 1 Error in 12 Trials, Color Only.[a]

	YES	NO
DICHROMATIC	Correct Detection $P_{CD} = 0.951$ N = 6,412	Missed Detection $P_{MD} = 0.049$ N = 330
NORMAL	False Detection $P_{FD} = 0.00065$ N = 51	Correct Rejection of Dichromaticy $P_{CR} = 0.99935$ N = 79,069

[a] N is the estimated number of locomotive engineers and conductors corresponding to the probabilities in each cell. Eight percent of the male population is dichromatic, and there are 86,000 locomotive engineers and conductors of whom 98% are male (Gertler and DiFiore, 2009).

3.1.4 Rationalizing the Criterion for a Failure

Determining the failure criterion for a field color vision test can be accomplished in many different ways. As discussed in previous sections, information about error probabilities can be used to generate probability distributions to help in setting the criterion. However, setting the criterion so the test does not miss more than 5% of dichromats is arbitrary and it is based on the current practice for describing the accuracy of medical diagnostic tests. In the absence of theory about the relationships between the probabilities in Table 7, any chosen criterion is arbitrary because it cannot set a decision goal upon which to base a decision rule for setting a criterion. Possible decision goals include maximizing percentage correct decisions, maximizing the expected value of decisions, and maximizing correct detections (P_{CD}) for a fixed value of false detections (P_{FD})[7] (see Egan, 1975, pp 15 – 24).

In the medical literature, sensitivity refers to the ability of a diagnostic test to determine the proportion of a population who have a condition (e.g., dichromaticity), and specificity refers to the ability of a diagnostic test to determine the proportion of a population who do not have that condition (Lilienfeld and Lilienfeld, 1980, p. 151). Taken together, sensitivity and specificity are thought to characterize the accuracy of medical tests. In the matrices presented above, such as

[7] This is current practice in testing statistical hypotheses. The Neyman-Pearson objective aims to keep Type I errors (P_{FD}) at a fixed low probability, usually 5%, while maximizing the ability to reject the null hypothesis.

Table 7, sensitivity is P_{CD}, and specificity is P_{CR}. From this perspective, all of the criteria considered in Tables 4 – 7 have very high specificity but vary in sensitivity. Which criterion is best? How accurate the outcomes are for each criterion is not easy to determine.

There is, however, a better way to characterize the performance of a diagnostic test using the exact same information as was presented in Tables 3 – 7 (Swets and Pickett, 1982; Swets, 1996). In signal detection theory (SDT) the accuracy of a test is indicated by its discrimination accuracy, or d', which is jointly determined by P_{CD} and P_{FD}. Figure 6 is a graphical representation of the outcome matrix in Table A1 using Gaussian distributions as an example.

