Digital Human Modeling and Applied Optimization, Vol. 76, 2023, 20–28
https://doi.org/10.54941/ahfe1003350

Development of Test and Evaluation Methodologies for Headborne Low Light Sensor Systems

Jose D. Villa[1], Edward Hennessy[1], Rebecca Skulsky[2], Reeve Goodenough[3], and Blake Mitchell[1]

[1]U.S. Army Combat Capabilities Development Command Soldier Center, Natick MA 01760 USA
[2]Edwards Lifesciences, 1 Edwards Way, Irvine, CA 92614, USA
[3]MathWorks, 2 Apple Hill Dr Ste 103, Natick, MA 01760, USA

ABSTRACT

The military is improving technical capabilities by developing increasingly advanced low-light sensor headborne systems. There are several known limitations of legacy night vision devices (e.g., narrow field of view, poor depth perception, monochrome display, monocular output), and as technology moves forward, the U.S. Army needs ways to test and evaluate visual performance between systems. The presented research sought to identify, modify and/or develop a series of test methodologies to assess the visual performance of users wearing these new technologies. Standard vision performance tests were examined, and, in some cases due to device constraints, modified and piloted. Both standard and modified tests were assessed for suitability as potential test and evaluation tools for developmental systems. The tests assessed included contrast sensitivity, depth perception, field of view, pegboard test, visual acuity, multi-target stepping task, and a novel stamping task. Results showed promise for these tasks as assessment tools; all but the depth perception task were sensitive enough to distinguish between devices. Follow-on work is needed to understand task interrelationships, refine procedures to improve reliability, determine additional tasks, and utilize methods in item assessments.

Keywords: Headborne systems, Night vision systems, Test methodologies, System assessment tools, Depth perception, Field of view, Dexterity, Visual acuity

INTRODUCTION

The military is improving technical capabilities across the board, enabling and modernizing our warfighters. One of the most technology-dynamic areas currently involves helmet mounted technologies and visual enhancement/night vision capabilities. There are several known limitations of legacy night vision devices (e.g., narrow field of view, poor depth perception, monochrome display, monocular output, etc.). (Redden, Turner and Carstens, 2006; Redden, 2002). Newer technologies are emerging to improve upon these legacy devices. The presented research sought to identify test methodologies to assess the visual capabilities of the new technologies in comparison to legacy devices as well as to inform the devices' research, development, and procurement.

In identifying potential methodologies, we sought standardized performance tests because of their acceptance and validation. There are several standardized and validated vision performance tests used to assess characteristics and visual capabilities of the human eye. Many have been in use for decades. These tests and methodologies have been validated through research and clinical usage, usually medically based, and are accepted as means of assessing human visual performance. However, we acknowledged that some tests cannot be used exactly as designed due to design or functional characteristics of night vision/visual enhancement devices (e.g., helmet mounting, monocular view, etc.). We assessed the potential of each chosen standard test to be used as designed and, when that was not feasible, the test was modified and assessed via pilot testing. Once a suitable, feasible modification was identified, a "standard operating procedure" for each of the selected methodologies was developed.

After finalizing methodologies, a group of U.S. Army Infantry Soldiers assessed two legacy night vision devices (NVD) and one prototype device using the selected methodologies. The test was conducted to both assess the methodologies and compare the legacy devices to a prototype device to determine any improvements in its performance over the legacy devices. Methodologies fell into two categories: 1) standard vision test and 2) operational or task-based test. The standard vision tasks included visual acuity (static and dynamic), contrast sensitivity, depth perception, and field of view. The operational tests were pegboard, stamping, and walking path tasks. The methodologies developed are presented in this paper, along with the results of the performance comparison between the tested night vision systems using the methodologies. Further research will be needed to continue validation of the methodologies, but the tests used show promise as tools to evaluate night vision devices using standard visual performance tests or modified versions of them.

