UC Berkeley
UC Berkeley Previously Published Works

Title

The eRDS v6 Stereotest and the Vivid Vision Stereo Test: Two New Tests of Stereoscopic Vision.

Permalink
https://escholarship.org/uc/item/1wf697pj

Journal
Translational Vision Science & Technology, 12(3)

ISSN
2164-2591

Authors
Denkinger, Sylvie
Antoniou, Maria-Paraskevi
Tarello, Demetrio
et al.

Publication Date
2023-03-01

DOI
10.1167/tvst.12.3.1

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

tvst

Pediatric Ophthalmology & Strabismus

The eRDS v6 Stereotest and the Vivid Vision Stereo Test: Two New Tests of Stereoscopic Vision

Sylvie Denkinger[1,6], Maria-Paraskevi Antoniou[1,2], Demetrio Tarello[1], Dennis M. Levi[3], Benjamin T. Backus[4], Daphné Bavelier[1,5], and Adrien Chopin[1,6]

[1] Psychology and Education Sciences, University of Geneva, Switzerland
[2] Institute of Information Systems, University of Applied Sciences & Arts Western Switzerland (HES-SO) Valais-Wallis, Sierre, Switzerland
[3] Herbert Wertheim School of Optometry and Vision Science, University of California, Berkeley, CA, USA
[4] Vivid Vision, Inc., San Francisco, CA, USA
[5] Psychology and Education Sciences, University of Geneva & Campus Biotech, Switzerland
[6] Sorbonne Université, INSERM, CNRS, Institut de la Vision, Paris, France

Correspondence: Adrien Chopin, Sorbonne Université, INSERM, CNRS, Institut de la Vision, 13 rue Moreau, 75012 Paris, France. e-mail: adrien.chopin@gmail.com

Received: August 23, 2022
Accepted: January 24, 2023
Published: March 1, 2023

Keywords: stereovision; stereoacuity; reliability; test-retest; validity

Citation: Denkinger S, Antoniou MP, Tarello D, Levi DM, Backus BT, Bavelier D, Chopin A. The eRDS v6 stereotest and the vivid vision stereo test: Two new tests of stereoscopic vision. Transl Vis Sci Technol. 2023;12(3):1, https://doi.org/10.1167/tvst.12.3.1

Purpose: To describe two new stereoacuity tests: the eRDS v6 stereotest, a global dynamic random dot stereogram (dRDS) test, and the Vivid Vision Stereo Test version 2 (VV), a local or "contour" stereotest for virtual reality (VR) headsets; and to evaluate the tests' reliability, validity compared to a dRDS standard, and learning effects.

Methods: Sixty-four subjects passed a battery of stereotests, including perceiving depth from RDS. Validity was evaluated relative to a tablet-based dRDS reference test, ASTEROID. Reliability and learning effects were assessed over six sessions.

Results: eRDS v6 was effective at measuring small thresholds (<10 arcsec) and had a moderate correlation (0.48) with ASTEROID. Across the six sessions, test-retest reliability was good, varying from 0.84 to 0.91, but learning occurred across the first three sessions. VV did not measure stereoacuities below 15 arcsec. It had a weak correlation with ASTEROID (0.27), and test-retest reliability was poor to moderate, varying from 0.35 to 0.74; however, no learning occurred between sessions.

Conclusions: eRDS v6 is precise and reliable but shows learning effects. If repeated three times at baseline, this test is well suited as an outcome measure for testing interventions. VV is less precise, but it is easy and rapid and shows no learning. It may be useful for testing interventions in patients who have no global stereopsis.

Translational Relevance: eRDS v6 is well suited as an outcome measure to evaluate treatments that improve adult stereodepth perception. VV can be considered for screening patient with compromised stereovision.

Introduction

Stereoscopic vision relies on binocular disparities created by the difference of viewpoints between the two eyes to extract depth information from the environment. Stereoacuity refers to the quantitative measure of stereoscopic vision and is most commonly assessed with clinical stereotests. Unfortunately, these tests often suffer from several of the following limitations: (1) failure to detect stereoblindness,[1] (2) a lack of precision, because of both classification of results in discrete values rather than on a continuous scale,[2,3] and inherent imprecision in the estimates of thresholds, (3) a high chance level (probability of obtaining a non-stereoblind result while responding randomly to the test),[1] and (4) contamination by non-stereoscopic cues that allow stereo-deficient patients to produce deceptively good results.[4] As a result, these tests have limited construct validity: they do not necessarily measure what they are supposed to.

An added challenge comes when evaluating intervention efficacy: are the measures truly reflecting intervention effects rather than possible test-retest learning effects? Indeed, tracking changes in performance throughout intervention implies multiple

Copyright 2023 The Authors

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

measurements and the measurement itself may induce improved performance, either through perceptual learning[5] or through task familiarity.[6,7] Thus good tests must show little or no learning through repeated measurements while also being precise enough to measure small improvements. In recent years, there have been several attempts at developing new computer-based stereoacuity tests.[2,8–16] However, studies have long since demonstrated that repeated testing with the same stimuli results in perceptual learning (i.e., improved perception of depth with an improvement in thresholds).[17,18] Some test-retest studies have found improvements in stereoacuity[19,20] despite a large interindividual variability.[3,15,16]

To address the limitations of current clinical stereotests, we developed and evaluated two new measures of stereoacuity: a dynamic Random Dot Stereogram (RDS) test called eRDS (version 6; Adrien Chopin, Paris)[1] and a stereotest developed in a virtual reality environment, the Vivid Vision Stereo Test version 2 (VV; Vivid Vision Inc., San Francisco, CA, USA).[3]

The eRDS v6 stereotest measures global stereopsis, which requires the participant to first resolve the binocular correspondence problem when there are numerous items in depth, as in Random Dot Stereograms (RDS). We designed the test to precisely follow the recommendations from Chopin and collaborators[1] for a "pure" measure of the sensibility to binocular disparity. Among other features:

(1) The eRDS v6 stereotest can measure a large range of stereoacuities, from fine stereopsis to what Chopin et al.[1] refer to as ecological stereoblindness (∼1300 arcsec, based on the distribution of environmental disparities at the fovea).
(2) The test uses a continuous scale and implements a new efficient Bayesian sampling method, psi-marg-grid, that avoids catastrophic test failures when participants show a specific profile called non-monotonic (see references[21,22] for further explanations and demonstrations).
(3) Using simulations, Chopin and colleagues[21] found the chance level for the sampling procedure to be near 0.2% (the lower the better).
(4) To prevent participants from deceiving the test by using monocular cues, the test presents dynamic random dot stereograms (dRDS),[23] and to prevent the use of binocular non-stereoscopic cues,[24] the test is designed around a near-far depth-sign task,[1] where the participant simply reports whether a strip of dots is closer or further than another strip of dots.

The VV test measures local stereopsis, based on matching the contours of isolated items with no correspondence ambiguity. It is available clinically, has a short test time, and was also designed to partially address the limitations above:

(1) VV can measure a large range of stereoacuities, from fine stereopsis to 2400 arcsec.
(2) It uses a continuous scale. As provided, the test assumes that performance increases monotonically with disparity. Because this assumption is sometimes incorrect, we have analyzed the test data here by using a non-monotonic psychometric function.
(3) It is designed to prevent monocular cues by adding binocular random jitter to the items to be compared in depth. However, it was not designed to eliminate binocular non-stereoscopic cues.[24]
(4) We performed the test 30 times using randomly generated responses, which always produced a "no stereo" result, suggesting a chance probability of passing the test without stereo ability below 3.3%.

