International Symposium on Aviation Psychology - 2015

International Symposium on Aviation Psychology

2015

Procedure Used for Establishing Screening Test Cut-Points Based on Aviation Occupational Task Performance

Nelda Milburn

Thomas Chidester

Kevin Gildea

Linda Peterson

Carrie Roberts

See next page for additional authors

Repository Citation

Milburn, N., Chidester, T., Gildea, K., Peterson, L., Roberts, C., & Perry, D. (2015). Procedure Used for Establishing Screening Test Cut-Points Based on Aviation Occupational Task Performance. *18th International Symposium on Aviation Psychology*, 213-218.
https://corescholar.libraries.wright.edu/isap_2015/71

Authors
Nelda Milburn, Thomas Chidester, Kevin Gildea, Linda Peterson, Carrie Roberts, and Deborah Perry

PROCEDURE USED FOR ESTABLISHING SCREENING TEST CUT-POINTS BASED ON AVIATION OCCUPATIONAL TASK PERFORMANCE

Nelda Milburn, Thomas Chidester, Kevin Gildea, and Linda Peterson
Federal Aviation Administration, Civil Aerospace Medical Institute
Oklahoma City, OK
Carrie Roberts and Deborah Perry
Xyant Technology, Inc.
Norman, OK

Previous research has shown that some individuals with color vision deficiencies (CVD) are capable of performing some aviation occupational tasks as well as those with normal color vision (NCV); implying that passing a screening test with a diagnosis of NCV may not be necessary for all aviation occupations. Our goal was to find *outcome consistency* between performance on occupational tasks and several screening tests; further, to compare those pass/fail outcomes to the Colour Assessment and Diagnosis (CAD) test for aviation certification. The strategy involved establishing a pass/fail cut-point separately for four occupational tasks at the 5th percentile of the NCV group. A scatterplot was constructed displaying the sum of correct screening test trials on the x-axis and the red/green threshold of the CAD test on the y-axis. By defining the markers according to pass/fail status on the occupational tasks, it was easy to evaluate multiple factors to arrive at an appropriate cut-point for each screening test.

All of the Federal Aviation Administration (FAA) approved color vision screening tests have reasonable pass/fail agreement (as measured by kappa scores; Cohen, 1960) for diagnosing normal or deficient color vision when compared to diagnoses with the Nagel anomaloscope (Mertens & Milburn, 1993). However, we have found that some individuals with color vision deficiencies (CVD) are capable of performing some aviation occupational tasks as well as individuals with normal color vision (NCV). Because the FAA allows different several color vision screening tests to ensure that those passing are capable of performing the necessary pilot color-decoding tasks, it is important to match the pass/fail cut-point to performance on aviation color-coded tasks. Furthermore, if an airman fails the initial screening, he/she can request secondary screening involving presentation of signal lights at an airport by FAA personnel, which is time-consuming and more expensive to conduct than in-office clinical screening tests. Therefore, it is prudent to maximize the sensitivity and specificity of the in-office screening tests by appropriately setting the pass/fail cut-points; and, it is also important from a safety standpoint to minimize the “false negative” numbers to ensure that airmen that pass the screening test are capable of accomplishing the requisite color tasks of modern aviation.

Purpose

Our goal was to find a decision point that reliably defined performance on occupational tasks matched to passing cut-points on clinical and precision tests; and further, to compare those cut-points to a valid and reliable criterion measure. The Colour Assessment and Diagnosis (CAD) certification standard was based on performance of aviation-related color tasks and has been successfully in use for pilot selection since 2008 (Barbur, Evans, & Milburn, 2009). The CAD conveniently provides individual threshold values in standardized normal units. By doing so, it essentially quantifies, on a linear scale, one’s ability to see color, which in turn can be linked more directly to percent correct performance on specific tasks, unlike traditional normal/deficient test outcomes.

Method

Participants

The CAD test was used for diagnoses of type and degree of color vision deficiency, and it has a high diagnosis agreement for red-green types of color vision deficiencies with the gold-standard, the Nagel anomaloscope (Barbur et al., 2009). However, the Nagel does not diagnose yellow-blue types of deficiencies, and it requires tedious, one-on-one screening, averaging about 20 to 30 minutes to complete. Study participants included 57 males and 38 females, with 89% between the ages of 18 and 31 to match the population of air traffic control applicants (a separate study). Ten adults over 31 years of age with CVD were recruited to equalize the NCV and CVD groups. Most subjects with CVD were male because the congenital deficiency results from a recessive trait on the X chromosome. All subjects met a screening requirement of at least 20/30 near and far visual acuity. The participants included 47 NCV and 48 with CVD, classified by type of deficiency: 16 protan, 20 deutan, 3 tritan, and 9 exhibiting both red-green (RG) and yellow-blue (YB) weaknesses. Table 1 shows participant CAD type classifications and thresholds.

