Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA-AM-85-7		

4. Title and Subtitle	5. Report Date
PASSING SCORES FOR THE FAA ATCS COLOR VISION TEST	6. Performing Organization Code
	8. Performing Organization Report No.
7. Author(s)	
John J. Convey	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Federal Aviation Administration Office of Aviation Medicine 800 Independence Avenue, SW. Washington, D.C. 20591	11. Contract or Grant No.
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	
Office of Aviation Medicine Federal Aviation Administration 800 Independence Avenue, S.W. Washington, D.C. 20591	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

In response to recommendations from the Office of Personnel Management for measures of normal color vision that reflect as closely as possible the functional color vision requirements for an air traffic controller (Christrup, 1981), subtests which simulated ATC tasks were developed in three content areas: (1) aircraft colors for fuselage and lights, (2) color weather radar displays, and (3) navigational chart terrain elevations. Pickrel and Convey (1983) performed an item analysis on these subtests using data obtained from 41 persons with normal color vision as determined by their performance on the Pseudoisochromatic Plates Test (PIP) from the American Optical Corporation and 22 persons with defective color vision according to the PIP. The item parameters and the internal consistency reliability estimates obtained were satisfactory; however, minimum passing scores were not established. The purpose of this research is to determine a minimum passing score for each of these tests and for a simple composite of all three.

17. Key Words	18. Distribution Statement
color vision tests minimum passing score terrain, elevations color normal, color defective	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	16	

Passing Scores for the FAA ATCS Color Vision Test

In response to recommendations from the Office of Personnel Management for measures of normal color vision that reflect as closely as possible the functional color vision requirements for an air traffic controller (Christrup, 1981), subtests which simulated ATC tasks were developed in three content areas: (1) aircraft colors for fuselage and lights, (2) color weather radar displays, and (3) navigational chart terrain elevations. Pickrel and Convey (1983) performed an item analysis on these subtests using data obtained from 41 persons with normal color vision as determined by their performance on the Pseudoisochromatic Plates Test (PIP) from the American Optical Corporation and 22 persons with defective color vision according to the PIP. The item parameters and the internal consistency reliability estimates obtained were satisfactory; however, minimum passing scores were not established. The purpose of this research is to determine a minimum passing score for each of these tests and for a simple composite of all three.

Theoretical Framework

An important problem in mastery testing is how to determine a passing score or standard which separates masters from nonmasters. Several methods for setting standards have been proposed, and the literature on standard setting has been reviewed extensively (Glass, 1978; Hambleton & Eignor, 1980; Meskaukas, 1976).

Because of its judgmental nature, all standard setting is arbitrary. Glass (1978) was particularly pessimistic about standard setting and recommended that the use of standard-setting techniques be abandoned. However, decisions concerning who can perform well enough and who can't, still need to be made. In order to avoid complete arbitrariness, like using 70% or 80% correct as a standard, research is needed to determine the behavior of standard-setting methods in applied settings.

Livingston and Zieky (1982) classify the common standard-setting methods as (1) methods based on judgments about test questions and (2) methods based on judgments about individual test-takers. Nedelsky (1954) and Ebel (1972) provide examples of methods based on judgments about test questions. In Nedelsky's method a panel of experts is asked to identify those distractors for multiple-choice items which members feel that a minimally competent person would be able to recognize as incorrect, or, at least, as not the best alternative present. Ebel's method is a two-stage procedure. First, each judge classifies the items into groups based on a judgment of the item's difficulty (easy, medium, hard) and relevance (essential, important, acceptable, questionable). Second, each judge estimates the expected performance, usually expressed as a percentage of the items answered correctly, of a minimally acceptable test-taker for all items in each category.

Examples of methods based on judgments about individual test-takers are the Contrasting-Groups method and the Borderline-Group method. The Contrasting-Groups method requires the establishment of two groups of examinees, a qualified group and an unqualified group, on the basis of some external

criterion or judgment about the skill and knowledge of each examinee. The passing score of the test is determined from the scores of the examinees by examining the distributions of the qualified group and the unqualified group (Livingston & Zieky, 1982) or by using a version of the discriminant function (Koffler, 1980). The Borderline-Group method requires the establishment of a group of examinees who are judged to be just barely qualified or just barely unqualified. The median score on the test for the members of this group usually is used as the passing score.