ASSESSED METHODOLOGIES

Participants completed visual acuity, contrast sensitivity, depth perception, field of view, pegboard, stamping, and stepping tasks in three device configurations consisting of two legacy NVDs (AN/PVS-7 and AN/PVS-14) and a prototype NVD. The tasks were also completed in a bareheaded/eyed configuration to understand the participants' natural performance. All NVD configurations were attached to a helmet with an appropriate mount. All participants tested all of the configurations in a quasi-randomized order.

The AN/PVS-7 is a biocular NVD. It has one optical tube (sensor) and that tube's image is reflected in an internal mirror to present the same scene to both eyes. The AN/PVS-14 is a monocular device—it has one optical tube and presents the image to one eye. The wearer chooses over which eye to wear the device. The prototype is a binocular system (2 optical tubes) that presents a stereoptic view to both eyes (i.e., it presents a slightly different view from each tube, similar to normal human eye function).

Visual Acuity Task

The visual acuity task (Bailey & Lovie-Kitchin, 2013) consisted of a static visual acuity eye exam (non-instrumented). The dynamic visual acuity movements were not feasible with the NVDs and thus were not included as part of the test. Bailey-Lovie paper eye charts were used for the task. The completed static task used a letter-row-by-letter-row measure where size is the only variable that changed to assessed visual acuity monocularly. The set distance for the task was 10 feet. The task score consisted of a Logarithmic Minimum Angle of Resolution (LogMAR) score (Bailey & Lovie-Kitchin, 2013).

Contrast Sensitivity

To understand the impact of systems on contrast, the Pelli-Robson test was used (Pelli, Robson, & Wilkins, 1988). The Pelli-Robson Contrast Sensitivity test measures the viewer's capability in identifying stimuli of varying degrees of contrast to the background. Stimuli are triplets of letters, with a total of 8 rows of 2 triplets each. Each triplet is printed in black ink, with each successive triplet printed more faintly than the preceding one. The participants were 3m from the stimulus and were asked to identify the letters in each triplet until they could no longer see any letters. The task's score is a logarithmic value called a "log contrast sensitivity" score derived from the last triplet identified. The scores are on an interval scale, beginning at 0.0 for the darkest triplet and progressing by 0.15 to the maximum of 2.25 for the faintest triplet. For night vision devices, contrast sensitivity affects several factors such as acuity, target recognition, and eye movements (saccades) (Parush, Gauthier, Arseneau & Tang 2011). This test was used as originally designed to assess the NVDs.

Depth Perception

One traditional method of assessing depth perception, or stereopsis, is the Howard-Dolman test, used for decades to assess stereopsis in both clinical and research settings (Howard, 1919). Additional methods of assessing stereopsis are also used clinically (using stereograms—charts with slightly differing depths of their stimuli per eye) (Kalloniatis and Luu, 2005) because those methods are less cumbersome and require less execution space than the Howard-Dolman test but provide similar results. The Howard-Dolman test has been used in the past by the US Army Aeromedical Research Laboratory (USAARL) as a means of determining visual issues in early models of night vision devices prior to conducting operational assessments of the systems (Wiley, 1989).

The Howard-Dolman test has a fixed rod and a moving rod. The participant aligns the movable rod with the fixed rod. The score for the trial is the number of millimeters away from the fixed rod the movable rod was placed. A perfect alignment equals a zero score. Negative scores indicate the rod was placed nearer to the viewer than (or ahead of) the fixed rod, and positive scores indicate placement farther from the viewer than (or beyond) the fixed rod. This test was used as originally designed without modifications.

Each participant conducted the test twice in each condition—once with the movable rod initially set behind the fixed rod, and once with the movable rod ahead of the fixed rod, determined at random. The data from these two trials were averaged, with the mean analyzed for differences between NVD and bare head conditions.