Table 1.
Participant Color Vision Classification and Threshold Values

		CAD Test	
Diagnosis	N	Red/Green Threshold	Yellow/Blue Threshold
Normal	47	.84 - 1.71	.67 - 1.62
Protan	16	11.67 - 29.84	.66 - 1.64
Deutan	20	2.82 - 29.31	.71 - 1.64
Tritan	3	1.35 - 1.68	1.79 - 1.98
RG & YB	9	1.76 - 30.77	1.83 - 15.06

Materials

Evaluation measurements were defined as clinical tests, computerized tests, or occupational tasks. The clinical tests and the computerized tests are available commercially; however, the occupational tasks were created in the laboratory (with the exception of the signal light gun) strictly to serve as work samples for validation purposes. Consequently, we had to determine an appropriate passing score for the *occupational* tasks. That process is described later in this paper. The *clinical tests* included: the Dvorine®, the Ishihara (-14, -24, and -38 plate versions), the Waggoner HRR®, the Waggoner PIPIC®, the Richmond Products HRR®, and the Stereo Optical 900® (OPTEC 900®). The *computerized tests* included: the Rabin Cone Contrast Test (RCCT®), ColorDx®, and the Colour Assessment and Diagnosis (CAD®) Test.
The *occupational tasks* included:

- Incandescent red (R) and white (W) precision approach path indicator lights (INC-PAPI), which included 26 pairs of lights, scored as 52 trials. Light pairs were presented with the following combinations: R-R, R-W, W-R, and W-W.
- Light-emitting diode red and white lights (LED-PAPI) that included 64 pairs of lights, scored as 128 trials. Light pairs were presented with the following combinations: R-R, R-W, W-R, and W-W.
- Signal light gun test (SLGT). The SLGT presents a single light of red, green, or white. A total of 6 lights were presented at 1,000 ft and 6 lights at 1,500 ft. To pass, FAA Order 8400 stipulates that no errors are allowed on the 12 trials to pass.

- Pilot cockpit display colors task was comprised of 10 targets (trials) for each of 8 colors. Trials were presented as colored text (red, white, green, blue, cyan, yellow, amber, or magenta). The total percent correct was scored using the number of correct selections minus false positives.

Strategies to achieve consistency

With our ultimate goal in mind—to find an appropriate cut-point on the screening test that ensures those passing are capable of performing critical, color-coded aviation occupational tasks—we tried several options. First, we explored setting pass/fail cut-points on the *occupational tasks* at 95% correct of all trials, separately for each task, which seemed like a reasonable standard generally accepted in academia to reflect good performance. The final decision was to use a performance equivalent to the 5th percentile of the normal color vision group because about 95% of all individuals have normal color vision. About 8-10% of men and less than ½ of 1% of women have a color vision deficiency. In other applications, using a pass/fail point of the 5th percentile of the normal color vision group is similar to setting a production goal for assembly-line workers based on a goal that 95% of the workforce meets or exceeds. Therefore, to set individual screening test cut-points, we took the following steps:

1. Determine the 5th percentile for the NCV group for each occupational task separately (NCV determined by CAD test)
2. Cross-tabulate the pass/fail of 5th percentile performance on the occupational tasks with CAD certification
3. Cross-tabulate 5th percentile occupational task pass/fail performance with each clinical test using the manufacturer's pass/fail criterion and evaluate their agreement
4. Graph the sum of correct trials (x-axis) by CAD RG thresholds (y-axis) and color-code points by pass/fail performance (at the 5th percentile of NCV group) on the composite of occupational tasks, separately, for each screening test
5. To guard against jeopardizing the integrity of screening tests and to prevent motivated examinees from memorizing a limited number of plates, the total trials administered must exceed 11 and the passing score must exceed a minimum of 7 trials correct
6. Using both the cross-tabulation tables and the graphs, we examined the FAA-defined, pilot cut-points (when available) for *clinical* tests to determine their effectiveness for passing those who performed the occupational tasks at the 5th percentile of the NCV group. When necessary, we altered the cut-point of the *clinical* tests to balance the false positive with the false negative cells to achieve optimal sensitivity and specificity scores (using composite *occupational* task performance as the criterion measure)
7. For those tests without FAA pilot cut-points, we followed the same procedure as previously described to set screening test cut-points
8. Once the pass/fail cut-points were selected, the burden on airmen could be evaluated—meaning the number of airmen that would be required to take medical flight tests (MFTs) as a result of failing their initial screening test