Research to date has indicated that different results can be expected when different methods are applied; however, the differences tend to be fairly consistent. Passing scores from Nedelsky's method generally are lower than those from Ebel's method (Andrew & Hecht, 1976; Convey & Appleton, 1984; Skakun & Kling, 1980). Mills (1983) found that the Contrasting-Groups method produced lower passing scores than did the Borderline-Group method; however, Convey and Appleton (1984) found the passing scores from these methods were virtually identical. When the methods based on judgments about test questions are compared to those based on judgments about individual test-takers, the Contrasting-Groups method has produced lower passing scores than has Ebel's method (Convey & Appleton, 1984), but not always lower than those from Nedelsky's method (Poggio, Glasnap, and Eros, 1981).

In summary, while some consistent patterns emerge when different methods are compared, the research has not identified one method that is clearly superior to the others. What is clear is that the behavior of a model is somewhat situation dependent, most likely due to the knowledge and expertise of the judges involved or to the suitability of any external criterion which is used. Thus, for any specific application, several methods for determining passing scores should be used, if possible, and criteria identified to select which of the resulting minimum passing scores will be used for making subsequent decisions.

Method

Of the techniques reviewed, only the Contrasting-Groups method is applied to the color vision data in this research. The methods of Nedelsky and Ebel are not suitable because the color vision tests are speeded; that is, the test is designed so that only a few examinees will normally complete the test in the allotted time. Since some items will be answered only by a fraction of examinees, these items may appear to be more difficult than if the examinees had an opportunity to answer them. Thus, any analyses based on the difficulty or relevance of an item, or on the appropriateness of the distractors for each item are not valid. The Borderline-Group method also will not be used to determine a passing score since the number of examinees in any reasonable borderline group is small.

In addition to the Contrasting-Groups method, a method based on the regression of the PIP score on the appropriate color vision subtest score and a method based on minimizing the probability of a classification error are used to determine a minimum passing score for each subtest and the composite. All methods are applied to the data obtained from the 41 color-normal examinees and the 22 color-defective examinees from Pickrel and Convey's (1983) research.

Contrasting-Groups Method

Livingston and Zieky's (1982) suggestions were used to calculate the passing score for the Contrasting-Groups method for each of the three content area subtests and the composite score. First, the 63 examinees were divided into a color-normal group and a color-defective group on the basis of the PIP test. Examinees scoring 10 or greater on the PIP formed the color-normal group; those scoring 9 or less formed the color-defective group. Second, for each subtest the percentage of examinees with scores in a specific interval who were in the color-normal group was plotted. The intervals for the aircraft colors subtest and the color weather radar test had widths of 5 points, those for the terrain elevations subtest had widths of 3 points, and those for the composite of all three subtests had widths of 10 points (see Table 1). Third, each percentage distribution was smoothed by plotting the weighted average of the percentages of color-normal examinees in the specific interval and in the two adjacent intervals. Fourth, the score on the smoothed distribution for which the percentage of color-normal examinees was 50% was used as the minimum passing score.

Regression Method

In order to provide a comparison to the passing scores produced by the Contrasting-Groups method, minimum passing scores were calculated using the regression of the PIP scores on the score of each of the three subtests and the composite. The score on each subtest which had a predicted value of 9, the highest failing score on the PIP, was obtained. The next integer higher than that score was used as the minimum passing score for each subtest and the composite.

Minimum-Error Method

A third passing score was obtained for each content subtest and the composite score by finding the score which minimized the amount of disagreement between the classification decisions made on the basis of the PIP test and those made on the basis of each content subtest. A correct classification occurs when either a color-normal examinee as determined by the PIP test passes a subtest or a color-defective examinee as determined by the PIP test fails a subtest. A false positive occurs when a color-defective examinee passes a content subtest. A false negative occurs when a color-normal examinee fails a content subtest. For each subtest and the composite, the probability of a false positive and the probability of a false negative were calculated using each score as a possible passing score. The score for which the sum of the two probabilities was the smallest was chosen as the minimum passing score.

Results

Table 2 to Table 5 show the results of the various analyses for aircraft colors, color weather radar, navigational chart terrain elevations, and the composite score, respectively. Each table contains the following statistics:

(1) number of questions in the subtest or composite;

(2) the average score for color-normal examinees and color-defective examinees;

(3) the slope, intercept, and multiple R squared for the regression of PIP on the appropriate subtest or composite score;

(4) for the Contrasting-Groups, Regression, and Minimum-Error methods:

(a) minimum passing scores;
(b) percent passing the subtest or composite;
(c) percentage of correct classifications;
(d) percentage of false positives and percentage of false negatives;
(e) probability of a false positive and probability of a false negative.