Field of View

Measurements of field of view (FOV) are often done clinically using a Ferree-Rand projection perimeter (Ferree & Rand, 1922; Sloan, 1939). For this study, a modernized version of a Ferree-Rand projection perimeter was used. The unit has LED lights mounted in a hemispheric arc rather than projecting a light onto a hemisphere, and light movement is automated. The chin rest and testing head position of the revised perimeter was designed to ensure that the NVDs and helmet cleared the arc with the participant seated properly. Each eye is tested separately with the other eye covered. The data output is the visual angle in degrees formed from the participant's eye to the center point of gaze of the perimeter and from the eye to the maximum limit of vision, on 8 different azimuths in a sphere around the participant's head. The data for each eye is combined to create the entire field of view available from both eyes. The visual angle from each azimuth is used to create a radar graph, with the area contained within the radar graph calculated for each test condition for each participant. Analyses compare the visual field area of each condition.

Night vision devices have several shortcomings in terms of visual performance, with restricted field of view being one of the most problematic (Patterson, Winterbottom & Pierce 2006). For NVDs, each participant's field of view was generated based on the view presented and visible through the eye piece. The prototype device, for example presents a two-eye, binocular view, though reduced from a bare-head field of view. The legacy devices present a monocular view, even though one of them has two eyecups that present the same view.

Pegboard Task

The ASTM F2010 pegboard task (ASTM, 2018) is typically used to assess hand-eye coordination and dexterity (Yancosek & Howell, 2009). Participants are tasked with placing as many pegs as possible into a pegboard. For this assessment, the pegboard consisted of a 5X5 layout with 25 holes placed over 5 rows and 5 columns. A bowl with 25 pegs was placed at the center top of the board. The participants completed the task using their dominant hand, starting at the top opposite side of the board. For example, right-handed participants grabbed pegs with their right hand and filled the holes in the top left of the pegboard, proceeding left-to-right and top-to-bottom. The test was modified from the standard in that the number of unfilled holes left on the board after 30 seconds was recorded as the task score (i.e., lower scores indicate better performance). The number of placed pegs is affected by the participant's ability to quickly perceive and locate the pegboard holes within the visual limitations of the NVD (e.g., poor depth perception).

Target Stamping Task

The Target Stamping Task (TST) is a novel task developed to evaluate hand-eye coordination as well as how changes in vision affect reach accuracy. The task used in this assessment was similar to the one used in Dunn et al. (2022). Participants sat in an adjustable chair with a table directly in front of them. The participants sat upright, with their head, neck, back, and hips in a generally straight line while holding a stamp in their dominant hand. When not in motion, participants rested their dominant hand on the table with their elbow at a 90-degree angle with their fist against the near table edge. A paper target was placed centrally in front of the participants at their measured thumb tip reach minus 18 inches. At the experimenter's signal, the participants stamped the marked center of the target paper and returned their hand to its original spot as quickly as possible. The task was repeated four more times without additional cuing. The task score consisted of the mean accuracy of the five stamp marks measured in millimeters (i.e., distance from the stamp center to the target center).

Multi-Target Stepping Task

The Multi-Target Stepping Task (or MTST) is a task typically used to assess limb control for the elderly (Yamada et al., 2011). It was adapted here to assess depth and contrast perception. The MTST developed in-house consisted of a plywood base with 5 rows of foam square blocks. Each foam block was 10" W x 10" L x 1.875" H. Each row was comprised of 3 block columns of varying heights and contrast. The column heights consisted of 1, 2, or 3 foam blocks. The top block on each column varied by color (white, gray, or black). The height and contrast orders were randomized for each row. One of the 3 colors was the designated target, while the other 2 were distractors. The participants were tasked with traversing the 5 rows forth and back by stepping on the block with the target color. The time to complete the task was recorded as the task score.

RESULTS

The participants in this study were 14 male, active-duty Infantry Soldiers (Age: 23.3 years ± 3.07; years in service 3.2 ± 2.40) stationed at Ft. Benning, GA. All had normal corrected vision (1 participant was red-green color blind). Four participants reported left eye dominance and 10 reported right eye dominance. Twelve reported a preference for wearing monocular NVDs over the left eye in the field and 2 over the right eye. Bare-head visual performance of the group was as follows: visual acuity mean = 0.2 (SD 0.3), contrast sensitivity mean = 1.51 (SD = 0.20), depth perception mean = 1.79 (SD = 4.42), and field of view (area) mean = 1016.6 (SD = 107.59).