Our study has two aims: (1) to describe the above mentioned new stereoacuity tests and (2) to assess the concurrent validity of these new stereotests, their reliability across multiple measures and their susceptibility to test-retest learning. Concurrent validity assesses whether a new test is measuring the construct that it is supposed to measure, by comparing the test to another test that has already been proven to be valid for that construct (the reference test). For this study, we used ASTEROID as a reference for measuring stereoacuity. This stereotest is a child-friendly clinical stereotest, developed with the goal of overcoming some of the above limitations.[2,25,26] It tests over a large continuous range of disparities, and we measured its chance level to be lower than 3.3%.[24] The test consists of a dynamic RDS, which does not have any useable monocular cues.[27] ASTEROID was recently evaluated and showed both good reliability and good concurrent validity compared to the Randot circles stereotest[26] and the Randot Preschool.[2]

In measuring the validity of the two new tests, we are measuring how well they agree with ASTEROID. Lower validity might be expected for VV because, by design, VV measures contour stereopsis rather than global stereopsis. However, if instead, VV had good validity for global stereopsis, then we would have a single, flexible tool for all patients. With the reliability analysis we measure the susceptibility to test-retest learning and the noise in repeating those measures. If we observe no learning between sessions, then we would consider that our tests could be reliably used for measuring specific treatment effects during interventions. At a group level, a reliable test allows to interpret any statistical differences between sessions in an intervention or between conditions as a real change. Evaluating the Bland-Altmann limit of agreement between 2 measures allows one to know precisely what can be considered a real change for a particular individual.[3,28] Indeed, if a test shows a limit of agreement of ±0.4 log units, it means that an intervention needs to reach at least an effect size of 0.4 log unit to be considered not simply test noise for that particular individual.

Methods

Participants

Sixty-six young adults (13 females, 53 males) from 18 to 35 years old (mean = 22.8, standard deviation [SD] = 3.71) were recruited from the University of Geneva and surrounding areas. The study was approved by the Ethics Committee of the Faculty of Psychology, University of Geneva. Participants were paid for participating and all gave written informed consent. Inclusion criteria were normal or corrected-to-normal binocular visual acuity of at least 20/40, measured on a SLOAN chart at 3m viewing distance with habitual optical correction, interocular acuity difference smaller than 0.2 logMAR, as well as failing the Butterfly stereoblindness test or the Frisby test at 20 arcsec (see Supplement S1 for more information). This last criterion was intended to increase the range of stereoacuities present in our sample. Another inclusion criterion was the ability to fuse without suppression when tested with the Diplopia-Suppression Test[29] during the first session. Two participants were excluded: one due to an episode of blurred vision and one stereoblind on all tests. Six participants dropped out after the first session (two for unknown reasons and four because of the COVID19 situation), leaving a total of 58 participants for reliability analysis and 64 for validity analysis.

Experimental Design

The full experiment called for 6 sessions spaced 10 to 15 days apart (Fig. 1). At the first and last sessions (T1 and T6), participants underwent a battery of visual measurements (see Supplementary information, S1). Here we report the stereotests of interest for the present study: eRDS, the Vivid Vision Stereo test version 2 (VV) and ASTEROID. To control for potential confounds, the order of the eRDS and VV tests in T1 and T6 was counterbalanced across participants. For the other sessions (T2, T3, T4 and T5) participants were pseudo-randomized into two groups: the eRDS-repeat group or the VV-repeat group. Experimenters were not masked as to the assigned group, but participants were masked to the existence of separate groups.

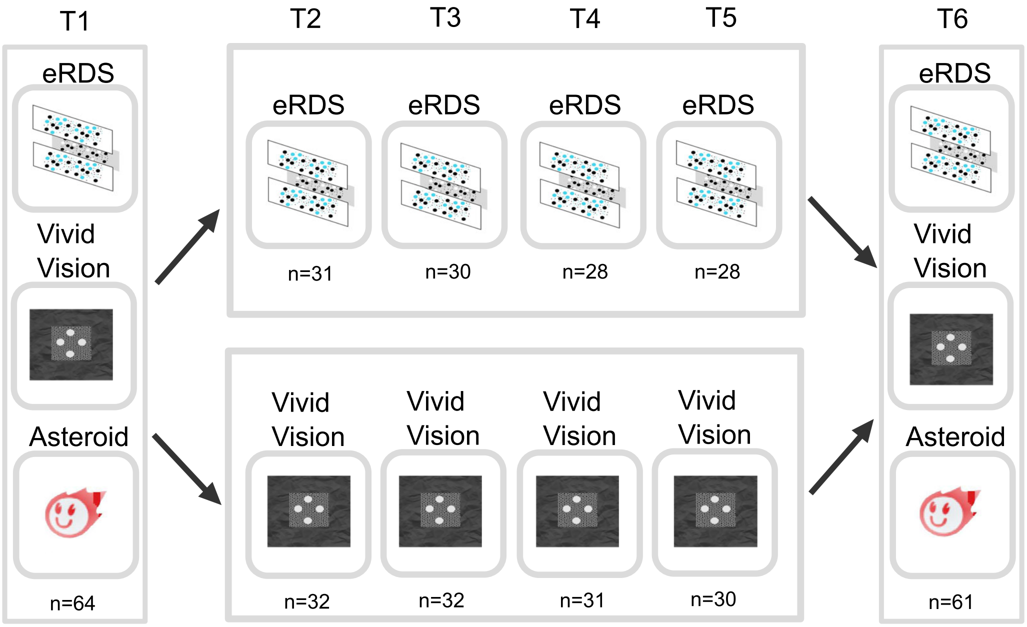

Figure 1. Experimental design. For sessions 2 to 5, participants were attributed either to the eRDS-repeat group or to the VV-repeat group; each session (T1 to T6) lasted between 15 to 30 minutes, depending on the group. n at T6 is higher than n at T5 because participants leaving after the start of the study were kindly asked to come for the post-test session.

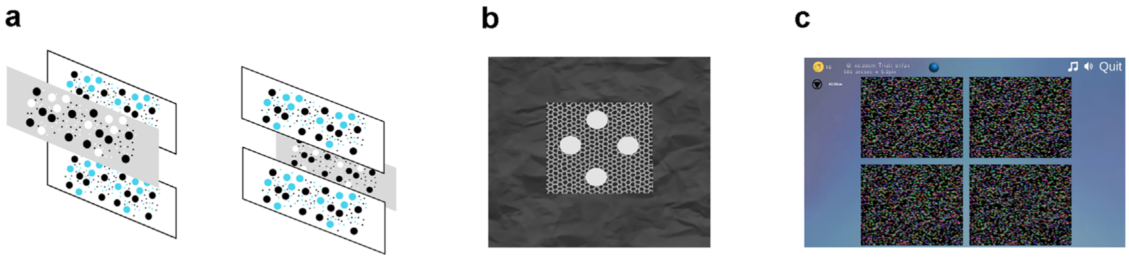

Figure 2. (a) eRDS stimuli. Participants have to decide whether the blue-and-black-dot strips are in front or behind the black-and-white-dot strip. Lines and white background do not exist and are presented here only for illustration purpose. The background of each strip is gray. (b) Vivid Vision Stereoacuity stimulus. Participants must detect which target appeared in front of the others c. ASTEROID stereotest. Participants must touch the panel where a square of dots appears in depth.

