Vision Research 164 (2019) 69–82

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

# Assessment of stereovision with digital testing in adults and children with normal and impaired binocularity

Juliane Tittes[a], Alex S. Baldwin[b], Robert F. Hess[b], Licia Cirina[a], Yaroslava Wenner[a], Claudia Kuhli-Hattenbach[a], Hanns Ackermann[c], Thomas Kohnen[a], Maria Fronius[a,*]

[a] Department of Ophthalmology, Child Vision Research Unit, Goethe University, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany
[b] McGill Vision Research Unit, Department of Ophthalmology, 1650 Ave Cedar, L11.403 Montreal, Canada
[c] Institute of Biostatistics, Goethe University, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany

ARTICLE INFO

*Keywords:*
Stereovision
Amblyopia
Visual development
Binocular vision
Digital stereo test
Strabismus

ABSTRACT

New digital approaches allow stereovision to be assessed with greater precision than current clinical stereo tests. Those current tests present a relatively narrow range of stimulus disparities in coarsely quantized steps. With dichoptic treatments for amblyopia emerging, more accurate assessment of especially coarse stereopsis becomes increasingly important for verifying their aim to improve 3D vision. We used digital testing in subjects of a large age range (4–59 years), with groups having both normal (n = 34) and impaired binocular vision due to unilateral amblyopia, with or without strabismus (n = 27). Random-dot stimuli were presented on a 3D monitor with shutter glasses. The test applies adaptive procedures to measure psychometric functions and provides thresholds with associated confidence intervals. Digital thresholds for controls (range 11–160 arcsec) and stereodeficient subjects (range 43–911 arcsec) were compared to the TNO, a standard clinical test which uses similar random-dot targets presented with anaglyph glasses. Agreement between digital and TNO thresholds varied with the level of stereopsis. Stereoacuity was measurable in several subjects who failed on the TNO. With the digital test we found good repeatability for both groups, with the indication of a small learning effect for subjects with coarse stereopsis. Thus, assessment of all target groups for new tests is important, and repeated testing before therapy may avoid confusing learning and treatment effects. Our digital approach supplies a large range of accurate stereo data in children and adults; together with its associated measure of variability, it will be useful in longitudinal treatment studies.

## 1. Introduction

The depth of a stimulus can be calculated through exploiting the two slightly different projections formed on the two retinae (binocular horizontal disparity). In humans this extraction of depth information from stereoscopic disparity involves the occipital striate cortex. This process relies on ocular alignment and the quality of the ocular images in each eye (Fortin, Ptito, Faubert, & Ptito, 2002). Failure to fulfill these requirements during visual maturation causes the development of amblyopia ("lazy eye"). This neuro-developmental disorder leads to deficits in visual acuity, position acuity, contrast sensitivity and also the perception of depth (Levi, 2006; Sireteanu, 2000). The most common causes of amblyopia are refractive errors, strabismus and sensory deprivation. Abnormal binocular interaction (chronic interocular suppression) is considered to be the mechanism through which vision is impaired (Hess, Thompson, & Baker, 2014; Kehrein, Kohnen, & Fronius, 2016; Sireteanu & Fronius, 1981).

The standard therapy for amblyopia is occlusion of the non-amblyopic eye. This monocular treatment is administered in children, with the aim of restoring function in the amblyopic eye (Fronius, 2016). However besides being distressing to the child, the concern was raised that it could reduce binocular vision abilities, e.g. stereopsis (Astle, McGraw, & Webb, 2011a; Stewart, Wallace, Stephens, Fielder, & Moseley, 2013). Recently, new therapies have been advocated. These aim to restore visual acuity, while also improving binocular vision by reducing interocular suppression and facilitating binocular fusion (Hess, Mansouri, & Thompson, 2010; Vedamurthy et al., 2015). Results regarding improvements in stereoacuity have been inconsistent. Hess

* Corresponding author at: Ophthalmology Department, Goethe University Hospital, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany.
*E-mail addresses:* tittes@med.uni-frankfurt.de (J. Tittes), alexander.baldwin@mail.mcgill.ca (A.S. Baldwin), robert.hess@mcgill.ca (R.F. Hess), YaroslavaC@gmx.de (Y. Wenner), hattenbach@med.uni-frankfurt.de (C. Kuhli-Hattenbach), h.ackermann@add.uni-frankfurt.de (H. Ackermann), kohnen@em.uni-frankfurt.de (T. Kohnen), fronius@em.uni-frankfurt.de (M. Fronius).

https://doi.org/10.1016/j.visres.2019.07.006
Received 2 August 2018; Received in revised form 10 July 2019; Accepted 24 July 2019
Available online 04 September 2019
0042-6989/ 

et al. (2010, 2012) and Long To et al. (2011) found significant improvement of stereoacuity even in amblyopic adults, whereas Birch et al. (2015) did not find significant improvement in amblyopic children whose visual acuity improved. Birch et al. (2015) ascribed these differences to the different clinical stereo tests that were used: random-dot based tests (Birch et al., 2015), vs. contour or hybrid tests (Hess et al., 2010). However, in a new study, the Birch group (Kelly et al., 2018) found an improvement in stereoacuity with the Randot Preschool Stereoacuity and Stereo Butterfly Test (Stereo Optical, Inc.) after treatment with new dichoptic methods.

Clinical stereo tests allow quick assessment even in young children and are adequate for general screening (Tomaç & Altay, 2000). However they have several shortcomings: some have monocular cues; in tests with fine features, thresholds may be limited by visual acuity (Simons, 1996); stereoacuity is tested stepwise from coarse to fine disparities, usually in large steps, with threshold estimates often based on a single judgement (Bömer, Dölp, & Kommerell, 1995). The clinical tests neither provide a measure of variability nor do they allow quantification of very fine or coarse disparities (Serrano-Pedraza, Vancleef, & Read, 2016; Vancleef et al., 2017), which would be required e.g. for studies comparing treatment effects.

To overcome these limitations, methods have been developed using modern display technology, often applying adaptive staircase procedures (Bach, Schmitt, Kromeier, & Kommerell, 2001; Foley & Tyler, 1976; Giaschi, Lo, Narasimhan, Lyons, & Wilcox, 2013; Giaschi, Narasimhan, Solski, Harrison, & Wilcox, 2013; Greenwood et al., 2012; Lindblom & Frisen, 1988; Serrano-Pedraza et al., 2016a; Tidbury, Brooks, O'Connor, & Wuerger, 2016). Hess et al. (2016) developed a method for measuring stereoacuity in adults using a handheld device with red-green anaglyph glasses. Participants had to decide which of two random-dot based disks was in front or behind the plane of the screen, thereby requiring the polarity of the depths to be calculated. Thresholds and standard errors were calculated by means of an adaptive staircase procedure, providing a measure of variability. Adults with presumed normal vision were tested. Their results were widely distributed, reaching outside of what is considered normal. Similar results were reported by Schmitt, Kromeier, Bach, and Kommerell (2002) in a stereo test involving the discrimination of depth by deciding whether a stimulus was perceived in front or behind a screen. Red-green filters were shown to reduce stereoacuity due to binocular color rivalry, affecting the accuracy of obtained stereo thresholds (Cornforth, Johnson, Kohl, & Roth, 1987). The use of frame-interleaving displays with synchronized shutter glasses should provide a better alternative for testing stereovision (Lindblom & Frisen, 1988).

Similarly to testing amblyopes of a large age range with a uniform visual acuity testing procedure (Fronius, Cirina, Ackermann, Kohnen, & Diehl, 2014), we were interested in assessing stereo vision digitally from child to adult with the same test, and in observers with both normal and impaired binocular vision. We used a stereo test developed at the McGill Vision Research unit which displayed random-dot stimuli on a 3D monitor. It provides a measure of variability, and assesses a larger range of quantifiable disparities than commonly available clinical tests. Compared to the previous stereo test by Hess et al. (2016), this test has some important modifications (see also Methods section). It is administered with shutter glasses, thereby avoiding the problem of binocular color rivalry. There is no critical dependence on normal visual acuity, as dot size, stimulus size and screen size are larger. Instead of more difficult judgment of polarity of depth, which could have led to the wide distribution of results, it uses a 4-alternative forced choice discrimination task with crossed disparities. Furthermore it has features which may motivate successful participation of children or subjects with little or no stereovision. Lastly, it allows a direct comparison with the TNO (Laméris Ootech), one of the clinical gold standard stereo tests, using a closely similar random-dot stimulus. The TNO uses red-green anaglyph glasses; assessment is quick, with no monocular cues and a low probability of false negatives.

Measurements were done in the Child Vision Research Unit in Frankfurt, Germany, involving participants of a large age range, with both (corrected-to) normal vision and stereodeficient subjects due to unilateral amblyopia. In our study, we investigated the test-retest reliability of our measurements. We studied the correlation of our results with those of the TNO. Furthermore we were interested in the correlation of stereoacuity thresholds with age in participants with (corrected-to) normal vision, and with interocular visual acuity difference in amblyopes. As young children (< 7 years) are the main target group for amblyopia therapy, we assessed if children from the age of 4 would be able to complete our measurements.

## 2. Methods

### 2.1. Participants

We tested 34 control subjects with normal or corrected-to normal vision and 27 subjects with impaired binocular vision due to unilateral amblyopia (referred to as "stereodeficient"). According to the literature, stereovision may not fully develop until early adulthood. Based on this, we decided to categorize participants up to the age of 15 years as "children" and those older than 15 as "adults". For adults in the control group, ages ranged from 22 to 49 years (mean 27.0 years). For children in the control group the range was 4 to 13 years (mean 7.5 years). In the stereodeficient group, 14 adults (5 anisometropic, 2 strabismic and 7 combined amblyopes) and 13 children (9 anisometropic, 1 strabismic and 3 combined amblyopes) were tested (Table 1). The stereodeficient adults' ages ranged from 16 to 59 years (mean 29.1 years), and children's ages ranged from 5 to 12 years (mean 8.5 years).

Amblyopia is usually defined as visual acuity difference of 0.2 logMAR between the eyes. Participants who did not fulfill the criterion of having a visual acuity difference of 0.2 logMAR had a history of amblyopia and previous occlusion treatment (see Table 1). For our analyses, we classified the stereodeficient subjects into a group "without strabismus" (anisometropic amblyopes, with a refractive interocular difference of > 1D sph and/or > 1.5D cyl without heterotropia) and a group "with strabismus" (strabismic and combined amblyopes). One of the subjects (no 5 in Table 1) had a fully accommodative esotropia and hence orthophoria when wearing the prescribed glasses. As this subject wore his glasses since the age of 1 and all tests were done with his correction, we categorized him as "without strabismus". Although it is assumed that the maximum angle of strabismus consistent with true stereopsis is 4PD (Leske & Holmes, 2004), we included a few patients with larger angles for two reasons. One reason was that our stimuli were designed to enable quantification of larger disparities than clinical stereo tests. The other reason was that we wanted to check for hidden monocular cues in the digital test stimuli which might not be noticed when testing subjects capable of seeing stereo.

