Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FAA-AM-75-1		

4. Title and Subtitle	5. Report Date
PREDICTIVE VALIDITIES OF SEVERAL CLINICAL COLOR VISION TESTS FOR AVIATION SIGNAL LIGHT GUN PERFORMANCE	January 1975
	6. Performing Organization Code
	8. Performing Organization Report No.
7. Author(s) Karen N. Jones, M.S., Jo Ann Steen, B.A., and William E. Collins, Ph. D.	
9. Performing Organization Name and Address FAA Civil Aeromedical Institute P.O. Box 25082 Oklahoma City, Oklahoma 73125	10. Work Unit No. (TRAIS)
	11. Contract or Grant No.
	13. Type of Report and Period Covered OAM Report
12. Sponsoring Agency Name and Address Office of Aviation Medicine Federal Aviation Administration 800 Independence Avenue, SW. Washington, D.C. 20591	14. Sponsoring Agency Code

15. Supplementary Notes

This research was conducted under Tasks AM-A-73-PSY-45 and AM-D-74-PSY-45.

16. Abstract

Scores on the American Optical Company (AOC) test (1965 edition), Dvorine test, Farnsworth Lantern test, Color Threshold Tester, Farnsworth-Munsell 100-Hue test, Farnsworth Panel D-15 test, and Schmidt-Haensch Anomaloscope were obtained from 137 men with color-defective vision and 128 men with normal color vision. The validity of each of these tests in predicting scores on the aviation signal light gun was assessed by using daytime and nighttime administrations of the light gun as the criteria. Two "best sets" of plates from the AOC and Dvorine tests were selected by calculating a multiple regression equation in a stepwise manner with the nighttime and then the daytime administration of the signal light gun test as the criteria. Based on a graphic presentation of the miss and false alarm rates for each test at various possible cut scores, suggestions were made regarding the use of each test and the selection of optimal pass/fail scores.

17. Key Words	18. Distribution Statement
Aviation Signal Light Gun Color Defect Color Vision Tests Test Standards	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22151.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	9	$3.00 PC $ .95 MF

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

PREDICTIVE VALIDITIES OF SEVERAL CLINICAL COLOR VISION TESTS FOR AVIATION SIGNAL LIGHT GUN PERFORMANCE*

I. Introduction.

To obtain an airman medical certificate (Classes II and III) issued by the Federal Aviation Administration (FAA), an airman must demonstrate the ability to discriminate aviation signal red, green, and white by passing an approved clinical color vision test.[2] If he fails the clinical color vision test, he can request a practical examination using the aviation signal light gun (SIGLIGUN).[9] Unless he takes and passes the practical test, he is issued a medical certificate with the limitation: "Not valid for night flight or by color signal control." Since an airman can appeal to take the practical test after failing the clinical test and since the instrument used for administering the practical test is the same as that used for directing air and ground traffic at airports, it would appear desirable to use clinical color vision tests that have high predictive validity for performance on the SIGLIGUN test.

A color vision test approved by the FAA should be reliable; i.e., it should give consistent results across different testing sessions. If a test has high reliability, the examiner can be fairly certain that if the test were readministered the examinee would make the same, or a very similar, score and would receive the same "color normal" or "color defective" classification. Although the results of a color vision test can greatly affect a person's life, reliability data on many of the devices are often inadequate (e.g., based upon a small sample size) or unavailable.

In a previous study,[9] several clinical color vision tests were evaluated in terms of predicting performance on the aviation SIGLIGUN. These tests included four pseudoisochromatic plate tests (both the 1940 and 1965 editions of the American Optical Company (AOC) test, the Dvorine test, and the color plate from the Titmus Optical Company (T/O) Vision Tester), the Farnsworth Panel D–15 test, the SAM Color Threshold Tester (CTT), the Farnsworth 100-Hue test, the Schmidt-Haensch (Nagel-type) Anomaloscope, and the Farnsworth Lantern test (given in both a lighted room and a darkened room). The purpose of the present study was to further analyze the data obtained by Steen *et al.*[9] in order to (a) estimate the reliability of the AOC (1965 edition), Dvorine, and SIGLIGUN tests; (b) select a "best set" of plates from the AOC and Dvorine tests for use in predicting performance on the SIGLIGUN and obtain an estimate of the cross-validation coefficient of this set; (c) compare the efficiency of seven tests independently of the "approved" cut score (i.e., the score representing the lowest number of errors for failure, as recommended by the FAA or the manufacturer); and (d) determine if we could select, on any of the tests, a different cut score that would act as a better predictor of a cut score on the SIGLIGUN test.