The eRDS Stereotest

The eRDS (version 6) is a dRDS test that measures stereoscopic vision performance using recommendations from Chopin and collaborators.[1] It is a 3D depth detection task where stimuli are presented on a computer screen and viewed through a stereoscope (see Supplementary Information S1 for more details). The stimuli are composed of three strips of RDS, with one strip having a different depth (Fig. 2a). The strips are 6.6° of visual angle long × 2.7° wide, with either black-and-white dots or black-and-blue dots. Participants have to detect whether the strips with black-and-blue dots are in front or behind the strip with black-and-white dots. Half of the dots are 0.5 degree in diameter, half are 0.1 degree, allowing optimal performance both for typical and amblyopic observers.[30] To increase stereoacuity, dots cannot overlap[31] and their minimal distance is 10 arcmin to prevent crowding.[32] The dots have no coherent motion, because coherent motion decreases stereoacuity.[33] Instead, a new stereogram configuration is generated every 400 ms, thereby avoiding monocular cues. Disparities are introduced as an equal shift of the dot locations in opposite directions between the two eyes. The disparity magnitude varies separately for near and far disparities, following a new Bayesian adaptive method, psi-marg-grid.[21] We extracted a single stereoacuity measure using the weighted sum of the posterior distribution with all thresholds capped at 1300 arcsec.

All participants performed a short training session prior to the test, then started the test with long presentation duration (2000 ms) followed by the test with short presentation duration (200 ms). Long presentations were expected to lead to better (lower) thresholds by allowing time for vergence and eye movements. Participants received meaningful feedback during the training and for the first 12 trials of each test. The total test duration was approximately 30 minutes (five minutes for stereoscope calibration, 15 minutes for 2000 ms presentation and five minutes for 200 ms presentation).

Vivid Vision Stereotest

VividVision (https://www.seevividly.com/) is a virtual reality computer-based application that was developed for assessing and treating different vision problems[34] (i.e. convergence insufficiency, strabismus, lack of binocular vision, or amblyopia) using a virtual-reality headset. The Vivid Vision Stereo test version 2 (VV) presents four filled discs on a slowly moving textured square background (Fig. 2b). The mean position of each disc is horizontally shifted in one eye, creating disparity. Measurable stereoacuities range from 15 arcsec to 2400 arcsec.

Once the participant was used to the headset positioning and the virtual environment, a stimulus was presented for 2000 ms and participants chose which target was in front of the others. To avoid motion parallax, the stimulus was locked to the headset, meaning that the targets and the background moved with head movements. Participants received no feedback on their responses. Thresholds were measured by a VV based on a staircase procedure. As this procedure did not capture non-monotonic psychometric functions found in stereovision[35] and could sometimes lead to nonconvergence of the staircase,[22,36] we extracted a 62% correct threshold from a psychometric function fitted to all the responses in a session, which is a departure from the commercial version. The test itself lasted between two to five minutes. Because of the use of reversals during the staircase procedure the time was dependent on the participant's responses (i.e., in our sample: between 23 to 176 trials, mean = 42.0, SD = 18.9).

ASTEROID Test

ASTEROID v1.0 (Accurate STEREOtest) is a battery of vision tests on an auto-stereogram tablet (Commander 3D). This test includes a disparity detection test presenting a dRDS composed of small and dense colored dots presented on a black background (Fig. 2c).[26] We administered the "stereoacuity standard test" (20 trials). Each trial of the test presents four square boxes where three are flat RDS and the fourth contains a squared shape which appears in depth. Measurable stereoacuities range from 12 arcsec to 1200 arcsec.

Participants' task was to detect the square and they performed the test three times. The 75% correct threshold was computed using a geometric mean of the three scores. We used a chin rest and a stand to rest the tablet at 40 cm distance. As imprecise head position may result in crosstalk between the eyes affecting disparities, if participants had trouble perceiving depth, they were instructed to move the device or their head slightly. As a result, the test may have included both static stereoscopic and motion-in-depth cues. Motion-in-depth perception relies both on stereoscopic cues and on motion cues,[37] which raises the possibility that participants may have used some binocular non-stereoscopic motion cues to pass the test. The stimulus remains on the screen until the participant answers. Following the authors' instructions, three measures were taken from which the geometric mean was extracted. One measure lasted approximately five minutes with the total test duration lasting between 10 to 15 minutes (three measurements).

Statistical Analysis

Data analysis and statistics were carried using Matlab R2018b (version 9.5) and Jamovi version 1.6.18.0. A minimum sample size of 30 was chosen, sufficient to interpret a true correlation above 0.35 as being greater than 0, at a confidence level $P < 0.05$, in each of our groups.[38]

Because the thresholds were not normally distributed, all disparity values were transformed into base-10 logarithms. All statistical tests were bilateral, and effects were considered significant at P values ≤ 0.05.

A large proportion of eRDS scores indicated stereoblindness when the eRDS duration was 200 ms (61.5% at T1, 71.8% at T2). Although a "stereoblindness" result could have diagnostic utility, this short presentation version did not allow us to quantify stereovision with sufficient resolution in individuals with likely weak stereovision for our analysis, so we did not perform any further analysis on those data.

Spearman correlation was used as part of the validity and reliability analyses, and we assessed *agreement, repeatability precision,* and *homogeneity* of the measures through the Bland-Altman method.[39–41] The Bland-Altman plot represents the differences between two measurements against the mean of those two measurements. We considered *agreement* between measures to be good when their mean difference was close to zero. The repeatability of a measure is represented by the 95% limits of agreement (LOA; the ±1.96 SD range of the differences). The *homogeneity* across the range of possible scores—ensuring the differences are independent from the magnitude of the means—was evaluated through Spearman correlations between the differences and the means of two measurements.

Validity analyses were performed at first session for both groups ($n = 64$), and at T2 ($n = 31$) and T3 ($n = 30$) for the eRDS-repeat group. A correlation <0.3 was considered to be weak, between 0.3 and 0.59 to be moderate, and >0.6 to be high.[42] For Bland-Altman analysis, as lower and upper limits of measurable stereoacuity were different for each test, the minimum threshold was capped at 12 arcsec and the maximum at 1200 arcsec when comparing eRDS to ASTEROID, and when comparing VV to ASTEROID, the minimum was capped at 15 arcsec and the maximum at 1200 arcsec. We interpreted significant mean differences as one method overestimating stereoacuity compared to the other method (poor *agreement*). This was estimated with Wilcoxon signed rank tests.

Reliability analyses were conducted on participants who completed the entire protocol ($n = 58$). A test-retest correlation lower than 0.5 was considered as poor, between 0.5 and 0.75, as poor-to-moderate reliability, between 0.75 and 0.9, as good and >0.9, as acceptable for clinical measures.[43] Bland-Altman mean differences between two sessions significantly lower than zero were considered as *learning* (perceptual learning or task familiarity). This was estimated with a non-parametric Friedman analysis of variance (ANOVA) and post-hoc analyses (Durbin-Conover) when the result was significant.