The pattern of the passing scores and the magnitude of the differences between the passing scores produced by the three methods varied across the three content areas and the composite. No method consistently produced either the highest or lowest passing score. The magnitudes of the differences between the highest and lowest passing score were 3 points for the aircraft colors subtest, 6 points for the color weather radar subtest and the terrain elevations subtest, and 11 points for the composite of the three subtests. Neither examination of the pattern of passing scores or the magnitude of their differences revealed a clearly best choice among the methods.

The passing scores agreed most for the aircraft colors subtest (see Table 2). Of the three content areas, this one yielded the most consistent data. This subtest correlated highest with the composite score ($r = .897$) and with PIP ($r = .508$). The Contrasting-Groups method yielded a highest passing score (25), while the Minimum-Error method produced the lowest passing score (22). All three methods resulted in correct classification decisions in excess of 76%, and the probability of a false positive for each was higher than the probability of a false negative. Excluding the Minimum-Error method, which by definition has the highest percentage of correct classifications and the fewest errors, the Regression method produced more correct classifications and fewer false negatives than did the Contrasting-Groups method.

For the color weather radar subtest (see Table 3), the Contrasting-Groups method and the Minimum-Error method had the same passing score which was lower than the passing score from the Regression method. The percentage of correct classifications for this subtest was lower than that for the the aircraft colors subtest, as were its correlations with the composite ($r = .884$) and with PIP ($r = .499$).

For the navigational chart terrain elevations subtest (see Table 4), the Minimum-Error method produced the highest passing score (29), while the passing scores from the Contrasting-Groups method and the Regression method were considerably lower, 18 and 17, respectively. The percentages of correct classifications for this subtest were lower than those in the other content subtests, as were its correlations with the composite ($r = .774$) and with PIP ($r = .435$). The Minimum-Error method has a very low percent passing (39.7%) and is the only example where the percentage of false negatives exceeds the percentage of false positives.

For the composite of the three content areas (see Table 5), the Minimum-Error method and the Regression method produced similar passing scores, 73 and 74, respectively. The passing score from the Contrasting-Groups method was 63. Despite the variation in passing scores, the percentage of correct classifications varied only slightly, ranging from 76.2% to 79.4%.

Discussion

In this study several factors may limit the ability of each method to produce valid minimum passing scores. The most important potential limiting factor is the validity of the PIP test for the sample of examinees used in the study. The validity of the Minimum-Error and Contrasting-Groups methods directly depend upon the ability of the PIP test to identify correctly examinees with normal color vision and those with defective color vision. For each content area subtest and the composite, a few examinees who failed the PIP test scored higher than the mean of the color-normal group, and a few examinees who passed the PIP test scored lower than the mean of the color-defective group. In addition, the Regression method depends upon the magnitude of the relationship between the PIP test and each content area test. These correlations ranged from .435 to .565 which are reasonable indexes of predictive validity. Any problems with the correct functioning of the PIP test could directly affect the calculation of the minimum passing score for each area test.

A second potential limiting factor in this study is the size of the sample used to obtain the minimum passing scores. A larger sample would produce more stable, but not necessarily different, results. A smaller sample size has the greatest possible impact on the Contrasting-Groups method since the choice of the width of the interval used in the smoothing process and the stability of the estimates of the percentage of qualified examinees in each interval depend upon the sample size. Larger samples will generally result in narrower intervals and more stable estimates of the percentage of qualified examinees in each interval. In this study the estimates of the probability of a false positive for each method are based on only 22 examinees. Thus, any one-unit change in the number of false positives results in a probability change of .05. The probability estimates of a false negative for each method are slightly more stable, being based on 41 examinees. However, the stability of the estimates of both types of probability would be enhanced with a larger sample size.

Because the tests are speeded and due to the nature of the task involved in some of them, a third potential limiting factor in this study is the cognitive ability of the examinees. Higher ability examinees are likely to work more quickly than lower ability examinees, and thus will answer more items. Also, some of the tasks involve learning on the part of a naive examinee so that higher ability examinees will have an advantage.

In the light of these limitations and the variation in the results from each method in terms of the passing scores produced and the magnitude of the differences between them, what passing score should be used for each content area subtest and the composite? Three strategies will be considered: (1) passing scores from the most theoretically appealing method; (2) the most conservative passing scores, i.e., the highest passing scores obtained for each test; and (3) the most consistent passing scores obtained. Each of these strategies is discussed below.