On entering the study, participants received a comprehensive ophthalmological and orthoptic examination (for details see Fronius et al., 2014), including: examination of the anterior and posterior segment of the eye, assessment of objective refraction, visual acuity, eye movements and eye alignment (cover test). Furthermore, the Bagolini striated glasses and stereo tests (Titmus, Stereo Optical Co., Inc.; Randot, Stereo Optical Co., Inc.; TNO) were administered. The pattern of fixation was determined with the Cüppers visuscope test. The participants were recruited from: the outpatient clinic of the Department of Ophthalmology, Goethe University, Frankfurt; several local ophthalmologists' offices; social network pages; local primary schools or students. Exclusion criteria were: deprivation amblyopia, any ocular morphological disorder, and impaired visual acuity due to medication, brain damage, trauma or neurological disorders. Control subjects were not included when having a refractive error exceeding ± 2.5D sph or ± 1.5D cyl or anisometropia of > 1D sph/cyl.

The research adhered to the Code of Ethics of the World Medical

**Table 1**
Baseline characteristics of stereodeficient subjects.

| Patient No. | Age [years] | Type of amblyopia<br>A (Anisometropic)<br>S (Strabismic)<br>C (Combined) | Eye | Refraction [D] | Angle of squint (near) [PD] | Visual acuity crowded Landolt [log MAR] | History: Occlusion [age]<br>Surgery [at the age of]<br>Family history |
|---|---|---|---|---|---|---|---|
| 1 | 5.5 | C† | RE | +1.0 −0.75/173° | Micro | 0.1 | Untreated |
| | | | LE* | +4.75 −0.75/2° | | 0.7 | |
| 2 | 5.7 | A† | RE* | −6.0 −1.5/35° | ± 0 | 0.4 | Occlusion 5 y till present; Family history |
| | | | LE | +1.5 | | 0.2 | |
| 3 | 5.9 | A | RE | +1.0 −0.5/176° | ± 0 | 0.7‡ | Untreated |
| | | | LE* | +2.75 −1.0/67 | | 0.7‡ | |
| 4 | 6.5 | A | RE | +1.25 −0.75/175° | ± 0 | 0.1 | Occlusion some hours; age unknown |
| | | | LE* | +3.0 −0.5/0° | | 0.5 | |
| 5 | 7.3 | A | RE* | +7.0 −2.5/10° | ± 0 | 0.5 | Occlusion 4 y till present |
| | | | LE | +5.5 −1.5/175° | | 0.1 | |
| 6 | 7.4 | A | RE | +2.0 −2.25/5° | ± 0 | 0.1 | Occlusion 2.5 y till present |
| | | | LE* | −4.25 −0.5/155° | | 0.2 | |
| 7 | 7.6 | A | RE | +2.5 −0.75/0° | ± 0 | 0.2 | Untreated |
| | | | LE* | +5.5 −3.0/150° | | 0.6 | |
| 8 | 8.3 | C† | RE | +3.75 −1.25/180° | +2 | −0.1 | Occlusion 4.8 y till present |
| | | | LE* | +5.25 −1.25/175° | | 0.2 | |
| 9 | 9.7 | A† | RE | +1.5 −0.5/5° | ± 0 | 0.1 | Untreated |
| | | | LE* | +7.0 −0.5/5° | | 0.9 | |
| 10 | 10.7 | C† | RE* | +6.75 −0.75/166° | Micro | 0.8 | Occlusion 4 y till present |
| | | | LE | +5.0 −0.5/171° | | −0.1 | |
| 11 | 11.9 | A | RE | +2.0 | ± 0 | 0.1 | Occlusion |
| | | | LE* | +3.5 | | 0.2 | 6–8 y |
| 12 | 12.2 | A | RE* | +5.0 1.5/160° | ± 0 | 0.7 | Untreated |
| | | | LE | +0.5 −0.5/5° | | −0.1 | |
| 13 | 12.5 | S | RE* | +5.25 −1.75/7° | +2 | 0.6 | Occlusion 3–9 y |
| | | | LE | +4.75 −1.75/174° | | 0.1 | |
| 14 | 16.0 | A | RE* | +5.00 −1.0/60° | ± 0 | 0.5 | Untreated |
| | | | LE | +1.75 | | −0.1 | |
| 15 | 16.5 | A | RE* | +2.25 −3.5/166° | ± 0 | 0.7 | Occlusion 9–12 y (patch and Bangerter filters) |
| | | | LE | 0.00 −0.25/35° | | −0.1 | |
| 16 | 20.3 | C† | RE* | +5.75 −2.75/164° | +1 | 0.5 | Occlusion 6–14 y; Family history |
| | | | LE | +0.25 | | −0.1 | |
| 17 | 23.9 | C | RE | 0.00 −0.25/19° | +1 | −0.1 | Occlusion 4–6 y; Family history |
| | | | LE* | +3.75 −1.75/174° | | 0.6 | |
| 18 | 24 | C | RE* | +1.0 −1.5/165° | +1 | −0.1 | Occlusion 3–10 y; Surgery at 6 y; Family history |
| | | | LE | −1.5 −1.5/175° | | −0.1 | |
| 19 | 24 | C | RE | −0.25 −0.5/177° | +1 | −0.1 | Occlusion 4–12 y; Family history |
| | | | LE* | +3.0 −1.0/4° | | 0.5 | |
| 20 | 26.2 | A | RE | −2.25 −1.25/178° | ± 0 | −0.1 | Occlusion 14–15.7 y |
| | | | LE* | −2.75 −1.75/178° | | 0.1 | |
| 21 | 26.7 | C† | RE* | +0.25 −1.75/162° | RE'fix −9 −VD3 | 0.0 | Occlusion 3–6 y; Surgeries at 3 and 5 y; Family history |
| | | | LE | +0.25 −0.25/172° | LE'fix −10 | 0.0 | |
| 22 | 26.9 | A | RE | −0.25 | ± 0 | −0.1 | Occlusion 5–6 y (Bangerter filters) |
| | | | LE* | +0.75 −0.75/80° | 0.1 | | |
| 23 | 27.7 | S† | RE* | −0.75 −0.75/2° | +1 | 0.8 | Occlusion 6–6.5 y |
| | | | LE | −0.25 −0.5/94° | | −0.1 | |

(*continued on next page*)

**Table 1** (*continued*)

| Patient No. | Age [years] | Type of amblyopia<br>A (Anisometropic)<br>S (Strabismic)<br>C (Combined) | Eye | Refraction [D] | Angle of squint (near) [PD] | Visual acuity crowded Landolt [log MAR] | History: Occlusion [age]<br>Surgery [at the age of]<br>Family history |
|---|---|---|---|---|---|---|---|
| 24 | 31.7 | A | RE | −1.25 −0.25/34° | ± 0 | 0.0 | Untreated; Family history |
| | | | LE* | +7.25 −2.0/46° | | 1.0 | |
| 25 | 33.9 | S | RE | −0.25 −1.0/91° | +2 −VD1 | 0.0 | Occlusion 3–12 y; Family history |
| | | | LE* | +0.25 −0.75/112° | | 0.0 | |
| 26 | 51 | C† | RE | −0.5 −1.25/109° | +10 +VD | −0.1 | Occlusion pre-scribed, not done reliably; Surgeries at 3.5 and 32 y; Family history |
| | | | LE* | +3.0 −1.0/58° | | 0.4 | |
| 27 | 59 | C† | RE | +1.25 −0.75/141° | +10 −VD8 | −0.1 | Untreated; Family history |
| | | | LE* | +2.75 −0.25/175° | | 1.3 | |

D = diopter, PD = prism diopter (manifest angle). The asterisk marks the amblyopic eye. RE = right eye; LE = left eye; VD = vertical deviation, NAE = nonamblyopic eye; AE = amblyopic eye.
†No stereovision.
‡Single optotypes: RE 0.2 logMAR; LE 0.4 logMAR.




**Fig. 1.** Digital test pie-shaped stimulus (A) binocularly visible figure and (B) monocularly visible figure. Both can be seen with red-green anaglyph glasses (right eye: green) in the online version, shown for demonstration purposes, not used in the experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Association (Declaration of Helsinki). The Ethics Committee of the University of Frankfurt approved the study protocol prior to initiation of the study. Written informed consent was obtained from all parents and from participants older than 7 years, and all participants gave assent, prior to inclusion in the study after explanation of the nature and possible consequences of the study.

### 2.2. Apparatus and stimulus

For our study, two different methods for testing stereovision were being used: our random-dot stimuli on a 3D screen with electronic shutter glasses and the TNO book test with anaglyph glasses. Both use a similar random-dot based “pie-shaped” stimulus. Diameter of the circular TNO stimulus is 8.5° and dot size < 480 to 1500 arcsec. Our digital stimuli (Fig. 1) were shown on an ACER GN246HL LCD screen using the ACER GN246HL 3D system with NVIDIA Vision 2 wireless shutter glasses and an infrared transmitter to achieve stereoscopic presentation. The mean background luminance was 101 cd/m$^2$ and mean illuminance 549 lx (measured with a Konica Minolta Illuminance Meter T-10 and a Luminance Meter LS-150). The temporal resolution was 120 Hz, and the spatial resolution was 1920 by 1080 pixels. Subjects viewed the stimuli from a distance of 0.76 m on a 24″ display. At this distance, the screen provided 48 pixels per degree of visual angle. Stimulus diameter was 10° and dot size 10 arc min (“effective size”). A binocularly presented black and white striped 16.5 degree of visual angle frame around the target stimulus was added to aid fusion. The

random striped pattern on the frame was recalculated on each trial. The experiment was programmed in Matlab R2017b using Psychtoolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). To improve the accuracy of sub-pixel disparities, a gamma correction was performed with a Konica Minolta Luminance Meter LS-150.

The stimuli were defined by disparity modulations of fields of spatially bandpass dots. They were generated from an isotropic log-Gabor function with a spatial frequency of 0.5c/deg and a bandwidth of 2.4 octaves. They have the appearance of a central light blob with a darkened surround. Presentation was at 80% Michelson contrast. The dots were initially arranged in a grid with an average peak-to-peak spacing between adjacent dots of 1575 arcsec. The actual position of each dot was randomly sampled from a uniform distribution with range of ± 675 arcsec both horizontally and vertically about their original grid location. This broke up any impression of a regular grid structure.