II. Method.

In a previous study,[9] 137 men with color-defective vision and 128 men with normal color vision were given a battery of clinical color vision tests and sets of flashes from an FAA SIGLIGUN in the daytime and at night. We selected for further analysis seven of the tests: the 1965 edition of the AOC test, the Dvorine test, the CTT, the Farnsworth Lantern test, the Panel D–15 test, the 100-Hue test, and the Schmidt-Haensch Anomaloscope. Each test was treated as a predictor of a pass/fail score on the SIGLIGUN test given in the daytime and the same test given at night. For both sets of the pseudoisochromatic plates (AOC and Dvorine),

* This paper is based on the first author's thesis, which completed her M.S. degree requirements in the Department of Psychology at the University of Oklahoma. Constructive criticism and guidance from Charles F. Gettys and W. Alan Nicewander, members of her thesis committee, are gratefully acknowledged.

we recorded the pass/fail score on each plate and the total number of errors on each test. For the Farnsworth Lantern test (given in a lighted room and a darkened room) and the 100-Hue test, the total number of errors was recorded. For the Panel D–15 test, the number of cross-overs was tabulated; for the anomaloscope, the absolute value of the range and midpoint of matches for neutral and chromatic adaptation were recorded; and for the CTT, the number of correct responses was recorded. Separate pass/fail scores were recorded for the daytime and the nighttime administrations of the SIGLIGUN test.

III. Results and Discussion.

As an estimate of test-retest reliability for the AOC and the Dvorine tests, we calculated coefficient alpha or KR–20[6] from the data obtained by Steen *et al.*[9] (coefficient alpha is a measure of internal consistency of the test). Our reliability estimate of 0.94 for the AOC test is in close agreement with Seefelt's[8] test-retest reliability coefficients of 0.94 when the test was given in a mass-testing situation (i.e., on a limited-time, assembly line basis) and of 0.98 when it was given in a clinical setting (N=183 color defectives and 226 color normals). Our reliability estimate for the Dvorine test was 0.98; we know of no other test-retest reliability data for this instrument.

Since the test-retest measures we obtained with the SIGLIGUN were for tests given in both daytime and nighttime hours (i.e., nonparallel measurements), we used Guttman's[3] Lambda four and obtained a reliability estimate of 0.64. Note that this is a lower bound estimate of the reliability; if the retest measurements for the nighttime SIGLIGUN condition had been obtained at night and those for the daytime condition had been obtained in the daytime, the test-retest reliability coefficients for both conditions of test administration would probably have been much higher than this reliability estimate. There appear to be no other published test-retest reliability data for the SIGLIGUN.

A. *An Equation-Derived Set of Pseudoisochromatic Plates.* The correlations of the individual plates from the AOC and the Dvorine tests with the pass/fail scores for the daytime and nighttime administrations of the SIGLIGUN test ranged from 0.26 to 0.79. Using the 28 plates from these two tests (demonstration plates were omitted) as the predictors and the pass/fail scores on the daytime and the nighttime SIGLIGUN tests as the criteria, we calculated two multiple regression equations in a stepwise manner to select a series of plates that would act as the best predictor of SIGLIGUN performance.

For daytime scores on the SIGLIGUN test, the best set of predictors consisted of eight AOC plates (plate numbers 1, 2, 3, 4, 6, 8, 12, and 13) and six Dvorine plates (plate numbers 2, 5, 6, 10, 11, and 14). The rationale for including only 14 plates was that the squared semipartial correlation had approximately reached asymptote when the 14th plate was added. Since prediction using a least-squares multiple regression equation is not generally recommended when prediction is based on individual test items rather than tests in a battery, a unit-weight multiple regression equation with the 14 selected plates was used.

