A Digital Alternative to the TNO Stereo Test to Qualify Military Aircrew

Bonnie N. Posselt; Eric Seemiller; Marc Winterbottom; Chris Baber; Steve Hadley

INTRODUCTION: Stereopsis is usually required in military aviators and may become increasingly important with reliance on newer technologies such as binocular Helmet-Mounted Displays (HMDs) and stereo displays. The current stereo test used to qualify UK military aircrew (TNO test) has many limitations. To address these limitations, two computer-based digital versions of a random dot stereogram (RDS) were developed: a static version (dRDS-S), and a version in which the dots appear to move dynamically within the depth plane (dRDS-D), both capable of measuring stereo acuity to threshold.

METHODS: There were 41 participants who performed all 3 stereo tests, TNO and both digital dRDS tests, on two separate occasions.

RESULTS: The best (lowest) mean stereo acuity threshold was measured with dRDS-S (33.79 arcseconds, range 12.64–173) and the worst mean stereo acuity thresholds were measured with the TNO test (91 arcseconds, range 60–240). Both dRDS tests were strongly correlated, but neither correlated with the TNO test. Both dRDS tests were more reliable, as indicated with tighter limits of agreement.

DISCUSSION: With a large floor effect at 60 arcseconds, the TNO test was unable to characterize any finer degree of stereo acuity. Both dRDS tests demonstrated better test-retest reliability and addressed many of the limitations seen with the TNO test. The dRDS tests were not correlated with the TNO test, which suggests that the TNO test does not provide the accuracy or reliability for use as a meaningful aeromedical screening test. The dRDS tests will enable research to investigate the relationship between stereo acuity and operational performance.

KEYWORDS: stereo acuity, stereo test, vision standards, aviators.

Posselt BN, Seemiller E, Winterbottom M, Baber C, Hadley S. *A digital alternative to the TNO stereo test to qualify military aircrew*. Aerosp Med Hum Perform. 2022; 93(12):846–854.

Since the dawn of powered flight, adequate vision has been considered vital in aviators, and numerous vision standards exist to qualify aircrew to fly. However, some vision tests used by aeromedical examiners today could be considered outdated and crude, with many limitations. There is a need to evaluate vision tests used for military aviators to assess whether they are fit for the purpose or if newer test methods could be more effective and appropriate.[24]

Stereopsis, in particular, is desired in military aviators for its link with binocular vision and depth perception, which, in turn, are thought to benefit flying performance.[31,43] As stereopsis is largely exercised for closer visual ranges, but up to 18 m,[2] adequate stereo acuity is most advantageous in situations operating in close proximity to other aircraft such as air-to-air refueling, taxiing, and formation flying. However, stereopsis may become increasingly important with the advent of newer visually demanding technologies such as binocular Helmet Mounted Displays (HMDs) and stereo displays.[24] In essence, stereopsis is the ability to perceive precise depth based on the difference in position of an image between the left and right retinas due to the slightly different perspective of each eye (binocular disparity). Stereo acuity is the smallest disparity that can be perceived in depth and is one way to measure binocular function. Stereo acuity varies significantly among individuals and in the general population ranges from a few arcseconds to more than

From the Operational Based Vision Assessment Laboratory, 711th Human Performance Wing, Wright-Patterson AFB, OH, USA.

This manuscript was received for review in April 2022. It was accepted for publication in September 2022.

Address correspondence to: Bonnie Posselt, M.B.Ch.B., RAF Centre of Aviation Medicine, RAF Henlow, Henlow, Bedfordshire SG16 6DN, United Kingdom; bonnie.posselt739@mod.gov.uk.

DOI: https://doi.org/10.3357/AMHP.6111.2022

hundreds of arcseconds,[11] with a proportion of people (5–30%) lacking stereopsis altogether.[4,27] For those with measurable stereo acuity, there is a bimodal peak in the general population at 96 and 699 arcseconds[11] which does not appear to be affected by age according to some studies,[11] but found to deteriorate over 50 yr of age in others.[12] A further 32% of people are stereo anomalous despite otherwise normal vision.[11]

As demonstrated by the significant proportion of people with deficient stereo acuity, stereopsis is not essential to daily life. One can rely solely on monocular cues, also known as pictorial or object-centered cues, to perceive depth. Such monocular cues are: relative size, interposition, linear perspective, aerial perspective, textural gradient, atmospheric shading, luminance, height in visual field, and motion parallax.[39] It is also possible to successfully pilot an aircraft without stereopsis, as evidenced by a few monocular pilots, and monocular vision is allowed in trained civilian pilots of all classes following a 6-mo adaptation period.[34] Despite this, it is generally thought that while not absolutely necessary, stereopsis complements and enhances flying abilities.[31,43] For example, landings performed monocularly are altered with steeper and higher descents[8] and, in some cases, aviation mishaps have been attributed to a lack of stereopsis.[21]

Across all three UK military services, aircrew must meet the required entry stereo acuity vision standards set out in AP1269A.[29] These standards and test methods are summarized in **Table I** and compared with U.S. military and civilian stereo acuity vision standards. Among the Five Eyes Air Force Interoperability Council (AFIC), all but Australia test their aircrew population for stereo acuity, while Canada tests for stereo acuity but does not enforce any stereopsis standard.