A pie-shape was introduced by modulating the disparity of the dots. A 45°angle "pie slice" was removed from one side of the stimulus (the top, bottom, left, or right side). The location of the missing piece defined the orientation of the target. The dots falling within the pie shape were given a crossed disparity, and those outside the shape were given an equal uncrossed disparity. The disparity was introduced by shifting the position of the dots in the left and right eye images. Where the disparity required a shift that was not an integer number of pixels the remainder was accounted for using subpixel interpolation. See the Appendix for further details regarding stimulus generation and properties.

In an earlier version, at large disparities (higher than 1000 arcsec) the shift in the position of the dots for the foreground and background could introduce monocularly-detectable gaps in the stimuli. In the final version of the experiment these gaps were filled in with zero-disparity dots (i.e. halfway between the foreground and background) after stimulus generation. We removed 10 subjects tested with the earlier version who had disparity thresholds over 1000 arcsec from the analysis and marked them as not achieving a testable disparity on the digital test. Five of them obtained an apparent valid psychometric function with a threshold slightly over 1000 arcsec. Four of them were available to be tested again with the newer version after noticing the monocular cues. With the new version, none of those subjects obtained a valid psychometric function. This indicates that they were making use of the monocular cues in the earlier version of the task.

### 2.3. Procedure

The examination always took place in the same room with the same normal interior light conditions, for the entire duration of the experiment and examination. Subjects performed a four-alternative forced choice task discriminating between the four possible orientations of the target. The disparity was controlled by a pair of interleaved staircase procedures (Baldwin, 2019, July 2). One staircase was 1-up-1-down, the other 2-up-1-down. Each staircase terminated after 70 trials or 9 reversals, whichever was reached first. The median number of trials was 52. The staircase traversed a list of disparities between 1 and 4096 arcsec, in ratio steps of √2. Before the first reversal the step size was doubled (ratio steps of 2).

Thresholds were obtained through maximum-likelihood fitting of a cumulative normal function in Palamedes (Prins & Kingdom, 2018). The guess rate was set to 25%, for our 4AFC task, and we assumed a fixed lapse rate of 2%. Ignoring the lapse rate, thresholds were calculated at the 62.5% correct point. Estimates of error were calculated from parametric bootstrapping (250 samples) using the PAL_PFML_BootstrapParametric function in Palamedes. The standard error was calculated as the standard deviation of the bootstrap sample thresholds. On occasion the psychometric function fitting in the bootstrap routine will return extreme values for the thresholds (as a result of a failure in the fitting). These outliers will have a large effect on the calculation of the standard deviation. To avoid this, before taking the standard deviation we apply a more conservative version of Tukey's outlier-removal algorithm (Emerson & Strenio, 1983; Tukey, 1977). We calculate the interquartile range of the bootstrap thresholds and remove any values that are more than three interquartile ranges below the first quartile or more than three interquartile ranges above the third quartile.

The initial offset was set to 800 arcsec for control subjects and 1200 arcsec for amblyopes and all young children. To ensure that participants understood the task and as encouragement and support for amblyopes and children, 5 monocularly discriminable stimuli were presented before the beginning of the stereo task (Fig. 1B). These declined from a contrast cue of 100%, where all background dots are gone, to a contrast cue of 50%, where background dots are at half the contrast of the target, in 5 steps (100%, 87.5%, 75.0%, 62.5%, 50.0%). During the subsequent experiment, every trial had a 5% chance to have the monocularly-visible contrast cue. Data from these trials were not included in measurements of subjects' stereo threshold.

Participants were seated in front of the monitor with their chin resting on a chin rest. Young children sat on one of their parent's laps, while they held their head and the shutter glasses. Participants wore their optical correction underneath the shutter glasses. Older participants had to indicate the direction of the gap by voice whereas children could engage in a game. At each possible direction, a figure (robot, monkey, girl, star) was placed at each side of the screen, and children had to state who of the four "ate the slice". In case they did not see any stimulus, they could decide who of the four "ate the slice", to facilitate forced choice. To ensure that they were paying attention, the experimenter occasionally was wearing another pair of shutter glasses to check the answers and motivate the children, without however giving feedback as to the correctness of the judgements. There was no time limit for responding.

For the TNO, assessment was started with plate I and continued until plate VII. For plates V to VII, to pass on to the next disparity, participants had to correctly identify the direction of the gap two out of two times. For comparison with the digital test, we only included the plates with quantifiable disparities. Previous modelling performed by Vancleef et al. (2017) has found that the TNO obtains thresholds at the 85% correct point. Therefore we added a corresponding threshold calculation for the digital test data.

### 2.4. Statistical analysis

Intraclass-correlation coefficient (ICR) was computed to assess how much repeated measurements resemble each other quantitatively. The weighted Cohen's Kappa coefficient was calculated to compare the digital test with the TNO. It measures inter-rater agreement for categorical items, taking into account the different intervals between disparities. The categories were formed according to the disparities presented on the TNO (in arcsec: 0–15, > 15–30, > 30–60, > 60–120, > 120–240 and > 240–480). The common method to compare different assessment methods is the Bland-Altman analysis. However, in most of our cases, assumptions were not met, e.g. because differences varied significantly with mean stereo values, or data distribution was not in accordance with requirements for Bland-Altman, or 95% limits of agreement could not be calculated due to small sample sizes. Thus, only where Bland-Altman analysis was adequate, corresponding results were presented for method comparison as well as for test-retest measurements. For method comparison in all cases TNO thresholds were subtracted from digital test thresholds. Simple regression analysis and correlation analysis was performed to assess the relationship of stereoacuity with its associated standard error, age and interocular visual acuity difference. If appropriate, Spearman's rank correlation analysis was used. Significance was defined as $p < 0.05$. Data was only included in the quantitative analyses where thresholds could be reliably computed. All statistical analyses were done with log arcsec.

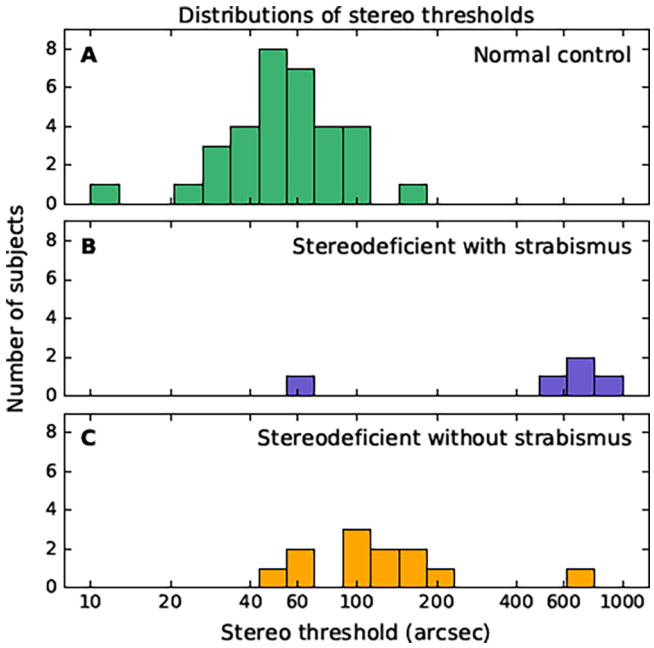


**Fig. 2.** Histogram of distribution of disparity thresholds for control subjects (A) and stereodeficient subjects with strabismus (B) and without strabismus (C) for the first trial of the digital test.

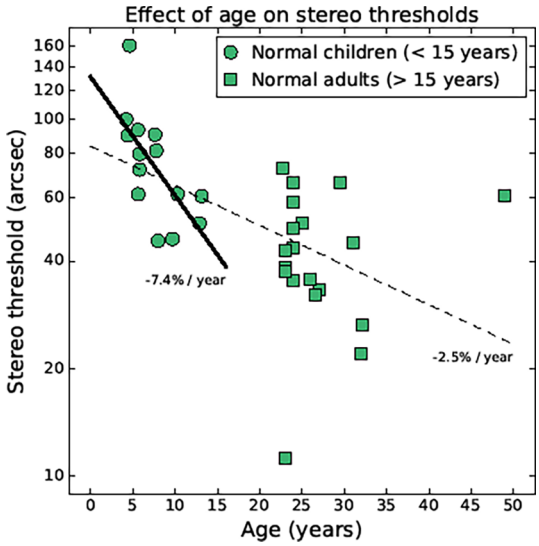


**Fig. 3.** Relationship between age and stereovision on the digital test for child (circles) and adult (squares) control subjects. Decrements are presented next to the corresponding regression lines with the dotted line representing the whole group and the solid line the child group.

## 3. Results

### 3.1. Stereoacuity ranges

Stereoacuity results for our digital measurements for the first trial are presented in Fig. 2 for control ($A$) and stereodeficient ($B + C$) subjects. One young adult control subject had to be excluded due to unusable results for our digital measurements. This subject had good visual acuity and good stereovision on the clinical tests; however a valid psychometric function was not obtained. Ten stereodeficient subjects (2 without strabismus and 8 with strabismus) also did not achieve a valid stereo threshold on our digital measurements. Therefore they do not appear in Fig. 2 and were labeled as "not having stereovision" (marked with † in Table 1). Only two of these subjects had fusion on either the Bagolini striated glasses or the Titmus fly (see Table 1, no 10 and no 16) and all had either strabismus, microstrabismus or very pronounced anisometropia.

For the control subjects, the median stereoacuity for our digital measurements was 51 arcsec (range 11–161). Their results on the TNO ranged from 15 to 240 arcsec. Stereodeficient participants without strabismus had a median stereoacuity of 113 arcsec (range 51–658). Participants with strabismus had a median stereoacuity of 626 arcsec (range 43–911). The subject reaching a stereo threshold of 43 arcsec on the second trial was a young adult with a microstrabismus who received occlusion treatment at an early age. This was also the only participant with strabismus (see Table 1, no 19) obtaining a quantitative stereo threshold on the TNO (120 arcsec). Results on the TNO for participants without strabismus ranged from 60 to 480 arcsec. There was only one subject without strabismus from whom we could not measure a threshold on the TNO (see Table 1, no 24); this was an adult with severe untreated anisometropic amblyopia.

### 3.2. Stereoacuity and age

We calculated a simple linear regression for control subjects to analyze the relationship between stereoacuity on the 3D monitor and age (Fig. 3). We found a significant relationship ($n = 33$, $r = -0.56$, $p < 0.001$) with the average threshold decreasing by 2.5% per year (indicating an improvement in stereoacuity). Separately, the regression equation was significant for children ($n = 14$, $r = -0.66$, $p = 0.01$), but not for adults ($n = 19$, $r = 0.1$, $p = 0.67$). Grubbs' test identified an adult control subject as a possible outlier, which is why we additionally calculated Spearman's rank correlation. Results remained similar for the whole group ($n = 33$, $\rho = -0.69$, $p < 0.001$) and the adult group ($n = 19$, $\rho = -0.14$, $p = 0.56$). The average child's threshold decreased by 7.4% per year. Adults' average threshold (40 arcsec) was finer than children's (74 arcsec), with a $t$-test showing a significant difference ($t(31) = 4.2462$, $p < 0.001$). The effect size was low ($d = 1.5$). These results confirm data in the literature that stereovision improves beyond the age of 4 years (Serrano-Pedraza, Herbert, Villa-Laso, Widdall, Vancleef, & Read, 2016).