Using the procedure explained in the next section, we chose optimal cut scores for each of the three plate tests (i.e., the AOC, Dvorine, and equation-derived tests). The cut scores selected for failure were seven or more errors on the AOC test, seven or more errors on the Dvorine test, and six or more errors on the 14-plate equation-derived test. The correlations for the total test score, currently accepted pass/fail score, and selected optimal pass/fail score for each of these tests with the pass/fail score on the SIGLIGUN test were calculated. These validity (phi) coefficients for the daytime administrations of the SIGLIGUN test are presented in Table 1. There were no significant differences ($F<1.0$) between tests or between types of scores on the tests (e.g., the difference between the correlation of 0.78 for the AOC total test score versus 0.77 for the accepted pass/fail score on the AOC test was not significant, nor was the difference between 0.77 for the currently accepted pass/fail score on the AOC test versus 0.75 for the accepted pass/fail score on the Dvorine test). This indicates that all three plate tests are equally effective as predictors of a pass/fail score for the daytime administration of the SIGLIGUN test.

TABLE 1.--Validity coefficients for the AOC, Dvorine, and equation-derived pseudoisochromatic plate tests (using total, accepted pass/fail, and optimal pass/fail scores for each) with the pass/fail score on the signal light gun (SIGLIGUN) test as the criterion. The values in parentheses are the estimated cross-validation coefficients.[1]

Plate Test	Daytime SIGLIGUN Administration			Nighttime SIGLIGUN Administration		
	Total Test Score for Plates	Accepted Pass/Fail Score	Optimal Pass/Fail Score	Total Test Score for Plates	Accepted Pass/Fail Score	Optimal Pass/Fail Score
AOC (1965)	0.78	0.77	0.78	0.50	0.44	0.46
Dvorine	0.79	0.75	0.77	0.48	0.47	0.47
Equation-derived by unit weight multiple regression	0.81 (0.81)	----	0.81	0.51 (0.51)	----	0.52

Note: Although not best suited for these data, validity coefficients for an equation-derived test using a least-squares multiple regression model provide a comparison that might interest some investigators. The computations yielded validity coefficients of 0.84 and 0.64 and cross-validation coefficients of 0.82 and 0.59 for daytime and nighttime criteria, respectively.

The same procedure was followed using the nighttime SIGLIGUN test scores as the criterion. The best set of predictors consisted of eight AOC plates (plate numbers 1, 2, 5, 6, 7, 9, 11, and 12) and seven Dvorine plates (plate numbers 5, 6, 7, 10, 11, 12, and 13). The correlation (phi) coefficients for each test are given in Table 1. There were no significant differences between any of the correlations, an indication that all three plate tests are equally good as predictors of nighttime performance on the SIGLIGUN test.

When the validity coefficients for the daytime and the nighttime administrations of the SIGLIGUN test were compared, the differences (e.g., 0.78 versus 0.50, the respective correlations of the AOC total test score with the daytime and nighttime administrations of the SIGLIGUN test) were significant for all plate tests. This finding agrees with previous research[9,10] and indicates that, although the pseudoisochromatic plate tests may be good predictors of daytime SIGLIGUN test performance, they may not be adequate predictors of nighttime performance. However, in making this type of comparison, it should be noted that the ratio of color normals to color defectives determines one set of the marginal totals and can thereby greatly affect the obtained phi coefficients. Therefore, caution is advised in making comparisons of phi coefficients (or product-moment coefficients on dichotomized data) when the ratios of color normals to color defectives differ for the groups being compared (for example, the difference in this study between the percentages of subjects failing the nighttime SIGLIGUN test and those failing the daytime test).