As shown, a number of different stereo tests are employed and not all are necessarily comparable.[16,26,36] The Howard-Dolman, Verhoeff, and Frisby tests use real life depth stimuli. The Titmus, Randot, and Armed Forces Vision Tester (AFVT) use circle contours, all of which by their nature have monocular cues. Random Dot Stereograms (RDS), as the name suggests, are comprised of small random dots whose positions differ between the two eyes and are the only stereo tests to isolate disparity without the use of contours or monocular cues. Because the stimulus can only exist as binocular disparity, RDS tests are often considered the best measure of pure stereopsis.[16]

Tests used to qualify military aircrew, listed in Table I, were all developed for clinical settings and largely for screening purposes. A thorough review of U.S. Air Force stereo test methods is provided by Winterbottom et al.[41] While some of the stereo tests listed use the 'gold standard' RDS method, none of the tests provides a true threshold measure of stereo acuity. Instead, these stereo tests rely only on the ability to detect a disparity with respect to zero and most do not asses the person's ability to differentiate between crossed and uncrossed disparities. Thus, it is eminently possible to score well on one test yet fail another.[14,23] The Operational Based Vision Assessment (OBVA) Laboratory aims to improve test methods to counter the shortfalls in paper-based analog tests, measure vision more reliably, and investigate relationships between vision and operational performance so that decisions regarding vision standards are rooted in evidence. Indeed, newer computer-based tests are continuously being developed and tested both by the OBVA Laboratory and others.[10,16,28] The two digital stereo tests investigated here are digital versions of an RDS (dRDS), which is compared against the Toegepast Natuurwetenschappelijk Onderzoek (TNO) test (**Fig. 1**) used to qualify UK military aircrew.

The TNO test is a seven-page booklet of randomly paired red and green dots which should be viewed at 40 cm using red-green colored glasses. The individual must identify the image and the orientation of the missing segment of the circle, which might be in one of four positions. TNO test stimuli are presented with crossed disparity and thus appear in front of the reference plane. There are two circles for each level of stereo acuity and a subject must identify both correctly to progress to the next level. The binocular disparity of targets is 480, 240, 120, 60, 30, and 15 arcseconds, although the UK military employs a version of the TNO test with only six pages, which means the best score that can be achieved is 60 arcseconds. Unless directly specified, the TNO test referred to in this work is the six-page version used by the UK military. The subjects' scores are then recorded manually by the examiner. The TNO test uses RDS stimuli, so monocular cues should not play a part in interpreting the orientation of the missing segment of circle. As an RDS test, with minimal monocular cues, it is not unexpected that subjects have higher (worse) thresholds on the TNO test compared to a stereo test which is contour based.[9] However, even taking this

Table I. Stereopsis Vision Standards Across UK and U.S. Militaries, as Well as Civilian Organizations.

	UNITED KINGDOM		U.S. AIR FORCE		U.S. NAVY	U.S. ARMY	FAA	CAA
	PILOT	WSO	FC I/II	FC III	CLASS I, II (EXCEPT FIXED WING), CLASS III (INCL. UAV OPERATORS & CRITICAL FLIGHT DECK PERSONNEL)	PILOT CLASS 1 (COMMERCIAL) AND 2 (PRIVATE)	CLASS 1/2/3/4	
Stereo acuity (arcseconds)	120	N/A	40 waivable to 120 on AO-V	N/A	25 (VTA-ND) or 40 (Randot/ Titmus/AFVT) or 8/8 Verhoeff; no waiver	Normal binocular vision	40	No standard
Test method	TNO		AFVT	N/A	As above	AFVT; Randot; Titmus	None specified	

FAA = Federal Aviation Authority, CAA = Civil Aviation Authority (UK), AFVT = Armed Forces Vision Test, AO-V = AO-Vectograph, VTA-DP = Vision Test Apparatus-Near and Distant, WSO = Weapon System Operator, FC = Flying Class.[35]

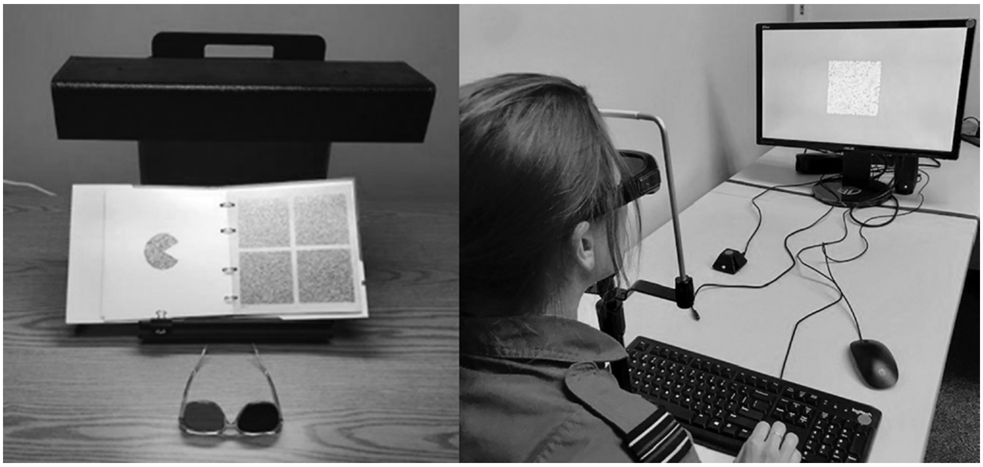

Fig. 1. Left: TNO test booklet with anaglyph glasses on an inclined stand. Right: Set up for dRDS-S and dRDS-D tests. Photograph taken by OBVA personnel.

difference into consideration, performance on the TNO test can be up to 25% worse than with any other stereo test.[20,37] A possible reason why the TNO test results in higher stereo thresholds than even other RDS tests could be that the different color filters in the glasses cause an imbalance in luminance transmittance and contrast.[37] This could be exacerbated further if lighting conditions are suboptimal. Another reason why stereo acuity thresholds are higher using the TNO test is that it is a more complex two-stage process; the user must first identify the circular shape and then indicate the orientation of the missing segment.[9] In comparison, simple detection tests such as the Randot test merely requires the sole stimuli in depth to be identified as the 'odd one out'.[37]

In addition to the TNO test yielding higher stereo acuity thresholds, there are other concerns with using the TNO test to assess stereo acuity. The TNO test has poor test-retest reliability, answers can be easily memorized, and there can be an unacceptable degree of variation between different test editions due to flaws in the printing process.[3] Van Doorn et al. demonstrated a statistically significant difference in stereo acuity results obtained using two separate editions of the seven-page TNO test, with mean measured stereo acuities of 30 arcseconds using one edition and 60 arcseconds using a different edition ($P < 0.001$). These differences were likely due to inconsistent image quality resulting from differences in the printing process.[5] Such profound limitations could result in human-machine technology mismatch with equipment such as stereo displays, which require users to have a minimum level of stereo acuity in order to be perceived and interpreted correctly. Furthermore, crude groups for stereo acuity scores make it impossible to closely track stereo acuity in an individual, as a marker of underlying pathology or effects of clinical treatment, or to monitor recovery to enable a return to flying duties. For example, traumatic brain injury and dementia are associated with worsening stereo acuity,[18,19] and could potentially be detected earlier and appropriately monitored with an accurate and reliable stereo acuity test.