At baseline, there were no significant differences between eRDS-repeat ($n = 28$) and VV-repeat ($n = 30$) groups in mean age (Mann-Whitney $U = 393$, $P = 0.11$) and stereoacuity levels for each of the three tests: eRDS (Mann-Whitney $U = 436$, $P = 0.31$), VV (Mann-Whitney $U = 486$, $P = 0.73$), and ASTEROID (Mann-Whitney $U = 473$, $P = 0.61$). The order in which the tests were given (eRDS first or VV first) did not affect stereoacuities (Mann-Whitney $U_{\text{eRDS}} = 111{,}5$ $P = 0.56$; $U_{\text{VV}} = 89$, $P = 0.15$; $U_{\text{ASTEROID}} = 106$, $P = 0.43$). For further analyses we pooled the different orders and groups together.

Results

Threshold Distribution of the Tests at T1

The eRDS stereotest labeled 15 (23%) participants "stereoblind" at T1 whereas VV scored three as stereoblind (5%) and ASTEROID two (3%). Half of the participants labeled stereoblind on eRDS at T1 had measurable thresholds on this test at T2. Note that only one of these individuals was labeled stereoblind on all tests (eRDS, VV and ASTEROID); this person was excluded from further analyses.

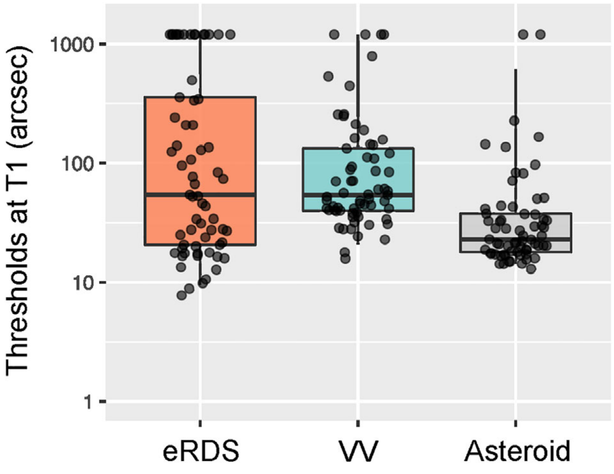

Figure 3. Box plots of stereo-thresholds with median line for eRDS, VV and ASTEROID tests at T1 (n = 65). Each dot is a participant.

eRDS was able to effectively capture 75%-correct thresholds smaller than 10 arcsec as shown by the distribution of stereoacuity thresholds at T1 (Fig. 3), while VV did not capture 62%-correct thresholds under 15 arcsec, as expected by design.

Validity: Comparison With ASTEROID at T1

We found a moderate correlation between eRDS and ASTEROID ($r = 0.46$, $P < .001$, $R^2 = 0.17$; Figure 1a in supplementary S2). We conducted a Bland-Altman analysis and found the LOA to be ±1.31 log unit. eRDS thresholds were significantly higher than ASTEROID thresholds (mean difference of −0.47 log unit, confidence interval [CI; −0.64; −0.31], Wilcoxon signed rank W = 1662, $P < .001$; Fig. 4a). Higher thresholds were also associated with higher ASTEROID-eRDS differences, indicating heterogeneity ($r = 0.70$, $P < .001$).

Comparing VV and ASTEROID scores at T1, we found a weak correlation ($r = 0.23$, $P = 0.07$,

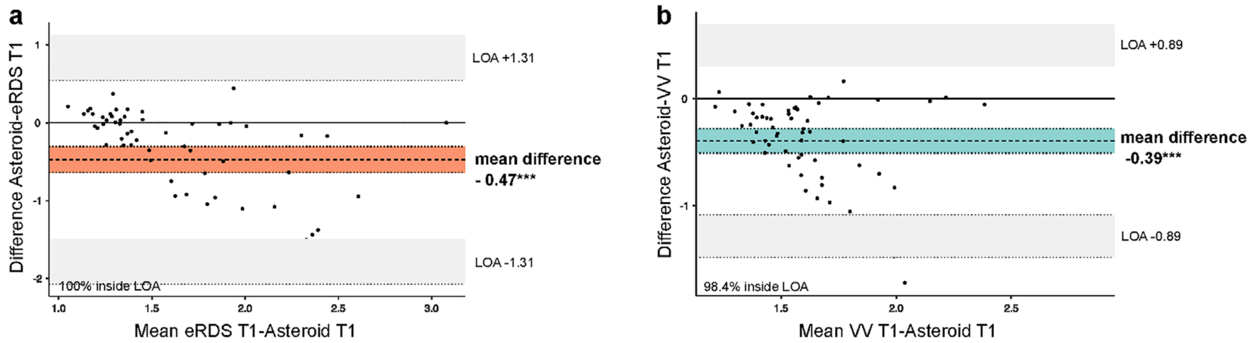

Figure 4. eRDS and VV concurrent validity with ASTEROID. All values are in log10(arcsec); ***$P > .001$; (n = 64) Bland-Altman plots with mean difference and LOA ±1.96 SD (*dotted lines*). Points below difference (*solid line*) represent participants with higher thresholds on eRDS or VV, compared to ASTEROID. The *shaded area* depicts the confidence intervals for the mean difference (*dark gray*), and the upper and lower LOA (*light gray*). Ninety-five percent of observations are expected to be within the LOA (points in CI were considered as inside the limits). (a) eRDS maximum thresholds were capped at 1200 arcsec. (b) VV maximum thresholds were capped at 1200 arcsec and ASTEROID minimum threshold were capped at 15 arcsec.

Table 1. ERDS Test-Retest Reliability (n = 28)

	Spearman Correlations			Measures of Agreement			Durbin-Conover Post-Hoc Analysis	
Sessions	*r df = 26*	*P*	Effect Size (R^2)	Mean Difference	95% CI	LOA	Stat	*P*
T1 vs. T2	0.83	<.001	0.65	−0.28	[−0.44, −0.13]	0.78	2.98	0.004
T2 vs. T3	0.89	<.001	0.90	−0.11	[−0.19, −0.03]	0.40	3.39	<.001
T3 vs. T4	0.90	<.001	0.74	−0.00	[−0.13, 0.14]	0.68	0.09	0.93
T4 vs. T5	0.87	<.001	0.70	−0.06	[−0.21, 0.08]	0.73	0.18	0.86
T5 vs. T6	0.88	<.001	0.80	−0.11	[−0.21, −0.01]	0.49	1.61	0.11

$R^2 = 0.12$; Figure 1c in Supplementary Material S2) and a LOA at ±0.89 log units. VV thresholds were significantly higher than ASTEROID thresholds (mean difference of −0.39 log unit, CI [−0.51; −0.28], Wilcoxon signed rank W = 1928, $P < .001$, Fig. 4b) and ASTEROID-VV differences were homogeneous across the different scores ($r = 0.22$, $P = 0.09$).