These strategies were used to find consistency between multiple screening tests, performance on several occupational tasks, and a linear scale of color vision ability in standardized normal units, which some might argue is a proven pilot certification test. This paper will focus on the techniques employed for setting cut-points rather than the cut-points assigned to each color vision screening test because we believe that when establishing cut-points, traditional methods may not allow the researcher, test developer, or test validator to fully see multiple comparisons that are essential to validating a test. Typically, when establishing a screening test cut-point, one seeks to find the point that maximizes the sensitivity and specificity of the test without sacrificing one for the other; but, our ultimate goal was to differentiate between those who can and cannot perform safely within a reasonable degree of accuracy and certainty. Several statistical tests can be used to measure the agreement between the screening test and the criterion measure (Milburn & Mertens, 2004). In medicine, the criterion measure may be whether or not a person has a particular disease and screening tests are used to predict the disease when determining the presence or absence of the disease is expensive or invasive. Sometimes, screening tests are used to predict performance, such as standardized tests (e.g., ACT, SAT, or GRE, which are used to predict readiness for college or graduate school). Likewise, a screening test can be used to predict ability to perform a specific task. The FAA uses color vision screening tests to predict one's ability to decode and interpret vital color-coded information used in signal lights.

In the current research, the purpose of the occupational tasks was to simulate actual work tasks required by pilots that necessitate good color perception to perform safely. One possible solution would have been to hire several seasoned pilots to perform the occupational tasks and set a pass/fail point based on their performance; however, we honed the tasks to the bare essential requirement of identifying, naming, differentiating, or matching colors—tasks that did not require any piloting experience or knowledge.

It may be important to explain why, if we are using the CAD aviation certification test as a benchmark or a reference point for performance, we don't simply use the CAD certification exclusively for all pilot screening. The answer is that the test is not conveniently available at national test sites or at aviation medical examiners' offices; additionally, it is very expensive (about $9,000 US dollars). Consequently, it is unlikely to be widely purchased by aviation medical examiners when the return on investment ratio is minimal, and likely requires many years to break-even. Furthermore, a typical flight physical costs about $200 with color vision screening as a very small part of the examination.

Results

Using the strategy described above, setting the cut-point was easy in some cases and more difficult in others. For example, The Waggoner PIPIC is a relatively new pseudoisochromatic plate test, which we categorized as a clinical test. It was not in production when the FAA first set pilot-specific cut-points (Mertens & Milburn, 1993). Because all NCV participants passed all screening tests with a strict (NCV) criterion; and, because 5% of NCV participants were sacrificed to establish the 5th percentile pass/fail point for the occupational tasks, we chose to isolate only the CVDs to examine the cost/benefit of setting a new cut-point. Using the manufacturer's criterion for determining NCV (12 of 14 correct), 13 CVD subjects failed the screening test but passed all of the occupational tasks with performance equivalent to, or exceeding the 5th percentile of the NCV group. Table 2 is a cross-tabulation of performance on the Waggoner PIPIC using the new (9 of 14) cut-point by pass/fail performance on the composite of occupational tasks for only the CVD participants and shows a gain of 7 additional subjects passing the screening test that also passed all of the occupational tasks. There were also 6 subjects that failed the Waggoner PIPIC who were also able to perform the occupational tasks. There was a substantial improvement in the number passing the screening test, but as you can see from the scatterplot in Figure 1, some CVD participants scored very poorly on the screening test but well on the occupational tasks. This evidence supports the FAA's occupational color vision test (OCVT) or the FAA developing such a test that can be administered in the office, such as the air traffic color vision test (ATCOV), which serves that purpose for air traffic control applicants (Chidester et al., 2011).

Figure 1 best presents the outcome of using the strategy we described previously to set individual screening test cut-points. Essentially, if an examinee passes at least one of the 12 screening tests, he/she should be highly likely to accomplish the color-coded aviation occupational tasks that we examined. Some points are overlapping—47 subjects passed all 12 screening tests and 31 subjects failed all 12 and that information is not readily apparent on the scatterplot. Therefore, it is essential that cross-tabulation tables are used in conjunction with the scatterplots. Not all screening tests measure the same color perception abilities; hence, some subjects do well on some tests and not on others. We recognize that the manufacturer's cut-point, based on separating NCV from CVD, is an unfair requirement for some CVD examinees. We found that regardless of which approved screening test the examinee takes and passes, it is highly likely that his/her performance on color-coded tasks will be essentially as good as the 5th percentile of the normal color vision group.