To address some of the limitations of a paper TNO test, two digital RDS tests were developed. This research aims to assess whether computer based RDS tests could offer a fairer, more accurate, more reliable, and repeatable alternative stereo test to qualify military aircrew. Additionally, having the same tests or even comparable tests employed by different countries would improve interoperability with regard to human resources, allowing each country to accept aircrew from allied partner nations without further vision testing. For this experiment, results obtained using three different stereo acuity tests were analyzed.

METHODS

Three stereo tests were used to measure stereo acuity: the TNO test (six-page version) currently used by the UK military to qualify aircrew, and two digital RDS tests: a version in which the dots appear to move dynamically within the depth plane (dRDS-D) and a static version (dRDS-S). All three stereo tests were taken together and repeated a second time on a separate day. Test order was randomized using Microsoft Excel software (Microsoft, Redmond, WA, USA). Participants wore their habitual visual correction for all stereo tests.

Subjects

Recruited from the OBVA subject database were 45 participants. Volunteers were provided with a written information sheet. All participants were required to sign informed consent, indicating permission for their unidentifiable data to be stored and used within the OBVA laboratory. This study was approved by the U.S. Air Force (USAF) Air Force Research Laboratory Institutional Review Board (FWR20170095H).

Equipment

The TNO test is a six-page version with stimuli composed of random dot stereograms viewed at 40 cm using red-green colored glasses (Fig. 1). Reflective luminance of the booklet pages under a broadband reading light (incandescent bulb) was 197 to 306 cd · m^{-2} measured with a handheld Konica Minolta LS-110 (Konica Minolta Sensing America, Ramsey, NJ, USA).

Luminance decreased when viewed through the colored lenses to 13–23 cd · m^{-2} through the red lens and to 3–8 cd · m^{-2} through the green lens. Verbal instructions were given for the TNO test and responses were recorded by the examiner.

Both digital tests use an RDS stimulus similar to the TNO test, a circle with a missing segment, displayed on a 53 × 23-cm 3D computer monitor and viewed through Nvidia 3D active shutter glasses (ASUS, Taiwan) synchronized via an infrared transmitter. The shutter glasses synch with the refresh rate of the screen to ensure that the two disparate images, displayed one after the other, are presented to each eye separately to create a stereoscopic image. The test is performed in a darkened room with the participants' head fixed at 1 m using a chin rest (Fig. 1). The stimulus is presented with crossed disparity, appearing to be forward in depth compared to the plane of the computer screen. The perceived orientation of the missing segment of the circle is entered directly by the subject using the directional arrows on a keypad. The two different versions of the digital RDS test are: 1) dRDS-S and 2) dRDS-D. In the dRDS-S, all dots remain stationary, presented for 8 s. In the dRDS-D, the position of each dot was randomized with each frame refresh to give the dots a dynamic moving appearance (in the x- and y-planes), completely eliminating any monocular cues, also presented for 8 s. The square box measured 14 × 14 cm and the diameter of the stimulus was 7 cm. The mean background luminance of the stimulus was 149 cd · m^{-2} measured with 100% monitor brightness using a Konica Minolta Luminance Meter LS-110. The temporal resolution was 120 Hz (60 Hz to each eye) and spatial resolution 1920 × 1080 pixels. The test was programmed in C# in visualbasic.net using Direct X and a 2.2 gamma correction applied to improve accuracy for subpixel shift. Each dot was defined by a Gaussian function with a sigma of 3.75 pixels and there were 4000 dots randomly placed within the box.

To create the disparate image, three regions were designated within the box: A, B, and C (**Fig. 2**). Region A was the shape of the stimulus at its origin and all dots within it were shifted horizontally to region B and duplicated in region C. The dots in region B are presented only to the right eye and the dots in region C are presented only to the left eye, thereby creating a stereoscopic image. Simply shifting dots from region A to regions B and C creates both gaps between dots and an overlap of dots when regions B and C are viewed binocularly relative to the background dots in the area outside of region A, which could create monocular cues. To correct for this, excess background dots (checked regions) were moved to fill the void areas (striped region) for the opposite eye (Fig. 2).

For both the dRDS tests, instructions were given on the screen. Using the four alternative forced choice method, participants were instructed to choose up/down/right/left orientation of the stimulus. The forced choice model is a more effective way of measuring a detection threshold than relying on signal detection by a subject, which is biased by individual decision criterion.[38] Using four choices is more efficient than using two, reducing the guess rate to 25% with fewer trials needed to reach the detection threshold.[38] Each stimulus was presented for a maximum of 8 s; if no response was given within that time it was logged as an incorrect response. Each participant had a 10-trial practice session using stimuli with disparities of 300 to 2000 arcseconds prior to the formal test to ensure that participants understood the test and to reduce practice effects. Each dRDS test consisted of 45 trials with stereo acuity threshold, standard error, and slope estimate results displayed in an Excel spreadsheet (Microsoft). The Psi paradigm [25] is an adaptive procedure that was used to fit the psychometric function and estimate detection threshold.[13] Disparity of the test stimuli was altered in 0.1 log arcsecond step sizes, based on the participant's previous prior responses (i.e., the test generally got easier if the participant answered incorrectly, but more difficult if the participant responded correctly). The design and thresholding method of the dRDS test enabled each participants' stereo acuity threshold to be measured from 5 to 8000 arcseconds. The lapse rate of the psychometric function was fixed at 2.5%, but the slope was allowed to vary to allow for greater accuracy of the threshold estimate.[25] The lapse rate, which is sometimes called the finger-error rate, accounts for psychophysical errors not directly related to the observer's perception of the stimulus, such as accidentally pressing the wrong response button on the keypad.