B. *Efficiency of Tests and Evaluation of Cut Scores.* Previous evaluations of color vision tests as predictors of SIGLIGUN test performance[e.g., 9] have used pass/fail categories based on only one cut score (the "recommended" one) for each test. However, it might be useful to evaluate the predictive validity of a clinical test independently of the cut score, as in signal detection theory (TSD) where such an evaluation of the sensitivity of a measure is a standard procedure.[4] To evaluate predictive validities of the tests in this study, we used the probability of a false alarm and the probability of a miss (i.e., the probabilities of making the two types of incorrect decisions). Other studies have evaluated color vision tests in terms of the percentage of misses and false alarms, where a miss occurred when a subject passed the clinical (predictor) test and failed the SIGLIGUN (criterion) test[5] and a false alarm occurred when the subject failed the predictor test and passed the criterion test.[10] In the present study, a miss rate and a false alarm rate are described in the

same way. The four possible TSD events in the present study are defined in Figure 1, where, for instance,

p(miss) = p(pass predictor | fail criterion).

The above equation is read as: "The probability of a miss equals the probability of passing the predictor test given that the criterion test will be failed." Other equations in Figure 1 are read in a similar manner.

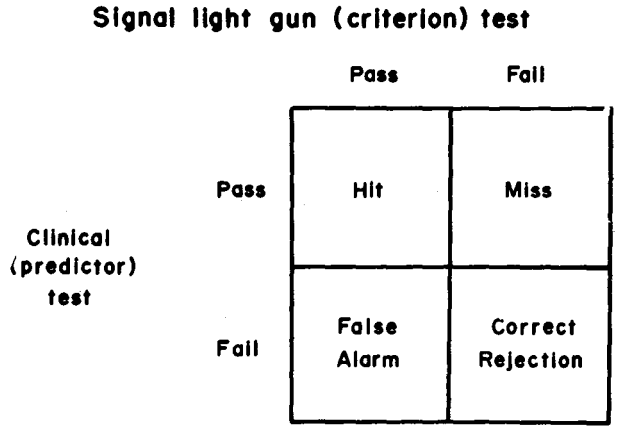

FIGURE 1. Predictor test-criterion test matrix and definition of decision theory events.

Each color vision test used in this study was evaluated against two criteria: the SIGLIGUN test given in the daytime and at night. Figures 2 through 7 show the probability of a miss (p(miss)) plotted against the probability of a false alarm (p(false alarm)). The cut score for a given miss rate and false alarm rate combination can be determined by referring to the scale for each test at the top of the appropriate figure. (Miss and false alarm rates were plotted after a p-to-z transformation[4]; no misses occurred for cut scores below the lowest value on the error scales in the figures.) For all tests, as the cut score is made more stringent, the miss rate decreases and the false alarm rate increases, and as the cut score is made more lenient, the miss rate increases and the false alarm rate decreases. In many practical testing situations (e.g., aviation), a miss is considered more serious than a false alarm. For example, if a miss occurs in aviation, a pilot may be allowed to fly under circumstances for which he may not be qualified; however, if a false alarm occurs, a pilot can request a "second chance" (i.e., a chance to demonstrate his ability to discriminate color in a practical situation). Therefore, in the comparison of tests, an "acceptable" miss rate (or the probability of a miss) should be selected and then the false alarm rates for the tests at that miss rate should be compared. The test with the lower false alarm rate for a given miss rate is the better predictor of performance on the SIGLIGUN test. (We recommend a miss rate of approximately 0.01 to a maximum of 0.15.) Since a miss rate of 0.50 (i.e., 50 percent) or greater in a practical testing situation is unsatisfactory, data for miss rates greater than 0.50 are not given.

C. *Pseudoisochromatic Plate Tests.* Figure 2 shows the miss and false alarm rates for the pseudoisochromatic plate tests with each of the two SIGLIGUN test administrations as the criterion. On the nighttime administration, all three plate tests were approximately equal in efficiency (i.e., for a given miss rate, the false alarm rates were approximately equal). On the daytime administration, the AOC and Dvorine tests were again approximately equal in efficiency; however, for miss rates below p=0.15, the false alarm rates for the equation-derived plate test were somewhat lower than the false alarm rates for the AOC and Dvorine tests. This suggests that the equation-derived test may be slightly preferable to the AOC or Dvorine test as a predictor of performance on the daytime SIGLIGUN test. A more definitive statement would require a cross-validation study with a different subject sample.

