DOT/FAA/AM-04/14

Office of Aerospace Medicine
Washington, DC 20591

Predictive Validity of the Aviation Lights Test for Testing Pilots With Color Vision Deficiencies

Nelda J. Milburn
Henry W. Mertens
Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

September 2004

Final Report

This document is available to the public
through the National Technical Information
Service, Springfield, Virginia 22161.

U.S. Department
of Transportation
**Federal Aviation
Administration**

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

Technical Report Documentation Page

1. Report No. DOT/FAA/AM-04/14	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Predictive Validity of the Aviation Lights Test for Testing Pilots With Color Vision Deficiencies	5. Report Date September 2004
	6. Performing Organization Code
7. Author(s) Milburn N, Mertens H	8. Performing Organization Report No.
9. Performing Organization Name and Address FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125	10. Work Unit No. (TRAIS)
	11. Contract or Grant No.
12. Sponsoring Agency name and Address Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	13. Type of Report and Period Covered
	14. Sponsoring Agency Code

15. Supplemental Notes

Work was accomplished under approved task HRR-522.

16. Abstract

Background. The color filters of the Farnsworth Lantern (FALANT) were changed to meet the Federal Aviation Administration's signal color specifications, thereby creating a job-sample color vision test called the Aviation Lights Test (ALT) that is used for secondary screening of air traffic control specialist applicants in the terminal option. The purpose of this experiment was two-fold: to determine whether the ALT could be used in place of the FALANT for testing pilots and whether the altered filters in the ALT (primarily, a more highly saturated red) improved its predictive validity with the criterion instrument called the signal light gun (SLG). The SLG is used by air traffic controllers to communicate with pilots in aircraft experiencing radio failure within the airport terminal area. **Method.** Participants were 145 individuals with moderate to strong red-green color vision deficiency, 10 individuals with minimal color vision anomalies, and 227 individuals with normal color vision, as classified by a Nagel anomaloscope. Participants identified 3 series of 9 pairs of colored lights of the FALANT and the ALT. A subset of 82 participants also identified the color-coded signals of the signal light gun test (SLGT). **Results.** The frequency of confusing white and green lights was similar for all tests; however, as predicted, errors involving red targets were reduced for the color deficient sample for the ALT relative to the FALANT. Compared with the FALANT, the use of signal colors in the ALT had little effect on cross-tabulated pass/fail outcomes with the SLGT, K(82)= .70 and .675. **Conclusions.** Results suggest that if the ALT is administered and scored with procedures identical to the FALANT, the incidence of passes and failures for pilots with color vision deficiencies will be essentially the same for the two tests.

17. Key Words FALANT, Color Vision, Signal Light Gun, Color Deficiency, Screening Tests, Lantern Test, Aviation Lights Test, Pilot Color Vision, Job-Sample Color Vision Test	18. Distribution Statement Document is available to the public through the National Technical Information Service Springfield, Virginia 22161

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 21	22. Price

Predictive Validity of the Aviation Lights Test for Testing Pilots With Color Vision Deficiencies

INTRODUCTION

The color vision requirements of the Code of Federal Regulations (CFRs) for aeromedical certification for Classes I, II, and III (Title 14 CFR 67.103, .203, and .303; respectively) are the same, i.e., "Ability to perceive those colors necessary for the safe performance of airman duties." That requirement applies to both commercial and private pilots and is based on the international use of a color-coded signal system used by air traffic control specialists in control towers to direct aircraft in case of radio failure (International Civil Aviation Organization, ICAO, 1988). Interpretation of the code requires correct identification of 3 colors (red, green, and white) presented as steady, flashing, or alternating colors. Consequently, a color perception error could lead to an erroneous decision by the pilot and an increased risk of conflict with other aircraft.

Aviation Medical Examiners (AMEs) administer the initial color vision screening tests to pilots. The current *Guide for Aviation Medical Examiners* (USDOT, FAA, 2003) lists the Farnsworth Lantern (FALANT) and a variety of other color vision screening tests that the Federal Air Surgeon has approved for use, based on empirical research (Mertens & Milburn, 1993). If the pilot fails the AME-administered color vision test, the pilot may request authorization (from the Aerospace Medical Certification Division or from the Regional Flight Surgeon) to demonstrate his/her ability to perceive the necessary aviation colors to obtain a medical certificate and color vision proficiency letter. Under the current standard, color vision ability is demonstrated by correctly identifying the color of three 5-sec duration steady lights with randomly ordered colors (red, green, or white) that are shown at 3-min intervals from a control tower at a distance of 1,000 ft and again at 1,500 ft with all 3 colors being displayed at least once at each distance before completing the test. If the applicant fails to name the color of each of the 6 lights correctly while the light is shown, the applicant fails. That test is called the Aviation Signal Light Gun Test (SLGT) and is administered by a Flight Standards District Office aviation safety inspector as outlined by USDOT, FAA Orders 8400 and 8700.1. (Both orders are currently being revised; therefore we recommend reviewing the latest update).

The FALANT has been used successfully as an occupational color vision test for civilian aviation, marine, and railway signal lights (Birch & Dain, 1999; Cole & Vingrys, 1982). It has been used by the U.S. Navy for more than 50 years as the selection test for all pilots, deck personnel, and other color vision sensitive occupations. In research it has demonstrated a high predictive validity (r= .79 reported in Mertens & Milburn, 1993; r= .79 reported in Steen, Collins, & Lewis, 1974) for identifying the safety-critical color-coded signals of the SLGT that are used by air traffic control specialists in control towers. To obtain a stronger correlation between the lantern and practical abilities, the FALANT was redesigned with the intent for the resulting Aviation Lights Test (ALT, Mertens, Milburn, & Collins, 2000) to serve as a practical color vision test for terminal option air traffic control specialist applicants. The primary objective for development of the ALT was to use test colors that conformed to specific occupationally relevant signal colors. In contrast, the original colors of the FALANT were selected because of their similar appearances to deficient observers resulting in color confusion errors among the deficient observers but not for normal color vision observers (Farnsworth & Foreman, 1946a).

This study compared performance on the ALT, the SLGT, and the FALANT as a function of red-green color vision deficiency for the purpose of validating the effectiveness of the ALT for screening pilots. The 3 tests share a similar purpose; notably, they each serve as work-sample color vision screening tests. The FALANT screens U.S. Navy personnel (Hackman, Holtzman, and Walter, 1992; Birch & Dain, 1999) including Navy Seals (www.sealchallenge.navy.mil/faqmedical.htm); the ALT screens air traffic control specialist applicants (Mertens, Milburn, & Collins, 2000); and the SLGT evaluates pilots (USDOT, FAA Orders 8400 and 8700.1). The latter test has the added distinction of using the actual instrument used on-the-job. The objective of comparing test performances was to determine whether the ALT could be used as a substitute for the FALANT. That issue was raised because the Farnsworth Lanterns belonging to the offices of FAA Regional Flight Surgeons were modified to produce the ALT that was needed for work-sample screening of air traffic control specialist applicants for the terminal work option. The ALT required modifying the color characteristics of the red and green lights to make all colors in the lantern met the specifications of both the Federal Aviation Administration (FAA, 1988) and International Civil Aviation Organization (ICAO, 1988) for the red, green, and white signal light colors on aircraft. (A graphic illustration of the FAA and ICAO signal color gamuts is available in Mertens, Milburn and Collins, 2000). Those modifications made it uncertain whether the tests were more or less difficult than the original FALANT and raised the question of whether it was appropriate to use the modified lanterns to screen pilots. (See Table 1 for the CIE 1931 color coordinates of the 3 tests.) Consequently, the more important question was whether the ALT accurately predicts, or predicts at least as well as the FALANT, performance on the colored lights of the signal light gun (because of its use in control towers).

In addition to the chromaticity modifications made to the red and green filters, the administration and scoring procedures of the ALT differ from the FALANT as follows:

1. The ALT procedure demonstrates each of the 3 test light colors prior to testing; FALANT does not.
2. The ALT always involves 3 random series of the 9 pairs of lights produced by the lantern and requires scoring of all 27 pairs of lights. FALANT passes individuals who identify the first 9 light pairs without error. However, if an error is made, the FALANT presents the 9 light pairs in 2 more series.
3. On the ALT, the participant fails if 2 or more incorrect responses occur during the 27 trials. Failure on the FALANT results if 3 or more errors are made during the 18 trials of series 2 and 3.
4. The ALT is given in a very dim room that approximates the light level of the air traffic control (ATC) tower cab at night. The FALANT is given in a normally lit room.

