Rheinisch-Westfälische Technische Hochschule Aachen
Institute of Biology II, Department of Zoology and Animal Physiology

The influence of stimulus configuration in visual hyperacuity tasks on test repeatability

Der Einfluss der Stimulus-Konfiguration auf die Testwiederholbarkeit in hypergenauen Sehaufgaben

Bachelor thesis

Anna-Marina van der Meer
Matriculation Number 287201

Aachen, September 2010

1st Reviewer: Univ.-Prof. Dr. rer. nat Hermann Wagner

2nd Reviewer: Univ.-Prof. Dr. med. Peter Walter

Supervisor: Dr. rer. nat. Wolf Harmening

Table of contents

1. INTRODUCTION 3
1.1 Hyperacuity in the human visual system 3
1.2 Psychophysical methods 7
1.3 Aim of the thesis 8
2. MATERIAL AND METHODS 9
2.3 Subjects and procedure 9
2.2 Stimuli and data analysis 12
2.3 FrACT 14
3. RESULTS 15
3.1 Acuity 17
3.2 Repeatability 19
3.3 Subjective scores 21
4. DISCUSSION 24
4.1 Acuity 24
4.2 Repeatability 26
4.3 Subjective scores 28
CONCLUSION 30
REFERENCES 31
DECLARATION 33
ACKNOWLEDGEMENTS 34

1. Introduction

When it comes to information pickup, orientation, movement, and social interaction, the visual sense is the most important sense in human perception. In order to visually communicate with other individuals we use facial expressions, body language and eye contact. Posture can tell us a lot about another person's feelings. As the visual sense plays such a major role in our everyday life, it evolved to a highly developed sensory system. The importance of the visual system is emphasized by the amount of neuronal machinery allocated to perform visual tasks: more than 50% of the human cortex is involved in visual processing (Snowden, Thompson, & Troscianko, 2006). Humans are usually active during daytime, thus, our visual system is very well adapted to day vision. We are able to see thousands of colors, perceive and discriminate biological movement effortless and have excellent spatial vision, enhanced by a sophisticated stereoscopic viewing system. If thresholds fall below the theoretical limit of resolution calculated by photoreceptor spacing and diameter this phenomenon is called hyperacuity. Hyperacute visual perception allows us to judge the relative position of objects to the most minute scale.

1.1 Hyperacuity in the human visual system

Visual tests are used to characterize the human visual system in clinical as well as in scientific contexts. Hyperacute thresholds suggest that the neural fraction of the observer's visual capabilities is addressed (Williams, Enoch, & Essock, 1984). Unfortunately, repeatability of hyperacuity thresholds is rather low. This may be due to the psychophysically disadvantageous 2-AFC paradigm in Vernier tasks, thus a higher number of trials may be advisable (Abbud & Cruz, 2002). On the other hand longsome testing procedures would probably have an effect on the subject's motivation, concentration and fatigue. Hence, the aim of this thesis is to find a reasonable combination of stimulus configuration and testing procedure with the intention to get better test results and to confirm this combination in psychophysical tests.

In the usual visual tests normal resolving power is estimated with help of Landolt-Cs (see Figure 1a), Snellen letter charts or tumbling-E charts where thresholds of 30 to 60 arcsec are achieved in emmetropic subjects (Carney & Klein, 1997). In hyperacuity tasks thresholds considerably fall below normal resolving power of 30 to 60 arcsec. The most famous stimulus in hyperacuity/Vernier tasks is the Vernier stimulus.

Vernier acuity is the ability to detect a slight offset between two vertically arranged lines (Figure 1b). Alternatively, a three-dot stimulus can be used. It consists of three vertically arranged dots with the middle dot slightly shifted to the right or left (Fig 1c, right). With this kind of stimulus, similar results are achieved (Garcia-Suarez, Barrett, & Pacey, 2004). Subjects are able to tell the direction of offset in such stimuli with supreme precision: the smallest offset usually seen is on the order of only a few seconds of arc, which equals the size of a 2-EURO coin spotted at a distance of about 2 kilometres. This example is estimated with help of the following calculation and experimental results of Klein and Levi (Klein & Levi, 1985):

$$\tan\frac{\alpha}{2} = \frac{d}{2D}$$

$$\Rightarrow D = \frac{26mm}{2\tan\frac{1}{7200}} = 5.3km \qquad \text{(26mm = 2 € coin diameter)}$$

Visual acuity thresholds are usually measured in minutes of arc (arcmin) of visual angle α and reciprocal angles, respectively. The angle α states the angular distance between two dots which is dependent on viewing distance. E.g. the two dots defining the gap in a Landolt-C are each represented on the retina. If the ‘real’ dots are connected with their equivalent dots represented on the retina two crossing lines are formed that enclose the angle α. To relate the viewing distance D to the gap between the dots d and the angle α, the following formula is used:

$$\tan\alpha = \frac{d}{D}$$

Because the result is given in degrees of visual angle, it has to be multiplied by 60 to get the result in minutes of arc (1° = 60’). The reciprocal angle of the result in arcmin (arcmin = ‘) is the value called visus which is the commonly used term for visual acuity (VA). 60 arcsec (arcsec = “) make one arcmin, so that one arcsec is the 3600th part of one degree of visual angle. Another commonly used scale is the logMAR/MAR (MAR = minimal angle of resolution) scale. As ‘minimal angle of resolution’ suggests, this is the gap size in arcmin and arcsec, respectively. logMAR is the logarithmical value of the minimal angle of resolution, which is equivalent to the negative logarithmic value of the reciprocal minimal angle of resolution.