The simple linear regression for stereovision on the TNO showed a small, but non-significant decline for the whole group ($n = 32$, $r = -0.27$, $p = 0.13$). Separately, results were neither significant for children ($n = 13$, $r = -0.34$, $p = 0.25$), nor for adults ($n = 19$, $r = 0.08$, $p = 0.74$). There was no significant difference between the child and adult group ($t(30) = 1.7311$, $p = 0.09$). Both had a median stereoacuity of 60 arcsec.

### 3.3. Reliability and association with standard error

Repeatability of our digital testing approach was computed for 25 control subjects with the intraclass-correlation coefficient (ICR). Correlation was strong and significant (Fig. 4A, $ICR = 0.77$, $p < 0.001$). The mean difference between the measurements for children was 0.05 log arcsec ($t(12) = 1.5416$, $p = 0.15$) and for adults 0.04 log arcsec ($t(11) = 0.9286$, $p = 0.37$). So a significant learning effect was observed for neither of the groups. An additional Bland-Altman analysis showed the mean difference to be 0.04 log arcsec and the 95% limits of agreement to be 0.5 log arcsec.

Results for the relationship between the stereo thresholds from the first and second trials and their associated standard errors for control subjects (Fig. 4B) were not normally distributed, which is why we performed a Spearman's rank-order correlation. We obtained a negative correlation of $\rho = -0.62$ which was significant ($p < 0.001$).

Fig. 5A displays the results for the first and second trial of our digital measurements for 15 stereodeficient participants. Results were

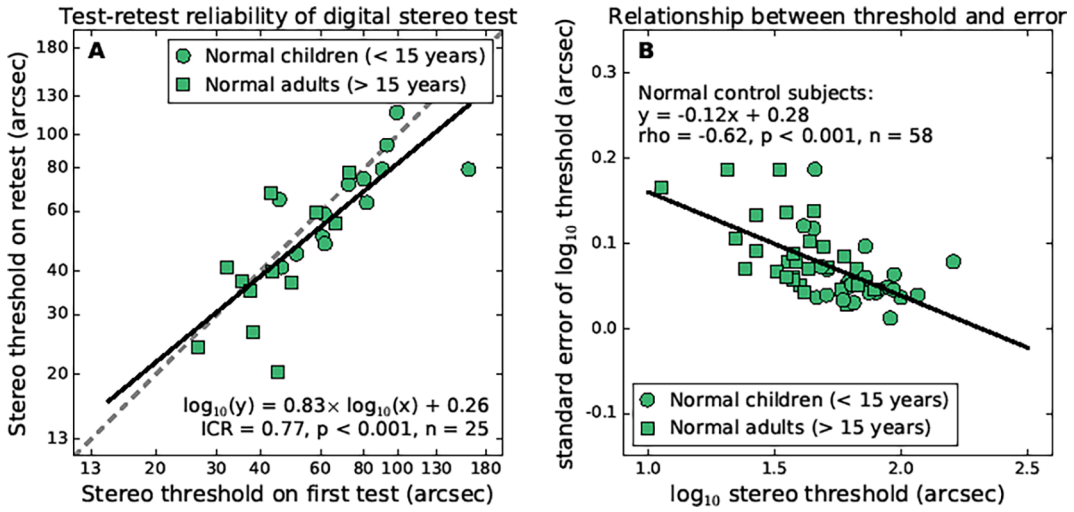


**Fig. 4.** (A) Test-Retest Reliability between first and second trial of the digital test for child and adult control subjects. The dotted line represents the bisection line and the solid line the regression line for the whole group with the corresponding regression equation displayed. (B) Relationship between stereo threshold on the first and second trial of the digital test and the corresponding standard errors for controls.

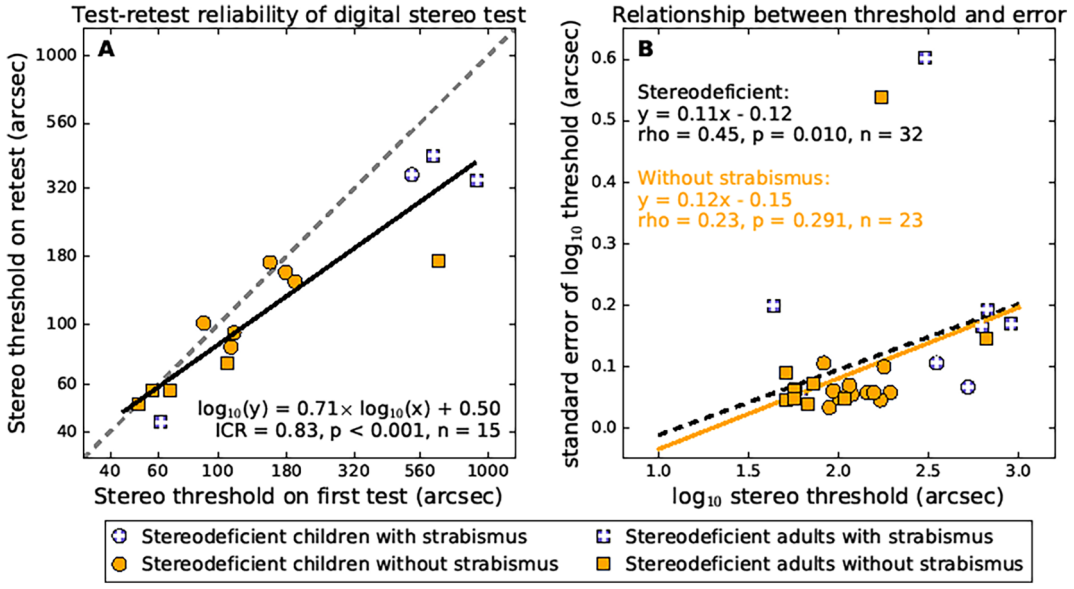


**Fig. 5.** (A) Test-Retest Reliability (ICR) between first and second trial of the digital test for stereodeficient participants with strabismus (circles/squares with a cross) and without strabismus (symbols without a cross). The solid line indicates the regression line with the associated equation displayed and the dotted line the bisection. (B) Relationship between stereo threshold on the first and second trial of the digital test and the corresponding standard errors for stereodeficient individuals, where the solid line describes the regression line for the stereodeficient participants without strabismus and the dotted line the regression for the entire group.

significantly correlated ($ICR = 0.83$, $p < 0.001$). We found a small learning effect with a mean difference of 0.14 log arcsec ($t(14) = 3.127$, $p = 0.01$) which was statistically significant; however according to Tsirlin, Colpa, Goltz, and Wong (2015), a difference considered clinically significant lies between 0.78 and 2.27 octaves for different clinical stereo tests, with a mean of 1.6 octaves (an octave is a halving/doubling of the score or 0.3 log arcsec). Furthermore this statistically significant learning effect was only present for subjects with strabismus: they had a mean difference of 0.23 log arcsec ($t(3) = 3.4515$, $p = 0.04$), whereas those without strabismus had a mean difference of 0.1 log arcsec ($t(10) = 1.9731$, $p = 0.08$). For the Bland-Altman analysis we excluded one subject without strabismus who had a threshold of 658 arc sec, which caused a violation of the assumptions. Excluding this subject, the mean difference for the stereodeficient subjects without strabismus was 0.06 log arcsec with 95% limits of agreement of 0.37 log arcsec. Assumptions for all other groups were not met.

For stereodeficient subjects, correlation between stereo thresholds of the first and second repetition and the corresponding standard errors was significant for the whole group, but not for stereodeficient subjects without strabismus (Fig. 5B). For every 10-fold increase in threshold there was around a 30% increase in standard error. As can be seen in Fig. 5B, most participants reaching a coarser threshold are subjects with strabismus, who have more difficulties in seeing 3D than subjects without strabismus.

To verify the way standard errors are calculated, we calculated over all of the thresholds for individual testing runs, the average standard error (standard deviation of sampling distribution) obtained from the bootstrap distributions. For amblyopes we got a mean of 0.11 log10

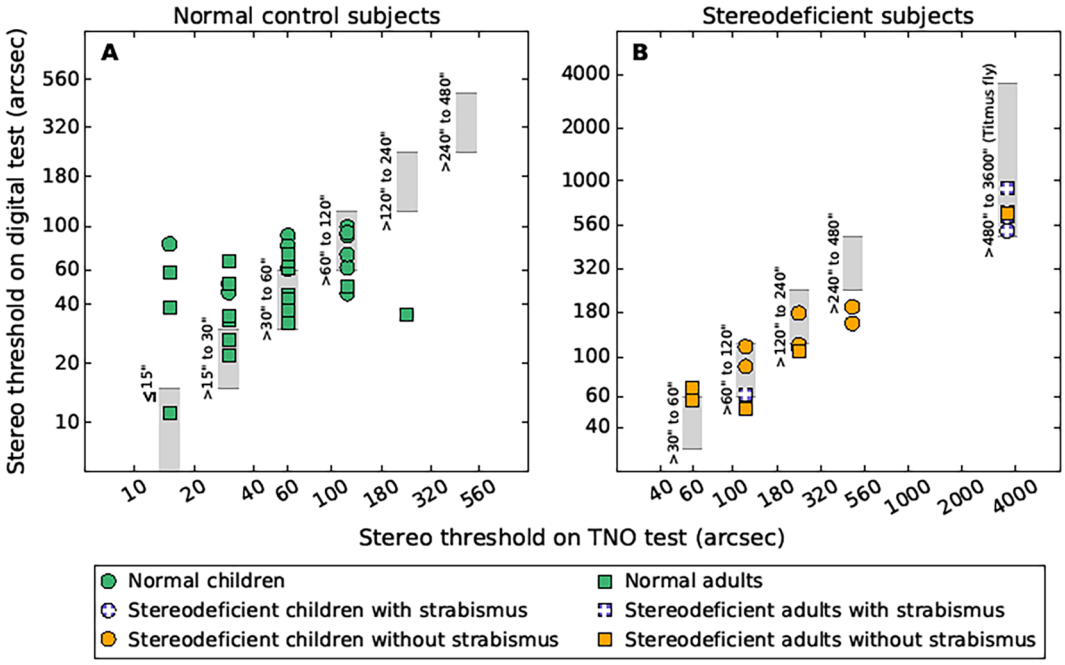


**Fig. 6.** Comparison between first-trial stereo thresholds on the TNO and on the digital test for (A) control subjects and (B) stereodeficient subjects with and without strabismus. TNO categories are shown by vertical grey rectangles with corresponding disparities in arcsec.

units. When two measurements are taken from that same distribution, we expect the standard deviation of the differences to be equal to 0.11 × sqrt(2) = 0.16. To test this, we took the test and retest measurements for each subject and first normalized them by subtracting 0.13 (mean of differences between test and retest) from the retest value. This factors out the average improvement we saw between test and retest. We then looked at the distribution of the differences between test and (normalized) retest. That distribution has a standard deviation of 0.15, which is similar to that we predict from the average standard error. For controls the standard deviation of the differences is 0.12 log10 units and the mean of the standard errors is 0.07. Our results showed that this is similar to 0.07 × sqrt(2) = 0.1.