Several decisions regarding the administration (including ambient illumination) and scoring must be addressed if the altered FALANTs (the ALTs) are considered an appropriate color vision screening test for pilots. For example, Birch and Dain (1999) reported that Schmidt (1951) found slightly better lantern performance in a darkened room, but that finding was not supported by Dain, Honson, and Ang's (1988) study on the effect of two lighting conditions on the performance of the FALANT. The findings of Dain et al. supported the FALANT designers' reports (Farnsworth & Foreman, 1946a and 1946b, cited in Birch & Dain, 1999) that room illumination was not critical but recommended 60 to 300 lux as a testing environment. Jones, Steen, and Collins (1975) also compared 2 similar ambient lighting conditions of the FALANT to test whether one was a better predictor of SLGT performance and concluded that there was no advantage to testing in the dark. Still, a slightly higher probability of a miss (miss rate) occurred when the FALANT was administered in a darkened room (compare probability of a miss) of < 0.12 to < 0.06 for lighted condition).

Similarly, Mertens, Milburn, and Collins (2000) found very few differences in performance of the ALT administered under low illumination (similar to the illumination at night in an ATC tower) and high illumination (similar to the illumination of an office). In that study, individuals with normal color vision passed both presentations of the

Table 1. Chromaticity of Red, Green, and White Lights in the FALANT, the SLGT, and the ALT

		CIE 1931 Coordinates	
	Color	x	Y
FALANT	Red	.61	.29
	Green	.20	.70
	White	.47	.41
SLGT	Red	.68	.27
	Green	.18	.35
	White	.41	.36
ALT	Red	.67	.32
	Green	.18	.67
	White	.45	.42

ALT, with the exception of one who made 1 error in the low illumination condition. Therefore, the manufacturer-recommended test condition of a normally lighted room was used to administer the FALANT; and the ALT was administrated in both dark and light testing room conditions. The order of the 3 test presentations was controlled. Also, the participant's ALT responses were scored using both the ALT and the FALANT guidelines.

Aside from the issues of ambient lighting and scoring was the more important concern of the effect on performance caused by the altered colored filters in the ALT and the extent to which the alterations affected the likelihood of passing the ALT compared with the FALANT. The color change to a more highly saturated ALT red was predicted to reduce the frequency of red color confusions when compared with the number on the FALANT, but green-white confusions should not be affected by the very small change in the ALT green. To test that hypothesis, errors on the FALANT and the ALT were compared separately for each color.

Although errors on red targets were predicted to be lower on the ALT compared to the FALANT, the poor performance on the green and white targets for participants with strong to severe color vision deficiencies was predicted to result in failure of both tests. It was also predicted that a few people with near-normal color vision would exhibit the greatest benefit from the increased red saturation. Based on those 2 premises, if the scoring method was held constant, it was predicted that pass-fail performance on the 2 tests for moderate-to-severe deficients would be little affected by the color changes in the ALT. Also, agreement (measured by *Kappa*) was expected to be high between the pass-fail decisions of the FALANT and the ALT, regardless of which scoring procedure was used for the ALT.

METHOD

Participants

Prior approval for all procedures and use of human participants was obtained from the Institutional Review Board of the Civil Aerospace Medical Institute. Volunteers were recruited and paid by an independent contractor. The informed consent was obtained prior to participation, and each participant was free to withdraw from the experiment without prejudice at any time.

Participants were 145 individuals with moderate to severe red-green color vision deficiency, 10 with minimal color vision anomalies and 227 with normal color vision. All volunteers had at least 20/30 corrected visual acuity in both near and distant vision as determined with the Bausch and Lomb Orthorater. Their observations for this experiment were conducted in conjunction with 2 unrelated experiments. Table 2 shows the distribution in each color vision classification of the 190 participants of Experiment 1 and the 192 participants in Experiment 2. A subset of 82 participants from Experiment 1 were administered the SLGT. Color vision classification was performed with the Nagel anomaloscope. Table 3 describes the classification criteria. Depending upon data availability and the appropriateness of analysis, the results are reported either for (a) Experiment 1 (n=190), (b) the subset (n=82) that received the SLGT, (c) only the moderate-to-severe color deficient groups (n=145), (d) only the participants with normal color vision (n=227), or (e) all participants (n=382), and the reported findings are distinguishable by the number of participants.

Apparatus

The Aviation Lights Test (ALT) makes use of the body and mechanisms of the Farnsworth Lantern (FALANT, Macbeth Corporation, Newburg, NY). The

Table 2. Participants by Experiment by Color Vision Classification

	Diagnosis [a]								
	N	M	SP	EP	DP	SD	ED	DD	Total
EXP 1	102	4	11	9	17	11	16	20	190
EXP 2	125	6	13	8	10	13	10	7	192
Total	227	10	24	17	27	24	26	27	382
SGLT Subset[b]	39	4	11	7	8	8	3	2	82

[a] See Table 3 for a description of the diagnostic categories and the anomaloscope classification procedure.

[b] A subset of participants from Experiment 2 were administered the SGLT.

Table 3. Anomaloscope Classification

Code	Diagnosis	Anomaloscope Criteria for Diagnosis
		Normal/Almost Normal Color Vision
N	Normal	Midpoint of color matches between 36 and 45 with a range less than 11 units
M	Minimal Anomaly	Midpoint of color matches between 33 and 48 with a range less than 16 units
		Moderate Degree of Deficiencies
SP	Simple Protanomalous	Midpoint of color matches greater than 40.5 with a range of 16 to 25 units or midpoint greater than 48 and range less than 16
SD	Simple Deuteranomalous	Midpoint of color matches less than 40.5 with a range of 16 to 25 units with little variation in matching brightness or midpoint less than 33 and range less than 16
		Strong/Severe Degree of Deficiencies
EP	Extreme Protanomalous	Midpoint of color matches greater than 40.5 and/or color matching range of 26 to 72 units with a systematic decrease in matching brightness as test color approaches red
ED	Extreme Deuteranomalous	Midpoint of color matches less than 40.5 and color matching range of 26 to 72 units with little variation in matching brightness
DP	Dichromat Protan (Protanope)	Color matching range of 73 units with a systematic decrease in matching brightness as test color approaches red
DD	Dichromat Deutan (Deuteranope)	Color matching range of 73 units with little variation in matching brightness

main differences between the two lanterns concern the specific chromaticities of the red and green lights. As discussed earlier, those filters of the Farnsworth Lantern were changed in the ALT to produce colors that met the specifications for the red and green signal colors, as given by both the FAA (2004) and the ICAO (1988). Custom red and green filters were designed and manufactured by Kopp Glass Company (Pittsburg, PA). The foremost difference was from a desaturated to a highly saturated red. The slight change in green to ensure that it met FAA specifications was probably not of practical significance. The original white lights of the FALANT were not changed, since they already met the signal color specifications for white. Both instruments show 9 different pairs of lights. Detailed descriptions of the FALANT (Cole & Vingrys, 1982), the SLGT (FAA Order 8700.1 and Steen, Collins, & Lewis, 1974), and the ALT (Mertens, Milburn, & Collins, 2000) are available elsewhere. An alternative for the FALANT, produced by the Stereo Optical Company called the OPTEC900, was evaluated by Laxar, Wagner, and Cotton (1998) by comparing performance on the alternative with the original lantern test and; subsequently, the OPTEC900 was recommended for use by the U.S. Navy.

Procedure

The illumination for the ALT *dark* room condition, measured at the position of the ALT, was approximately 1 lux, which was recommended as the upper limit for illumination at the windows in the ATC tower (Kaufman & Christensen, 1987). This nominal illumination level was achieved by placing a small desk lamp fixture (with a clear, incandescent 7-W, 120 V a.c. bulb) approximately .91 m (3 ft) behind the participant, with the lamp pointed toward the ceiling.

The SLGT was presented according to FAA Order 8700.1 and was the final test for the morning groups and the first test for the afternoon groups. Participants were escorted to locations 1000 feet and 1500 feet from the building and observed the signal light gun presented from a third-floor window. Order of presentation of the dark (1 lux) and light (300 lux) ALT and the FALANT

(300 lux) testing was controlled. Observations from both instruments were made from a 2.438 m (8 ft) viewing distance, and the resultant size of all light points was 3.5 arc min. The constant vertical separation of the 2 apertures was 13.0 mm, or 18.3 arc min at the recommended viewing distance. That separation simulates the case of an aircraft with a 25-ft (7.62 m) wingspan at a distance of approximately 1,432 m (4,700 ft). Both lanterns presented pairs of lights involving the colors red, green, or white. A 0.1-inch aperture that subtended approximately 3 arc min of visual angle created each light of a pair. The light-pairs within each series of 9 were always presented in random order with both the FALANT and ALT lanterns.

An exception to the administration procedure was that the ALT test administrator demonstrated light pairs numbered 1 and 2 while saying "This is green over red" and "This is white over green." The ALT presented 3 random series of the 9 pairs of lights and required scoring all 27 pairs, with a pass criterion of not more than 1 error. As with the FALANT, only the color names red, green, and white were allowed. The instructions asked the participant to identify the colors of the lights, naming the top color first.