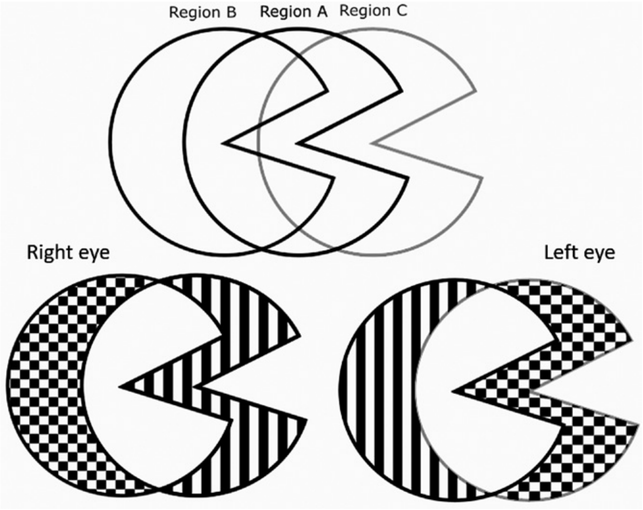

Fig. 2. Top: Region A outlines the original stimulus area. To create the disparate stereo image, the dots in region A are shifted horizontally leftwards to region B, viewed only by the right eye. The dots of region A are shifted an equal distance horizontally rightward to region C, viewed only by the left eye. Bottom: Shifting dots from region A to B, as viewed by the right eye, and to region C, as viewed by the left eye, will result in areas with an excess of dots (checkered area) overlapped onto the background dots, and a void of dots (striped area), creating monocular cues. To correct for this, excess background dots (checkered regions) are moved to fill the void areas (striped regions) in the opposite eye. Image created by OBVA personnel.

RESULTS

Using the OBVA subject database, 45 participants were recruited. One participant did not complete both sessions and

Table II. Mean, Median, Range, and Standard Deviation (SD) of the Three Stereo Acuity Tests.

	dRDS-S	dRDS-D	TNO
Mean – Log arcseconds	1.53 (33.79)	1.78 (59.6)	1.96 (91.56)
Median – Log arcseconds	1.51 (32.63)	1.78 (60.9)	1.78 (60)
Range – Log arcseconds	1.10–2.24 (12.64–173.78)	1.38–2.41 (23.87–197.83)	1.78–2.38 (60–240)
SD (arcseconds)	±0.25	±0.24	±0.22

another participant's data file was corrupted. A further two participants were essentially stereo blind, with scores of 4 and 3.8 log arcseconds as measured with the dRDS-S test. These scores are at the uppermost limit of the tests' capability and are likely unreliable. The final sample size was $N = 41$ (24 men; 37.5 mean age, SD ± 10.1 yr, range 21–70). Analyses were conducted using IBM SPSS Version 26 (IBM Corp., Armonk, NY, USA).

Recruitment was targeted to ensure participants with a wide spread of stereo acuities were included; thus, the sample population is broader than traditional military aircrew populations, who are required to meet the stereo acuity selection standard of 120 arcseconds as measured with the TNO stereo test in the UK and 40 arcseconds as measured by the AFVT for the USAF. Mean and median stereo acuity values are given in **Table II** with a histogram illustrating frequency of results in **Fig. 3**.

Statistical Analysis

A Friedman two-way analysis of variance (ANOVA) by ranks test showed there was a statistically significant difference in the distribution of stereo acuity scores for the three tests [$\chi^2 = 53.12$, df (2), $P < 0.01$]. A post hoc Wilcoxon signed-rank test was performed to determine if there was a median difference between each pair of the different stereo acuity tests. Using a Bonferroni-corrected *P*-value of 0.012, all test pairs differed significantly (**Table III**).

Bland-Altman plots were used to quantify the reliability of a repeated test, or agreement between two measures, by comparing mean differences and calculating limits of agreement. The closer the mean difference is to zero, the better the agreement between two measures, and the smaller the standard deviation, the more repeatable and reliable a test is. It is considered a better way to understand comparability between two measures than simple correlation analyses, which evaluate only linear association and used on their own could be misleading.[7] It should be noted that while the Bland-Altman method calculates limits of agreement, it is unable to determine whether these are acceptable or not. That is a separate task entirely depending on the user's appetite for risk and need for reliability.

On both attempts of the TNO test, 27 participants achieved the same score (66%), with some participants able to simply remember their answers from the previous session. With Bland-Altman analysis the mean difference between first and second sessions was 0.03 log arcseconds (95% CI = ± 0.05) with limits of agreement ± 0.34 log arcseconds, a combined total of 0.69 log arcseconds (**Fig. 4A**). These limits of agreement are artificially narrowed, since there are only four possible stereo acuity values for the RAF six-page version of the TNO test; thus, a large degree of variance has already been removed. When translated into a real-life example, subjects scoring 120 arcseconds with the TNO may actually have a true stereo acuity varying anywhere between 70.6 to 203.8