Unlike in the digital test, we did not find a statistically significant correlation between the first and second test on the TNO for control subjects ($\rho = 0.35$, $p = 0.21$, $n = 15$) or the stereodeficient participants ($\rho = 0.53$, $p = 0.12$, $n = 10$). As sample sizes for both groups are small, we calculated the proportions for participants obtaining a smaller, equal or higher threshold on the second test. For control subjects, 47% reached the same threshold on the second test, whereas 33% reached a finer and 20% a coarser threshold. 70% of stereodeficient participants reached the same threshold on the second test, with 10% reaching a finer threshold and 20% a coarser threshold. An additional Bland-Altman analysis for control subjects showed the mean difference to be 0.06 log arcsec with 95% limits of agreement of 1.53 log arcsec. For stereodeficient subjects, we obtained a mean difference of −0.09 log arcsec with 95% limits of agreement of 1.51 log arcsec.

### 3.4. TNO and digital testing comparison

To compare stereo thresholds from digital testing with results on the TNO, weighted Cohen's $\kappa$ was run. There was fair agreement (Landis & Koch, 1977) for control subjects ($\kappa = 0.34$, $p = 0.13$, $n = 32$). For individual measurements, stereo thresholds tended to be finer on the TNO than for digital testing (Fig. 6A). This effect decreases for subjects with higher thresholds. One of the youngest subjects could not be included in the analysis, as no quantifiable result was achieved on the TNO. We found moderate agreement for stereodeficient participants ($\kappa = 0.61$, $p = 0.09$, $n = 17$). For finer disparities (< 240 arcsec) TNO and digital testing results corresponded, whereas for coarser disparities participants did somewhat better on digital testing (Fig. 6B). Four stereodeficient individuals with strabismus and one without strabismus, who did not succeed on TNO plate V (480 arcsec), but obtained a measurable stereo threshold on digital testing, are included in Fig. 6B. Their results are not included in the quantitative analysis, but compared to the Titmus Fly which all participants were able to see. Seeing stereo might be facilitated for stereodeficient participants by repeated presentation of disparity stimuli in the digital test. This could be a reason why amblyopes who did not fall in one of the TNO categories or those who reached a coarser threshold on the TNO, obtained better results for digital testing.

For a Bland-Altman analysis results obtained on the digital test were sorted into the matching TNO categories. The mean difference for control subjects was 0.16 log arcsec with 95% limits of agreement of 1.23 log arcsec. For stereodeficient subjects we obtained a mean difference of −0.13 log arcsec with 95% limits of agreement of 0.92 log arcsec. As can be seen in Fig. 6, agreement was dependent on threshold categories.

To account for the difference in level of performance between the digital test (62.5%) and the level of performance found by Vancleef et al. (2017) for the TNO (85%), we reanalyzed all of our subjects to obtain the thresholds for a level of performance of 85% (Fig. 7). Taking the average difference of all trials from all subjects between the thresholds for a level of performance of 62.5% and 85% we get a difference of 0.2 log arcsec. Agreement was poorer than with our initial threshold calculation for control subjects ($\kappa = 0.17$, $p = 0.43$, $n = 32$) and similar for stereodeficient subjects ($\kappa = 0.67$, $p = 0.07$, $n = 17$). After sorting the new thresholds into the matching TNO categories, we performed a Bland-Altman analysis. For stereodeficient subjects we obtained a mean difference of −0.07 log arcsec with 95% limits of agreement of 0.95 log arcsec, similar to our initial threshold calculation. For control subjects, assumptions were not met.

Durations of the digital tests were recorded for all subjects. On average, the digital stereo test took 8:02 min for control, and 6:50 min for stereodeficient subjects. For some participants durations were

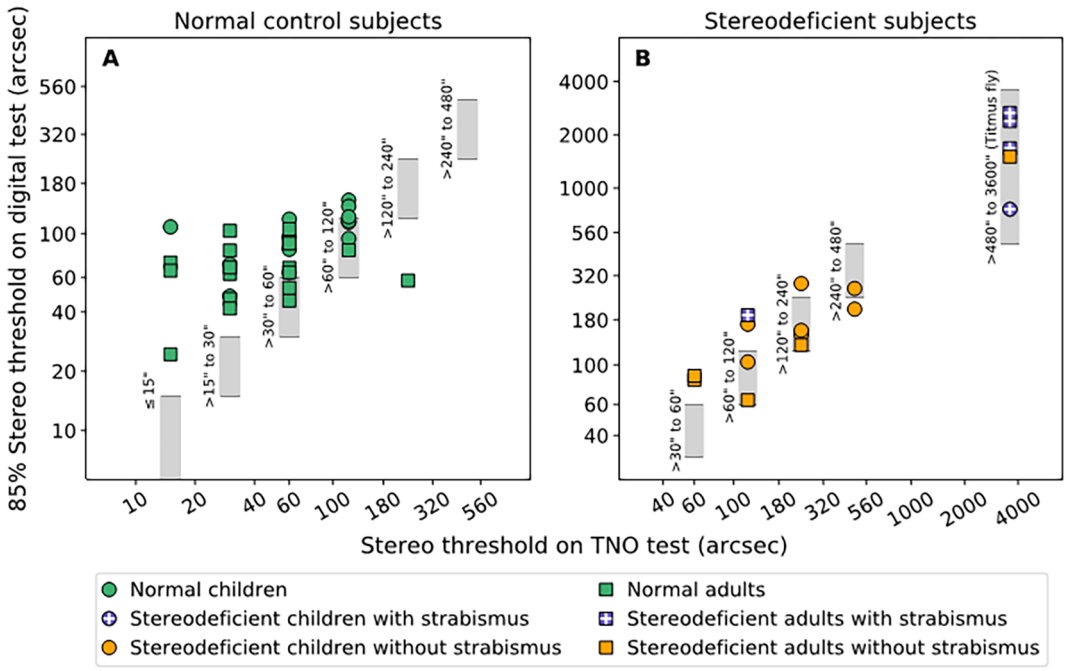


**Fig. 7.** Comparison between first-trial stereo thresholds on the TNO and on the digital test for a performance level of 85% for (A) control subjects and (B) stereodeficient subjects with and without strabismus. TNO categories are shown by vertical grey rectangles with corresponding disparities in arcsec.

recorded for the TNO test, which was always faster than the digital stereo test (3:20 min for controls, 3:36 for stereodeficient subjects).

### 3.5. No correlation with interocular visual acuity difference

Interocular visual acuity difference is often assumed to be related to the degree of stereovision (Greenwood et al., 2012; Lee & Isenberg, 2003). Regression analysis showed no significant correlation ($r = -0.02$, $p = 0.93$, $n = 17$) for all stereodeficient participants who obtained measureable stereovision in the digital test. Correlation for stereodeficient participants without strabismus was higher ($r = 0.37$) than for the whole group, but did not reach significance ($p = 0.24$, $n = 12$). For the correlation between interocular visual acuity difference and stereoacuity results on the TNO we calculated Spearman's rank correlation as we had different scales. No significant correlation was found ($\rho = 0.25$, $p = 0.43$, $n = 12$).

## 4. Discussion

### 4.1. Evaluation of our methods and results

Our representative sample, including participants with and without a stereodeficiency, and a large age and visual acuity range, provides an overview of thresholds measured with this type of digital stereo test. The disparity ranges obtained were in line with the expected results obtained by clinical stereo tests (Piano, Tidbury, & O'Connor, 2016). The digital testing was able to measure over a large and continuously sampled range of disparities. This is unlike any of the available clinical tests, which measure a highly quantized and restricted disparity range. Also, unlike any of the current clinical stereo tests we were able to calculate a measure of variability for each measurement we made.

Our study shows that by turning the task into a game, children at young ages (4–7 years) were able to perform the task required for our digital testing. This sample covers the age range where amblyopia therapy is very effective (Fronius, 2016; Fronius et al., 2014; Holmes et al., 2011; Stewart, Stephens, Fielder, Moseley, & Cooperative, 2007). We found that stereodeficient subjects with coarse stereovision had less difficulty obtaining a threshold on the digital test than on the clinical test. Some child control subjects, who either did not understand the task or were not motivated to give answers on the TNO, obtained valid thresholds on our digital test easily. A possible explanation for this could also be the problem of binocular color rivalry, which leads to reduced stereoacuity (Vancleef et al., 2017). One adult subject with amblyopia (see Table 1, no 24) with pronounced anisometropia was able to see TNO plate V (480 arcsec) only on the second trial, after having performed the digital stereo measurements. Such "learning effects" deserve future study in order to understand the underlying mechanisms. It is known that neural plasticity may exist in the visual system of adults (Astle, McGraw, & Webb, 2011b; Fronius, Cirina, Cordey, & Ohrloff, 2005; Levi, 2006; Levi & Polat, 1996, Thompson, Chung, Kiorpes, Ledgeway, & McGraw, 2015), and that amblyopia treatment can still be effective outside the conventional treatment age. A digital approach enables better monitoring of quantitative improvement in stereovision during amblyopia therapy.

The five monocularly visible trial figures (Fig. 1B) at the beginning of the experiment were especially helpful to introduce children and amblyopes to the task and stimulus. These trials with the monocularly visible contrast cue, which were randomly interspersed between the stereo trials, helped to maintain concentration and motivation for participants who did not have stereovision, and for the forced-choice judgements close to the threshold.

Results were reliable for our digital measurements, both for individuals with and without stereodeficiency. This shows that thresholds can be measured with a high degree of accuracy in patients with amblyopia. Nevertheless, a small learning effect was present especially for participants who had coarser disparity results. To avoid obtaining an improvement in stereovision in prospective treatment studies due to a learning effect rather than through therapy, digital testing should be applied repeatedly before the beginning of treatment (Schmitt et al., 2002). Adams, Leske, Hatt, and Holmes (2009) analyzed the test-retest measurements with the 95% limits of agreement for the preschool Randot (0.59 log arcsec), near Frisby (0.24 log arcsec), FD2 (0.68 log arcsec) and distance Randot (0.46 log arcsec) in subjects with strabismus. For the digital test, the 95% limits of agreement from our

control subjects (0.5 log arcsec) and stereodeficient subjects without strabismus (0.37 log arcsec) agree with the results found by Adams et al. (2009). The 95% limits of agreement on the TNO did not show high agreement, neither for controls, nor for amblyopes. Still, 47% of control subjects and 70% of subjects with stereodeficiencies achieved the same threshold on the retest. The results from the 95% limits of agreement for controls (1.53 log arcsec) and stereodeficient subjects (1.51 log arcsec) could be due to small sample sizes. This is why we provided the percentages of participants that scored equal, better or worse on the second trial.