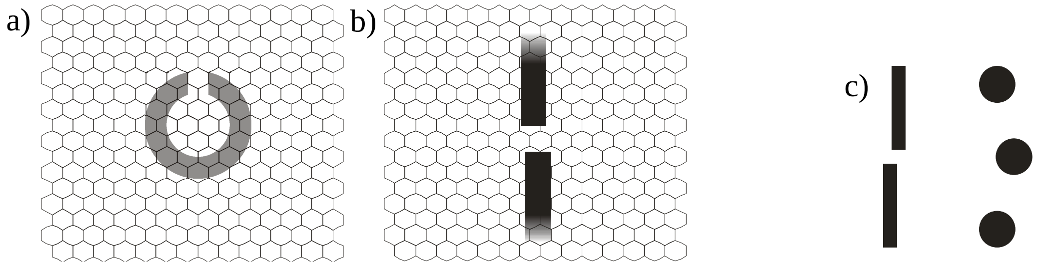

Figure 1: a) In conventional acuity tasks, receptor spacing limits acuity. b) Hyperacute thresholds are largely independent of the receptor lattice, possibly by neural summation and other integration processes. c) The classical Vernier stimulus (left) and three-dot stimulus (right) used in Vernier acuity tasks. Psychophysical thresholds in such tasks are only a fraction of the eye’s resolving power

In Vernier acuity tasks approx. 5 arcsec – a fraction of the eye’s normal resolving power – can be obtained under optimal viewing conditions (Glaser, 1999; Westheimer, 1987; Williams, Enoch, & Essock, 1984). The increased performance of observers in Vernier acuity tasks demonstrates that hyperacuity is fundamentally different from normal visual acuity. It is largely a product of the brain’s and not the eye’s resolving power. This is further illustrated by the fact that hyperacute resolving power is largely unaffected by optical interferences, such as amblyopia and cataracts (Enoch, Essock, & Williams, 1984; Essock, Williams, Enoch, & Raphael, 1984). Even optical blur does not have an effect on Vernier acuity that is as large as on resolution acuity.

In Figure 1a, the difference in mapping of Landolt stimuli and Vernier stimuli on the retina is illustrated. As one hexagon represents one photoreceptor in the retina, or more accurately in the fovea, one can see that maximum resolution in Landolt-C tests is consistent with photoreceptor diameter and photoreceptor spacing, respectively (Campbell & Green, 1965). In contrast to that, Vernier acuity thresholds correspond to approx. one fifth to one third of photoreceptor diameter and spacing, respectively (Beck & Schwartz, 1979). Two-dot resolving power is limited by the retinal photoreceptor lattice (Campbell & Green, 1965), whereas hyperacute resolving power is largely independent of photoreceptor arrangement and diameter, respectively (Fahle, 1991).

Vernier thresholds not only increase exponentially with decreasing contrast (< 40%) (Wehrhahn & Westheimer, 1990), but also increase with diminishing light intensity (von Campenhausen, 1993). Hyperacute resolution is possible under monocular as well as binocular viewing conditions. Performance under binocular viewing conditions is a factor $\sqrt{2}$ (≈ 1.4) better than under monocular viewing conditions (Banton & Levi, 1991). Thus, in an experimental setup light intensity and contrast should be kept on a high level in order to achieve good hyperacuity thresholds. Furthermore, it is advisable to run hyperacuity tests under binocular viewing conditions.

Not separation but eccentricity seems to be the variable limiting thresholds in hyperacuity tasks (Mckee, Welch, Taylor, & Bowne, 1990), which is a measure of shifting the stimulus out of the *fovea centralis*. Cellular density in the retina decreases with increasing eccentricity – that means that visual acuteness decreases with increasing eccentricity as well. (Mckee, Welch, Taylor, & Bowne, 1990). There is no stationary retinal image needed to achieve good acuity. It seems that image motion in a defined scope is no problem in hyperacuity tasks (Westheimer & McKee, 1975). Especially translation (parallel shifting) has no noteworthy effect on Vernier acuity. In contrast to that changes in orientation cause a remarkable increase in Vernier thresholds (Carney, Amnon Silverstein, & Klein, 1995), which shows that orientation is important in Vernier acuity tasks.

Vernier acuity seems to be a product of processing in the magnocellular (MC) pathway as experiments with macaques as well as with humans show (Lee, Wehrhahn, Westheimer, & Kremers, 1995; Wehrhahn & Westheimer, 1990). At low contrasts the MC pathway delivers more advantageous signal/noise ratio than the parvocellular (PC) pathway does. At high contrasts both, the MC and PC pathway, provide sufficient information for Vernier acuity. By means of further experiments with MC and PC pathways one can assume that the MC pathway is the more important part in Vernier acuity processing.

Adults achieve lower thresholds in Vernier acuity than children, what may have to do with the children's shorter attention spans. The training effect in Vernier tasks is not unimportant (Mckee & Westheimer, 1978) particularly with regard to extensive studies with many experimental runs.

As Abbud and Cruz showed a trial number of approx. 200 to 450 in an age group of 21 to 30 years old subjects would be advisable in order to get reasonable test-retest reliability (Abbud & Cruz, 2002). Test-retest reliability (= repeatability) is the variation in measurements under the same conditions:

- the same subject
- the same location
- the same measuring instruments, used under the same conditions
- the same measurement procedure
- repetitions over a short period of time

(http://www.statistics.com/resources/glossary/r/repeatbty.php; September, 11th, 2010)

A usual measure for repeatability is the coefficient of variation (CV).