Scoring

Although 3 series of the 9 color-pairs were also given with the FALANT, the scoring procedure initially involved scoring only the first series of the 9 color-pairs. If no error was made, the observer passed the test. If one or more errors occurred during the first series, then series 1 was ignored and series 2 and 3 of the 9 light-pairs were scored. If the *average* number of errors for series 2 and 3 was greater than 1 (i.e., more than 2 total errors on series 2 and 3), the observer failed the test. If the color of either or both lights of a light-pair was incorrectly identified, it was counted as 1 error.

RESULTS AND DISCUSSION

Although the ALT and FALANT are similar in many ways, distinct differences exist between the 2 tests such as the recommended ambient illumination during testing, the saturation of the red color presented, the pre-test training and the instructions given, and the scoring techniques. The effects of each of those factors were examined, with the exception of the pre-test demonstration of colors in the ALT instructions. That procedural difference between the ALT and the FALANT was not isolated and addressed in this study. Finally, measurement of the extent of agreement between passing and failing the ALT, the FALANT, and the SLGT is also included.

Analysis Methodology

For many years, Cohen's (1960) Kappa has been the statistic of choice for determining the strength of agreement between two raters or two tests and is preferred over the observed proportion of overall agreement because of its correction for chance. However, the scientific community is experiencing an ongoing debate concerning the appropriateness of the use of Kappa (Maxwell, 1977; Spitznagel & Helzer, 1985; Uebersax, 1987a, 1987b, 1988; Cicchetti & Feinstein, 1990) and is questioning whether other statistics may provide a better index of agreement. The use of Kappa is criticized for a variety of reasons; specifically, Kappa is not comparable across studies when the proportions of cases belonging to trait or diagnostic categories vary (Thompson & Walter, 1988; Feinstein & Cicchetti, 1990). Furthermore, the Kappa values cannot be generalized to a broader population unless the categorical composition of cases within a single study match those found in the general population. Typically, color vision studies include larger sample proportions of color deficient participants than are found in general populations; further, the distribution of diagnostic classifications is not matched to known population proportions. Hence, the Kappa statistics may be sample specific. One solution to the categorical composition of cases issue is to measure agreement separately as a function of the presence or absence of a trait. That is not always possible because criterion measures, especially of some latent traits, are not always obtainable. Fortunately, such an analysis is possible for color vision studies because criterion measures are obtainable. The analysis requires calculating a separate Kappa for participants in each of the anomaloscope-classified normal/deficient categories to determine agreement between the two screening tests. However, because very few participants with normal color vision fail color vision screening tests, the resulting Kappa would be extremely low and would not give a true picture of the agreement between tests (that issue will be discussed later in reference to specific analyses).

Some have criticized Kappa because it is too conservative in its estimates of agreement (Feinstein & Cicchetti, 1990; Guggenmoos-Holzmann, 1993); as an alternative, Uebersax (1988) suggested concentrating on sensitivity and specificity ratings. Still others presented the summation of those indices forming an efficiency index that measures the usefulness of a screening test (Birch & McKeever, 1993). The latter defined specificity as the percentage of normal trichromats a test correctly identified and sensitivity as the percentage of correctly identified color deficient observers. In further explanation of those conditional probabilities (the probability that cases will be classified correctly, given that the trait, condition, or disease is present), sensitivity and specificity can be interpreted as the proportion of positive and negative cases correctly classified (Uebersax, 1988). Conversely, when comparing two *screening* tests for agreement using the same contingency table formula for sensitivity and specificity without a known criterion classifier (such as positive or negative pathology results) with an arbitrary arrangement of screening tests, the resulting disparate sensitivity/specificity values cannot be interpreted with the same meaning as originally defined but, rather, should be interpreted as the predicted performance on one test from performance on the other. Spitzer and Fleiss (1974) and Cicchetti and Feinstein (1990) recommend using *proportions of specific agreement* to interpret the estimated conditional probabilities; which, for this study, means that given that one of the randomly selected screening tests passes a participant, the other screening test will do so also. That agreement between the two tests, based on passing the same individuals, is referred to as the proportion of specific agreement for a positive rating (Ps+). Likewise, if a large proportion of participants who fail one screening test also fail the other, the proportion of specific agreement for negative ratings will be high (Ps-).

In summation, because Kappa is not used in this study for testing the null hypothesis (i.e., independence of the 2 tests) but rather for descriptive purposes such as gauging comparisons between tests as a function of administration or scoring methods while using the *same* sample of participants as a whole, and, in some instances, separately for the normal and deficient color vision groups, Kappa (K) is considered appropriate. Especially since several additional measures such as overall agreement (OA), proportions of specific agreement (Ps+ and Ps-), Cochran-Mantel-Haenszel's odds ratios (CMH), sensitivity and specificity (Sn and Sp), test efficiency, and Spearman's rho (r) are also presented for their supportive and complementary descriptive values, Kappa is reported as an additional measure of agreement.

Ambient Illumination and Scoring Procedures

As stated earlier, guidelines for administration of the FALANT and the ALT specify different scoring procedures and ambient illumination levels during testing. The FALANT was administered in a lighted room in accordance with the manufacturer's recommendations, but because the ALT is given in a very dim room that approximates the level of the air traffic control (ATC) tower cab at night, the first analysis must examine the effect of ambient illumination on color identification performance for the ALT.

Ambient Illumination

Agreement was very high for pass/fail performance (using the ALT scoring criterion) comparing dark and light presentations of the ALT for participants in Experiment 1, *K* (190)= .933 with only 6 people who had a different ALT outcome under the two testing conditions. Four individuals passed the lighted room but failed the darkened room condition (3 moderate deutans and 1 strong deutan), and 2 individuals (a moderate protan and a moderate deutan) passed the darkened room but failed the lighted room condition. Kappa was also computed between the lighted, *K* (190)= .798 and darkened, *K* (190)= .789 ambient conditions of the ALT (using the ALT scoring procedure) crosstabulated with the FALANT, and good agreement was found.

Pass/fail agreement for the lighted and darkened room conditions of the ALT (using the FALANT scoring criterion) was computed, *K* (190)= .942 and then the 2 conditions were separately compared with performance on the FALANT *K*(190)= .931 and .918, light and dark, respectively. See Table 4.

Scoring

It is apparent that the ambient lighting had little effect on the pass/fail status of participants when the scoring procedure was held constant. When screening pilots, the question remains whether the *altered*-FALANT (i.e., ALT) should be scored using the original FALANT procedure or the ALT-designed procedure. It should be noted that a more stringent disposition criterion is used when the Aviation Lights Test is used for screening air traffic control specialist applicants (fail with a total of 2 or more errors in all 3 series) than when the FALANT is used to test pilots (fail with 3 or more errors in series 2 and 3). Comparing scoring procedures for the ALT produced an agreement statistic of *K*(190)= .966 in the lighted room condition and .887 when administered in a darkened room. The most notable differences between the Kappas were related to the scoring criterion for pass/fail when the ALT was compared with the FALANT. Under both room lighting conditions, higher Kappas were produced using the FALANT scoring of performance on the ALT instrument when compared with performance on the FALANT than were found using the ALT scoring procedure. In general, the Kappas were less than .80 using the ALT scoring procedure and were greater than .91 for the same conditions when the scoring methods were matched to the FALANT.

Table 4. Crosstabulation of Pass/Fail Performance on the ALT under Light and Dark Ambient Conditions with the FALANT when All Tests Were Scored Using the FALANT Pass/Fail Criterion

Kappa=.942	Dark ALT		
Light ALT	Pass	Fail	Total
Pass	121	0	121
Fail	5	64	69
Total	126	64	190

Kappa=.931	FALANT		
Light ALT	Pass	Fail	Total
Pass	119	2	121
Fail	4	65	69
Total	123	67	190

Kappa=.918	FALANT		
Dark ALT	Pass	Fail	Total
Pass	121	5	126
Fail	2	62	64
Total	123	67	190

Errors as a Function of Stimulus Color Differences Between Tests

Errors in color identification were rare on both the ALT and the FALANT for normal color vision observers or individuals with minimal color vision anomalies (Figure 1). Those with moderate and more severe color vision deficiencies of both protan and deutan types made frequent color confusions on both tests. In addition, total error scores increased with severity of deficiency.