For control subjects, reaching a finer threshold on the digital test seems to be related to more insecurity which is reflected by the spread of the individual results (Fig. 4B). On the other hand, the majority of stereodeficient participants who obtained finer results had a smaller associated variability, and those who obtained coarser results, had a larger associated variability. For individuals with strabismus, even when the angle of squint was small, stereo thresholds tended to be higher and so were the associated standard errors. By providing the variability of the threshold, not only the significance of a change in stereoacuity can be judged, it also creates the possibility to monitor whether reliability of thresholds varies or remains stable.

Our results show the importance of testing methods for measuring stereo in all target groups with respect to age and clinical conditions to avoid generalization, as results can differ. We became aware of the presence of monocular cues in the early version of our setup when testing amblyopes with presumed very coarse or no measurable stereovision. They would have remained undetected if the method was assessed only in subjects with fine stereopsis. Serrano-Pedraza et al. (2016a) identified some rules for avoiding monocular artifacts: use of dynamic stimuli, keeping the position constant and the use of a depth detection rather than discrimination task.

Although this was not a primary aim of our study, subgroup analysis revealed a significant correlation between age and digital stereo thresholds for child control subjects, but not for adult controls. Similarly, Serrano-Pedraza et al. (2016b) found an improvement of stereovision with digital testing between the ages of 3 and 13 years, but no improvement for the age group 18–32 years. There is no clear consensus for the development of stereopsis. Depending on the assessment methods used, some studies found an improvement to near adult level during the first seven months (Birch & Petrig, 1996), whereas other studies showed that stereoacuity improves up to 18 months (Birch, Morale, Jeffrey, O'Connor, & Fawcett, 2005), does not reach the adult level until 5 years (Drover et al., 2017; Simons, 1981) or continues to improve up to the age of 10 (Bohr & Read, 2013) or 14 years (Giaschi et al., 2013b). We found a median stereoacuity of 74 arcsec for our child control subjects, which goes in line with other results found for ages 3–9 years (Afsari et al., 2013; Ciner et al., 2014; Greenwood et al., 2012). Anketell, Saunders, and Little (2013) found a difference in stereovision between primary school age groups (6–7 years, 9–10 years) and post primary school age groups (12–13 years, 15–16 years). On the Frisby stereo test, the younger group had a median stereoacuity of 20 arcsec, whereas the older group achieved a median stereoacuity of 10 arcsec. Our digital test shows slightly higher results for both our child and adult groups. On the TNO, both groups from Anketell et al. (2013) achieved median stereoacuity of 60 arcsec, which goes in line with our results. Similarly to our results, the authors did not find a significant difference between age groups for the TNO, which suggests that the quantized nature of the TNO causes less sensitive results. Additionally Cooper, Feldman, and Medlin (1979) could not find a correlation between stereoacuity and age with the TNO.

### 4.2. Comparison with the clinical stereo test TNO and other digital stereo tests

On the TNO, many control subjects with good stereopsis obtained lower thresholds than on our digital tests. These differences could be due to the different spatial resolutions of the two tests. The disadvantage of using a test with very fine spatial resolution (e.g. TNO) is that it will be especially susceptible to artefacts due to uncorrected refractive errors. Stereodeficient subjects on the other hand obtained finer results on our digital testing. This is especially true for coarse disparities that cannot be measured by the TNO. Vancleef et al. (2017) compared TNO results to other clinical stereo tests and a digital test implementing a staircase procedure in 4–16 years old control subjects. They found that the TNO always overestimated threshold values. They found low agreement in their Bland-Altman analysis for the comparison between the TNO and their digital test. Binocular color rivalry which can be induced by different colors could be an influencing factor for differences in the results (Vancleef et al., 2017). Additionally, Vancleef et al. (2017) found the level of performance of the TNO to be at 85%, whereas the digital test we assessed has a level of performance of 62.5%. To account for these differences we redid the analyses of thresholds to have them at the same performance level (85%) as Vancleef et al. (2017), and compared the results to the TNO. Agreement between both tests was slightly better for stereodeficient subjects, but worse for control subjects in comparison to the previous analysis with a performance level of 62.5%. Both Vancleef et al. (2017) and our data suggest that the level of performance of the TNO depends on the stereo thresholds, which makes general conclusions about comparison with other tests difficult. Lindblom and Frisen (1988) compared the results of the same digital stereo test using anaglyph glasses and shutter glasses. Their results show that especially inexperienced adult subjects without any known eye abnormalities have difficulties performing a stereo test with red-green anaglyph glasses, but no problems in performing the stereo test with shutter glasses. We found that especially for young children who might have problems understanding the concept of depth, obtaining thresholds on the digital test was easier than on the TNO, thereby confirming Lindblom and Frisen (1988). Another influencing factor can be the size of the random-dots. As visual acuity is limited, especially in amblyopes and young children, a larger stimulus with bigger dots could facilitate the task. However, Vancleef et al. (2017) found stereo thresholds to be finer for smaller stimuli than for bigger ones. Furthermore different results can be due to different assessment methods. While on the TNO, disparities are tested only twice, our digital test implements a forced choice paradigm and two staircase procedures. Repeated stimulus presentations on our digital testing, although without feedback, could facilitate threshold detection for amblyopes.

### 4.3. Conclusions and future prospects

Digital tests of the type applied here allow detailed quantification of coarse to fine stereo thresholds. To our knowledge, there are no other studies testing adults and children with and without stereodeficiencies while also applying an adaptive staircase procedure. By assessing repeatability of the test, correlation of stereo thresholds with age and interocular visual acuity difference and the agreement with a clinical gold standard test for measuring stereovision, we aimed at giving a detailed overview of the method we used. The improved measurements come at some cost of time, as digital test times were longer than those for administering the TNO. The advantage however is that the digital test allows prospective following of patients in studies investigating how stereovision improves alongside visual acuity during amblyopia therapy. This has become especially important recently as new dichoptic therapy methods have been developed with the aim of improving binocular vision.

## Declaration of Competing Interest

ASB & RFH: The two authors from McGill University (ASB, RFH) are inventors on a patent application filed by McGill University on the measurement of stereovision. Should McGill University commercialize

the invention, the authors would receive a share of the proceeds. All other authors: None.

**Acknowledgements**

This work was supported by an ERA-NET NEURON grant (JTC2015) to MF and RFH, the German Ministry for Education and Research [BMBF, grant number 01EW1603B] and by the association "Augenstern-e.V." (non-profit association supporting research in pediatric ophthalmology) from Frankfurt, Germany. The authors are grateful to Peggy Feige for excellent orthoptic assessment of study participants and research assistance. Thanks are due to the orthoptists and ophthalmologists from the Department of Ophthalmology of the Goethe University for helping with the recruitment of subjects, and to the study volunteers and families for participation in the study.

## Appendix. Further information on stimulus generation

The stimuli used in this study are generated from hundreds of bandpass dots. These dots are generated in the Fourier domain as isotropic log-Gabors (Fig. A1A). They consist of a central bright positive peak, with a slightly darker negative surround (Fig. A1B). The structure of the central peak and surrounding negative regions results in the dots being spatially bandpass (Fig. A1C). They have a peak in their spatial frequency content at 0.5c/deg, with less energy at higher or lower frequencies. In our rendering of the dots however, the darker surround region extends past the edges of the image size in which we generate them. If not dealt with, this would result in the edges of these dot images being visible when we paste together hundreds of these dots to form our stimuli. To prevent this, we apply a softening window that linearly ramps the four pixels at the edge of the stimuli from a gain of unity down to a gain of zero. The effects of this windowing can be seen by comparing the blue (solid) and orange (dotted) curves in Fig. A1C. The solid blue curve gives the original dot, with its peak spatial frequency content at 0.5c/deg. The jump discontinuity at the edge of the dot image results in spectral leakage that gives a sinc-function profile to the fall-off. The dotted orange curve shows the spectrum after the softening window has been applied. The dots are now a little less bandpass, and the central frequency is slightly increased.

We introduced disparity to our stimuli by shifting the horizontal positions of the dots presented to the two eyes. Simply translating the location of these dots results in regions of increased density in the region dots are moved into, and decreased density in the region they are moved away from (Fig. A2A). These density changes will result in non-stereo cues that may allow a subject to identify disparity-defined edges in our task. These cues are visible monocularly, and so do not even require binocular vision to be useful. Deleting dots that overlap from just the eye in which that overlap occurs results in monocular dots (Fig. A2B). These monocular dots appear different from binocular dots, giving a different potential non-disparity cue (note however that perceiving this cue requires the integration of information from both eyes). To prevent the use of these monocular dots in solving the task, we also remove the corresponding dot in the other eye (Fig. A2C). However, this then leaves even larger gaps in the stimuli. To account for this, we fill in any large gaps with dots at zero disparity (Fig. A2D). Our stimuli feature a target in crossed disparity and a background in equal uncrossed disparity, so zero disparity is halfway between the two. In cases where that zero-disparity dot would overlap with another dot in just one eye, we do place the dot monocularly.

Fig. A3 shows how many filled-in dots were added to our stimuli as a function of target disparity. Up to 256 arc seconds the need to fill any gaps was minimal. On average, only two filling-in dots were added to a stimulus. Both of those dots were binocular. Beyond 256 arc seconds, the number of filling-in dots that were required increases, and some of the dots then become monocular. Note that for large disparities, all dots in the display would appear monocular to a hypothetical "stereoblind" subject who was not able to resolve positional offsets from disparity to link the corresponding dots (but who was still sensitive to dichoptic differences). We therefore do not expect our subjects to be relying on these occasional monocular dots to perform the task.