1.2 Psychophysical methods

Psychophysics relates subjective perception to physical stimuli. Principally, psychophysical experiments are aimed at finding an observer's threshold for a given perceptual task. There are two types of thresholds that have to be distinguished: Absolute thresholds and differential thresholds (JND = just noticeable difference). An absolute threshold is the smallest detectable stimulus, whereas a differential threshold is the smallest detectable difference between a reference and a secondary stimulus level. In the present study, only absolute thresholds were measured.

Because a single subject reacts differently to stimuli of the same amplitude and under the same viewing conditions, psychophysical thresholds have to be calculated with help of statistical methods. In order to relate stimulus intensity to the percentage of perceived stimuli − which is equivalent to the percentage of correct responses in an experiment – psychometric functions are used. In an ideal case, a psychometric function shows a sigmoid characteristic with asymptotes of perceptual performance at very high and very low stimulus intensities. The threshold is then defined as the stimulus intensity at the most dynamic region of that function – at the inflection point. The psychometric function (and hence the threshold), is influenced by the decision behaviour of the subject. If the subject favours one response possibility over another, this so-called response *bias* has a most notably effect on the decision behaviour in small stimulus amplitudes near the subject's threshold. In order to detect a subject's bias in detection tasks with yes/no paradigms, *catch trials* with zero stimulus amplitude are performed. For threshold estimation, the bias can then be taken into account to get more precise results. Another way to circumwent response bias is to perform n-alternative forced choice experiments. To keep up the subject's motivation, *easy trials* can be embedded that are well above threshold.

To measure a subject's threshold, different types of tests can be performed. In so-called *n-AFC* (n-alternative forced choice) paradigms the subject has to choose one out of n stimulus alternatives. Landolt-C tests are usually 8-AFC tasks that provide very exact results with very low coefficients of variation (CV) of about 0.1 to 0.15 (CV is calculated by the standard deviation divided by the average of repeated measurements) (Bach, 2007). The likelihood to make a correct guess in a single trial is 12.5%, which is a quite low rate in comparison to other tests, with less response possibilities. One can conclude that this low guessing rate is responsible for low CVs between repeated experimental runs. On the other hand, in 2-AFC paradigms, the trial number could be increased to achieve eligible repeatability (Abbud & Cruz, 2002) – but also note that a high trial number could have an effect on the subjects' motivation and attention and could also introduce training and learning effects (Poggio, Fahle, & Edelman, 1992). The commonly used Vernier task is by design a 2-AFC (2-alternative forced choice) task. Thus, the subjects have a fifty percent chance to make a correct guess in each trial. As Jäkel and Wichmann pointed out higher numbers of alternatives in n-AFC paradigms are more efficient, in general. Furthermore, 4-AFC tests are most efficient in naive observers. In this context efficiency 'is the time needed to collect sufficient data per psychometric function given a desired precision target' (J<unicode>228</unicode>kel & Wichmann, 2006).

1.3 Aim of the thesis

The aim of the thesis was to create a stimulus configuration and experimental procedure which provide reliable results with hyperacute thresholds after a maximum of 36 trials. I chose a maximum of 36 trials, because I would like to keep trial numbers on a low level in order to avoid concentration loss or training effects, but on the other hand intended to achieve reliable results.

The first aim was to create a stimulus that provides at least four (but eight at the most) response possibilities. It was important that the test could be done effortless. If a test required a high degree of concentration, attention could have decreased during a run – that would have distorted the outcome of the test. Thus, the second goal was to create a stimulus configuration and testing procedure that could be conducted with the highest subjective easiness possible. After a phase of pilot testing, centroid stimuli seemed to be promising with regard to the three criteria:

- Subjective easiness
- Hyperacute thresholds
- High repeatability

The created test with highest repeatability on the one hand and lowest thresholds and subjective easiness on the other hand might help and facilitate evaluating hyperacuity in a clinical daily routine.

2. Material and methods

2.3 Subjects and procedure

In our experiments 17 subjects, thereof 10 male and 7 female, have been tested. Most of them were students and untrained in doing hyperacuity/visual tests. Subjects were at the age of 21-26 years.

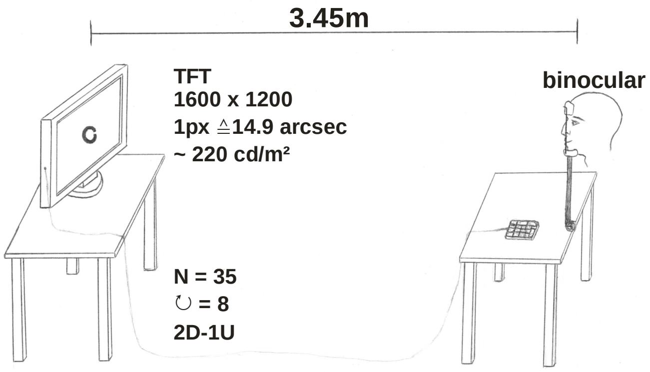

Figure 2: Sketch of the experimental setup. Subjects were placed in front of a TFT display with a resolution of 1600 x 1200 pixels - 1 pixel corresponds to approx. 14.9 arcsec. Heads were fixed with help of a forehead and chin rest. All experimental runs were performed under binocular viewing conditions and with 3.45m viewing distance. Luminance was measured before each subject started his/her experiments. Average luminance was 220cd/m².

The subjects' heads were fixed with help of a head and chin rest to ensure constant viewing distance of 3.45m throughout all experimental runs. Experiments were run under binocular viewing conditions. The blinds in the experimenting room were drawn down and the lights were switched on to assure as constant as possible lighting conditions. Before an experiment with another subject was started luminance was measured with a luminance meter. Average luminance was measured to be 219.4cd/m², ranging from 200.1 to 241.9 cd/m².