Table 2.

Composite Occupational Tasks for Pilot Cockpit, LED, and Incandescent Precision Approach Path Indicator (PAPI) Systems, and Signal Light Gun Test by the Waggoner PIPIC Pass/Fail Cross-tabulation[a]

		Waggoner PIPIC		Total
		Fail	Pass	
Composite Occupational Tasks for Pilot Cockpit, LED/Incandescent PAPI, and Signal Light Gun Test	Fail ANY	34	1	35
	Pass ALL	6	7	13
Total		40	8	48

a. Color Deficient subjects only

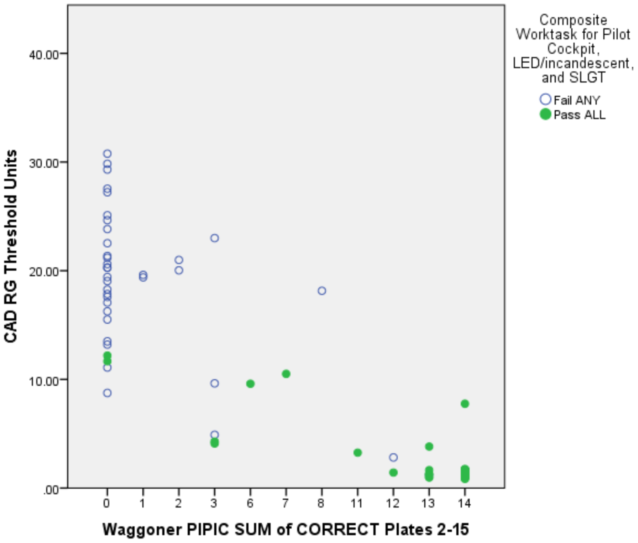

Figure 1. Scatterplot of the Waggoner PIPIC sum of correct responses on plates 2-15 by the Colour Assessment and Diagnosis Test red-green threshold, with the markers coded by passing or failing the occupational tasks.

One example of the improvement in Kappa agreement scores between screening test and occupational performance was the Waggoner HRR that changed from $K_{(92)}$= .62 to .72. Using this methodology to assign screening test cut-scores (rather than using the screening tests' designation of NCV to determine pass/fail), we found agreement scores, for the tests we examined, ranging between $K_{(92)}$=.70 to .78 and averaging .72. Agreement scores improved from those obtained using the previous cut-points—but more importantly, based on our validation analyses, we are assured that those cleared based on accepted screening tests will be highly likely to perform well on aviation color-tasks.

Conclusions

By using coded markers on a scatterplot with a diagnostic quantitative measure in standard normal units on the Y-axis and screening test scores on the X-axis, the decision points can be easier to determine. In contrast, one must explore several different pass/fail points by coding variables for analysis, and then create multiple cross-tabulations with the criterion variable to examine the resulting kappa agreement scores—to check for improvements. We believe that the methods we used and presented here may be a helpful strategy that is applicable to other test developers, researchers, and test validators.

Acknowledgements

Research reported in this paper was conducted under the Flight Deck Program Directive / Level of Effort Agreement between the Federal Aviation Administration Headquarters and the Aerospace Human Factors Division of the Civil Aerospace Medical Institute sponsored by the Office of Aerospace Medicine and supported through the FAA NextGen Human Factors and Engineering Division.

References

Barbur, J., Evans, S., & Milburn, N. (2009). Minimum Color Vision Requirements for Professional Flight Crew, Part III: Recommendations for New Color Vision Standards. (Report No. DOT/FAA/AM-09/11). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

Chidester, T., Milburn, N., Lomangino, N., Baxter, N., Hughes, S., & Peterson, L. (2011). *Development, Validation, and Deployment of an Occupational Test of Color Vision for Air Traffic Control Specialists.* (Report No. DOT/FAA/AM-11/8). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

Cohen, J.A. (1960). Coefficient of Agreement for Nominal Scales. *Educational & Psychological Measurement*, 20:37-46.

Mertens, H., & Milburn, N. (1993). *Validity of FAA-Approved Color Vision Tests for Class II and Class III Aeromedical Screening.* (Report No. DOT/FAA/AM-93/17). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

Milburn, N., & Mertens, H. (2004). *Predictive Validity of the Aviation Lights Test for Testing Pilots With Color Vision Deficiencies*. (Report No. DOT/FAA/AM-04/14). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.