The previous section demonstrated that the use of signal colors in the ALT had little effect on pass/fail outcome when compared with performance on the FALANT and that some differences were attributable to scoring procedures. The results shown in Table 4 indicate that performance was highly related on the two tests. However, the normal color vision group comprised a large portion of the total participants, and that group rarely made errors. Consequently, the next analysis examined the effects of the altered color filters related only to the deficient color vision group (those with moderate to severe deficiencies, n=145) to determine whether the more saturated red aided in the identification/discrimination of colors among that group. The within-groups factors were test (ALT and FALANT) and color (red, green, and white), and the between groups variables were type (protans and deutans) and degree (simple— includes simple protan (SP) and simple deutan (SD), moderate—includes extreme protan (EP) and extreme deutan (ED), or severe— includes dichromat protan (DP) and dichromat deutan (DD)) of color vision deficiency. The 10 participants classified with minimal anomaly were not included in this analysis. The ANOVA that compared accuracy of performance of identifying the colored lights using only participants with defective color vision (n=145) revealed significant differences between tests, F (1,139) = 21.37, p < .001. There were significant interactions between test and color, F (2,138) = 36.16, p < .001. Errors on red targets using the ALT were fewer than in the FALANT for all classifications of red-green defects, particularly in those classifications involving strong/severe defects. However, the frequency of errors on white and green lights was similar in both tests.

Significant differences were also found between the colors, F (2, 138) = 107.44, p < .001, with the fewest mean errors occurring on red targets (4.9), followed by green targets (11.2), and most errors involving the target color white (15.4). Although the fewest number of errors involved misnaming red targets, some improvement was noted in performance on the ALT, compared with the FALANT, presumably because of the increased saturation of the color red. Significant differences were found between tests as a function of color by deficiency type, F (2,138) = 3.07, p = .049. The Extreme Protan group was most aided by increasing the saturation of red filters.

Significant differences were found between the FALANT and the ALT as a function of degree, F (2, 139) = 4.05, p = .019, but not between tests by type or between tests as a function of type by degree. Extreme and

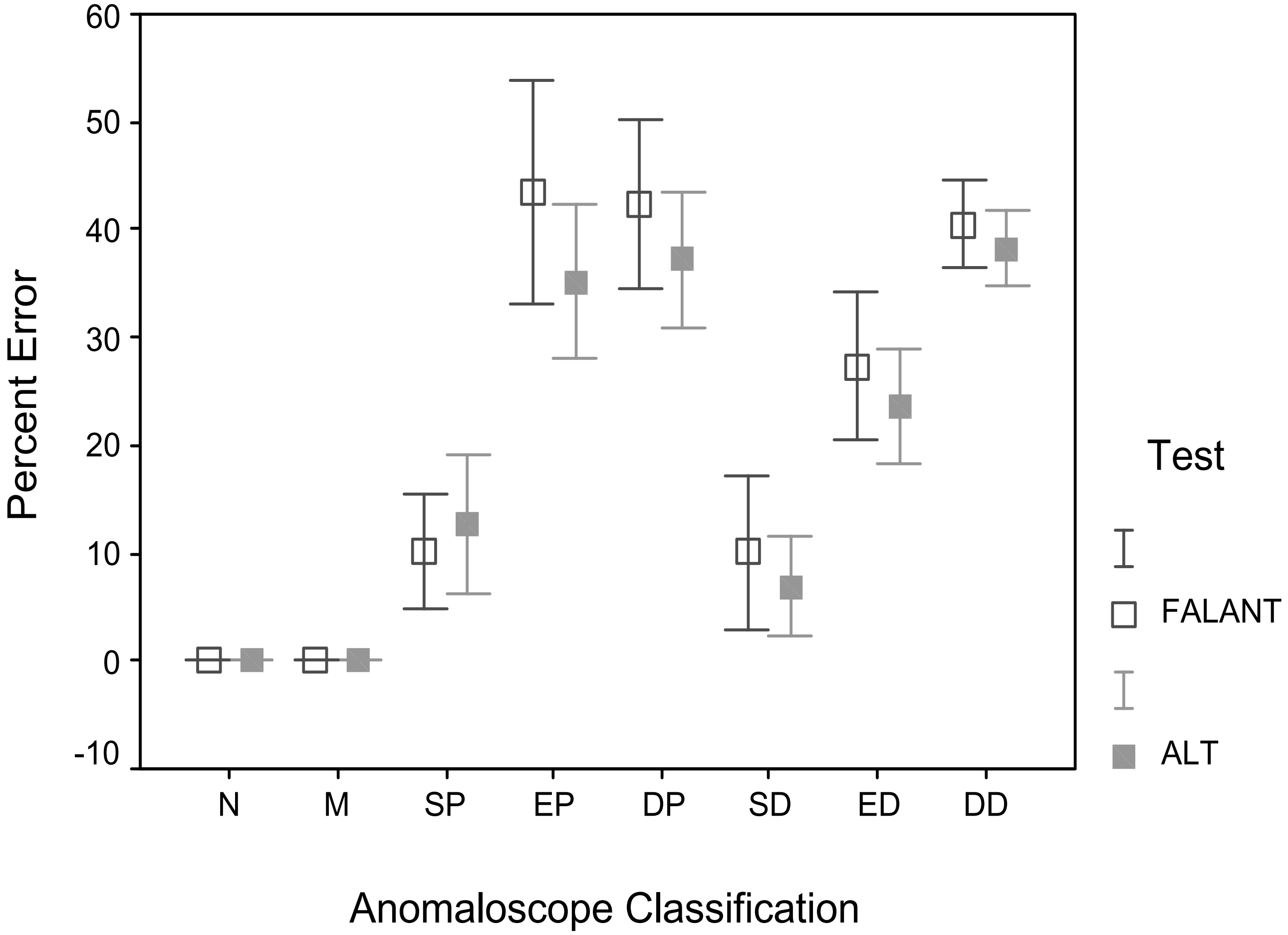

Figure 1. FALANT vs. ALT: Error Rates Over Series 2 and 3

Dichromat deficient participants experienced the most notable improvement as shown on Figure 2. Likewise, the significant difference previously noted between the colors was evident also as a function of degree, F (4, 278) = 7.49, p < .001, but not color by type or color by type and degree. Significant performance differences were evident between the colors of the FALANT and the ALT as an interaction with degree, F (4, 278) = 4.62, p = .001.

Color Confusions as a Function of Type and Degree of Color Vision Deficiency

A separate ANOVA analyzed errors to determine whether the same color confusions were made on both the FALANT and the ALT with similar frequencies and also for the purpose of isolating the qualitative types of color confusions. Errors were categorized by target color and response. For example, participants could misidentify red targets either as green or white responses. Likewise, confusion errors on white and green targets were counted separately and used as the dependent variable in the ANOVA. The analyses of interest were those involving test as a function of color by specific confusion and that interaction with type, degree, and type by degree of deficiency. The ANOVA results indicate that significant differences in performance occurred between the FALANT and the ALT as a function of the specific color confusion made, F (2, 138) = 20.05, p < .001, and as an interaction with degree of deficiency, F (4, 278) = 4.51, p = .002. Figures 3 and 4 show that compared with the FALANT, the percentage of errors involving red targets was greatly reduced in the ALT, mainly the result of fewer red targets being called white. Green was called white more often on the ALT than on the FALANT, but the percentage of misidentified green targets was about the same on both tests. Interactions involving type of deficiency were not significant.

Comparison of the ALT and FALANT with the SLGT Pass/Fail Performance

Recall that the SLGT uses the actual instrument that is used on the job by air traffic control specialists in the control tower to communicate with pilots without work-