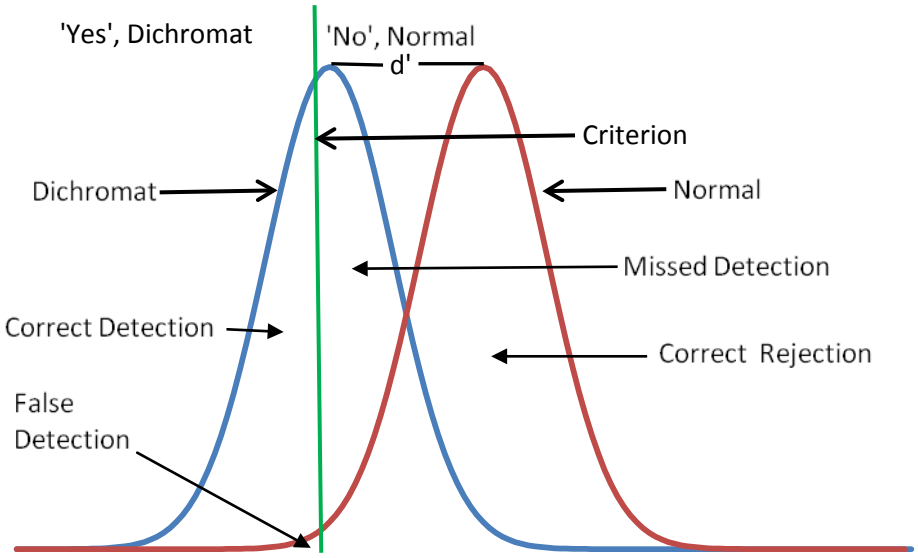


Figure 6. Decision Outcomes And Dichromat And Normal Probability Density Functions (Pdfs). Here P_{CD} Corresponds To The Area Under The Dicromate PDF And Left Of The Criterion Line (Green Vertical Line). P_{FD} Corresponds To The Area Under The Normal PDF And Left Of The Criterion Line. P_{MD} Corresponds To The Area Under The Dichromat PDF And Right Of The Criterion Line. P_{CR} Corresponds To The Area Under The Normal PDF And Right Of The Criterion Line. See Text For Details.

In SDT the observer's task is to detect a signal against a background of noise. In this application of SDT, the observer is detecting dichromats against a background of normals. The distributions for dichromats and normals overlap, so regardless of the criterion (green vertical line) used there are always False Detections. Discrimination accuracy (d') is the distance between the means of the dichromat and normal distributions in standard deviation units (see Macmillan and Creelman, 2005). This can be mathematically expressed as

$$d' = z_{CD} - z_{FD}, \tag{3}$$

where z_{CD} is the z-transform of P_{CD}, etc. As the criterion (green vertical line) is changed systematically from left to right in Fig. 6, changes occur in both P_{CD} and P_{FD}. This is shown in Fig. 7 for different values of d'.

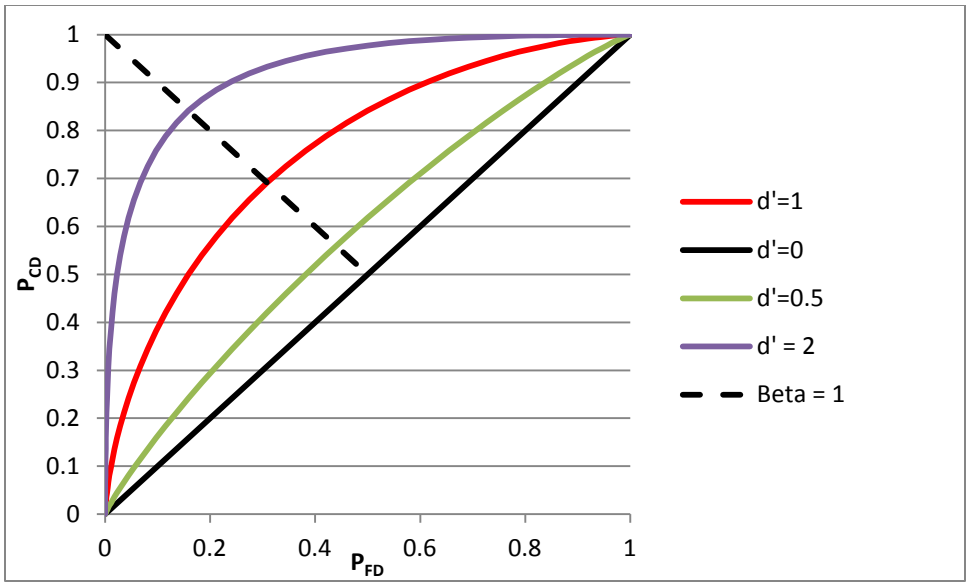


Figure 7. ROC Curves for Various d' Values.

As the value of d' increases, the separation between dichromat and normal distributions increases and the observer's ability to detect the dichromats increases. The line labeled d' = 0 is the case in which the dichromat and normal distribution are identical and dichromats are detected at chance. When P_{CD} and P_{FD} are close to 1, the criterion is at the extreme right in Fig. 6, and there is a bias for the observer to report the presence (YES) of dichromats. When P_{CD} and P_{FD} are close to 0, the criterion is at the extreme left in Fig. 6, and there is a bias for the observer to report the absence (NO) of dichromats. When the criterion is at the point where the distributions cross, there is no bias. Bias is often measured by β. A bias to the presence (YES) of the signal (dichromat) is indicated by values of $\beta < 1$. A bias for the observer to report the absence (NO) of the signal (normal) is indicated by values of $\beta > 1$. When $\beta = 1$ there is no bias.