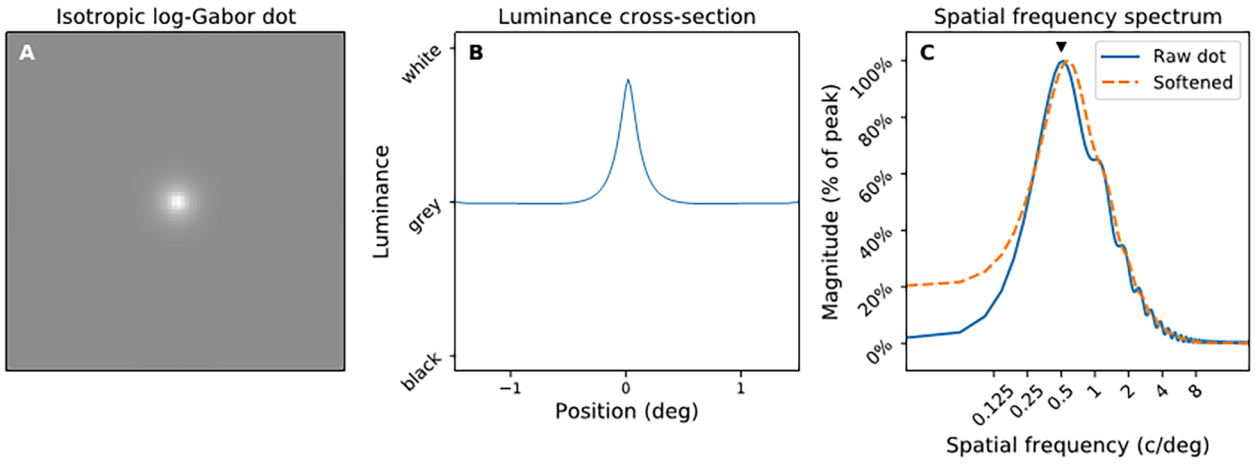


**Fig. A1.** Panel A shows a single bandpass dot (isotropic log-Gabor). These dots were used to generate our stimuli. The luminance cross-section of the dot is shown in Panel B. The bright centre of the dot corresponds to the central peak of this curve. There are also negative side-lobes adjacent to that central peak where the luminance decreases slightly below the mean grey. This results in a bandpass spectrum, as illustrated in Panel C. In Panel C, the solid blue curve gives the spatial frequency spectrum of the dot (taken from the Fourier transform) before the softening window is applied to the edges of the rendered image. The dotted orange curve gives the spectrum after that window is applied. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

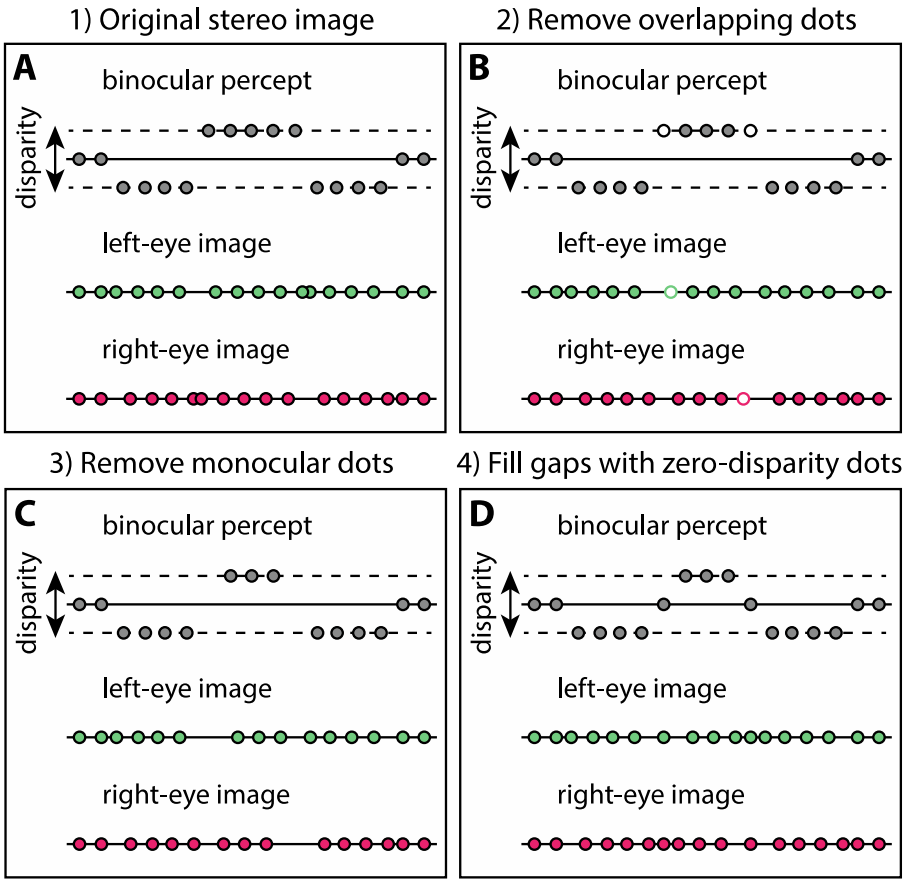


**Fig. A2.** showing the process by which the "gaps" in the monocular images are filled in. Panel A shows a row of dots from an idealised stimulus (dot spacing in the experiment stimuli is not uniform) containing crossed disparity in the centre flanked by uncrossed disparity. The grey dots at the top show the binocular percept, which is generated by the left and right eye dot positions shown below. Note the overlapping dots and relatively sparse regions caused by the position shifts in the left and right eye images. Panel B shows the effect of removing the overlapping dots from each eye's image. The white-filled dots are now monocular, with no corresponding dot in the other eye (therefore their true disparity is now undefined). Panel C shows the effect of removing those monocular dots. There are now large gaps in the resulting image. Panel D shows the effect of filling in any gaps with dots at zero disparity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

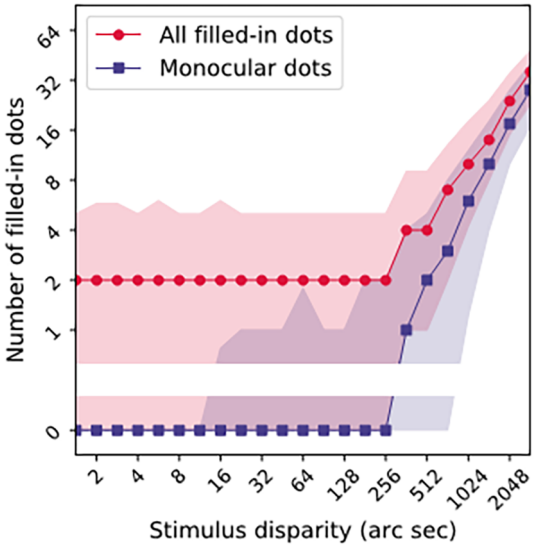


**Fig. A3.** shows how the expected number of "filled-in" dots at zero disparity increases as a function of the target disparity. The red curve (circles) was created by generating 250 example stimuli at each disparity and taking the median number of filled-in dots. The shaded region gives the range within which 95% of the 250 sample stimuli fell. The purple curve (squares) indicates the subset of those filled-in dots which were monocular. The median total number of dots across all of our generated stimuli was 622 dots per stimulus. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

## References


Adams, W. E., Leske, D. A., Hatt, S. R., & Holmes, J. M. (2009). Defining real change in measures of stereoacuity. *Ophthalmology, 116*(2), 281–285.

Afsari, S., Rose, K. A., Pai, A. S. I., Gole, G. A., Leone, J. F., Burlutsky, G., & Mitchell, P. (2013). Diagnostic reliability and normative values of stereoacuity tests in preschool-aged children. *British Journal of Ophthalmology, 97*(3), 308–313. https://doi.org/10.1136/bjophthalmol-2012-302192.

Anketell, P. M., Saunders, K. J., & Little, J. A. (2013). Stereoacuity norms for school-age children using the Frisby stereotest. *Journal of AAPOS, 17*(6), 582–587. https://doi.org/10.1016/j.jaapos.2013.08.012.

Astle, A. T., McGraw, P. V., & Webb, B. S. (2011a). Can human amblyopia be treated in adulthood? *Strabismus, 19*(3), 99–109. https://doi.org/10.3109/09273972.2011.600420.

Astle, A. T., McGraw, P. V., & Webb, B. S. (2011b). Recovery of stereo acuity in adults with amblyopia. *BMJ Case Reports, 7–10*. https://doi.org/10.1136/bcr.07.2010.3143.

Bach, M., Schmitt, C., Kromeier, M., & Kommerell, G. (2001). The Freiburg stereoacuity test: Automatic measurement of stereo threshold. *Graefe's Archive for Clinical and Experimental Ophthalmology, 239*(8), 562–566. https://doi.org/10.1007/s004170100317.

Baldwin, A. S. (2019, July 2). alexsbaldwin/MatlabStaircase: v0.9.0 (Version v0.9.0). Zenodo. http://doi.org/10.5281/zenodo.3266142.

Birch, E. E., Li, S. L., Jost, R. M., Morale, S. E., De La Cruz, A., Stager, D., ... Stager, D. R. (2015). Binocular iPad treatment for amblyopia in preschool children. *Journal of American Association for Pediatric Ophthalmology and Strabismus, 19*(1), 6–11. https://doi.org/10.1016/j.jaapos.2014.09.009.

Birch, E. E., Morale, S. E., Jeffrey, B. G., O'Connor, A. R., & Fawcett, S. L. (2005). Measurement of stereoacuity outcomes at ages 1 to 24 months: Randot® Stereocards. *Journal of American Association for Pediatric Ophthalmology and Strabismus, 9*(1), 31–36. https://doi.org/10.1016/j.jaapos.2004.11.013.

Birch, E., & Petrig, B. (1996). FPL and VEP measures of fusion, stereopsis and stereoacuity in normal infants. *Vision Research, 36*(9), 1321–1327.

Bohr, I., & Read, J. C. A. (2013). Stereoacuity with Frisby and revised FD2 stereo tests. *PLoS One, 8*(12), https://doi.org/10.1371/journal.pone.0082999.

Bömer, T., Dölp, R., & Kommerell, G. (1995). Psychometric function of stereo disparity in normal persons. *Der Ophthalmologe: Zeitschrift Der Deutschen Ophthalmologischen Gesellschaft, 92*(2), 120–124.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436.

Ciner, E. B., Ying, G. S., Kulp, M. T., Maguire, M. G., Quinn, G. E., Orel-Bixler, D., ... Huang, J. (2014). Stereoacuity of preschool children with and without vision disorders. *Optometry and Vision Science, 91*(3), 351–358. https://doi.org/10.1097/OPX.0000000000000165.

Cooper, J., Feldman, J., & Medlin, D. (1979). Comparing stereoscopic performance of children using the Titmus, TNO, and Randot stereo tests. *Journal of the American Optometric Association, 50*(7), 821–825.

Cornforth, L. L., Johnson, B. L., Kohl, P., & Roth, N. (1987). Chromatic imbalance due to commonly used red-green filters reduces accuracy of stereoscopic depth perception. *American Journal of Optometry and Physiological Optics, 64*(11), 842–845.