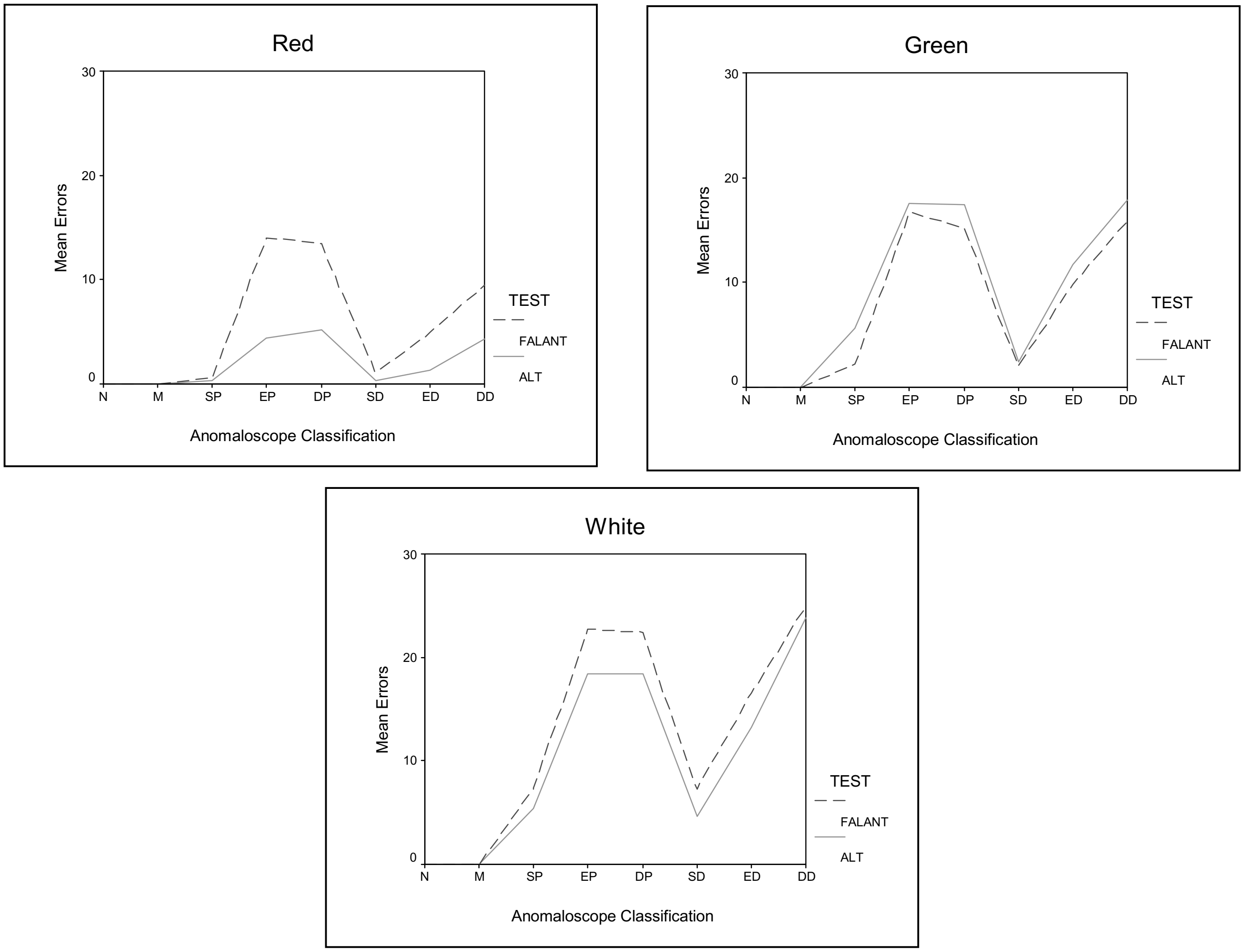

Figure 2. Percent Error on Red, Green, and White Lights on the FALANT and the ALT as a Function of Anomaloscope Diagnosis

ing radios. Consequently, it is considered the criterion measure of job performance. Ideally, the screening test should accurately predict performance on the job. However, passing the SLGT cannot be predetermined strictly by one's diagnostic category. Table 5 supports that claim and, notably, it is even evident in the relatively small subset of 82 participants. Therefore, the most important comparisons for these data are between the FALANT, the ALT (using the FALANT scoring criterion), and the SLGT and are shown in Figures 5 and 6, which compare the performance on the lights by qualitative responses as a function of diagnostic type. Notice that performance was generally better both for the protans and the deutans on the SLGT, compared with the other instruments. The most likely factor contributing to improved performance was the greatly increased brightness of the SLGT.

ALT Validity, Sensitivity, and Specificity

Using the SLGT as the criterion measure, validity was calculated for the ALT and the FALANT as screening tests and was assessed by the statistic Kappa. A test is considered valid if it measures what it purports to measure. The validity of the ALT pass/fail performance, compared to the SLGT performance (ability to identify color-coded signal lights in the aviation environment), was moderately high, $K(82) = .675$.

The sensitivity of a selection test is the probability that individuals who cannot pass the criterion (SLGT) will also fail the selection test (the ALT or the FALANT). Sensitivity of the ALT (when used with the FALANT testing and scoring procedures) was high, correctly predicting failure on the SLGT for 90% of the 82 participants. The specificity of a test is the probability that individuals who can pass the criterion

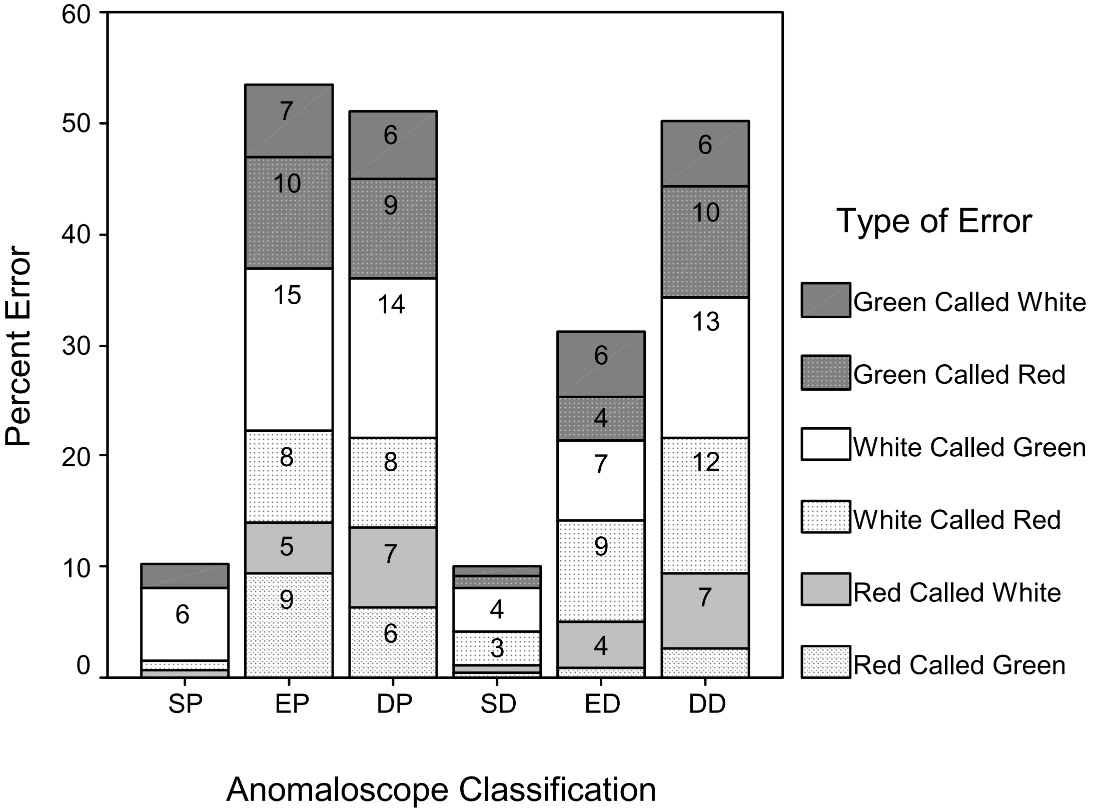

Figure 3. FALANT Color Confusions as a Function of Anomaloscope Diagnosis

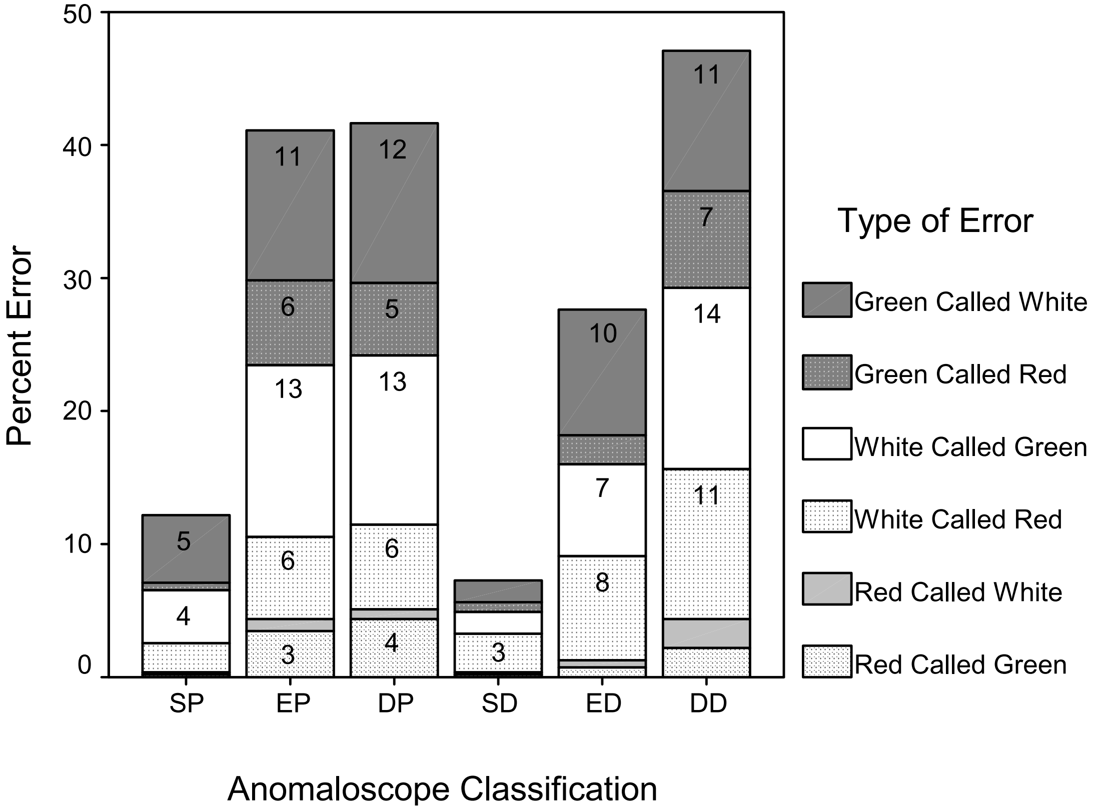

Figure 4. ALT Color Confusions as a Function of Anomaloscope Diagnosis

Table 5. Crosstabulation of Signal Light Gun Test Pass/Fail Performance by Anomaloscope Diagnosis