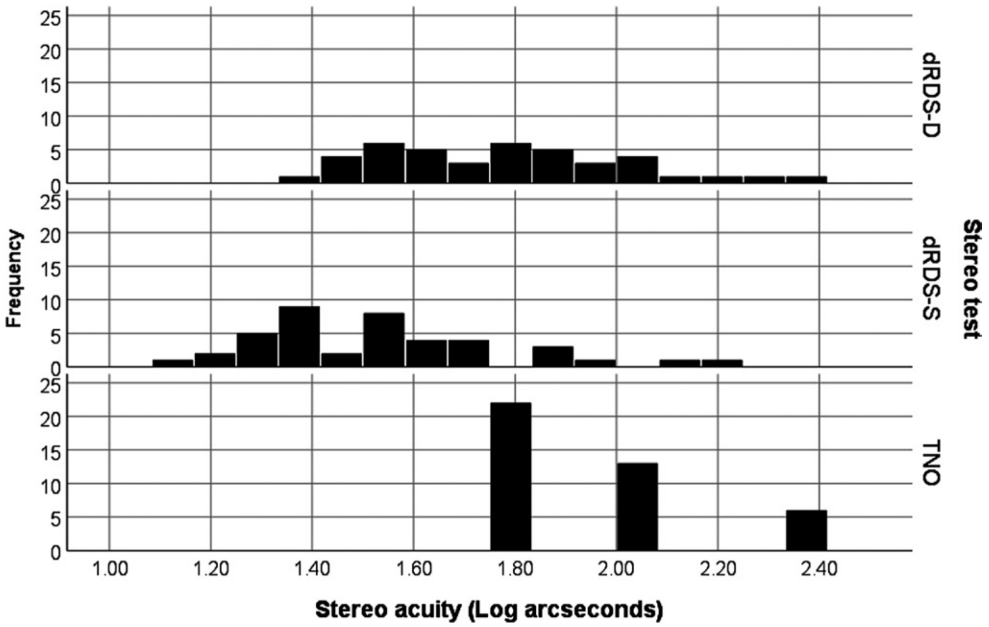

Fig. 3. Histogram of stereo acuity score (log arcseconds) frequencies measured with dRDS-D, dRDS-S, and TNO.

Table III. Wilcoxon Signed Rank Test Between Pairs.

COMPARISON	z-SCORE	SIGNIFICANCE
dRDS-S vs. dRDS-D	−5.18	*P* < 0.001
dRDS-S vs. TNO	−5.40	*P* < 0.001
dRDS-D vs. TNO	−3.58	*P* < 0.001

Significance level: *P* < 0.012.

arcseconds. These results indicate better test-retest reliability than previous reports of a difference of 0.06 log arcseconds and 95% limits of agreement of 1.53 log arcseconds.[32] It is noted that in their test-retest reliability analysis of the TNO test, Tittes et al.[32] used the seven-page version, which incorporated an additional two levels, able to measure stereo acuity down to 15 arcseconds. In our subsequent analyses, the first stereo acuity score was used.

With Bland-Altman analysis the mean difference between first and second dRDS-S sessions was 0.04 log arcseconds (95% CI = ±0.04) with limits of agreement ±0.25 log arcseconds, a combined total of 0.50 log arcseconds (**Fig. 4B**). With Bland-Altman analysis the mean difference between first and second dRDS-D sessions was 0.05 log arcseconds (95% CI = ±0.04) with limits of agreement ±0.23 log arcseconds, a combined total of 0.47 log arcseconds (**Fig. 4C**).

Each stereo test assesses stereopsis using a different method, giving significantly different results. It is important to compare levels of agreement between these tests to aid interpretation and relatability. A simple scatter plot between the two dRDS tests show that they correlate strongly (r = 0.73, *P* < 0.001) (**Fig. 4D**). The results of Bland-Altman analysis for stereo acuity scores measured with the two dRDS tests are shown in **Fig. 4G**. The mean difference between them is 0.24 log arcseconds (95% CI = ±0.05), with limits of agreement ±0.35 log arcseconds. There is no significant relationship between the TNO test and either dRDS test (**Fig. 4E** and **Fig. 4F**). With Bland-Altman analysis, the mean difference between the TNO test and dRDS-S is 0.43 log arcseconds (95% CI = ±0.08), with limits of agreement ±0.55 log arcseconds (**Fig. 4H**). With Bland-Altman analysis, the mean difference between the TNO test and dRDS-D is 0.11 log arcseconds (95% CI = ±0.14), with limits of agreement ±0.56 log arcseconds (**Fig. 4I**).

DISCUSSION

With a large floor effect at 60 arcseconds, the six-page paper TNO test was unable to characterize any finer degree of stereo