The used TFT-display ('DELL') had a screen resolution of 1600x1200. Thus, one pixel corresponded to 14.9 arcsec. Because of a fixed pixel size and the fact that a pixel can only be switched on or off the method of antialiasing was required for adequate stimulus presentation.

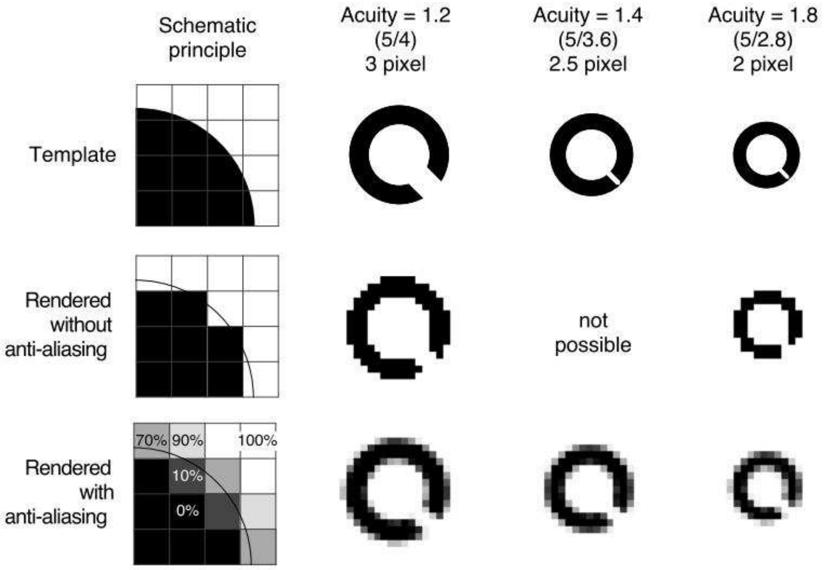

Figure 3: Antialiasing. Top = ideal shape; Center = shape rendered on a pixel raster without antialiasing; Bottom = smoother shapes with antialiasing; (Figure from (Bach, 1996)).

The problem in stimulus presentation on a computer screen is that a pixel that would be 60% black in a schematic condition with perfectly smooth edges (see 'template' condition in Figure 3) is actually 100% black in a condition without antialiasing. If a pixel would be black less than 50%, it is white. This is why the shape would have a jaggy outer appearance (stair shape – see 'rendered without anti-aliasing' condition in Figure 3). To avoid a coarse display of stimuli antialiasing is used: The pixels' grey value is adjusted to gain more spatial information. The grey value is then equivalent with the percentage of black area in one pixel (compare with bottom part of Figure 3).

In my experiments the antialiasing effect was very important, because I had relatively small stimuli and relatively short viewing distance. As the threshold for hyperacuity tasks is in the range of 5 to 7 arcsec (approx. one third of a pixel) presentation of stimuli with small amplitudes would not have been possible without antialiasing. Especially, if small figures (e.g. the dots in the FiveDot configuration) are created it is important to have an as large as possible viewing distance.

Each subject was tested for his/her normal visual acuity in three runs of Landolt-C tests with the FrACT to be sure that no visual interferences exist that could distort the results. Ametropic subjects wore their corrective lenses throughout all experimental runs so that I could consider them to be emmetropic. In the next step the other four stimuli were tested in eight runs with 35 trials, each. The order of the presented stimuli was varied to rule out that loss of concentration has a remarkable influence on the outcome of the test. The first group was asked to respond to the tasks in the following order: Landolt, Vernier, FiveDot, Circle, Square. The second group responded to stimuli in the order: Landolt, Square, Circle, FiveDot, Vernier. Note that each of the two groups performed tests in only one of the two sequences.

The subjects' information regarding subjective simplicity for each stimulus have been documented. They were asked to evaluate each stimulus after all experimental runs with help of a chart, by ticking the box that fitted their evaluation best. The scale reached from 1, representing 'very easy' to 6, representing 'very hard'

Table 1: Evaluation chart

	easy					hard
Rating **Stimulus**	**1**	**2**	**3**	**4**	**5**	**6**
Landolt						
Vernier						
FiveDot						
Square						
Circle						

Subjects were allowed to pause whenever they wanted – in case they get tired or they would recognize their performance to get worse due to dreary testing procedure.

The entire testing procedure took about 30 to 50 minutes, very much depending on the subjects themselves. Eventually occurring risks have been explained to all our subjects. None of our subjects aborted the test, after we had explained that they could break off whenever they want and without any reason. All applied experimental procedures met the tenets of the Declaration of Helsinki.

2.2 Stimuli and data analysis

The new stimuli have been created with the Open GL Utility Kit (GLUT) in an ANSI-C computer language environment. Specifically, four stimuli have been created (Figure 4).