	Signal Light Test		
Anomaloscope Classification	Pass	Fail	Total
N	40	1	41
M	4	0	4
SP	7	3	10
EP	4	3	7
DP	1	7	8
SD	5	2	7
ED	0	3	3
DD	1	1	2
Total	62	20	82

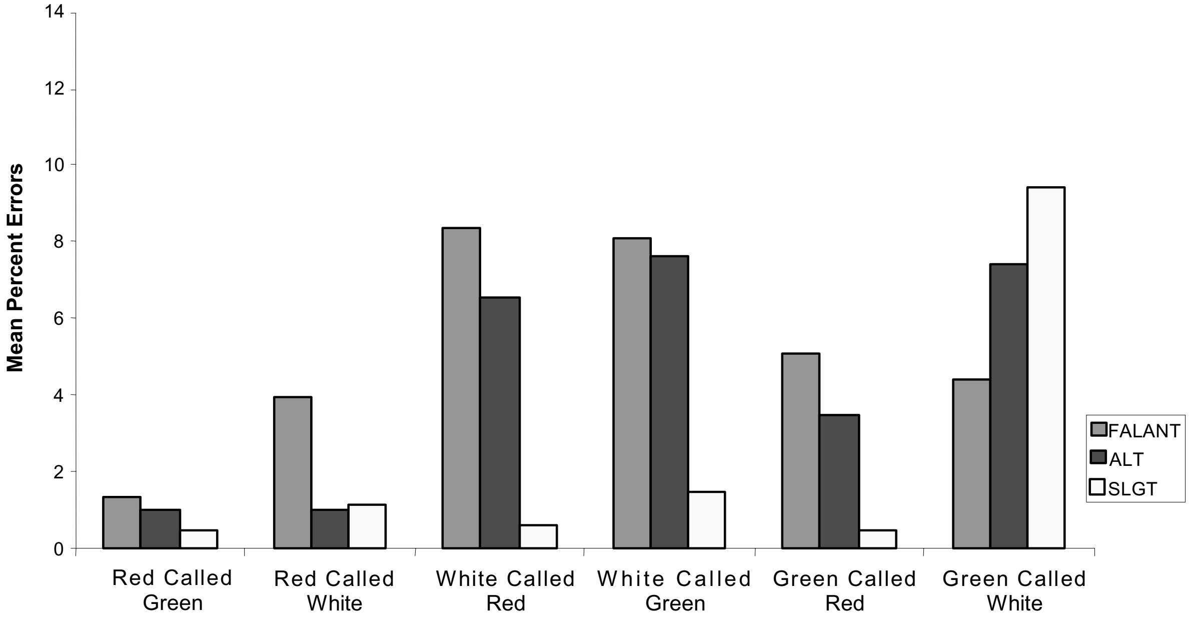

Figure 5. A Comparison of the SLGT, ALT, and FALANT by Qualitative Responses for Deutans

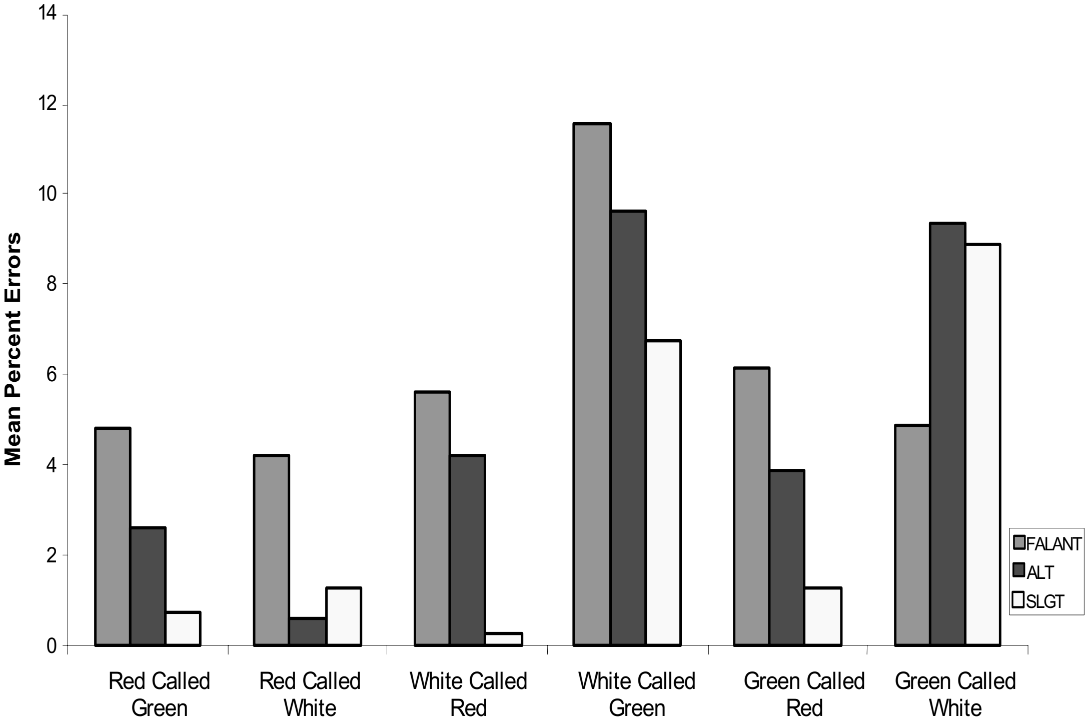

Figure 6. A Comparison of the SLGT, ALT, and FALANT by Qualitative Responses for Protans

(SLGT) will also pass the selection test (ALT). Notice on Table 6 that specificity was also high at 85.5%. Kappa, sensitivity, and specificity values were similar in magnitude to the better aeromedical screening tests evaluated in Mertens and Milburn (1993) for predicting performance on the daytime SLGT using the pilot disposition criterion. In light of the earlier discussion that cautioned against comparing Kappas across studies when the distribution among/between trait categories is either unknown or disparate, this comparison is considered reasonable because the ratio of normal to deficient participants within the sub-set who received the SLGT is approximately equal to the normal/deficient ratio (but not the diagnostic distribution) reported in the 1993 study. Therefore, it is probably safe to make comparisons in terms of "magnitude" between the Kappas obtained.

FALANT Validity, Sensitivity, and Specificity

As with the ALT, the same 3 indices (Kappa, sensitivity, and specificity) were moderately high, K (82) = .70, 90% and 87.1%, when calculated for the FALANT using the SLGT as the criterion measure. Additionally, in an analysis comparing pass/fail performance of the ALT to the FALANT, Kappa (K (190)= .931) supported the conclusion that the ALT, *when administered and scored as the FALANT*, will give results highly similar to the FALANT. Figure 7 summarizes the various Kappa comparisons between the ALT (as a function of ambient lighting conditions and scoring procedures) and the FALANT with pass/fail performance on the SLGT.

Although the original purpose of this experiment was to determine whether the ALT could be administered in lieu of the FALANT for pilot color vision screening, the broader question addressed whether the chromaticity shift, made to meet the signal color specifications of the FAA and ICAO, significantly improved the validity of the instrument (ALT) as a screening test. Comparing the Kappas for the FALANT and the ALT (*administered and scored as the FALANT and using the same sample of participants)* for predicting performance on the SLGT, K (82)=.70 and .675, no significant differences were found using the procedure described by Terry (1987). Upon closer examination of the differences between the pass/fail outcomes of the ALT and FALANT with regard to pass/fail performance on the SLGT (see Table 6), only 1 dichromat protan had a different outcome.