From a decision framework, the Minimum-Error method is the most theoretically appealing since it minimizes the percentage of classification errors, false positives and false negatives. In applying this method to the data obtained in this study, no differentiation was made between the relative seriousness of a false positive (an individual with defective color vision who is judged to be color normal) and a false negative (an individual with normal color vision who is judged to be color defective). From a public perspective, a false positive may be perceived as more serious than a false negative because safety may be jeopardized by hiring an individual who does not have normal color vision. Assuming the errors to be equally serious, the resulting passing scores are: 22 for the aircraft colors subtest; 23 for the color weather radar subtest; 23 for the terrain elevations subtest; and 73 for the composite of the three subtests. If the relative seriousness of each error is taken into account as well as the presence of a higher incidence of false positives in this sample than false negatives, a higher passing score for each content area and composite would be needed.

The most conservative passing score for each subtest and composite is the highest score produced by any of the three methods. The advantage of using this score is that the probability of a false positive is lowest for this score when compared to the scores produced by the other methods. The highest passing scores for each subtest are: 25 for aircraft colors, 29 for color weather radar, 23 for terrain elevations, and 74 for the composite of the three subtests. Another possible value for the composite is 77 which is the sum of the three passing scores for each of the content area subtests.

For each content area subtest and the composite, two of the passing scores from the three methods were either identical or differed by one point. A third strategy is to select as the passing score for each subtest and composite the score which is identically produced by two methods or the higher of the two scores which differ by one point. Use of this strategy results in the following passing scores: 25 for the aircraft colors subtest, 23 for the color weather radar subtest, 18 for the terrain elevations subtest, and 74 for the composite of the three subtests.

Because information from all three of the methods is used, the consistency criterion of the third strategy above is recommended as the basis for establishing a minimum passing score for each subtest and the composite. Table 6 shows the results of applying this strategy to the data from this study. For the 22 examinees who were classified as color defective because they failed the PIP, 8 (36.4%) passed the aircraft colors subtest, 14 (63.6%) passed the color weather radar subtest, 18 (54.5%) passed the terrain elevations subtest, and 10 (45.6%) passed the composite of the three subtests. The passing rates for the 41 examinees who were color normal on the basis of the PIP were 82.9%, 95.1%, 78.0%, and 87.8%, respectively.

The value of the ATCS Color Vision Test is enhanced as a screening instrument to identify color-defective examinees if it is used in conjunction with other screening tests like the PIP. A reasonable multi-stage screening strategy is as follows:

(1) If an examinee passes the PIP, consider that person to have normal color vision.

(2) For examinees who fail the PIP, administer the ATCS Color Vision Test. If an examinee fails either the composite or one of the subtests, consider that person as defective in color vision.

For this study, 16 of the 22 examinees who failed the PIP also failed either the composite or one subtest of the ATCS Color Vision Test. Thus, six examinees who were classified as color defective on the basis of the PIP would be considered color normal using the ATCS Color Vision Test.

The validity of the recommended minimum passing scores as well as the reasonableness of the two-stage screening strategy should be monitored carefully. Additional data for cross validating and refining the minimum passing scores for the ATCS Color Vision Test are needed.

Table 1

Grouped Distributions of Color-Normal

and Color-Defective Examinees

Interval	Defective	Normal	% Qualified	Smoothed
		Aircraft Colors		
46-50	0	4	100.0	
41-45	2	8	80.0	89.3
36-40	1	13	92.9	82.8
31-35	3	8	72.7	78.5
26-30	2	1	33.3	68.4
21-25	1	4	80.0	41.2
16-20	7	2	22.2	40.0
11-15	1	0	00.0	18.7
6-10	5	1	16.7	
		Color Weather Radar		
46-50	1	1	50.0	
41-45	2	12	85.7	85.7
36-40	1	11	91.7	78.4
31-35	5	6	54.5	71.4
26-30	4	8	66.7	59.3
21-25	2	2	50.0	52.6
16-20	3	0	00.0	25.0
11-15	4	1	20.0	
		Navigational Chart Terrain Elevations		
34-36	0	1	100.0	
31-33	0	3	100.0	100.0
28-30	0	7	100.0	100.0
25-27	0	7	100.0	77.8
22-24	6	7	53.8	63.6
19-21	6	7	53.8	53.1
16-18	3	3	50.0	48.0
13-15	4	2	33.3	50.0
10-12	1	3	75.0	46.0
7- 9	2	1	33.3	