Since P_{CR} (specificity) equals $1 - P_{FD}$, P_{CD} (sensitivity) and P_{CR} (specificity) also covary with changes in the criterion. Since there is no knowledge of the separation of the normal and dichromat distributions, it is not possible to know how to interpret a sensitivity – specificity pair. For instance, "Is system A with sensitivity of 0.90 and specificity of 0.30 more or less accurate than system B with values of 0.70 and 0.50?" (Swets and Pickett, 1982, p. 25-26). In SDT, the indices d' and β provide complete information concerning the relative locations of the dichromat and normal distributions and the decision criterion.

The criterion can be chosen rationally within the SDT framework if a decision goal has been established. Common decision goals, as noted previously, include maximizing the percentage correct decisions, maximizing the expected value of decisions, maximizing correct detections (P_{CD}) for a fixed value of false detections (P_{FD}), and maximizing a weighted combination of Correct Detections and Correct Rejections (see Egan, 1975, pp. 15 – 24; Macmillan and Creelman, 2005, pp. 42 – 44). In each case, a value of β can be calculated that will meet the decision goal:

$$\beta = \frac{V(CR)+V(FD)}{V(CD)+V(MD)} \times \frac{P(normal)}{P(dichromat)}, \quad (4)$$

where V(CR) is the value of a Correct Rejection, V(FD) is the value of a False Detection, V(CD) is the value of a Correct Detection, V(MD) is the value of a Missed Detection, P(normal) is the probability of having normal color vision, and P(dichromat) is the probability of having dichromatic vision. The ratio

$$\frac{P(normal)}{P(dichromat)}$$

is also called "prior odds." If the values of all the decision outcomes are equal, β is determined by the prior odds alone. The probability of having dichromatic vision is approximately 0.08 (Geldard, 1972), so the prior odds and β equals 11.5, and there is a very high bias to say 'No'. Since β can also be defined as the likelihood ratio,

$$\beta = \frac{y(CD)}{y(FD)} \tag{5}$$

where

$$y(CD) = 0.3989e^{-z[P(CD)]^2/2}, \tag{6}$$

and

$$y(FD) = 0.3989e^{-z[P(FD)]^2/2}, \tag{7}$$

equation 4 can be used with equation 5 to determine the location of the criterion on the decision axis[8].

In the examples from this report, an implicit and arbitrary emphasis is placed on a particular decision goal: maximize correct detections while keeping missed detections below 5%. (the Neyman-Pearson objective, see footnote 7) and minimizing the cost of field tests. This decision goal is used for illustrative purposes only. Other decision goals can, and should, be considered. The point of the discussion above is to make explicit the costs and benefits that drive such decision goals and to indicate that different decision goals can result in the same decision outcomes. For instance, one could attain 5% missed detections by a particular setting of values of the decision outcome matrix (when values are assigned to decision outcomes it becomes a payoff matrix) and equation 4. Consider the payoff matrix in Table 8[9].

Here missed detections are very costly, and β = 0.004 which indicates a large bias to detect dichromats. It so happens that this also corresponds to a missed detection rate of 5%. If our decision goal was to maintain missed detections at ≤ 5%, it should be irrelevant that a missed detection would cost $12,000,000. However, there are obvious economic consequences that follow from the Neyman-Pearson objective. Eight percent of the male population is dichromatic, and there are 86,000 locomotive engineers and conductors of whom 98% are male (Gertler and DiFiore, 2009). The number of missed detections, using the criterion applied in Table 7 would

[8] In general, the likelihood ratio is the ratio of the ordinates of the PDFs for signal and noise. Equations 6 and 7 are the formulas for the ordinates if the distributions are Gaussian. See Egan (1975).

[9] The dollar values in Table 8 are for illustration only. They are totally fictitious..

be 86,000 x 0.98 x 0.08 x 0.049 = 330. The total economic consequence of allowing a 5% missed detection rate is $3,960,000,000[10].

Table 8. Payoff Matrix for Table 7.

	YES	NO
DICHROMATIC	Correct Detection V(CD) = $100	Missed Detection V(MD)=$12,000,000
NORMAL	False Detection V(FD)= $4,000	Correct Rejection of Dichromaticy V(CR)= $100

The values of outcomes in the payoff matrix are usually not known or are known imprecisely. The values may involve monitizing political costs and benefits, assessing reputational costs and benefits, and estimating many other intangibles. Consequently, criterion setting, while a rational process in SDT, remains highly subjective. Nevertheless, it is the case that consideration of all the costs and benefits accruing to different decision goals will result in a better decision than setting a criterion arbitrarily. Moreover, the resulting criterion will be defensible.