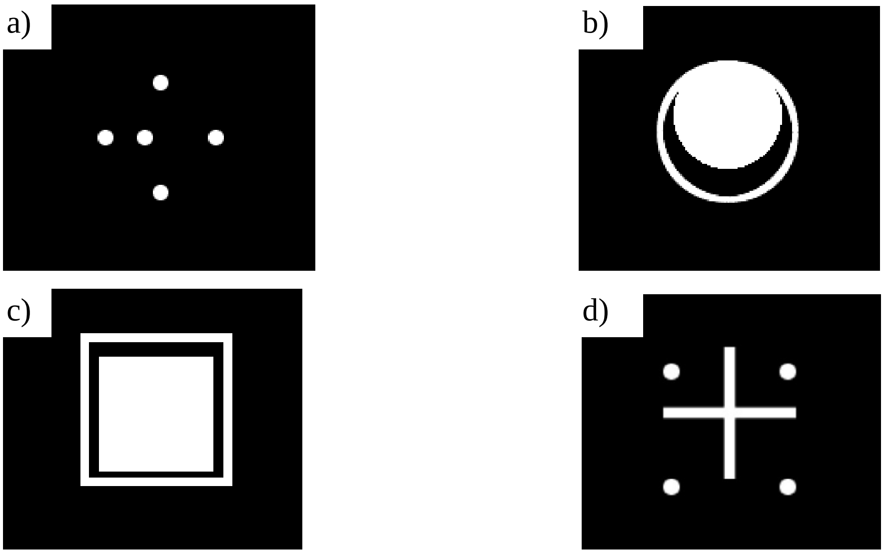

Figure 4: a) Screenshot FiveDot stimulus; the four outer dots remained static, while the dot in the middle was shifted from trial to trial. b) Screenshot Circle stimulus; the inner, filled circle was shifted from trial to trial, while the outer circle remained static. c) Screenshot Square stimulus; the inner, filled square was shifted from trial to trial, while the outer square remained static. d) Screenshot Cross stimulus; the four outer dots remained static, while the cross in the middle was shifted from trial to trial.

In my tests I presented the stimuli ‘white on black’ as illustrated in Figure 4. Only three of these four stimuli have been used for our experiments (FiveDot, Circle, Square), because thresholds in the Cross stimulus were not in a hyperacute range during pilot testing. In the FiveDot stimulus task the four outer dots remained static – only the position of the middle dot was varied (shifted up, down, left or right). In the Circle configuration the outer outlined circle remained static and the inner filled circle was shifted up, down, left or right – equivalently in the Square stimulus. In the Cross stimulus the four dots remained static and the cross was shifted up, down, left or right. The Cross stimulus turned out not to be qualified for hyperacuity tests after some pilot testing with four subjects, because the obtained thresholds were not in the hyperacute range. Figure 5 illustrates the specific spatial parameters of the different stimuli that were used in our tests.

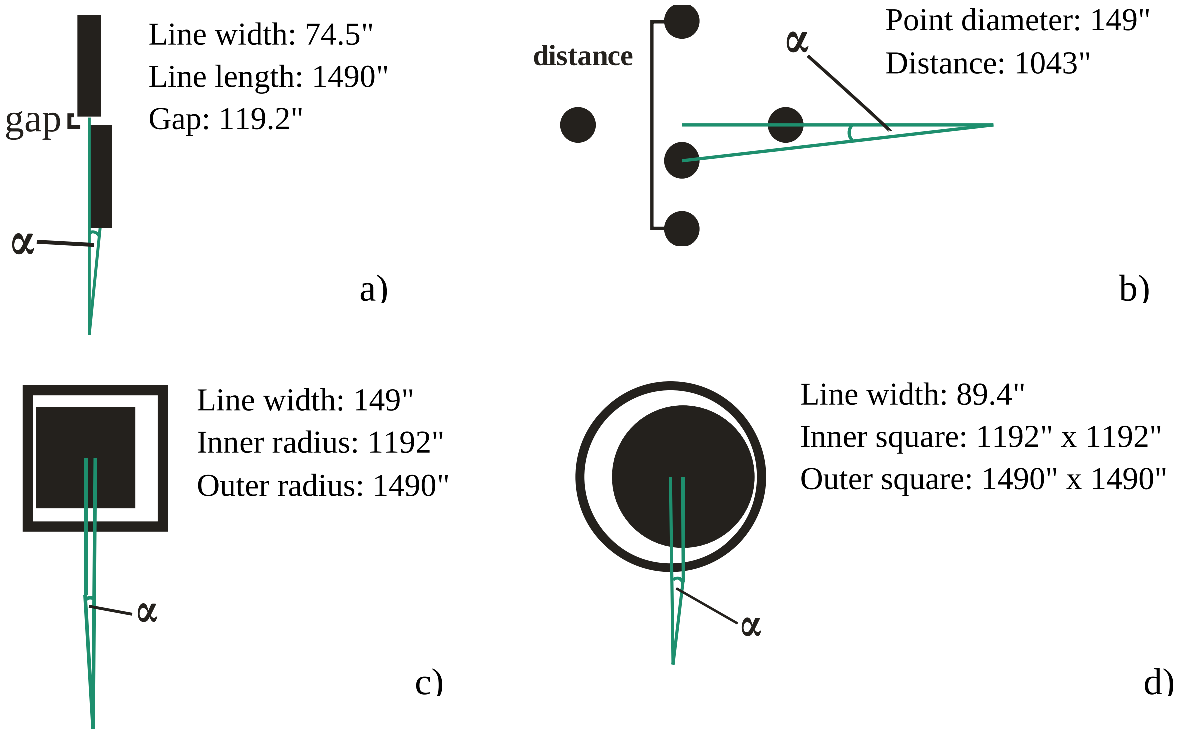

Figure 5: Distances, line widths and radii are stated in arcsec (") to be able to better evaluate the stimulus amplitudes in comparison to the dimension of the stimuli. The green lines have their starting points in the centres of the static and the shifted parts of the stimuli, respectively. The angle α states the shifting of the moveable part of the stimulus from the static part of the stimulus. a) Parameters of Vernier stimulus. b) Parameters of FiveDot stimulus. c) Parameters of Square stimulus. d) Parameters of Circle stimulus.

Visual C++ has not only been used for programming stimulus configurations, but also for the testing procedure itself. The program uses a similar procedure as the FrACT (introduced in 2.3 FrACT). We used a 2up-1down staircase function. That means that the stimulus amplitude was decreased if the subject gave correct responses in two successive trials, but increased if the subject gave an incorrect response.