Drover, J. R., Cornick, S. L., Hallett, D., Drover, A., Mayo, D., & Kielly, N. (2017). Normative pediatric data for three tests of functional vision. *Canadian Journal of Ophthalmology, 52*(2), 198–202. https://doi.org/10.1016/j.jcjo.2016.08.016.

Emerson, J. D., & Strenio, J. (1983). Boxplots and batch comparisons. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.). *Understanding robust and exploratory data analysis* (pp. 58–96). New York: John Wiley & Sons.

Foley, J. M., & Tyler, C. W. (1976). Effect of stimulus duration on stereo and vernier displacement thresholds. *Perception & Psychophysics, 20*(2), 125–128. https://doi.org/10.3758/BF03199443.

Fortin, A., Ptito, A., Faubert, J., & Ptito, M. (2002). Cortical areas mediating stereopsis in the human brain: A PET study. *Neuroreport, 13*(6), 895–898.

Fronius, M. (2016). Occlusion treatment for amblyopia. Age dependence and dose-response relationship. *Der Ophthalmologe: Zeitschrift Der Deutschen Ophthalmologischen Gesellschaft, 113*(4), 296–303. https://doi.org/10.1007/s00347-016-0235-7.

Fronius, M., Cirina, L., Ackermann, H., Kohnen, T., & Diehl, C. M. (2014). Efficiency of electronically monitored amblyopia treatment between 5 and 16 years of age: New insight into declining susceptibility of the visual system. *Vision Research, 103*, 11–19. https://doi.org/10.1016/j.visres.2014.07.018.

Fronius, M., Cirina, L., Cordey, A., & Ohrloff, C. (2005). Visual improvement during psychophysical training in an adult amblyopic eye following visual loss in the contralateral eye. *Graefe's Archive for Clinical and Experimental Ophthalmology, 243*(3), 278–280. https://doi.org/10.1007/s00417-004-1014-8.

Giaschi, D., Lo, R., Narasimhan, S., Lyons, C., & Wilcox, L. M. (2013a). Sparing of coarse stereopsis in stereodeficient children with a history of amblyopia. *Journal of Vision, 13*(10), 17. https://doi.org/10.1167/13.10.17.

Giaschi, D., Narasimhan, S., Solski, A., Harrison, E., & Wilcox, L. M. (2013b). On the typical development of stereopsis: Fine and coarse processing. *Vision Research, 89*, 65–71. https://doi.org/10.1016/j.visres.2013.07.011.

Greenwood, J. A., Tailor, V. K., Sloper, J. J., Simmers, A. J., Bex, P. J., & Dakin, S. C. (2012). Visual acuity, crowding, and stereo-vision are linked in children with and without amblyopia. *Investigative Ophthalmology & Visual Science, 53*(12), 7655–7665. https://doi.org/10.1167/iovs.12-10313.

Hess, R. F., Ding, R., Clavagnier, S., Liu, C., Guo, C., Viner, C., ... Zhou, J. (2016). A robust and reliable test to measure stereopsis in the clinic. *Investigative Ophthalmology & Visual Science, 57*(3), 798–804. https://doi.org/10.1167/iovs.15-18690.

Hess, R. F., Mansouri, B., & Thompson, B. (2010). A new binocular approach to the treatment of amblyopia in adults well beyond the critical period of visual development. *Restorative Neurology and Neuroscience, 28*(6), 793–802. https://doi.org/10.3233/RNN-2010-0550.

Hess, R. F., Thompson, B., & Baker, D. H. (2014). Binocular vision in amblyopia: Structure, suppression and plasticity. *Ophthalmic and Physiological Optics, 34*(2), 146–162. https://doi.org/10.1111/opo.12123.

Hess, R. F., Thompson, B., Black, J. M., Maehara, G., Zhang, P., Bobier, W. R., & Cooperstock, J. (2012). An iPod treatment of amblyopia: An updated binocular approach. *Optometry (St. Louis, Mo.), 83*(2), 87–94.

Holmes, J. M., Lazar, E. L., Melia, B. M., Astle, W. F., Dagi, L. R., Donahue, S. P., Frazier, M. G., Hertle, R. W., Repka, M. X., Quinn, G. E., Weise, K. K., & Pediatric Eye Disease Investigator Group (2011). Effect of age on response to amblyopia treatment in children. *Archives of Ophthalmology, 129*(11), 1451. https://doi.org/10.1001/archophthalmol.2011.179.

Kehrein, S., Kohnen, T., & Fronius, M. (2016). Dynamics of interocular suppression in amblyopic children during electronically monitored occlusion therapy: first insight. *Strabismus, 24*(2), 51–62. https://doi.org/10.3109/09273972.2016.1170047.

Kelly, K. R., Jost, R. M., Wang, Y.-Z., Dao, L., Beauchamp, C. L., Leffler, J. N., & Birch, E. E. (2018). Improved binocular outcomes following binocular treatment for childhood amblyopia. *Investigative Opthalmology & Visual Science, 59*(3), 1221. https://doi.org/10.1167/iovs.17-23235.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception, 36*, 1–16.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*. https://doi.org/10.2307/2529310.

Lee, S. Y., & Isenberg, S. J. (2003). The relationship between stereopsis and visual acuity after occlusion therapy for amblyopia. *Ophthalmology, 110*(11), 2088–2092. https://doi.org/10.1016/S0161-6420(03)00865-0.

Leske, D. A., & Holmes, J. M. (2004). Maximum angle of horizontal strabismus consistent with true stereopsis. *Journal of AAPOS, 8*(1), 28–34. https://doi.org/10.1016/S1091853103002568.

Levi, D. M. (2006). Visual processing in amblyopia: Human studies. *Strabismus, 14*(1), 11–19. https://doi.org/10.1080/09273970500536243.

Levi, D. M., & Polat, U. (1996). Neural plasticity in adults with amblyopia. *Proceedings of the National Academy of Sciences of the United States of America, 93*, 6830–6834.

Lindblom, B., & Frisen, L. (1988). Measuring stereo acuity with liquid crystal shutters and computer graphics. *Neuro-Ophthalmology, 8*(6), 283–287. https://doi.org/10.3109/01658108808996056.

Long To, L., Thompson, B., Blum, J. R., Maehara, G., Hess, R. F., & Cooperstock, J. R. (2011). A game platform for treatment of amblyopia. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 19*(3), 280–289. https://doi.org/10.1109/TNSRE.2011.2115255.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Piano, M. E. F., Tidbury, L. P., & O'Connor, A. R. (2016). Normative values for near and distance clinical tests of stereoacuity. *Strabismus, 24*(4), 169–172. https://doi.org/10.1080/09273972.2016.1242636.

Prins, N., & Kingdom, F. A. A. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the palamedes toolbox. *Frontiers in Psychology, 9*, 1250. https://doi.org/10.3389/fpsyg.2018.01250.

Schmitt, C., Kromeier, M., Bach, M., & Kommerell, G. (2002). Interindividual variability of learning in stereoacuity. *Graefe's Archive for Clinical and Experimental Ophthalmology = Albrecht von Graefes Archiv Fur Klinische Und Experimentelle Ophthalmologie, 240*(9), 704–709. https://doi.org/10.1007/s00417-002-0458-y.

Serrano-Pedraza, I., Vancleef, K., & Read, J. C. A. (2016a). Avoiding monocular artifacts in clinical stereotests presented on column-interleaved digital stereoscopic displays. *Journal of Vision, 16*(14), 13. https://doi.org/10.1167/16.14.13.

Serrano-Pedraza, I., Herbert, W., Villa-Laso, L., Widdall, M., Vancleef, K., & Read, J. C. A. (2016b). The stereoscopic anisotropy develops during childhood. *Investigative Ophthalmology and Visual Science, 57*(3), 960–970. https://doi.org/10.1167/iovs.15-17766.

Simons, K. (1981). A comparison of the Frisby, Random-Dot E, TNO, and Randot circles stereotests in screening and office use. *Archives of Ophthalmology, 99*(3), 446–452. https://doi.org/10.1001/archopht.1981.03930010448011.

Simons, K. (1996). Preschool vision screening: Rationale, methodology and outcome. *Survey of Ophthalmology, 41*(1), 3–30.

Sireteanu, R. (2000). The binocular visual system in amblyopia. *Strabismus, 8*(1), 39–51.

Sireteanu, R., & Fronius, M. (1981). Naso-temporal asymmetries in human amblyopia: Consequence of long-term interocular suppression. *Vision Research, 21*(7), 1055–1063.

Stewart, C. E., Stephens, D. A., Fielder, A. R., Moseley, M. J., & Cooperative, M. O. T. A. S.

(2007). Modeling dose-response in amblyopia: toward a child-specific treatment plan. *Investigative Opthalmology & Visual Science, 48*(6), 2589. https://doi.org/10.1167/iovs.05-1243.

Stewart, C. E., Wallace, M. P., Stephens, D. A., Fielder, A. R., & Moseley, M. J. (2013). The effect of amblyopia treatment on stereoacuity. *Journal of AAPOS, 17*(2), 166–173. https://doi.org/10.1016/j.jaapos.2012.10.021.

Thompson, B., Chung, S. T. L., Kiorpes, L., Ledgeway, T., & McGraw, P. V. (2015). A window into visual cortex development and recovery of vision: Introduction to the Vision Research special issue on Amblyopia. *Vision Research, 114*, 1–3. https://doi.org/10.1016/j.visres.2015.06.002.

Tidbury, L. P., Brooks, K. R., O'Connor, A. R., & Wuerger, S. M. (2016). A systematic comparison of static and dynamic cues for depth perception. *Investigative Opthalmology & Visual Science, 57*(8), 3545. https://doi.org/10.1167/iovs.15-18104.

Tomaç, S., & Altay, Y. (2000). Near stereoacuity: Development in preschool children; normative values and screening for binocular vision abnormalities; a study of 115 children. *Binocular Vision & Strabismus Quarterly, 15*(3), 221–228.

Tsirlin, I., Colpa, L., Goltz, H. C., & Wong, A. M. F. (2015). Behavioral training as new treatment for adult amblyopia: a meta-analysis and systematic review. *Investigative Ophthalmology & Visual Science, 56*(6), 4061–4075. https://doi.org/10.1167/iovs.15-16583.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Vancleef, K., Read, J. C. A., Herbert, W., Goodship, N., Woodhouse, M., & Serrano-Pedraza, I. (2017). Overestimation of stereo thresholds by the TNO stereotest is not due to global stereopsis. *Ophthalmic & Physiological Optics, 37*(4), 507–520. https://doi.org/10.1111/opo.12371.

Vedamurthy, I., Nahum, M., Huang, S. J., Zheng, F., Bayliss, J., Bavelier, D., & Levi, D. M. (2015). A dichoptic custom-made action video game as a treatment for adult amblyopia. *Vision Research, 114*, 173–187. https://doi.org/10.1016/j.visres.2015.04.008.