In a comparison of the day and night administrations of the SIGLIGUN test, two factors should be noted concerning the nighttime condition: (a) for a given miss rate the false alarm rate was much higher, and (b) for a given cut score there was a general trend for the p(miss) to be lower and the p(false alarm) to be higher. These findings are in agreement with the results of previous research on the AOC and Dvorine tests[9,10] and indicate that, for a selected cut score,

each test was a stricter predictor of performance on the nighttime as compared to the daytime SIGLIGUN test. This finding is not unexpected, since previous research[10] has indicated that it is easier to pass the SIGLIGUN test at night than in the daytime.

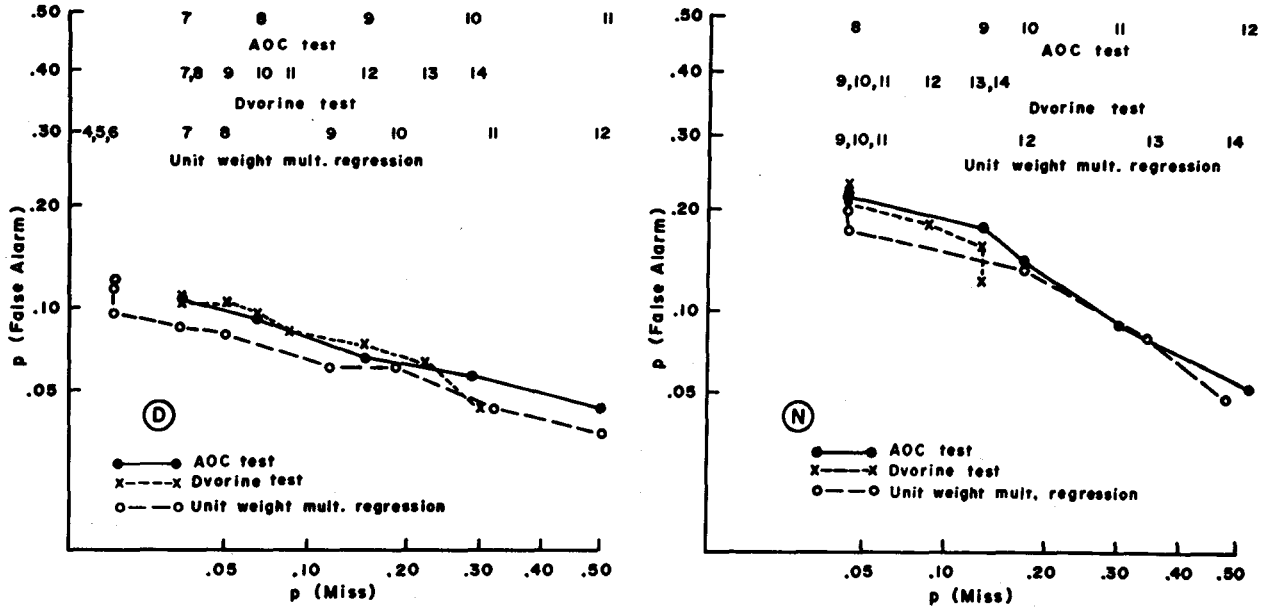

FIGURE 2. Miss and false alarm rates for the pseudoisochromatic plate tests as predictors of performance on the daytime (D) and nighttime (N) administrations of the SIGLIGUN test.