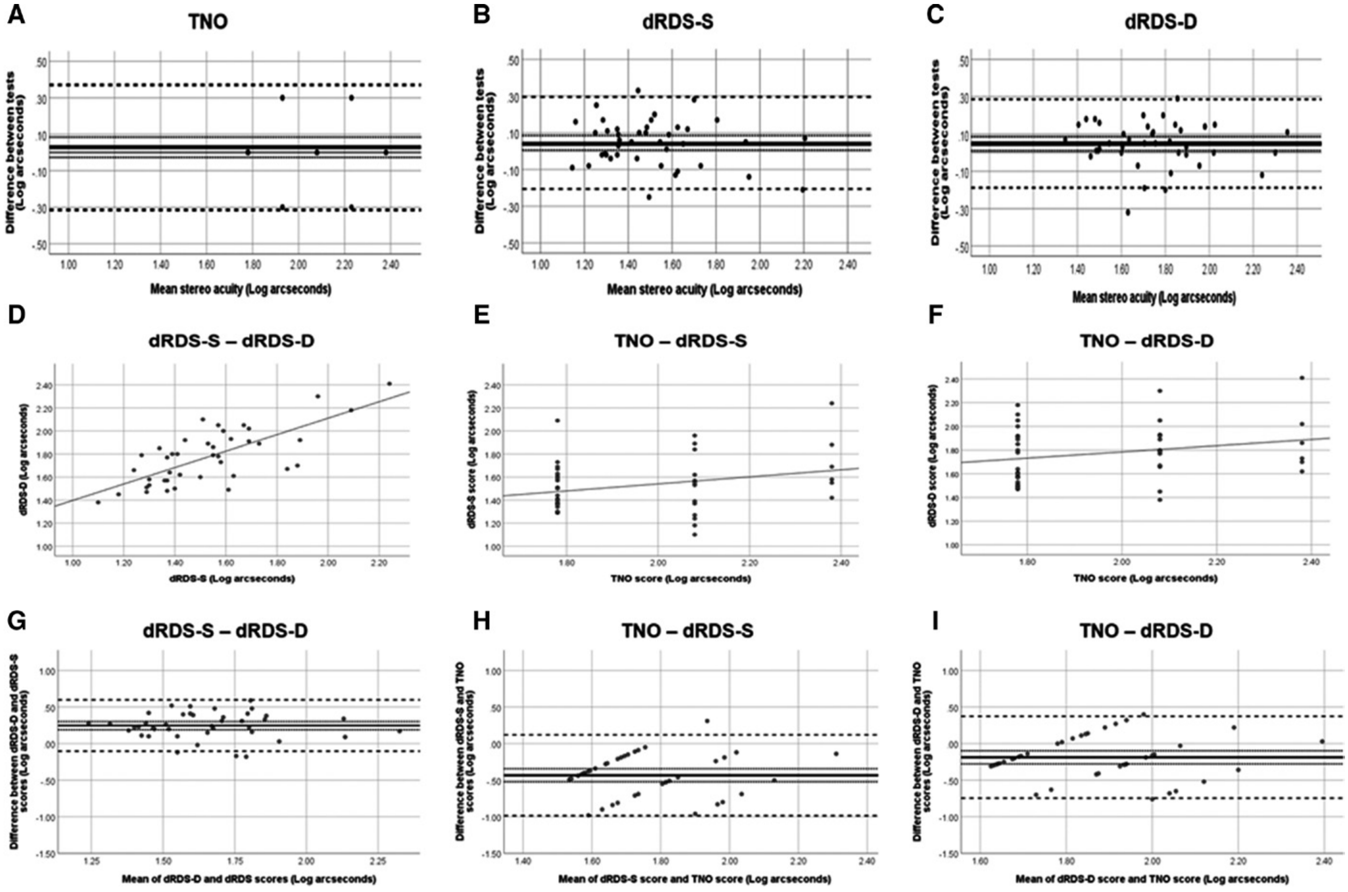

Fig. 4. A-C) Bland-Altman plots assessing agreement between first and second attempts of each test; D-F) correlation analyses between tests; and G-I) Bland-Altman plots of agreement between the three different stereo tests. The thick black line is the mean difference (bias), with the dashed lines indicating the upper and lower limits of agreements. Dotted lines show 95% confidence intervals.

acuity, which, by comparison, was possible using both the dRDS-S and dRDS-D tests. The lowest thresholds (best stereo acuity) were measured using the static version of the RDS (dRDS-S), while the worst scores were reported using the TNO test. Both digital RDS tests were more reliable than the TNO test, as demonstrated with tighter limits of agreement for Bland-Altman analyses. The tightest limits of agreement and, thus, the most reliable stereo test, were seen with the dynamic RDS version (dRDS-D) of the digital tests. With regard to the significant difference in results between the digital tests, a reason why the higher thresholds were recorded using the dynamic version of the dRDS when compared to the static dRDS could be that the dRDS-D is least likely to have any monocular cues and participants found it more difficult. It has been suggested that static and dynamic disparities are processed in a different manner, giving different results when measuring individual stereo acuity,[33,44] and as such these tests may not be directly comparable. However, the dynamic motion applied to the dots in the dRDS-D test is confined to the same plane of presentation and is not a dynamic change in depth. Notably, there was no significant correlation between either of the dRDS tests and the TNO test. An individual scoring 60 arcseconds on the TNO could obtain a score ranging anywhere from approximately 30 to 160 arcseconds on the dRDS-S. This suggests that the TNO test does not reliably measure stereo acuity.

Both computer-based threshold tests eliminated many of the limitations identified with the TNO test. Crucially, the random order of stimulus presentation makes it impossible to cheat or memorize answers, reducing the incidence of false positive results. Furthermore, there is no chromatic imbalance using active shutter glasses, and printing or illumination discrepancies are removed using a standardized computer screen. In addition, examiner interference is minimized, removing the possibility of transcription errors or human bias when giving instructions or recording results. The chief disadvantage of computer-based tests is that they require more expensive hardware resources, in the form of a computer, 3D monitor, and active shutter glasses, to operate. The value of evidence-based medical standards is difficult to quantify. However, given that the estimated cost to fully train a pilot on a 5th generation fast jet aircraft, in which a binocular HMD is critical, is over $10 million,[15] using an operationally relevant stereo test to accurately identify pilot candidates as either medically fit/unfit could significantly reduce the number of pilots unable to complete the intensive training programs, resulting in significant cost savings.

Furthermore, as the computer-based tests are more precise and repeatable, they would be better able to identify more reliably any relationship to operational performance if one exists. Such research is crucial in providing evidence to support aircrew vision standards. Another benefit of more precise vision screening tests is their ability to detect smaller changes in stereo acuity; thus, they are better able to identify medical situations that warrant further investigation at an earlier stage. Currently, no accurate baseline data exist to either better diagnose disease or injury requiring treatment, or to quantify recovery to support return to flying decisions.

There is no clear answer as to which test is the best and most appropriate to use. Such a decision will depend on the tests' ability to predict operational performance and further research is needed in this area. Measuring stereo acuity more accurately could enhance the effectiveness of qualifying standards and our understanding of human performance, as indicated by findings that lower stereo acuity thresholds predict superior performance in an aerial refueling task,[22,40,42] depth related surgical tasks,[1,30] and object placement tasks mediated by stereo displays.[17] Notably, a computer-based stereo acuity test was predictive of simulated air refueling performance in previous research while the AFVT stereo test was not.[42] Further research should also be conducted into stereo acuity measured using frontal dynamic motion-in-depth as this may be a better indicator of overall binocular vision because it includes a time and spatial component.[6]

In addition to these benefits, more reliable threshold estimation tests could support interservice and international co-operation. Currently, the TNO test (UK six-page version) is unable to measure stereo acuity to the vision standards required by the USAF (40 arcseconds). As there are pilots from both the RAF and USAF embedded in each other's flying operations, as part of the enduring exchange programs between allied countries, it is important to have tests that are reliable and clear. We would advocate for aligning aeromedical policy and vision standards to further aid interoperability. Research such as this, aimed at developing vision performance models that predict operational performance, will assist in providing evidence to set vision standards and drive aeromedical policy which can be shared with allied nations.