Test-Retest Reliability

The correlations between eRDS scores at times T and T+1 were all ≥0.83 (Table 1, Fig. 2 in supplementary information S2), suggesting good test-retest reliability. However, we found a significant effect of session on eRDS scores (repeated-measures Friedman ANOVA $Q_{(5)} = 54.6$, $P < 0.001$), so we conducted a post-hoc analysis. Scores were significantly better at T2 than at T1 (Durbin-Conover = 2.98, $P = 0.003$) and better at T3 than at T2 (Durbin-Conover = 3.39, $P < 0.001$), indicating learning between those sessions (Fig. 5a; see Supplementary Information S2 for Bland-Altman plots). Indeed, stereo-thresholds improved by 62.6% between T1 and T2, and 27.0% between T2 and T3. No learning occurred at the next sessions (all $P \geq 0.13$). The Bland-Altman LOA varied from ±0.4 to ±0.78 log units and score differences between sessions (learning) did not depend on score magnitudes ($r_{(T1-T2)} = -0.09$, $r_{(T2-T3)} = 0.08$, $r_{(T3-T4)} = -0.05$, $r_{(T4-T5)} = -0.20$, $r_{(T5-T6)} = -0.22$; all $P > 0.05$), indicating homogeneity across the possible thresholds.

Test-retest correlations for VV were poor or poor-to-moderate depending on the session (all ≥0.35, Table 2, Fig. 2 in Supplementary Information S2). All correlations were significant except for the correlation between T1 and T2, that was close to significance ($P = 0.06$). We found no significant effect of session on the scores (repeated-measures Friedman ANOVA $Q_{(5)} = 9.61$, $P = 0.09$), indicating an absence of learning (Fig. 5b; see Supplementary Information S2 for Bland-Altman plots). The Bland-Altman LOA varied from ±0.62 to ±1.07 log units (Table 2), with homogeneity of the test-retest differences across scores ($r_{(T1-T2)} = 0.06$, $r_{(T2-T3)} = 0.36$, $r_{(T3-T4)} = -0.15$, $r_{(T4-T5)} = 0.01$, $r_{(T5-T6)} = -0.25$).

Because VV intra-individual scores were highly variable between sessions and no learning was observed between sessions, we considered reducing measurement error by pooling the data from sessions T1–T2–T3 together and T4–T5–T6 together (Fig. 5c). The correlation between those pooled sessions (T1–T2–T3 vs. T4–T5–T6) was good ($r = 0.79$; $P < 0.001$, $R^2 = 0.66$) and no learning between sessions was present (mean difference of −0.01, CI [−0.11; 0.09], Wilcoxon signed rank W = 175, $P = 0.53$; Figure 4 in Supplementary Information S2). The Bland-Altman LOA was reduced (±0.52 log units) and the homogeneity of test-retest differences across scores was preserved ($r = 0.03$, $P = 0.88$).

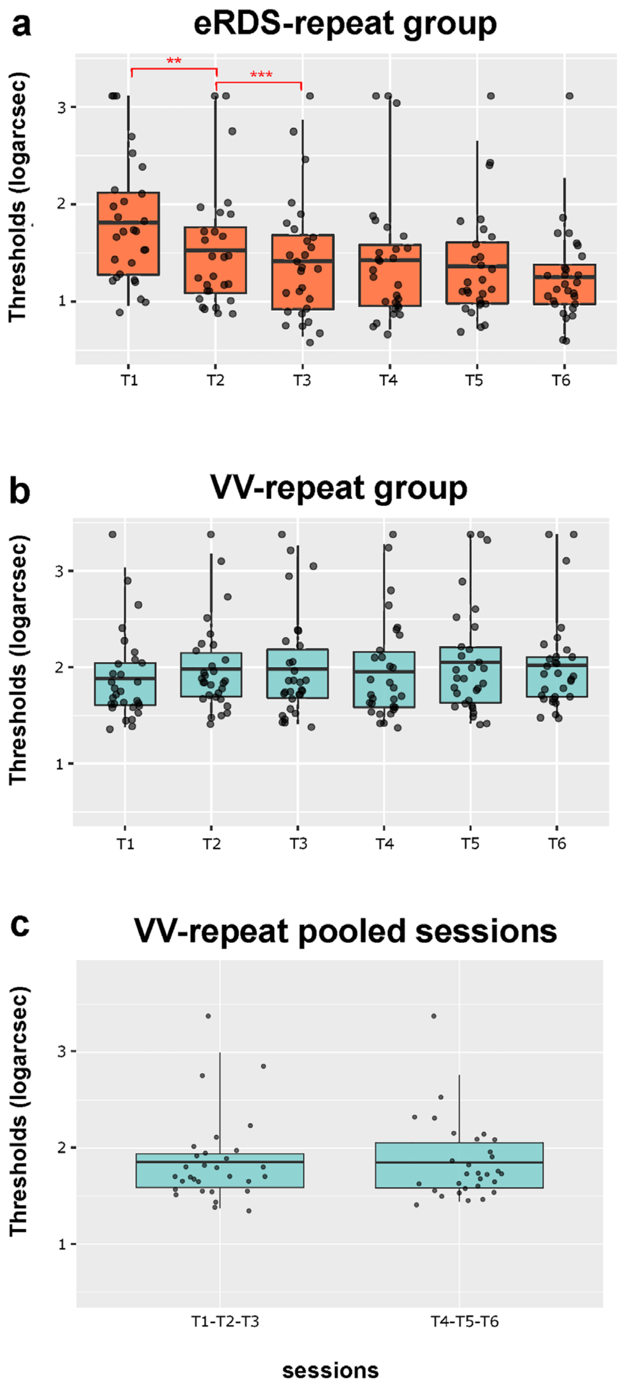

Figure 5. Test-retest reliability: thresholds at each session in boxplots with mean line and 95% CI. All values are in log10(arcsec). **$P < .01$, ***$P < .001$. (a) reliability for eRDS (n = 28). (b) reliability for VV (n = 30). (c) reliability for VV pooled sessions (n = 30).

translational vision science & technology

Table 2. VV Test-Retest Reliability (n = 30)

	Spearman Correlations			Measures of Agreement		
Sessions	*r df* = *28*	*P*	Effect Size (R^2)*	Mean Difference	95% CI	LOA
T1-T2	0.35	0.06	0.31	0.10	[−0.06, 0.26]	0.85
T2-T3	0.59	<.001	0.66	0.00	[−0.12, 0.12]	0.62
T3-T4	0.74	<.001	0.68	−0.03	[−0.15, 0.09]	0.62
T4-T5	0.48	0.007	0.24	0.10	[−0.11, 0.31]	1.07
T5-T6	0.52	0.004	0.50	−0.03	[−0.19, 0.12]	0.81

eRDS Validity Revisited

Because we found significant learning between T1, T2, and T3 with eRDS, the concern that participants might need to learn how to do the eRDS task led us to check our validity analysis on T2 (n = 31) and T3 (n = 30) measures. The correlation between eRDS at T2 and ASTEROID at T1 was higher ($r = 0.58$, p < .001, $R^2 = 0.17$, Fig. 5a in supplementary information, S2), with a LOA reduced to ± 0.89 log unit. The agreement between scores was good with the ASTEROID-eRDS differences not significant (−0.06, CI [−0.23; 0.11], Wilcoxon signed-rank W = 227, p = 0.69; Fig. 5b in supplementary information, S2), and with homogeneity in these differences across scores ($r = 0.27$, p = 0.14). For T3, the correlation with ASTEROID was high ($r = 0.60$, p < .001, $R^2 = 0.33$), and agreement between eRDS and ASTEROID measures remained good (mean difference = 0.01, CI [−0.15; 0.18], not significantly different from zero, W = 181, $P = 0.30$; LOA = ±0.88 log units), with homogeneity of these differences across scores ($r = 0.23$, $P = 0.23$)