Within each task, the data were checked for normality using a Shapiro-Wilk test. Normal data were compared by condition using repeated measures analyses of variance and post hoc tests (with Bonferroni correction), with Greenhouse/Geisser corrections as needed. Non-normal data were assessed via non-parametric Friedman Signed-Ranks tests and post hoc Wilcoxon tests.

Visual Acuity Task Results

A Friedman test was significant, $X^2(3) = 30.3$, $p<.001$, indicating differences between the bare head and the 3 NVD types. Post-hoc testing showed that the bare head score (mean = -0.2, SD = 0.13) was significantly higher (i.e., permitted greater visual perception) than with the PVS-7 (mean = 0.13, SD = 0.20, $Z = 4.28$, $p<.001$); with the PVS-14 (mean = 0.06, SD = 0.11, $Z = 4.07$, $p<.001$); or with the prototype (mean = 0.5, SD = 0.06, $Z = 3.06$, $p<.01$). Also, the prototype mean score was significantly lower than the scores for both the PVS-7 ($Z = 2.99$, $p<.05$) and the PVS-14 ($Z = 3.06$, $p<.01$). The PVS-7 and PVS-14 did not differ in acuity scores from one another.

Contrast Sensitivity Task Results

A Friedman test was significant ($X^2(3) = 38.6$, $p<.001$), indicating differences between the bare head and the 3 NVD types. Post hoc testing showed that the bare head score (mean = 1.51, SD = 0.20) was significantly higher (i.e., permitted greater contrast; viewers could see more faint triplets) than with the PVS-7 (mean = 0.88, SD = 0.30, $Z = -4.08$, $p<.001$), with the PVS-14 (mean = 0.88, SD = 0.26, $Z = -4.18$, $p<.001$), and with the prototype (mean = 0.11, SD = 0.22, $Z = -3.32$, $p<.001$). Also, the prototype mean score was significantly lower (i.e., permitted significantly lower contrast) than the score for the PVS-7 ($Z = -3.31$, $p<.001$) and the PVS-14 ($Z = -3.32$, $p<.001$). The PVS-7 and PVS-14 did not significantly differ in contrast from one another.

A decrement in the ability to perceive contrast when viewed through night vision devices is not unexpected, so the results overall are unsurprising. However, the prototype does not demonstrate improvement over the current NVDs, and in fact, performs significantly worse for contrast sensitivity.

Depth Perception Task Results

There were no significant differences between conditions (at $p<.05$) on a Friedman test. This was likely caused by the large variance in the data for the legacy systems. Interestingly, the PVS-14 had a negative (rearward, "far-sighted") mean score (mean = −22.68, SD = 45.01) while the PVS-7 (mean = 4.61, SD = 41.94) and the prototype (mean-16.36, SD = 25.81) had positive (frontward, "near-sighted") mean scores. The bare head score was close to zero/neutral (mean = 1.79, SD = 4.42). Even without significance, this indicates that the PVS-14 appears to provide visual information differently than the other device types.

Field of View Task Results

A Greenhouse-Geisser correction was required and applied to the data in a repeated measures ANOVA test. The results were significantly different, $F(1.1) = 707.4$, $p<.001$. All of the conditions significantly differed except for the PVS-7/ PVS-14 pair (bare head mean = 1016.6, SD = 107.59; PVS-7 mean = 145.29, SD = 10.04; PVS-14 mean = 137.29, SD = 15.41; prototype mean = 309.57, SD = 21.58). Significance on this analysis is not unexpected, as bare head field of view will be much larger than that of any NVDs at this point in technology development.