Sensitivity, Specificity, and Supporting Statistics as a Function of Color by Test

Table 7 presents several notable findings as a result of analyzing pass/fail performance separately for each color for the FALANT, ALT, and SLGT by categorical diagnoses of normal or deficient color vision participants. First, specificity (Sp) ratings were similar (97% or higher) for all colors of all tests—meaning that a high percentage of normal color vision participants passed all colors of all tests. Second, the sensitivity (Sn) values varied widely between colors for the same test (e.g., compare 36.8% for red and 72.5% for white targets on the ALT) and also

Table 6. Crosstabulation of Pass/Fail Performance on the FALANT and ALT (using the FALANT Scoring Criterion) with the SLGT

	Signal Light Gun Test		
FALANT	Pass	Fail	Total
Pass	54	2	56
Fail	8	18	26
Total	62	20	82

Kappa = .700
Sensitivity = 90% (18 of 20)
Specificity = 87.1% (54 of 62)

	Signal Light Gun Test		
ALT	Pass	Fail	Total
Pass	53	2	55
Fail	9	18	27
Total	62	20	82

Kappa = .675
Sensitivity = 90% (18 of 20)
Specificity = 85.5% (53 of 62)

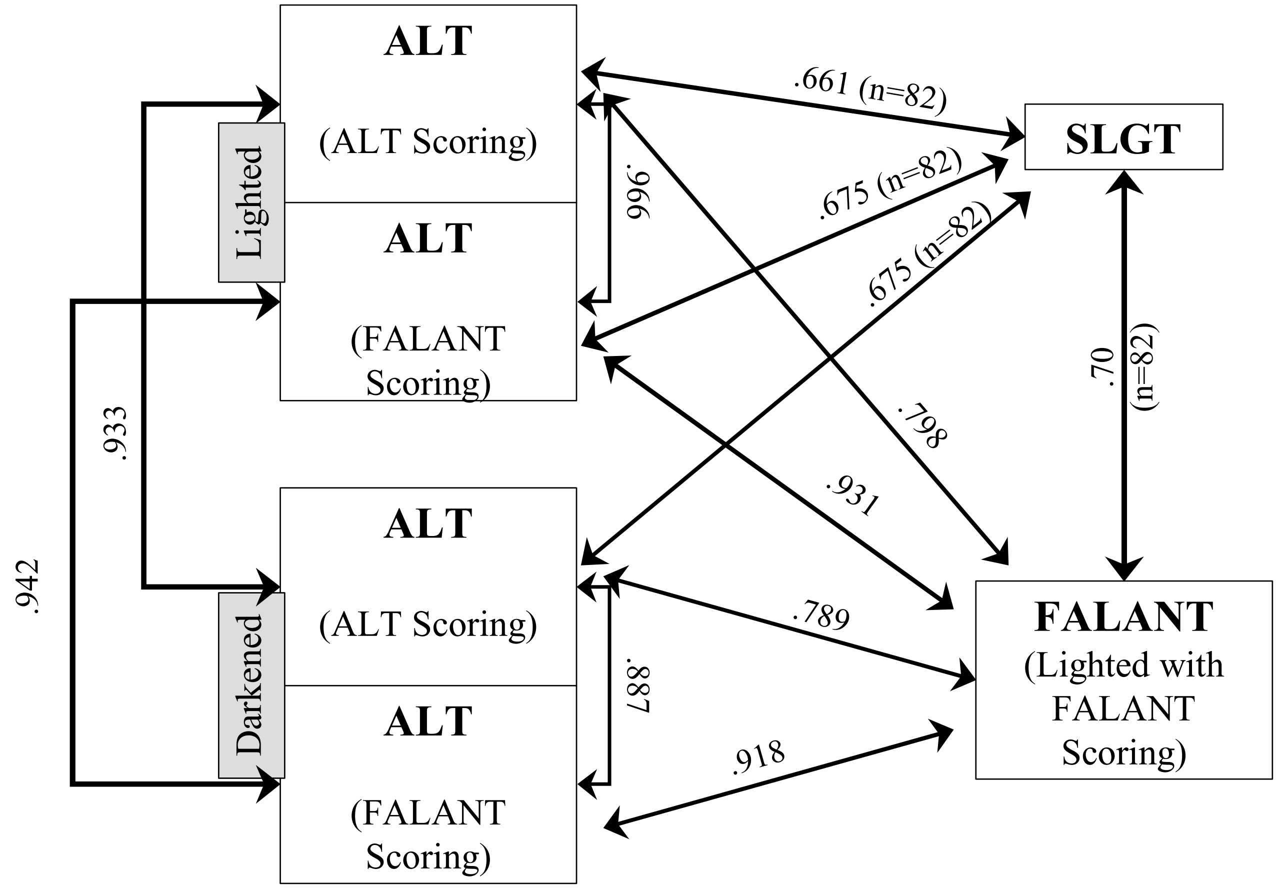

Figure 7. Kappas of Pass/Fail Performance on the ALT, FALANT, and the SLGT

Table 7. Sensitivity (Sn) and Specificity (Sp) Cochran-Mantel-Haenszel (CMH) Test of Odds Ratios, Spearman's Correlations (r), Proportions of Specific Agreement (Ps+ & Ps-), Overall Agreement (OA), and Kappa (K) Calculated from Pass/Fail Performance on the FALANT, ALT, and SLGT as a Function of Diagnosis and Color

Color	FALANT	Diagnosis		Statistics							
		Normal	Deficient	Sn	Sp	CMH	Ps+	Ps-	r	OA	K
Red	Pass	227	77	50.3	100	----	.669	.855	.613	.798	.546
	Fail	0	78								
Green	Pass	226	61	60.6	99.6	348.3	.752	.879	.684	.837	.641
	Fail	1	94								
White	Pass	225	44	71.6	99.1	283.8	.828	.907	.761	.879	.739
	Fail	2	111								

Color	ALT	Diagnosis		Statistics							
		Normal	Deficient	Sn	Sp	CMH	Ps+	Ps-	r	OA	K
Red	Pass	224	98	36.8	98.7	43.4	.530	.816	.478	.735	.393
	Fail	3	57								
Green	Pass	224	47	69.7	98.7	171.5	.812	.899	.739	.869	.716
	Fail	3	108								
White	Pass	220	42	72.5	97.8	116.3	.825	.903	.748	.875	.732
	Fail	5	111								

Color	SLGT	Diagnosis		Statistics							
		Normal	Deficient	Sn	Sp	CMH	Ps+	Ps-	r	OA	K
Red	Pass	40	35	14.6	97.6	6.8	.250	.689	.218	.561	.122
	Fail	1	6								
Green	Pass	40	23	43.9	97.6	31.3	.600	.769	.491	.707	.415
	Fail	1	18								
White	Pass	40	30	26.8	97.6	14.6	.415	.720	.345	.621	.244
	Fail	1	11								

The same individual with normal color vision failed the SLGT red, green, and white lights but did not make any errors on any other color vision test.

Table 8. Crosstabulations of Performance on the ALT and FALANT for All Participants and Separate Crosstabulations for Normal and Deficient Color Vision Participants with Error(s)/No Errors on Red, Green, or White Targets including Cochran-Mantel-Haenszel (CMH) Test Odds Ratios, Spearman's Correlations (r), Proportions of Specific Agreement (Ps+ & Ps-) and Overall Agreement (OA), and Kappa (K)

ALL PARTICIPANTS

		ALT		Statistics					
Color	FALANT	No Errors	Error(s)	CMH	Ps+	Ps-	r	OA	K
Red	No Errors	293	11	45.00	.71	.94	.66	.89	.648
	Error(s)	29	49						
Green	No Errors	268	19	432.56	.89	.96	.86	.94	.854
	Error(s)	3	92						
White[a]	No Errors	252	14	183.60	.89	.95	.85	.94	.849
	Error(s)	10	102						

NORMAL PARTICIPANTS

		ALT		Statistics					
Color	FALANT	No Errors	Error(s)	CMH	Ps+	Ps-	r	OA	K
Red	No Errors	224	3	-----	0.0	.99	---	.99	-----
	Error(s)	0	0						
Green	No Errors	223	3	-----	0.0	.99	---	.98	.006
	Error(s)	1	0						
White[a]	No Errors	218	5	-----	0.0	.98	---	.97	.012
	Error(s)	2	0						

DEFICIENT PARTICIPANTS

		ALT		Statistics					
Color	FALANT	No Errors	Error(s)	CMH	Ps+	Ps-	r	OA	K
Red	No Errors	69	8	14.57	.73	.79	.54	.76	.52
	Error(s)	29	49						
Green	No Errors	45	16	129.37	.91	.83	.76	.88	.747
	Error(s)	2	92						
White[a]	No Errors	34	9	48.17	.92	.80	.72	.89	.723
	Error(s)	8	102						

[a] Responses to the color white for 2 normal and 2 deficient participants were coded as *missing* due to an administrative error.

varied for the same color for different tests, (e.g. compare red targets 50.3%, 36.8%, and 14.6% for the FALANT, ALT, and SLGT, respectively). The disparity between the consistently high specificity values for all colors, in contrast with the lower and more variable sensitivity values for the 3 colors, reflects the accuracy with which the various colors classify normal versus abnormal color vision participants. Typically, the goal of most screening tests is to have both high sensitivity *and* specificity values with regard to pathological diagnoses. (Keep in mind that may not be the goal when screening tests are used to predict performance on the job, regardless of diagnostic classification.) However, using the rationale of summing the Sn and Sp values to form a test efficiency (TE) score (Birch & McKeever, 1993) the FALANT and ALT produced very similar values for the colors green (TE=160.2 and 168.4) and white (TE=170.7 and 170.3), but notable differences were apparent for the color red (TE=150.3 and 135.5). The FALANT obtained a somewhat higher sensitivity rating for red targets, meaning that the more saturated red filters of the ALT allowed more participants with deficient color vision to pass those red targets. In general, the sensitivity ratings for the SLGT were lower for all colors than were found for the FALANT or the ALT, once again reflecting the higher pass rate for individuals with abnormal color vision.