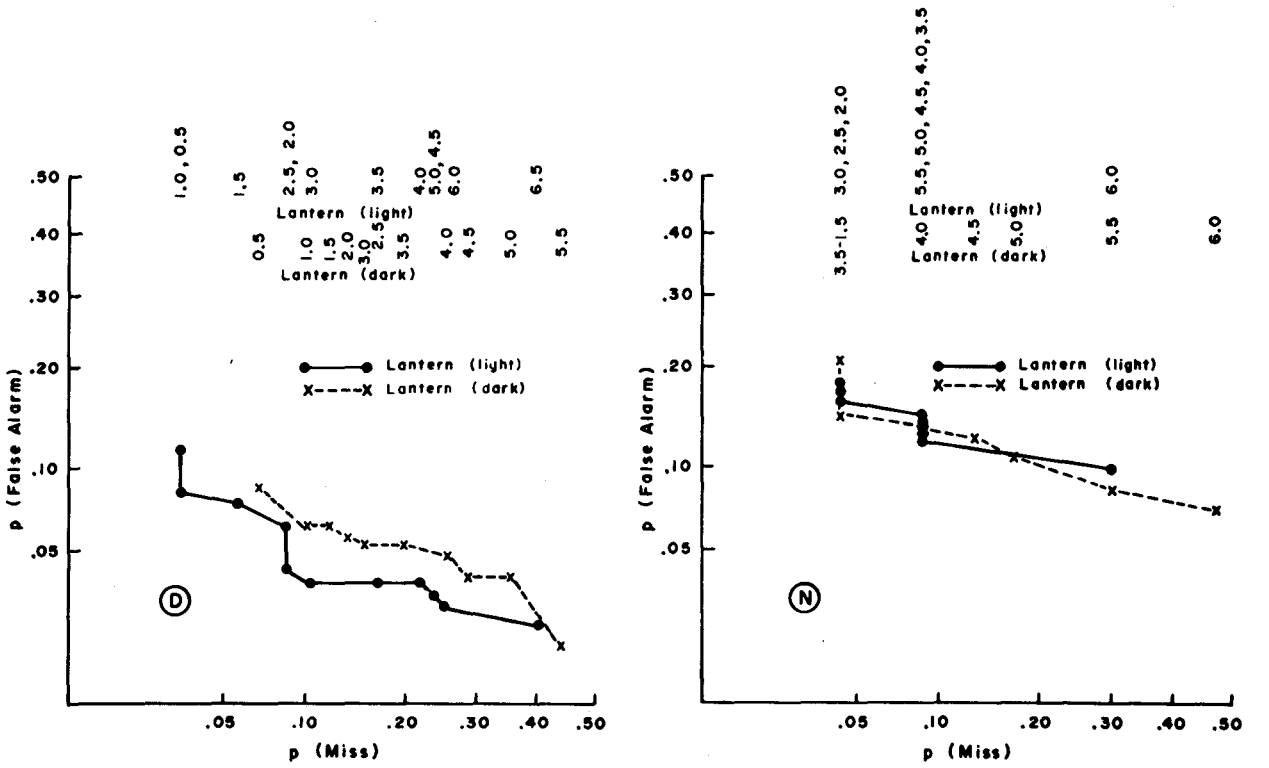

FIGURE 3. Miss and false alarm rates for the Farnsworth Lantern as a predictor of performance on the daytime (D) and nighttime (N) administrations of the SIGLIGUN test.

As the tests are presently used by the FAA, scores of 5 or more errors on the AOC and 12 or more errors on the Dvorine are classified as failing. Our data indicate that the cut score presently used for the AOC test may be somewhat too strict while the cut score for the Dvorine test appears to be too lenient ($p(miss) = 0.18$); a lower cut score for the Dvorine test (e.g., nine or more errors classified as failing, thus lowering p(miss) to about 0.07) could be adopted; however, since a change on the AOC test would involve allowing only one additional error, no modification of this test appears warranted.

D. *Farnsworth Lantern Test.* Figure 3 presents the false alarm and miss rates with the Farnsworth Lantern test given in both a darkened room and a lighted room as the predictor. For both conditions of administration of the Lantern test, with the daytime administration of the SIGLIGUN as the criterion, the miss rates are low, and for a given miss rate, the false alarm rate is low, which are indications that this test is a good predictor of performance on the daytime SIGLIGUN test. These data also indicate that the presently used cut score (an average of more than one error per trial is failing) is appropriate when the Farnsworth Lantern test is given in a lighted room (at this cut score, the $p(miss) < 0.06$ and the $p(false\ alarm) < 0.08$). The miss rate was slightly higher ($p(miss) < 0.12$) when the Farnsworth Lantern test administered in a darkened room was the predictor test. Since the present instructions recommend administration in a lighted room and since the present data do not indicate any advantage to testing in the dark, no change in the administration procedure for the Farnsworth Lantern appears useful.