The limitations of the paper TNO test have been clearly highlighted, with computer-based threshold tests addressing many of these and offering a feasible alternative solution. Digital tests are able to measure individual stereo acuity to a finer degree than the TNO test and do so in a manner that reduces examiner interference or bias and eliminates the possibility of cheating (false positives). For the two versions of the dRDS tests, the static version of the dRDS (dRDS-S) gave the lowest stereo acuity thresholds, but the dynamic version (dRDS-D) was more reliable with the tightest limits of agreement. While the computer-based stereo acuity tests produce significantly different scores, their results are strongly correlated. Neither of the computer-based tests correlates with the TNO test, which suggests that the TNO test does not provide either the accuracy or reliability needed in aeromedical screening for an increasingly digital cockpit environment. The greater granularity achieved with digital tests will enable us to investigate the relationship between stereo acuity and operational performance, which in turn will inform stereo acuity vision selection standards or display requirements. This will be increasingly important for military aviators for whom stereoscopic displays and HMDs are becoming more prevalent and critical to flying operations.

ACKNOWLEDGMENTS

The views expressed are those of the authors and do not necessarily reflect the official policy or position of the Air Force, the Department of Defense, or the U.S. Government. This research forms part of Dr. Posselt's Ph.D. thesis investigating human performance with binocular HMDs and examining associated stereo test methods and stereo acuity standards for military aviators. Many thanks to Alex Van Atta for writing the code to create both digital RDS stereo tests. Many thanks to Carrie Drake for her assistance in running participants to collect data.

Financial Disclosure Statement: This work was supported in part under USAF contract FA-8650-18-C-6932 and has been cleared for public release (AFRL-2022-0224). The authors have no competing interests to declare.

Authors and Affiliations: Bonnie N. Posselt, M.B.Ch.B., RAF Centre of Aviation Medicine, RAF Henlow, Bedfordshire, United Kingdom, University of Birmingham, Birmingham, United Kingdom, and Operational Based Vision Assessment, 711th Human Performance Wing, U.S. Air Force, Wright-Patterson AFB, OH, United States; Eric Seemiller, Ph.D., Eagle Integrated Services, Rockville, MD, United States; Marc Winterbottom, Ph.D., and Steve Hadley, M.D., Operational Based Vision Assessment, 711th Human Performance Wing, U.S. Air Force, Wright-Patterson AFB, OH, United States; and Chris Baber, Ph.D., University of Birmingham, Birmingham, United Kingdom.