Discussion

The purpose of this study was to describe and evaluate the validity, reliability and test-retest learning of two new stereotests. At stake is the issue of being able to measure stereoacuity and determine stereoblindness with valid and reliable tests showing minimal contamination by monocular or binocular non-stereoscopic cues and triggering little learning through multiple testing sessions. It is important to underline that our two new tests and the ASTEROID differed in length and duration (for more details on the tests differences, see the test comparison table in Supplementary Information S3): eRDS had 170 trials (15–20 minutes), VV presented between 28 to 176 trials (2–5 minutes) and ASTEROID had 3 × 20 trials (10–15 minutes).

Validity

At T1, we found that eRDS had moderate concurrent validity, and overestimated thresholds compared to ASTEROID. The overestimation was heterogeneous, with larger differences observed for larger scores. Indeed, eRDS attributed a large number of stereoblindness scores at T1 (13) to participants that were not classified as such by ASTEROID (thresholds between 14.3 arcsec and 136.4 arcsec). eRDS and ASTEROID are both global stereotests but moving the ASTEROID tablet laterally (which is allowed according to the manufacturer's instructions) could have introduced motion-in-depth which is based both on a stereoscopic cue (change of disparity across time) and a non-stereoscopic binocular cue to depth (interocular velocity difference). eRDS is free of these non-stereoscopic cues by design and can therefore be considered as a more "pure" test of stereopsis, which might explain why participants could detect depth at lower disparity levels in ASTEROID. Other possible reasons are discussed in the reliability section.

Because participants improved on eRDS between T1 and T3, we also performed validity analyses at T2 and T3. For both measures, we found good concurrent validity, with good agreement between eRDS and ASTEROID, and a reduced LOA, compared to T1. We note that the T2 and T3 analysis compared thresholds taken on different days: T1 for ASTEROID and T2 or T3 for eRDS. This could be an issue if the underlying ability measured with ASTEROID is not stable across time and indeed, our participants slightly improved their thresholds on ASTEROID between T1 and T6, independent of which group they were assigned to (Wilcoxon signed rank: W = 1332, $P = 0.002$, Supplementary Analysis, S2). This is in line with McCaslin et al.[2] who also reported slight learning between ASTEROID measures taken on different days. Still, the most plausible interpretation of these results is that eRDS at T2 and T3 agreed better than eRDS at T1 with ASTEROID at T1 because of the learning necessary to pass eRDS. We discuss this in the reliability section below.

translational vision science & technology

VV showed weak concurrent validity for global stereopsis. Being a local stereotest, we did not expect a high correlation between VV and ASTEROID, a global stereotest. More generally, the correlation between eRDS and ASTEROID was notably higher than the correlation between VV and ASTEROID. This difference is in line with a study that failed to find significant correlation between local and global stereopsis,[44] suggesting that the ability to perceive global stereopsis (eRDS and ASTEROID) is not well correlated with the ability to perceive local stereopsis (VV). Indeed, some people who cannot see stereodepth in RDSs can see it in other (local) stimuli with binocular disparity.[45–50]

We had expected that the presence of recognizable monocular shapes in VV would lead to better stereoacuity than when measured with RDS.[49,51] However, our data showed the opposite pattern, with VV estimating higher thresholds compared to ASTEROID. This discrepancy could be explained by other differences between those two tests. First, VV presented stimuli for 2000 ms only, while ASTEROID used unlimited presentation, allowing more time for vergence and scanning eye movements. Second, the differences in thresholds may be related to the 3D headset. Informally, we observed that many participants preferred VV to the other two tests, but virtual reality can generate fatigue or visual discomfort related to the conflict between vergence and accommodation distances.[52,53] Also, the pixel size in the headset significantly limits the smallest disparity that can be presented.[39]

Reliability

Although eRDS had good test-retest reliability, learning occurred in the first sessions. The large improvement between T1 and T2 together with the proportion of participants found stereoblind at T1 and yet showing measurable stereopsis in the following sessions (or when using ASTEROID and VV at T1), suggests that the initial practice session was not sufficient for participants to fully understand and accomplish the task at T1. Large and rapid improvements occurring at the start of a task are usually explained by learning of the task and material (task familiarization), while slow and gradual improvement reflects perceptual learning.[54–57] In particular, the observed improvement could be due to learning how to see depth in the stereoscope, e.g., how to cope with the accommodation-vergence conflict in the stereoscope. We minimized the conflict as much as possible by equalizing the accommodation and vergence distances of the screen, but there is still a small conflict when stimuli are presented in front or behind the screen. The stereoscope has been demonstrated to be responsible for a decrease in the initial precision of depth perception that can be improved with training.[58,59] Therefore we believe that the improvement from T1 to T2 was due to task familiarization (e.g., learning how to use the stereoscope). Residual learning was also observed between T2 and T3, the reasons for this later improvement being less clear. eRDS sessions accumulated 340 trials per session and 1020 by the end of T3 (counting trials with 2000 ms and 200 ms presentations). If perceptual learning was at play, it would also have been expected between all subsequent sessions. Yet our post-hoc analyses revealed no improvement between those last sessions, although this result might have been limited by our relatively small sample size. It is possible that the early fast phase of perceptual learning accounts for the progression between T1 and T3, with T4 to T6 being too few trials to allow for the slower phase of perceptual learning to be expressed. Alternatively, additional task familiarization may have occurred between T2 and T3.

Learning effects after multiple testing have been observed frequently in the literature. Gardiner and colleagues,[60] in a paradigm testing the visual field of patients with early glaucoma once a year, found a learning effect, with most of the improvement occurring over the first testing sessions. In another study, McCaslin et al.[2] reported a small learning effect on a third testing session of the ASTEROID test, although they found good test-retest reliability between the first two sessions. However, it has to be noted that their first two sessions were performed on the same day, whereas the third session was taken 14 days later. We underline that in our study, we also observed some learning between T1 and T6 on the ASTEROID test. This result is difficult to interpret, as we observed this improvement in both of our groups, eRDS-repeat and VV-repeat. We therefore cannot exclude the possibility of test-retest learning on ASTEROID. Other computer-based stereotests reported no learning on their retest session.[9–11,61] Looking more in detail, those studies repeated their testing sessions on the same day, which might be a potential explanation for this difference, as sleep can act to consolidate learning.[62,63] However, Tittes and collaborators[12] observed learning in subjects with poor stereopsis, although they repeated their testing on the same day. Literature approaching this issue of multiple testing is very sparse, and studies exploring the reliability of new tests often take two measures on the same day. This does not reflect clinical situations, where patients are tested on different days or months, and underlines the importance for a better understanding of the reliability of the used tests under multiple testing situations.

translational vision science & technology

VV showed no learning between sessions. However, single sessions had poor/poor-to-moderate test-retest reliability, with a test-retest correlation only marginally significant between T1 and T2, and a large test-retest LOA (between 0.62 and 1.08) compared to eRDS (between 0.40 and 0.78) and ASTEROID (0.46). This relatively high variability across sessions can be expected given that each VV test was much shorter than the eRDS test, with just 28 to 176 trials. Pooling 3 sessions together reduced this variability, producing a good test-retest correlation and an improved LOA (0.52). This result must be interpreted with caution because we pooled sessions over different days. That said, given that no learning occurred across the six sessions, the assumption of stability seems valid.