Since bare head FOV is much larger than that of any of the NVDs tested, an additional analysis was conducted on the NVD conditions alone. The purpose was to determine whether the significant differences between the conditions on the first analysis (above) remained when the bare head condition was omitted. The results were the same, with the prototype providing significantly larger FOV (mean = 309.57, SD = 21.58) than either the PVS-7 (mean = 145.29, SD = 10.04) or the PVS-14 (mean 137.29, SD = 15.41), which did not differ from one another, $F(2) = 524.9, p<.001$. This result was likely a result of the prototype using two sensors together to create the image seen by the user, rather than a single sensor's field presented to one (PVS-14) or both (PVS-7) eyes.

Peg Board Task Results

A Friedman test was significant, $X^2(3)=30.2$, $p<.001$, indicating differences between the bare head and the 3 NVD systems. Post-hoc testing showed that the bare head score (mean = −5.0, SD = 2.16) was significantly higher (i.e., participants had better hand-eye coordination) than in the PVS-7 (mean = −9.6, SD = 3.10), $Z = -4.27$, $p<.001$; the PVS-14, (mean = −8.6, SD = 2.02), $Z = -4.14$, $p<.001$; and the prototype (mean = −15.2, SD = 1.24), $Z = -3.20$, $p<.001$. Also, performance in the prototype was significantly lower (i.e., worse hand-eye coordination) than in the PVS-7, $Z = -3.12, p<.01$, and PVS-14, $Z = -3.19, p = .001$. The PVS-7 and PVS-14 did not differ in dexterity performance from one another.

Target Stamping Task Results

An ANOVA test result indicated significant differences between the bare head and the 3 NVD types, $F(3) = 12.5$, $p<.001$. Post-hoc testing, with Bonferroni correction, showed that bare head accuracy (mean = 6.8, SD = 2.54) was significantly higher than in the PVS-14 (mean = 11.7, SD = 3.38) and prototype (mean = 17.9, SD = 6.14), but not different from the PVS-7 (mean = 11.3, SD = 5.76). Accuracy in the PVS-14 was significantly higher than in the prototype. Performance in the PVS-7 was not significantly different from any of the other systems, although there was a marginally significant difference between the PVS-7 and prototype ($p = 0.097$).

MTST Results

A Friedman test was significant, $X^2(3) = 31.8$, $p<.001$, indicating differences between the bare head and the 3 NVD types. Post-hoc testing showed that the bare head score (mean = 6.9, SD = 1.14) was significantly higher (i.e., permitted greater depth perception and agility, or eye-foot coordination) than the PVS-7 (mean = 13.0, SD = 5.72, $Z=4.35, p<.001$); the PVS-14 (mean = 10.3, SD = 3.14, $Z=4.10, p<.001$); and the prototype (mean = 19.1, SD = 6.26, $Z=3.18$, $p<.001$). Also, the prototype mean score was significantly lower than mean scores for both the PVS-7 ($Z = 2.34$, $p<.05$) and the PVS-14 ($Z = 3.18$, $p<.001$). There was only a marginally significant difference between the PVS-7 and PVS-14 ($p = 0.083$).

Test-Retest Reliability

Each of the tests used was assessed for reliability. The test design included one additional test condition for each participant—a second trial with one of the NVDs or bare headed conditions, assigned at random. Not every participant was able to run the second trial due to time constraints and/or device malfunctions. The additional data were used to assess whether the tests were reliable with NVDs. For this analysis, a Pearson's Correlation Coefficient (R) was computed for each of the tests. If the two data sets are correlated with an R equal to or greater than 0.80, the test reliability is good (0.8 to 0.9) to excellent (>0.9). An R of 0.7 to 0.8 indicates acceptable reliability. An R below 0.7 indicates questionable or unacceptable reliability.

The contrast sensitivity, $r(15)=.74$, $p<.001$, depth perception, $r(15) = .75$, $p<.001$,visual acuity, $r(6) = .77$, $p<.05$, and MTST, $r(6) = .72$, $p<.05$, tests were reliable. The pegboard, $r(5) = -.09$, $p>.05$, and TST, $r(8) = .39$, $p>.05$, were not reliable. The field of view data was generated from 8 individual data points (azimuths) per eye per participant per condition. FOV data were reliable for both eyes, on all azimuths, with R values ranging from $r(15) = .90$ to $r(15) = .99$, p<.005.