3.2 Validity, Reliability and Fairness of Field Color Vision Tests

This discussion of validity, reliability and fairness in field color vision tests includes extensive references to the *Standards for Educational and Psychological Testing* (Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014). However, this report does not examine every issue that applies to field color vision test; its intent is to highlight the most important issues that need to be considered and addressed. The reader is referred to the *Standards* for additional information.

3.2.1 Validity

"Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014, p. 11). Given this definition of validity, the question that immediately comes to mind with regard to a field color vision test is: "What are the proposed uses of such a test?" If a clinical test (e.g., the Ishihara color plates) has already established that an individual has a color deficiency, why have a field test? There are several possibilities, which we list for consideration but do not endorse or disavow:

[10] This assumes that all dichromats fail the clinical test (per 49 CFR 240) and request the field test.

(1) The field test is intended to demonstrate that an employee has sufficient color vision to safely operate under normal operating conditions with signals that are ordinarily in use on that railroad.

(2) The field test is intended to confirm the results of the clinical test that an employee does not have sufficient color vision per the definition of 49 CFR 240 and 49 CFR 242 using signals that are ordinarily in use on that railroad and under normal operating conditions.

(3) The field test is intended to avoid problems with regard to labor relations and the Americans with Disability Act, by demonstrating that an employee does not have sufficient color vision to safely operate under normal operating conditions with signals that are ordinarily in use on that railroad.

(4) The field test is intended to allow continued employment of employees who would otherwise not meet the requirements of 49 CFR 240 and 49 CFR 242. This is possible if the field test that has face validity (see Guilford, 1954, p. 400) but lacks the rigor of clinical tests. For instance, if a signal system uses color and positional aspects, a dichromat would be expected to perform better in a field test (see discussion in section 1). Similarly, if signal colors are not corrected for differences in brightness (Cornsweet, 1970, p. 234-236), a dichromat might be able to identify different signal colors on the basis of brightness differences.

(5) The field test is intended to meet the minimal requirements of 49 CFR 240 and 49 CFR 242.

(6) The field test is intended to satisfy the NTSB's (2013) recommendation concerning a color vision field test.

Use Case 1: 49 CFR 240 Appendix F(4) states that its intent is "...to provide an examinee with at least one opportunity to prove that a hearing or vision test failure does not mean that the examinee cannot safely operate a locomotive or train." , and 49 CFR 242 Appendix D(4) states that its intent is "...to provide an examinee with at least one opportunity to prove that a hearing or vision test failure does not mean that the examinee cannot safely perform as a conductor." Given that color vision tests are designed specifically to detect color vision deficiencies, this is the statistical equivalent of proving the null hypothesis (i.e., that the person vision is sufficiently normal). The problem with proving the null hypothesis is that one does not know how much evidence is sufficient. What does it mean to safely operate a locomotive or to safely perform as a conductor in terms of color vision? What evidence is needed to support this interpretation of test scores? If railroad signals are redundant with regard to positon and/or brightness, a dichromat might be able to pass a field that that uses railroad signals, but, as has been demonstrated in Section 2, the dichromat will make more errors (be less safe) than a person with normal color vision. This is very different from simply determining that a person has normal color vision with a clinical test.

Use Case 2: If this is the intended purpose of the field test, the signals and conditions of testing must be as rigidly controlled as the stimuli in the clinical test. It is not clear that this would be easy to do: the stimuli should only vary in color (i.e., be controlled for brightness, position and any other extraneous cues), and testing conditions should always be the same (i.e., controlled for

ambient light, distance from signals, etc.). What evidence and theory would support this interpretation of test scores?

Use Case 3: This intended purpose must meet the requirements demanded by the definition of validity for a color vision test and meet the legal and political requirements necessary to satisfy labor relation agreements and laws concerning disability. What evidence and theory would support this interpretation of test scores?