Limitations, Future Directions and Recommendations

The eRDS stereotest could capture low thresholds (under 10 arcsec) which makes it particularly efficient for measuring changes in people with good stereovision. However, we observed substantial learning in the first sessions. Moreover, because of its use of a stereoscope and long duration (15 minutes), eRDS has limited utility with children. The Vivid Vision stereo task is easier to understand and can be rapidly performed, at the cost of lower reliability. To reduce inter-session variability, future studies might consider taking 3 measures of VV stereoacuity and pooling the data. Limitations of this test include its inability to capture threshold under 15 arcsec and the potential presence of binocular non-stereoscopic cues. On the other hand, ASTEROID, which we used as our reference test, is easy to use and captures threshold as low as 12 arcsec. This test is limited by the potential presence of binocular non-stereoscopic cues under the conditions that we used, and by the fact that it relies on the ZEST sampling method, which can be catastrophically sensitive to non-monotonic profiles.[21]

In addition, we did not specifically recruit participants with stereoblindness. More people with stereoblindness are likely to participate in future treatment studies, so a limitation of the current study is that our tests have not yet been fully characterized in this population during treatment.

In this work, we document two new tests and how they relate to a reference test. Therefore our conclusions only cover these three tests. The choice of which test to use, among the two new tests and ASTEROID, may vary depending on the measurement purpose (screening versus quantitative evaluation) and do not preclude the use of other tests. Considering the three tests used in this study, we recommend the following: (1) for testing children, ASTEROID is best suited; (2) for screening patients with compromised stereovision, VV can be considered given that it is quick, and it can be complemented by ASTEROID which targets a different aspect of stereovision (global stereopsis); (3) for precise measures or intervention studies targeting stereovision improvements in adults, eRDS could be advantageous as it captures small changes in stereoacuity and low thresholds, and is not contaminated by any binocular non-stereoscopic cues. eRDS appears best used by administering three measures before training and using T3 as the baseline to minimize test-retest learning effects that may contaminate any improvement due to the intervention. It comes with the burden of long measurement times (around 45 minutes) and the need for a stereoscope. In future versions, that burden could be limited through the use of a tablet version with anaglyph glasses or through a virtual reality version.

Acknowledgments

The authors thank Jennifer Read for providing ASTEROID to them The ASTEROID stereotest is independent research commissioned by the Health Innovation Challenge Fund (HICF-R8-422), a parallel funding partnership between the Wellcome Trust (http://www.wellcome.ac.uk/Funding/Innovations/Awards/Health-Innovation-Challenge-Fund/index.htm) and the Department of Health (http://www.hicfund.org.uk/). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the Department of Health.

Supported by the Chair SILVERSIGHT ANR-18-CHIN-0002, by the IHU FOReSIGHT ANR-18-IAHU-01, by the LabEx LIFESENSES (ANR-10-LABX-65), and the Swiss National Foundation grant number 100014_178814 (D.B.).

Disclosure: **S. Denkinger,** None; **M.-P. Antoniou,** None; **D. Tarello,** None; **D.M. Levi,** None; **B.T. Backus,** Vivid Vision, Inc. (E, I, O); **D. Bavelier,** None; **A. Chopin,** None

References

1. Chopin A, Bavelier D, Levi DM. The prevalence and diagnosis of “stereoblindness” in adults less than 60 years of age: a best evidence synthesis. *Ophthalmic Physiol Opt.* 2019;39:66–85.

2. 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):1–14.
3. Adams WE, Leske DA, Hatt SR, Holmes JM. Defining real change in measures of stereoacuity. *Ophthalmology.* 2009;116:281–285.
4. Chopin A, Chan SW, Guellai B, Bavelier D, Levi DM. Binocular non-stereoscopic cues can deceive clinical tests of stereopsis. *Sci Rep.* 2019;9:1–10.
5. Seitz AR. Perceptual learning. *Curr Biol.* 2017;27(13):R631–R636.
6. Lemay S, Bédard M-A, Rouleau I, Tremblay P-L. Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *Clin Neuropsychol.* 2010;18:284–302.
7. Wilson BA, Watson PC, Baddeley AD, Emslie H, Evans JJ. Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. *J Int Neuropsychol Soc.* 2000;6:469–479.
8. Kim J, Hong J-Y, Hong K, et al. Glasses-free randot stereotest. *J Biomed Opt.* 2015;20(6):1–9.
9. Smith KA, Damarjian AG, Molina A, Arnold RW. Calibrated measurement of acuity, color and stereopsis on a Nintendo 3DS game console. *Clin Optom*. 2019;11:47.
10. Hess RF, Ding R, Clavagnier S, et al. A robust and reliable test to measure stereopsis. *Clin Invest Ophthalmol Vis Sci.* 2016;57:798–804.
11. Portela-Camino JA, Martín-González S, Ruiz-Alcocer J, Illarramendi-Mendicute I, Garrido-Mercado R. An evaluation of the agreement between a computerized stereoscopic game test and the TNO stereoacuity test. *Clin Optom.* 2021;13:181.
12. Tittes J, Baldwin AS, Hess RF, et al. Assessment of stereovision with digital testing in adults and children with normal and impaired binocularity. *Vision Res.* 2019;164:69–82.
13. Zhao L, Wu H. Stereoacuity measurement using an auto-stereoscopic smartphone. *Ann Transl Med.* 2019;7(16):390.
14. Wu H, Jin H, Sun Y, et al.. Evaluating stereoacuity with 3D shutter glasses technology. *BMC Ophthalmol.* 2016;16(1):1–8.
15. Bonfanti S, Gargantini A, Esposito G, et al. Evaluation of stereoacuity with a digital mobile application. *Graefes Arch Clin Exp Ophthalmol*. 2021;1:3.
16. Liu F, Zhao J, Han T, et al. Screening for stereopsis using an eye-tracking glasses-free display in adults: a pilot study. *Front Med.* 2022;8: 814908.
17. O'toole A, O'toole AJ, Kersten DJ. Learning to see random-dot stereograms. *Perception.* 1992;21:227–243.
18. Ramachandran VS. Learning-like phenomena in stereopsis. *Proc R Soc.* 1976;262:382–384.
19. Tittes J, Baldwin AS, Hess RF, et al. Assessment of stereovision with digital testing in adults and children with normal and impaired binocularity. *Vision Res.* 2019;164:69–82.
20. Schmitt C, Kromeier M, Bach M, Kommerell G. Interindividual variability of learning in stereoacuity. *Arch Clin Exp Ophthalmol.* 2002;240:704–709.
21. Chopin A. Adaptive methods to quickly estimate psychometric functions: the case of Psimarg-grid and the effect of non-monotony. *J Vis*. 2022;22:3302.
22. García-Pérez MA. Adaptive psychophysical methods for nonmonotonic psychometric functions. *Atten Percept Psychophys.* 2014;76:621–641.
23. Breitmeyer B, Julesz B, Kropfl W. Dynamic random-dot stereograms reveal up-down anisotropy and left-right isotropy between cortical hemifields. *Science*. 1975;187(4173):269–270.
24. Chopin A, Chan SW, Guellai B, Bavelier D, Levi DM. Binocular non-stereoscopic cues can deceive clinical tests of stereopsis. *Sci Rep.* 2019;9(1):1–10.
25. Vancleef K, Serrano-Pedraza I, Sharp C, et al. ASTEROID: a new clinical stereotest on an Autostereo 3D Tablet. *Transl Vis Sci Technol.* 2019;8(1):25.
26. Read JCA, Wong ZY, Yek X, et al. ASTEROID stereotest v1.0: lower stereo thresholds using smaller, denser and faster dots. *Ophthalmic Physiol Opt*. 2020;40:815–827.
27. Serrano-Pedraza I, Vancleef K, Read JCA. Avoiding monocular artifacts in clinical stereotests presented on column-interleaved digital stereoscopic displays. *J Vis.* 2016;16(14):1–14.
28. Myles PS, Cui J. I. Using the Bland-Altman method to measure agreement with repeated measures. *Br J Anaesth.* 2007;99:309–311.
29. Chopin A, Silver MA, Sheynin Y, Ding J, Levi DM. Transfer of perceptual learning from local stereopsis to global stereopsis in adults with amblyopia: a preliminary study. *Front Neurosci.* 2021;15:719120.
30. Ding J, Levi DM. Recovery of stereopsis through perceptual learning in human adults with abnormal binocular vision. *Proc Natl Acad Sci.* 2011;108(37):E733–E741.
31. Read JCA, Cumming BG. A stimulus artefact undermines the evidence for independent ON and OFF channels in stereopsis. *BioRxiv.* 2018:295618.