When SIGLIGUN test scores from the nighttime administration of the test were used as the criterion, the Farnsworth Lantern test given in a darkened room and a lighted room yielded similar results (Figure 3). For any Lantern test cut score suggested in the literature (i.e., up to three errors[7]), both test conditions yielded very low miss rates. The false alarm rates were slightly higher for nighttime than for daytime administration, an indication that the Lantern test, like the pseudoisochromatic plates, is a stricter predictor of performance on the nighttime SIGLIGUN test than on the daytime test (i.e., there is an increase in the false alarm rate and a decrease in the miss rate).

E. *Anomaloscope.* In Figure 4, the miss rates and false alarm rates are presented with the absolute value of the matching range on the anomaloscope as the predictor to allow a comparison of two methods for obtaining a matching range score on the anomaloscope; i.e., neutral and chromatic adaptation. The data suggest that the use of the matching range value obtained under chromatic adaptation is a more appropriate predictor, since it yielded false alarm and miss rates that were lower than those obtained with the matching range under neutral adaptation. We suggest that, as a predictor of performance on the daytime SIGLIGUN test, a range of less than 30 scale units is preferable ($p(miss) < 0.05$) and a range between 31 and 40 scale units is acceptable ($0.05 < p(miss) < 0.15$). As a predictor of performance on the nighttime test, a range of 43 scale units or less appears preferable ($p(miss) < 0.05$) and a range between 44 and 48 units is acceptable ($0.05 < p(miss) < 0.15$).

F. *Color Threshold Tester.* In Figure 5, miss and false alarm rates are presented for the CTT as the predictor. Unlike the results of other tests in this study, CTT test scores reflect the number of correct responses rather than the number of errors. When daytime SIGLIGUN test performance is used as the criterion, for the cut score presently used (i.e., a score of 50 or better is passing) and for adjacent cut scores, the false alarm rates are comparable with the other tests but the miss rates are slightly higher. If a stricter cut score were to be selected (e.g., one so that the $p(miss) = 0.05$), then the false alarm rate would be higher than for the other tests (e.g., $p = 0.20$ as compared with $p = 0.10$ for the pseudoisochromatic plate tests). With nighttime SIGLIGUN test performance as the criterion and with the presently used cut score (or adjacent cut scores) for the predictor test, the $p(miss) = 0$ and the $p(false\ alarm) = 0.20$ (these data values are not given in Figure 5). Therefore, as a strict test for predicting performance on the nighttime SIGLIGUN test, the CTT with its presently used cut score is satisfactory because its false alarm rate is lower than that of the other tests at this miss rate.

G. *100-Hue Test.* The 100-Hue test, a very sensitive clinical device that is not currently used by the FAA in testing pilot applicants, might