Use Case 4: There may be circumstances under which employees could work as a locomotive engineer or conductor in a locomotive environment without normal color vision. For instance, there is evidence that in some color vision tests, decreasing the viewing distance decreases the number of errors made by color-defective individuals (Hovis and Ramaswamy, 2006). Such individuals might work in rail yards safely because the viewing distances for signals are shorter. Other considerations, which would require empirical justification, might include working only where positional signals were in use, etc. However, attempting to define "safe" leads to the same problems that are found in Use Case 1.

Use Case 5: In this case, a railroad wants to simply say that they have a field test that can be used to satisfy the regulations. Unless FRA audits such tests, there is no guarantee that the tests measure color vision in the sense that clinical tests measure color vision. Would a test protocol and materials for a field test provide sufficient evidence to support this intended use of a color vision test?

Use Case 6: This case would allow FRA to simply say that they have satisfied the NTSB recommendation. Unless NTSB audits the FRA requirements for a field color vision test, there is no guarantee that the FRA requirements meet the NTSB's recommendation. It should be noted that the NTSB recommendation to FRA does not specify the intended purpose of a field color vision test. Consequently, NTSB is not stating what would constitute a valid field test of color vision. What evidence and theory are sufficient for this interpretation to support this intended use of a color vision test?

Since there may be multiple intended uses of a color vision test, many of these intended uses may be combined. The use cases were separated in order to be clear about the issues that are raised in each case and which issues are involved in establishing the validity of a field color vision test.

3.2.2 Reliability

Generally speaking, the term reliability refers "...to the consistency of scores across replications of a testing procedure..." (Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014, p. 35). Test reliability is rarely perfect because there are multiple sources of variability that affect the conditions under which a test is administered, how a test is administered, and how the test is scored. Moreover, individuals vary over time or may react differently to a test on different occasions, and this also degrades reliability. Reliablility affects validity because low reliability decreases the ability of a test to provide a trustworthy measure of the attribute that is being measured. Reliability can be increased by imposing strict controls on test administration, test

conditions, and test scoring, which is considered in more detail in the next section. The questions here are

(1) How should the reliability of a field test of color vision be established?

(2) What is the acceptable level of reliability?

(3) How should reliability be measured?

3.2.3 Fairness

Test standardization, imposing strict control of test administration, test conditions, and test scoring, are all important aspects of test fairness.

"Regardless of the purpose of testing, the goal of fairness is to maximize, to the extent possible, the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to measure. Traditionally, careful standardization of tests, administration conditions, and scoring procedures have helped to ensure that test takers have comparable contexts in which to demonstrate the abilities or attributes to be measured. ..." (Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014, p. 51).

Since fairness affects reliability, tests should be standardized to assure that all individuals have the same testing protocol, scoring and conditions of testing in place. If this is not the case, test score consistency and reliability will suffer.

Fairness also has a direct impact on the validity of a test. If a test is administered or scored in a way that provides an advantage to one individual (or class of individuals) over another, the validity of the test has been fundamentally violated because it is no longer measuring what it was intended to measure. If a test is not standardized, which ensures that all individuals have the same testing protocol, scoring and conditions of testing, some individuals may have disadvantages when taking the test versus other participants. For instance, color signals are easier to distinguish if the viewer is closer to the signal. Thus, if railroad A sets the viewing distance for a field color vision test at 500 feet and railroad B sets the viewing distance at 1000 feet, employees at railroad B will be at a disadvantage in taking the test relative to employees at railroad A.

The questions here are:

(1) Should a field color vision test be standardized for the railroad industry, or should each railroad be allowed to have its own standard test?

(2) What aspects of a field test need to be standardized? This can include requirements for types of signals to be used (actual signals, prototypes (e.g., Christmas lights), stimuli that are physically specified with regard to chromaticity and luminance, etc.), test conditions (viewing distance, from locomotive, weather conditions, environmental luminance, etc.), scoring criteria (see section 3.1), personnel responsible for test administration, etc.

4. Summary and Conclusions

This report affirms that normal color vision is necessary for certain railroad employees, even if the signal system is completely redundant with regard to signal color and signal orientation. Railroad employees with defective color vision have a much higher relative error risk than employees with normal color vision when viewing redundant signals (relative risk of an error is nearly 8,000,000 times higher for individuals with defective color vision). Moreover, when employees with normal color vision encounter signal indications based on signal orientation alone, they are at greater risk of misjudging those signals relative to redundant signals or color alone signals.