Table 8 completes the comparison of performance on the ALT and the FALANT as a function of color--first for all participants, then separately for the normal and deficient groups. Notice that agreement is high (K= .854 and .849) between the 2 tests for all participants for green and white colored targets but, as noted earlier, fewer errors were made on red targets on the ALT than on the FALANT, hence resulting in a lower agreement between the 2 tests (K= .648). Next, notice that overall agreement on the 2 tests was very high (>.97) for all colors for participants with normal color vision, yet the Kappa statistics revealed virtually no agreement (<.012). Kappa statistics reported on Table 8 provide empirical evidence demonstrating why Kappa is not comparable across studies when the proportions of cases belonging to trait/diagnostic categories vary (Thompson & Walter, 1988; Feinstein & Cicchetti, 1990). Notice also that Spearman's rho, proportions of specific agreement (Ps+ and Ps-), odds ratios (CMH), and overall agreement (OA) vary greatly depending upon the distribution of the trait categories within the sample.

CONCLUSIONS

Given that first, performance on the ALT was essentially the same under both ambient lighting conditions, second, that agreement was highest with FALANT when scoring methods were matched, and third, that the original objective of this experiment was to determine whether the ALT *instrument* can be used at the Regional Flight Surgeons' offices in *place* of a FALANT, then it follows that the testing conditions and scoring procedures used should match that of the FALANT to be most consistent with other FALANT testing.

Both the FALANT and the ALT (*administered and scored as the FALANT)* predicted performance on the criterion SLGT with about the same accuracy, K (82)=.70 and .675, and the pass/fail agreement between the FALANT and the ALT was very similar, K(190)=.931. Therefore, the ALT can be administered with the FALANT procedures and will give a similar outcome. Also, it is highly probable that the ALT will perform well in color vision testing of pilots. Because its lights meet the signal color specifications of both the Federal Aviation Administration and the International Civil Aviation Organization, the ALT has greater face validity than the FALANT. The broader issue addressed in this study was whether the signal colors of the ALT would improve its predictive validity with performance on the SLGT, to which the answer was *no*; however, SLGT data were only available for 82 participants.

One final concern is the placard attached to each ALT that delineates the administration and scoring procedures established for that test. If the modified FALANT is accepted for pilot testing, then to avoid any confusion, the information on its placards should reflect proper administration and scoring procedures relevant to testing pilots *or* air traffic control specialists if separate methodologies are adopted.

REFERENCES

Birch J, Dain SJ. Performance of red-green color deficient subjects on the Farnsworth Lantern (Falant). *Aviat Space Environ Med* 1999; 70:62-7.

Birch J, McKeever LM. Survey of the accuracy of new pseudoisochromatic plates. *Ophthalmic Physiol Opt* 1993 Jan;13(1):35-40.

Cohen J. A coefficient of agreement for nominal scales. *Educ & Psychol Meas* 1960; 20:37-46.

Cole BL, Vingrys AJ. A survey and evaluation of lantern tests of colour vision. *Am J Optom & Physiol Opt* 1982; 59(4):346-74.

Cicchetti DV, Feinstein AR. High agreement but low Kappa: II. Resolving the paradoxes. *Clin Epidemiol* 1990; 43:551-8.

Dain SJ, Honson V, Ang J. The effect of two lighting conditions on performance of the Farnsworth Lantern color vision test. *Aviat Space Environ Med* 1988; 59:371-3.

Farnsworth D, Foreman P. Development and trial of the New London Navy Lantern as a selection test for serviceable color vision. New London, CT: US Naval Submarine Base, Medical Research Lab, 1946a; Report No 105.

Farnsworth D, Foreman P. A brief history of lanterns for testing color sensation and a description of basic principles. New London, CT: US Naval Submarine Base, Medical Research Lab, 1946b; Report No 104. Cited in: Birch and Dain, 1999.

Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clinical Epidemiol* 1990; 43(6):543-9.

Guggenmoos-Holzmann I. How reliable are chance-corrected measures of agreement? *Stat Med* 1993 Dec 15;12(23):2191-205.

Hackman RJ, Holtzman GL, Walter PE. Color vision testing for the U.S. Naval Academy. *Mil Med* 1992;157(12):651-7.

International Civil Aviation Organization. Technical airworthiness manual (2nd ed.—1987 Document 9051-AN/896), Amendment 10. Montreal, Quebec, Canada: 1988.

Jones KN, Steen JA, Collins WE. Predictive validities of several clinical color vision tests for aviation signal light gun performance. *Aviat Space Environ Med* 1975; 46(5) 660-7.

Kaufman JE, Christensen JF, eds. IES lighting handbook: Applications Volume. New York: Illuminating Engineering Society of North America; 1987.

Laxar KV, Wagner SL, Cotton TC. Evaluation of the Stereo Optical Co. Farnsworth Lantern (FALANT) color perception test: A specification and performance comparison with the original FALANT. Groton, CT: 1998; NSMRL Report No 1209.

Maxwell AE Coefficients of agreement between observers and their interpretation. *Br J Psychiatry*, 1977, 130, 79-83.

Mertens HW, Milburn NJ. Validity of FAA-approved color vision tests for Class II and Class III aeromedical screening. Washington DC: Department of Transportation/Federal Aviation Administration 1993; FAA report no. FAA/AM-93/17.[1]

Mertens HW, Milburn NJ, Collins WE. Practical color vision tests for air traffic control applicants: En route center and terminal facilities. *Aviat Space Environ Med* 2000;71:1210-17.

Schmidt I. Comparative evaluation of the New London Navy Lantern for testing color perception. Randolph Field, TX: US School of Aviation Medicine, 1951; Report of project No 21-20-009. Cited in Birch and Dain, 1999.

Spitzer R, Fleiss J. A re-analysis of the reliability of psychiatric diagnosis. *Br J Psychiatry Res* 1974; 341-7.

Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the Kappa statistic. *Arch Gen Psychiatry* 1985 Jul; 42(7):725-8.

Steen JA, Collins WE, Lewis MF. Utility of several clinical tests of color defective vision in predicting daytime and nighttime performance with the aviation signal light gun. *Aerospace Med* 1974;467-72.

Terry R. Generating Kappa statistics and testing useful hypotheses with PROC CATMOD. Proceedings of the SAS users group annual meeting; 1987 1149-53.

Thompson WD, Walter SD. A reappraisal of the Kappa coefficient. *J Clin Epidemiol* 1988; 41:969-70.

Uebersax JS. Diversity of decision-making models and the measurement of interrater agreement. *Psychol Bull* 1987a;101:140-6.

Uebersax JS. Measuring diagnostic reliability. [Reply to Spitznagel and Helzer letter]. *Arch Gen Psychiatry* 1987b;44:193-4.

Uebersax, JS. Validity inferences from interobserver agreement. *Psychol Bull* 1988; 104:405-16.

U.S. Department of Transportation - Federal Aviation Administration. Federal aviation regulations Part 23, Sec 23.1397. Washington, DC: 1988.

[1] This publication and all Office of Aerospace Medicine technical reports are available in full-text from the Civil Aerospace Medical Institute's publications Web site: http://www.cami.jccbi.gov/aam-400A/index.html

U.S. Department of Transportation - Federal Aviation Administration. Guide for aviation medical examiners. Washington, DC: 2004. Also available online at http://www1.faa.gov/avr/aam/Game/Version_2/03amemanual/home/home.htm

U.S. Department of Transportation - Federal Aviation Administration. Air Transportation Operations Inspector's Handbook. Order 8400, Vol 2, Chapter 27, 14 CFR Part 61. Retrieved April 23, 2004, from the World Wide Web: http://www.faa.gov/avr/afs/faa/8400/8400_vol5/5_009_07.pdf.

U.S. Department of Transportation - Federal Aviation Administration. Flight Standards Service, General Aviation Operations Inspections Handbook. Order 8700.1, Vol 2, Chapter 27, 14 CFR Part 61. Retrieved December 11, 2003, from the World Wide Web: http://www2.faa.gov/avr/afs/faa/8700/.