The data obtained were processed with help of *MATLAB* which delivered thresholds, psychometric functions, staircase functions, coefficients of variation and Bland-Altman-plots. The staircase function showed the increment/decrement of the stimulus intensity in logarithmical plotting. Additionally the answering behaviour of the subject could be retraced. The coefficient of variation was needed in order to evaluate test repeatability, thus, it was the most informative value in my experiments. The coefficient of variation is the quotient $\frac{s}{\bar{x}}$, with s = standard deviation and x = average of repeated measurements. A low CV represents a low deviation from the mean (for positive means only), thus, a low CV is what we aim for.

To get a visualisation of repeatability, a ‘Bland-Altman-plot’ (BA plot; difference plot) is useful. In this kind of plot, two consecutive runs are paired and their mean value is plotted against their difference.

2.3 FrACT

The FrACT (Freiburg Visual Acuity and Contrast Test) is a computer program that generates visual stimuli to measure Landolt-C acuity, Vernier acuity and contrast thresholds (Bach, 1996). Because it provides good results in an easy way rather quickly, we used it to compare our results with.

The Landolt-C task is an 8-AFC test, whereas Vernier acuity is estimated via a 2-AFC Vernier task. The program works with the *bestPEST* algorithm (PEST = Parameter Estimation by Sequential Testing). The stimulus intensity decreases in a specific magnitude if the subject gives a correct answer, but increases if the subject gives an incorrect answer. In the beginning of a run, the difference between the tested stimulus amplitudes is rather high, but typically decreases in order to approximate the threshold (*bracketing*). With help of this method, the threshold is narrowed down from above and from beneath. After each trial the most likely threshold and the resulting stimulus amplitude that has to be tested is calculated with the aim to create stimulus intensities near the threshold to get an as precise as possible estimation of the subject’s threshold. Every sixth trial is an easy trial to keep up the subject’s motivation. The psychometric function is estimated via a maximum-likelihood fit. The threshold is read off at the so-called *inflection point* which is the point on the y-axes (probability of correct report) exactly in between of the guessing rate and 100%. The threshold has to be read off at the point on the x-axes (stimulus amplitude) where the psychometric function and the inflection point cross. E.g. the inflection point in an 8-AFC experiment would be at 56.25% correct responses, in a 4-AFC experiment at 62.5%.

4-AFC: $25\% + (\frac{100\% - 25\%}{2}) = 62{,}5\%$ (25% represents the guessing rate)

The results of Landolt-C tests are displayed in *logMAR* which is the negative logarithm of the visual acuity (logMAR = -log (VA)). Visual acuity (VA) is the reciprocal value of the angle α.

3. Results

Each subject did 72 trials in Landolt tasks and 280 trials in each of the other tasks, thus, every subject did 1192 trials, overall. In Landolt tasks there were 1224 trials performed on the whole. In the remaining stimulus configurations 4760 trials have been performed, each. This is a total number of 20264 trials (compare Table 2).

Table 2: Number of trials across all experimental conditions

	Trials per person	**Trials per stimulus**
Landolt	72	1224
Vernier	280	4760
FiveDot	280	4760
Square	280	4760
Circle	280	4760
Total	1192	20264

The raw data were processed in MATLAB. Three graphs were generated for every run and every subject. In the first kind of graph the number of trials of one stimulus amplitude was plotted as a bar graph. The second type of graph was a psychometric function in which thresholds were included and the third type was a staircase function in which one could reconstruct the response behavior of the subject. The bar graph was used to see how often one specific stimulus amplitude has been tested. In an ideal case this graph has the shape of a pyramid. That means that one or two stimulus amplitudes (those near the threshold) have been tested more often than those stimulus amplitudes to the right or left (compare Figure 6). A typical psychometric function has a sigmoidal run. For every psychometric function 300 simulations have been performed by MATLAB (compare Figure 7). The staircase function tells about the subject’s answering behavior and stimulus amplitudes during an experimental run. Stimulus amplitude was plotted logarithmically on the y-axis, thus, one can see that stimulus amplitude decreased at the beginning, if the subject responded correctly. As a result of the used 2-down-1-up staircase function, stimulus amplitude decreased, if the subject made two correct responses in a row, but increased, if only one incorrect response occurred. The threshold was estimated by evaluating how often a correct or incorrect response was made for each stimulus amplitude, respectively.

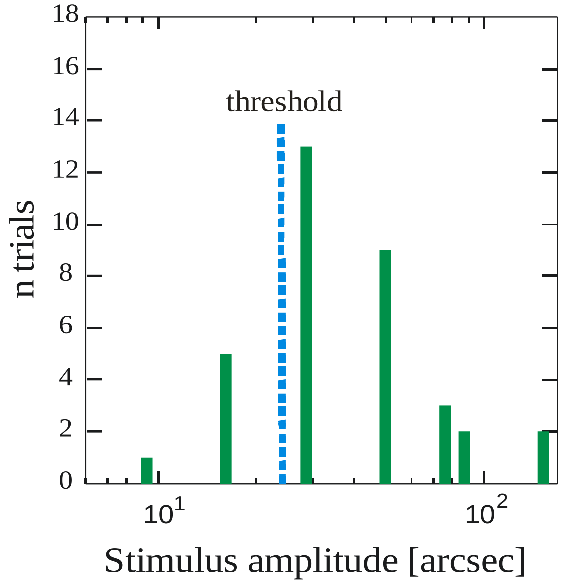

Figure 6: Exemplary histogram of numbers of trials performed at each stimulus amplitude; Subject VP, FiveDot task, run7; Note that most trials have been performed near threshold.