If FRA establishes a field test for color vision for railroad employees who fail standard tests of color vision (such as pseudoisochromatic plate tests, as described in 49 CFR 240, Appendix F and 49 CFR 242, Appendix D), there are four criteria which need to be considered in designing that test: statistical power, validity, reliability and fairness.

A color vision field test should be statistically capable of distinguishing individuals with normal color vision from individuals with defective color vision (i.e., statistical power). The problem can be broadly stated as setting a criterion for deciding how many errors in a number of trials are sufficient to detect an individual with defective color vision. The negative binomial distribution can be used to model this situation for dichromats and color vision normals viewing redundant and color only signals. Different criteria (e.g., 3 errors in 43 trials) result in a variety of decision outcomes. Decision outcomes include the proportion of correct detections of dichromats, missed detections, false detections and correct rejections. However, there is no single solution to this criterion setting decision since there are multiple decision goals that could guide it (e.g., maximize percent correct decisions, maximize expected value of decisions, maximize correct detections for a fixed value of false detections, etc.). A signal detection theory (SDT) framework is suggested as a means to rationally set a criterion if a decision goal has been established. This will result in a defensible criterion that meets an explicit decision goal. Otherwise, the criterion will be set arbitrarily and can result in unintended decision outcomes.

A color vision field test should also be valid, reliable and fair. "Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014, p. 11). The report focuses on six possible uses of a field test of color vision and the evidence that would be needed to support the interpretation of test scores for each use.

Reliability refers "...to the consistency of scores across replications of a testing procedure..." (Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014, p. 35). The report asks, "How should the reliability of a field test of color vision be established?"; "What is the acceptable level of reliability?"; and "How should reliability be measured?".

Test standardization, imposing strict control of test administration, test conditions, and test scoring, are an important aspects of test fairness. Should a field color vision test be standardized for the railroad industry, or should each railroad be allowed to have its own test?

What aspects of a field test need to be standardized? This can include requirements for types of signals to be used (actual signals, prototypes (e.g., Christmas lights), stimuli that are physically distinct or distinguishable with regard to chromaticity and luminance, etc.), test conditions (viewing distance, from locomotive, weather conditions, environmental luminance, etc.), scoring criteria (see section 3.1), personnel responsible for test administration, etc.

5. References

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin, 78,* 266-278.

Cornsweet, T.N. (1970). *Visual perception*. New York: Academic Press.

Egan, J.P. (1975). *Signal detection theory and ROC analysis.* New York: Academic Press.

Geldard, F.A. (1972). *The human senses* (2nd ed.). New York: Wiley.

Gertler, J., & DiFiore, A. (2009). *Work schedules and sleep patterns of railroad train and engine service workers*. (DOT/FRA/ORD-09/22). Washington, DC: Federal Railroad Administration. (Available at http://www.fra.dot.gov/eLib/details/L01507)

Hovis, J.K., and Ramaswamy, S. (2006). The effect of test distance on the CN lantern results. *Visual Neuroscience, 23,* 675-679.

Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Leibowitz, H.W., Meyers, N.A., & Grant, D.A. (1955). Radial localization of a single stimulus as a function of luminance and duration of exposure. *Journal of the Optical Society of America, 45,* 76-78.

Macmillan, N.A., & Creelman, C.D. (2005). *Detection theory. A user's guide.* (2nd ed.). Mahwah, NJ: Laurence Erlbaum.

National Transportation Safety Board. (1988). Safety Recommendation R-88-1 through -9. http://www.ntsb.gov/doclib/recletters/1988/r88_1_9.pdf

National Transportation Safety Board. (2013). Head-on collision of two Union Pacific Railroad freight trains. NTSB Number RAR-13-02. http://www.ntsb.gov/investigations/summary/rar1302.html

Stevens, S.S. (1961). Psychophysics of sensory function. In W.A. Rosenblith (Ed.), *Sensory communication* (pp. 1-34). New York: MIT Press and John Wiley.

Swets, J.A. & Pickett, R.M. (1982). *Evaluation of diagnostic systems. Methods from signal detection theory.* New York: Academic Press.

Taylor, M.M. (1963). Visual discrimination and orientation. *Journal of the Optical Society of America, 53,* 763-765