DISCUSSION

Initial testing shows these methodologies are promising as tools to evaluate night vision devices. Other than the Howard-Dolman depth perception test, all were sensitive enough to statistically differentiate between the tested items. Expected follow-on steps include conducting model analysis to understand the relationship between the standard vision performance (i.e., bareheaded/eyed) and degradations in the operational tasks from the equipment worn (i.e., how does a decrement in contrast impacts the ability to place pegs in the proper holes); develop the assessed methodologies further to improve test reliability; develop and identify other potential new tests and tools to use as evaluation methods; and utilize the finalized methodologies in any applicable system evaluations.

REFERENCES

ASTM International (2018). Standard Test Method for Evaluation of Glove Effects on Wearer Finger Dexterity Using a Modified Pegboard Test. Standard F2010-M18.

Bailey, I. L., & Lovie-Kitchin, J. E. (2013). Visual acuity testing. From the laboratory to the clinic. *Vision research*, *90*, 2–9.

Dunn, J. A., Taylor, C. E., Wong, B., Henninger, H. B., Bachus, K. N., & Foreman, K. B. (2022). Testing Precision and Accuracy of an Upper Extremity Proprioceptive Targeting Task Assessment. *Archives of Rehabilitation Research and Clinical Translation*, *4*(3), 100202.

Ferree, C. E., & Rand, G. (1922). A New Laboratory and Clinic Perimeter. *Journal of Experimental Psychology*, *5*(1), 46.

Howard, H. J. (1919). A test for the judgment of distance. *Transactions of the American Ophthalmological Society*, *17*, 195.

Kalloniatis, M and Luu, C. (2005). *The Perception of Depth*. In Webvision: The Organization of the Retina and Visual System [Internet], National Center for Biotechnology Information, National Library of Medicine., US National Institutes of Health. https://pubmed.ncbi.nlm.nih.gov/21413376/

Parush, A., Gauthier, M. S., Arseneau, L., & Tang, D. (2011). The human factors of night vision goggles: perceptual, cognitive, and physical factors. *Reviews of human factors and ergonomics*, *7*(1), 238-279.

Patterson, R., Winterbottom, M. D., & Pierce, B. J. (2006). Perceptual issues in the use of head-mounted visual displays. *Human factors*, *48*(3), 555–573.

Pelli, D. G., Robson, J. G., & Wilkins, A. J. (1988). The design of a new letter chart for measuring contrast sensitivity. Clinical Vision Sciences, 2(3), 187–199.

Redden, E. S. (2002). Safety Assessment of Wearing the AN/PVS-14 Monocular Night Vision Device (MNVD) and AN/AVS-6 Aviators' Night Vision Imaging System (ANVIS) During 5-Ton and HMMWV Night Driving. Aberdeen, MD: US Army Research Laboratory. Report ARL-TR-2580.

Redden, E. S., Turner, D. D., and Carstens, C. B. (2006). The Effect of Future Forces Warrior Planned Sensor Offset on Performance of Infantry Tasks: Limited User Evaluation. Aberdeen, MD: US Army Research Laboratory. Report ARL-TR-3764.

Sloan, L. L. (1939). Instruments and technics for the clinical testing of light sense: III. An apparatus for studying regional differences in light sense. *Archives of Ophthalmology, 22*(2), 233–251.

Wiley, R. W. (1989). Visual Acuity and Stereopsis with Night Vision Goggles. USAARL Report No. 89–9. Ft. Rucker, AL: US Army Aeromedical Research Laboratory. AD-A211552.

Yamada, M., Higuchi, T., Tanaka, B., Nagai, K., Uemura, K., Aoyama, T., & Ichihashi, N. (2011). Measurements of stepping accuracy in a multitarget stepping task as a potential indicator of fall risk in elderly individuals. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, *66*(9), 994–1000.

Yancosek, K. E., & Howell, D. (2009). A narrative review of dexterity assessments. *Journal of Hand Therapy, 22*(3), 258–270.