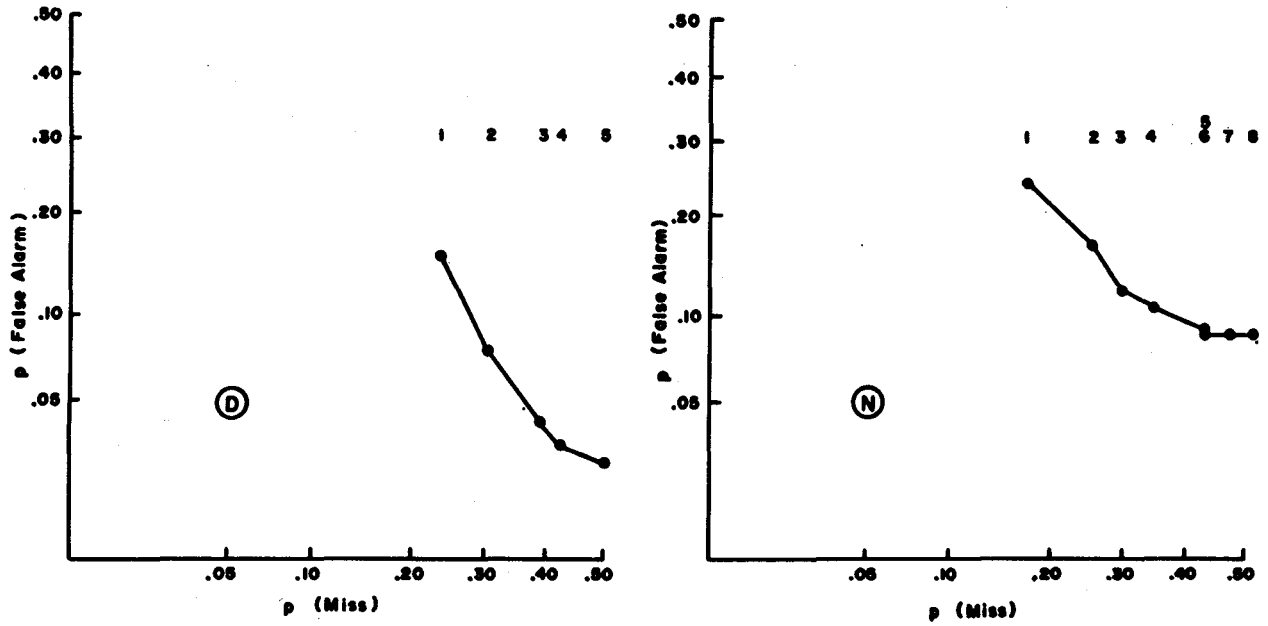

FIGURE 7. Miss and false alarm rates for the D–15 test as a predictor of performance on the daytime (D) and nighttime (N) administrations of the SIGLIGUN test.

4. As a predictor of daytime SIGLIGUN test performance, the Farnsworth Lantern test administered in a darkened room (the instructions specify a lighted room) affords no advantage (and perhaps some loss of predictive efficiency).

5. The 100-Hue and Panel D–15 tests (not currently used by the FAA) are not good predictors of performance on the SIGLIGUN test.

6. When the absolute value of the matching range on the anomaloscope is used as a predictor of performance on the SIGLIGUN test, the matching range obtained under chromatic adaptation has greater predictive efficiency than the range obtained under regular adaptation.

7. All the tests cited above were stricter predictors of SIGLIGUN test performance at night than during the day.

The findings and suggestions made in the present study apply only to the situation in which an attempt is being made to predict performance on tests using the SIGLIGUN (or a similar device).

REFERENCES

1. Browne, M. W.: Precision of Prediction. (Interoffice memorandum) Research Bulletin RB–69–69, Princeton, New Jersey, Educational Testing Service, 1969.
2. Federal Aviation Administration: *Guide for Aviation Medical Examiners*, Washington, D.C., FAA Office of Aviation Medicine, 1970.
3. Guttman, L.: A Basis for Analyzing Test Retest Reliability, PSYCHOMETRIKA, 10:255–282, 1945.
4. Green, D. M., and J. A. Swets: *Signal Detection Theory and Psychophysics*, New York, John Wiley and Sons, Inc., 1966.
5. Lewis, M. F., and J. A. Steen: Colour Defective Vision and Recognition of Aviation Colour Signal Light Flashes. In *The Perception and Application of Flashing Lights: Proceedings of an International Symposium Held at Imperial College, London, in April 1971*, London, Adam Hilger Ltd., 1971.
6. Lord, F. M., and M. R. Novick. *Statistical Theories of Mental Test Scores*, Reading, Massachusetts, Addison-Wesley Publishing Company, 1968.
7. Schmidt, I.: Comparative Evaluation of the New London Navy Lantern for Testing Color Perception. Project Number 21–29–009, Randolph Field, Texas, USAF School of Aviation Medicine, 1951.
8. Seefelt, E. R.: An Evaluation of the Validity and Reliability of the AOC 15-Plate Pseudo-Isochromatic Test in Routine Testing, AMERICAN JOURNAL OF OPTOMETRY AND ARCHIVES OF THE AMERICAN ACADEMY OF OPTOMETRY, 41:371–381, 1964.
9. Steen, J. A., W. E. Collins, and M. F. Lewis: Utility of Several Clinical Tests of Color Defective Vision in Predicting Daytime and Nighttime Performance With the Aviation Signal Gun, AEROSPACE MEDICINE, 45:467–472, 1974.
10. Steen, J. A., and M. F. Lewis: Color Defective Vision and Day and Night Recognition of Aviation Color Signal Light Flashes, AEROSPACE MEDICINE, 43:34–36, 1972.