32. Westheimer G, McKee S. Stereogram design for testing local stereopsis. *Invest Ophthalmol Vis Sci.* 1980;19:802–809.
33. Hadani I, Vardi N. Stereopsis impairment in apparently moving random dot patterns. *Percept Psychophys.* 1987;42:158–165.
34. Backus BT, Dornbos BD, Tran TA, Blaha JB, Gupta MZ. Use of virtual reality to assess and treat weakness in human stereoscopic vision. *J Electron Imaging.* 2018;2018(4):109-1–109-6.
35. Kane D, Guan P, Banks MS. The limits of human stereopsis in space and time. *J Neurosci.* 2014;34:1397–1408.
36. Levitt H. Transformed up-down methods in psychoacoustics. *J Acoust Soc Am.* 1971;49:467.
37. Nefs HT, O'Hare L, Harris JM. Two independent mechanisms for motion-in-depth perception: Evidence from individual differences. *Front Psychol.* 2010;1:155.
38. Bonett DG, Wright TA. Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika.* 2000;65(1):23–28.
39. Berchtold A. Test–retest: agreement or reliability? *Method Innov.* 2016;9:1–7.
40. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet.* 1995;346(8982):1085–1087.
41. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet.* 1986;327(8476): 307–310.
42. Andresen EM. Criteria for assessing the tools of disability outcomes research. *Arch Phys Med Rehabil.* 2000;81:S15–S20.
43. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, NJ: FA Davis Co.;2009.
44. Gantz L, Bedell HE. Transfer of perceptual learning of depth discrimination between local and global stereograms. *Vision Res.* 2010;50(18):1891–1899.
45. Hess RF, Mansouri B, Thompson B, Gheorghiu E. Latent stereopsis for motion in depth in strabismic amblyopia. *Invest Ophthalmol Vis Sci.* 2009;50:5006–5016.
46. Slreteanu R. Binocular vision in strabismic humans with alternating fixation. *Vision Res.* 1982;22:889–896.
47. McColl SL, Ziegler L, Hess RF. Stereodeficient subjects demonstrate non-linear stereopsis. *Vision Res.* 2000;40(9):1167–1177.
48. Kiraoji H, Toyamat K, Kitaoji H. Preservation of position and motion stereopsis in strabismic subjects. *Invest Ophthalmol Vis Sci.* 1987;28(8):1260–1267.
49. Simons K. A comparison of the frisby, random-dot E, TNO, and Randot circles stereotests in screening and office use. *Arch Ophtalmol.* 1981;99:446–452.
50. Giaschi D, Lo R, Narasimhan S, Lyons C, Wilcox LM. Sparing of coarse stereopsis in stereodeficient children with a history of amblyopia. *J Vis.* 2013;13(10):17–17.
51. Yildirim C, Altinsoy HI, Yakut E. Distance stereoacuity norms for the mentor B-VAT II-SG video acuity tester in young children and young adults. *J AAPOS.* 1998;2:26–32.
52. Lambooij M, IJsselsteijn W, Fortuin M, Heynderickx I. Visual discomfort and visual fatigue of stereoscopic displays: a review. *J Imaging Sci Technol.* 2009;53(3):030201–1–030201–14.
53. Banks MS, Kim J, Shibata T. Insight into vergence/accommodation mismatch. *Proc SPIE Int Soc Opt Eng.* 2013;8735:59–70.
54. Green CS, Banai K, Lu Z-L, Bavelier D. Perceptual learning. In: *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. Hoboken, NJ: John Wiley & Sons. 2018;1–47.
55. Ahissar M, Hochstein S. The reverse hierarchy theory of visual perceptual learning. *Trends Cogn Sci.* 2004;10:10.1117/12.2019866.
56. Fine I, Jacobs RA. Comparing perceptual learning tasks: a review. *J Vis.* 2002;2:190–203.
57. Zhang JY, Zhang GL, Xiao LQ, Klein SA, Levi DM, Yu C. Rule-based learning explains visual perceptual learning and its specificity and transfer. *J Neurosci.* 2010;30:12323–12328.
58. Hartle B, Wilcox LM. Depth magnitude from stereopsis: assessment techniques and the role of experience. *Vision Res.* 2016;125:64–75.
59. McKee SP, Taylor DG. The precision of binocular and monocular depth judgments in natural settings. *J Vis.* 2010;10(10):5
60. Gardiner SK, Demirel S, Johnson CA, Gardiner S. Is there evidence for continued learning over multiple years in perimetry? *Optom Vis Sci.* 2008;85:1043–1048.
61. Kim J, Yang HK, Kim Y, Lee B, Hwang JM. Distance stereotest using a 3-dimensional monitor for adult subjects. *Am J Ophthalmol.* 2011;151:1081–1086.e1.
62. Sasaki Y, Nanez JE, Watanabe T. Advances in visual perceptual learning and plasticity. *Nat Rev Neurosci.* 2009;11:53–60.
63. Lu ZL, Yu C, Watanabe T, Sagi D, Levi D. Perceptual learning: functions, mechanisms, and applications. *Vision Res.* 2009;49(21):2531–2534.