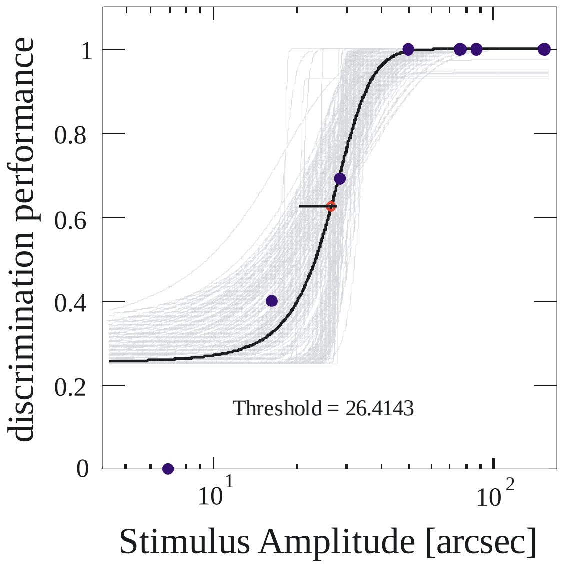

Figure 7: Exemplary psychometric function; Subject VP, FiveDot task, run 7; the black line presents the actual psychometric function, while the grey lines depict simulated psychometric functions performed by MATLAB. The blue dots show how much percent correct responses were measured at a specific stimulus amplitude. The red dot represents the inflection point, where the threshold is read up on the x-axis. The horizontal bar at the inflection point is the confidence interval.

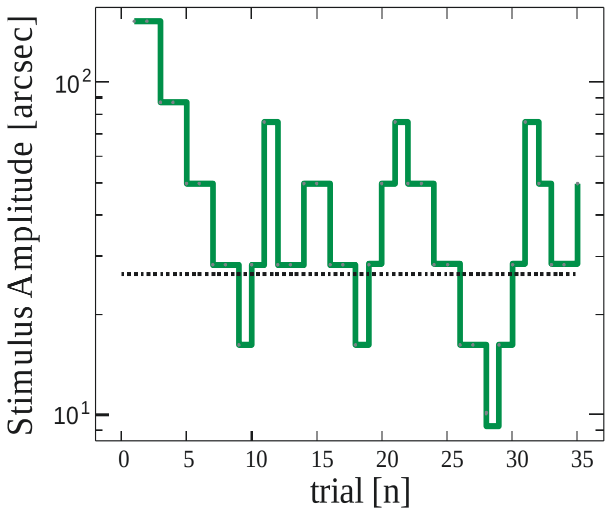

Figure 8: Exemplary staircase function; Subject VP, FiveDot task, run 7; The dotted line marks the position of the threshold on the y-axis. As a 2-down-1-up staircase procedure was used, a decreasing stimulus amplitude means that two correct responses were given. Contrary, an increasing stimulus amplitude stands for an incorrect response.

3.1 Acuity

All subjects were asked to wear their corrective lenses (contact lenses or spectacles) during all experimental runs, thus, they had normal or corrected-to-normal vision (mean = 34.7arcsec). All subjects could participate in the further experiments.

Three runs of Landolt-C tests were run per subject, thus, there are much less data points in this case. Thresholds in the Landolt configuration reached from 24.3 to 58.3 arcsec, in Vernier tasks results were in a range from 1.0 to 32.6 arcsec. In the FiveDot task the lowest value was 13.2 arcsec, the highest value 56.9 arcsec. In the Circle task thresholds reached from 3.5 to 20.3 arcsec, whereas thresholds in the Square task were in a range from 5.2 to 20.1 arcsec. The scatter in the FiveDot task was higher than in the other tasks (compare Figure 9). In the Landolt task the median was 33.7 arcsec, the lower and upper quartiles (Q_1 and Q_3) were at 29.3 and 39.1 arcsec, respectively. The distance between the upper and lower quartiles is the interquartile range (IQR), thus, the IQR in the Landolt task was 9.8 arcsec. The median in Vernier acuity tasks was at 9.0 arcsec. The IQR reached from 5.2 (Q_1) to 14.4 (Q_3) arcsec – that is equivalent with 9.2 arcsec. In the FiveDot task the median was at 25.8 arcsec, while the first quartile and the third quartile were at 20.2 and 33.4 arcsec, respectively, resulting in an IQR of 13.2 arcsec. The median in the Square configuration was at 9.2 arcsec, while the IQR reached from 7.6 (Q_1) to 12.8 (Q_3) arcsec – that is 5.2 arcsec. The IQR in the Circle test was in a range from 8.5 (Q_1) to 13.0 (Q_3) arcsec – that is equivalent with 4.5 arcsec. The median in this task was at 9.8 arcsec. There were no outliers in the Landolt acuity task (compare Figure 10). There was significant difference between thresholds of Vernier and Circle, highly significant difference between Vernier and FiveDot and no significant difference between Vernier and Square stimulus according to a Wilcoxon signed-rank test.