Table 1 (continued)

Interval	Defective	Normal	% Qualified	Smoothed
		Composite		
121-130	0	2	100.0	
111-120	0	7	100.0	94.4
101-110	1	8	88.9	96.0
91-100	0	9	100.0	73.3
81- 90	7	5	41.7	65.6
71- 80	4	7	63.6	52.0
61- 70	1	1	50.0	58.8
51- 60	2	2	50.0	33.0
41- 50	3	0	00.0	18.2
31- 40	4	0	00.0	

Table 2

Summary Statistics for Aircraft Colors Test

Number of Items:	54	Slope:	.174
Color-Normal Average:	34.17	Intercept:	4.94
Color-Defective Average:	22.41	R^2:	.260

Statistic	Contrasting-Groups	Regression	Minimum-Error
Passing Score	25	24	22
% Passing	66.7	69.8	74.6
% Correct Classified	76.2	79.4	80.9
% False Positive	12.7	12.7	14.3
% False Negative	11.1	7.9	4.8
Prob (False Positive)	.364	.364	.409
Prob(False Negative)	.171	.122	.073

Table 3

Summary Statistics for Color Weather Radar Test

Number of Items:	54	Slope:	.219
Color-Normal Average:	36.61	Intercept:	2.73
Color-Defective Average:	28.77	R^2:	.250

Statistic	Contrasting-Groups	Regression	Minimum-Error
Passing Score	23	29	23
% Passing	84.1	74.6	84.1
% Correct Classified	77.8	71.4	77.8
% False Positive	20.6	19.0	20.6
% False Negative	1.6	9.5	1.6
Prob(False Positive)	.591	.545	.591
Prob(False Negative)	.024	.146	.024

Table 4

Summary Statistics for Navigational Chart Terrain Elevations Test

Number of Items:	45	Slope:	.268
Color-Normal Average:	22.87	Intercept:	4.40
Color-Defective Average:	17.77	R^2:	.189

Statistic	Contrasting-Groups	Regression	Minimum-Error
Passing Score	17	18	23
% Passing	79.4	77.8	39.7
% Correct Classified	66.7	65.1	71.4
% False Positive	23.8	23.8	1.6
% False Negative	9.5	11.1	27.0
Prob(False Positive)	.681	.681	.045
Prob(False Negative)	.146	.171	.415

Table 5

Summary Statistics for Composite Test

Number of Items:	153	Slope:	.096
Color-Normal Average:	93.65	Intercept:	2.00
Color-Defective Average:	68.95	R^2:	.319

Statistic	Contrasting-Groups	Regression	Minimum-Error
Passing Score	63	74	73
% Passing	82.5	74.6	76.2
% Correct Classified	76.2	77.8	79.4
% False Positive	20.6	15.9	15.9
% False Negative	3.2	6.3	4.8
Prob(False Positive)	.591	.455	.455
Prob(False Negative)	.049	.098	.073

Table 6

Passing Scores for ATCS Color Vision Test

and Resulting Group Performance

	Aircraft Colors	Color Weather Radar	Terrain Elevations	Composite
Passing Score	25	23	18	74
		Color-Defective Group		
Pass	8	14	12	10
Fail	14	8	10	12
% Pass	36.4	63.6	54.5	45.6
		Color-Normal Group		
Pass	34	39	32	36
Fail	7	2	9	5
% Pass	82.9	95.1	78.0	87.8

References

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 35-50.

Christrup, H. (1981). Unpublished memorandum from the Assistant Director for Personnel Research and Development at the Office of Personnel Management to Donald B. Rock, Director of Personnel and Training, Federal Aviation Administration, October 16, 1981.

Convey, J. J., & Appleton, L. P. (1984, April). A comparison of methods for setting standards on Coast Guard competency tests. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs: Prentice-Hall.

Glass, G. V (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Hambleton, R. K., & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. Jaeger & C. Title (Eds.), Minimum competency testing. Berkeley, CA: McCutchan.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton: Educational Testing Service.

Meskaukas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 45, 133-158.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Journal of Educational Measurement, 20, 283-292.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Pickrel, E. W., & Convey, J. J. (1983). Color perception and ATC job performance (Report No. FAA-AM-83-11). Washington, DC: Federal Aviation Administration.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981, April). An empirical investigation of the Angoff, Ebel, and Nedelsky standard setting methods. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.

GPO 918-532

12675723