REFERENCES

1. Alhusuny A, Cook M, Khalil A, Treleaven J, Hill A, Johnston V. Impact of accommodation, convergence and stereoacuity on perceived symptoms and surgical performance among surgeons. Surg Endosc. 2021; 35(12): 6660–6670.
2. Allison RS, Gillam BJ, Vecellio E. Binocular depth discrimination and estimation beyond interaction space. J Vis. 2009; 9(1):10.1–14.
3. Antona B, Barrio A, Sanchez I, Gonzalez E, Gonzalez G. Intraexaminer repeatability and agreement in stereoacuity measurements made in young adults. Int J Ophthalmol. 2015; 8(2):374–381.
4. Ding J, Levi DM. Recovery of stereopsis through perceptual learning in human adults with abnormal binocular vision. Proc Natl Acad Sci USA. 2011; 108(37):E733–E741.
5. van Doorn LLA, Evans BJW, Edgar DF, Fortuin MF. Manufacturer changes lead to clinically important differences between two editions of the TNO stereotest. Ophthalmic Physiol Opt. 2014; 34(2):243–249.
6. Dunlop DB, Neill RA, Dunlop P. Measurement of dynamic stereoacuity and global stereopsis. Aust J Ophthalmol. 1980; 8(1):35–46.
7. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015; 25(2):141–151.
8. Grosslight JH, Fletcher HJ, Masterton RB, Hagen R. Monocular vision and landing performance in general aviation pilots: Cyclops revisited. Hum Factors. 1978; 20(1):27–33.
9. Hall C. The relationship between clinical stereotests. Ophthalmic Physiol Opt. 1982; 2(2):135–143.
10. Hess RF, Ding R, Clavagnier S, Liu C, Guo C, et al. A robust and reliable test to measure stereopsis in the clinic. Invest Ophthalmol Vis Sci. 2016; 57(3):798–804.
11. Hess RF, To L, Zhou J, Wang G, Cooperstock JR. Stereo vision: the haves and have-nots. Iperception. 2015; 6(3):2041669515593028.
12. Lee SY, Koo NK. Change of stereoacuity with aging in normal eyes. Korean J Ophthalmol. 2005; 19(2):136–139.
13. Leek MR. Adaptive procedures in psychophysical research. Percept Psychophys. 2001; 63(8):1279–1292.
14. Leske DA, Birch EE, Holmes JM. Real depth vs randot stereotests. Am J Ophthalmol. 2006; 142(4):699–701.
15. Mattock MG, Asch BJ, Hosek J, Boito M. The relative cost-effectiveness of retaining versus accessing Air Force pilots. Santa Monica (CA): RAND Corporation; 2019.
16. McCaslin AG, Vancleef K, Hubert L, Read JCA, Port N. Stereotest comparison: efficacy, reliability, and variability of a new glasses-free stereotest. Transl Vis Sci Technol. 2020; 9(9):29.
17. McIntire JP, Wright ST, Harrington LK, Havig PR, Watamaniuk SNJ, Heft EL. Optometric measurements predict performance but not comfort on a virtual object placement task with a stereoscopic 3D display. Dayton (OH): 711th Human Performance Wing, Wright-Patterson AFB; 2014.
18. Miller LJ, Mittenberg W, Carey VM, McMorrow MA, Kushner TE, Weinstein JM. Astereopsis caused by traumatic brain injury. Arch Clin Neuropsychol. 1999; 14(6):537–543.
19. Mittenberg W, Choi EJ, Apple CC. Stereoscopic visual impairment in vascular dementia. Arch Clin Neuropsychol. 2000; 15(7):561–569.
20. Momeni-Moghadam H, Kundart J, Ehsani M, Gholami K. Stereopsis with TNO and titmus tests in symptomatic and asymptomatic university students. Journal of Behavioral Optometry. 2012; 23(2):35–39.
21. Nakagawara VB, Véronneau SJH. A unique contact lens-related airline aircraft accident. Washington (DC): Office of Aviation Medicine; 2000.
22. O'Keefe E, Ankrom M, Seemiller ES, Bullock T, Winterbottom M, et al. The relationship between vision and simulated remote vision system air refueling performance. In: Proceedings of the IS&T International Symposium on Electronic Imaging: Stereoscopic Displays and Applications. Springfield (VA): Society for Imaging Science & Technology; 2022:289-1–289-6.
23. Pageau M, de Guise D, Saint-Amour D. Comparison of local and global stereopsis in children with microstrabismus. J Vis. 2009; 9(8):284.
24. Posselt BN, Winterbottom M. Are new vision standards and tests needed for military aircrew using 3D stereo helmet-mounted displays? BMJ Mil Health. 2021; 167(6):442–445.
25. Prins N, Kingdom FAA. Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the Palamedes Toolbox. Front Psychol. 2018; 9:1250.
26. Read JCA. Stereo vision and strabismus. Eye (Lond). 2015; 29(2):214–224.
27. Richards W. Stereopsis and stereoblindness. Exp Brain Res. 1970; 10(4):380–388.
28. Rodriguez-Vallejo M, Llorens-Quintana C, Montagud D, Furlan WD, Monsoriu JA. Fast and reliable stereopsis measurement at multiple distances with iPad. [Abstract 1609.0]. Computing Research Repository; 2016.
29. Royal Air Force Manual. Assessment of medical fitness - AP1269A, 3rd ed. London (UK): Defence Council, Ministry of Defence; 1998.
30. Sakata S, Grove PM, Hill A, Watson MO, Stevenson ARL. Impact of simulated three-dimensional perception on precision of depth judgements, technical performance and perceived workload in laparoscopy. Br J Surg. 2017; 104(8):1097–1106.
31. Snyder QC. Assessment of two depth perception test to predict undergraduate pilot training completion. Report number: AFIT/CI/CIA-91-057. Wright-Patterson AFB (OH): Air Force Institute of Technology; 1991.
32. Tittes J, Baldwin AS, Hess RF, Cirina L, Wenner Y, et al. Assessment of stereovision with digital testing in adults and children with normal and impaired binocularity. Vision Res. 2019; 164:69–82.
33. Tittle JS, Rouse MW, Braunstein ML. Relationship of static stereoscopic depth perception to performance with dynamic stereoscopic displays. Proc Hum Factors Soc Annu Meet. 1988; 32(19):1439–1442.
34. UK Civil Aviation Authority. Visual system guidance material. [Accessed Oct. 13, 2022]. Available from https://www.caa.co.uk/Aeromedical-Examiners/Medical-standards/Pilots-(EASA)/Conditions/Visual/Visual-system-guidance-material-GM/.
35. U.S. Air Force School of Aerospace Medicine. Air Force waiver guide. U.S. Air Force School of Aerospace Medicine; 2020.
36. Vancleef K, Read JCA. Which stereotest do you use? A survey research study in the British Isles, the United States and Canada. Br Ir Orthopt J. 2019; 15(1):15–24.

37. Vancleef K, Read JCA, Herbert W, Goodship N, Woodhouse M, Serrano-Pedraza I. Overestimation of stereo thresholds by the TNO stereotest is not due to global stereopsis. Ophthalmic Physiol Opt. 2017; 37(4):507–520.
38. Vancleef K, Read JCA, Herbert W, Goodship N, Woodhouse M, Serrano-Pedraza I. Two choices good, four choices better: for measuring stereoacuity in children, a four-alternative forced-choice paradigm is more efficient than two. PLoS One. 2018; 13(7):e0201366.
39. Wickens CD, Hollands JG, Banbury S, Parasuraman R. Engineering Psychology and Human Performance, 4th ed. Oxford (UK): Taylor and Francis; 2012.
40. Winterbottom MD. Individual differences in the use of remote vision stereoscopic displays. Dayton (OH): Wright State University; 2015.
41. Winterbottom M, Gaska J, Wright S, Hadley S, Lloyd C, et al. Operational based vision assessment research: depth perception. J Aust Soc Aerosp Med. 2014; 9(November):33–41.
42. Winterbottom M, Lloyd C, Gaska J, Wright S, Hadley S. Stereoscopic remote vision system aerial refueling visual performance. In: Proceedings of the IS&T International Symposium on Electronic Imaging: Stereoscopic Displays and Applications XXVII. Springfield (VA): Society for Imaging Science & Technology; 2016; 28:art00028.
43. Wright S, Gooch JM, Hadley S. The role of stereopsis in aviation: literature review. Wright-Patterson AFB (OH): Air Force Research Laboratory; 2013. Report No.: AFRL-SA-WP-TP-2013-0001.
44. Zinn WJ, Solomon H. A comparison of static and dynamic stereoacuity. J Am Optom Assoc. 1985; 56(9):712–715.