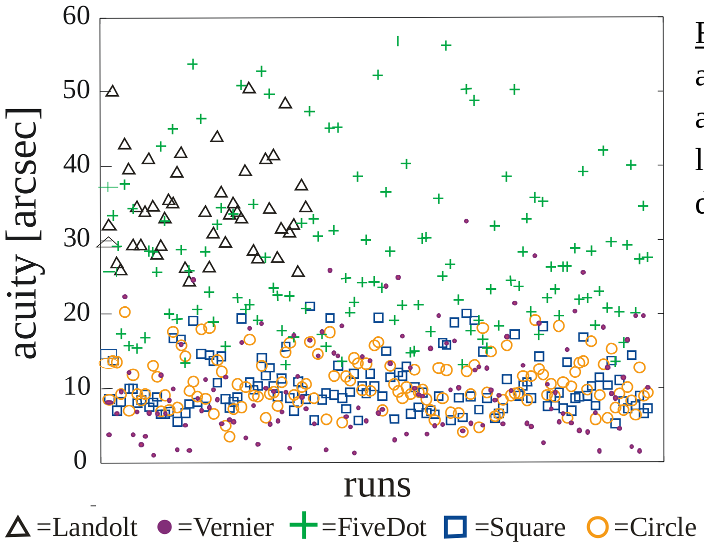

Figure 9: Thresholds of all runs across all subjects and experimental conditions are plotted. The different symbols/colours represent the experimental conditions.

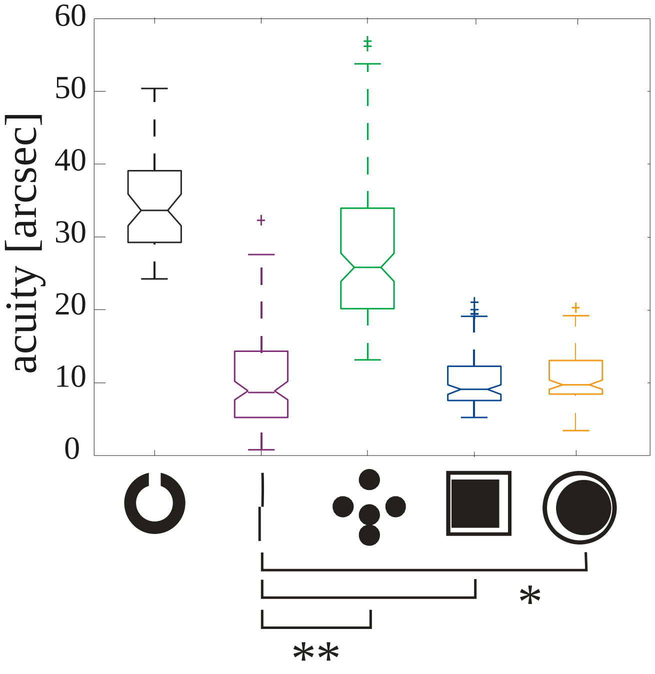

Figure 10: Box-whisker plots of thresholds across all experimental runs; in box-whisker plots the lines in the middle of the boxes represent the median, while the upper and lower borders of the box are the upper and lower quartiles, respectively. The lower (or first) quartile (= Q_1) states the 25% border and the upper (or third) quartile (= Q_3) states the 75% border. The interval between the Q_1 and the Q_3 is the interquartile range (IQR), which is a measure of spread. The whiskers are defined as the 1.5 fold IQR maximum – but not the exact value of the 1.5 fold IQR, but the last value that is within this range is taken as the end of the whisker. Data points lying outside the whiskers are called 'outliers'. Asterisks mark significant differences ($p \leq 0.05$ *; $p \leq 0.01$ **)

Table 3: Average thresholds across subjects; note that there were two groups and that each group performed tests in either sequence 1 or sequence 2. Asterisks denote statistical significance of a Wilcoxon signed-rank test across corresponding conditions (pairs marked in color); $p \leq 0.05$ * $p \leq 0.01$ **

	Landolt	**Vernier**	**FiveDot**	**Square**	**Circle**
Average all [arcsec]	34.7	10.3	28	10.6	10.6
Average female [arcsec]	34.5	11	**29.8**	10.2	10.6
Average male [arcsec]	34.8	9.7	**26.8***	10.9	10.7
Average sequence 1 [arcsec]	35.8	11.8	**31.9**	**12.3**	11.3
Average sequence 2 [arcsec]	33.7	8.9	**24.6****	**9.1****	10.1

Sequence 1: Landolt, Vernier, FiveDot, Circle, Square
Sequence 2: Landolt, Square, Circle, FiveDot, Vernier

Mean threshold in Landolt tasks was 34.7 arcsec while means in males and females were 34.8 and 34.5 arcsec, respectively. In sequence 1 a mean threshold of 35.8 arcsec was achieved, whereas the mean in sequence 2 was 33.7 arcsec. In Vernier acuity tests the mean threshold was 10.3 arcsec, thus, on average, Vernier thresholds were around 3.4 times lower (better) than Landolt acuity thresholds. In females and males mean thresholds of 11.0 and 9.7 arcsec were measured in Vernier acuity tasks, respectively. In sequence 1 the average threshold was 11.8 and in sequence 2 8.9 arcsec. In the FiveDot configuration mean threshold was 28.0

arcsec. Average in female subjects was 29.8 arcsec, whereas men had an average threshold of 26.8 arcsec. Mean threshold in sequence 1 and sequence 2 were 31.9 and 24.6 arcsec, respectively. In the Square configuration mean threshold was 10.6 arcsec. Females had an average threshold of 10.2 arcsec, males had an average threshold of 10.9 arcsec. In sequence 1 a mean threshold of 12.3 arcsec was measured, whereas the mean in sequence 2 was 9.1 arcsec. In the Circle acuity test a mean result of 10.6 arcsec was achieved. Women had an average acuity of 10.6 arcsec, while men had an average acuity of 10.7 arcsec. In sequence 1 the mean threshold was 11.3 arcsec and in sequence 2 mean threshold was measured 10.1 arcsec. Hyperacuity ratios of 1.2, 3.3 and 3.3 in the FiveDot, Square and Circle configuration, respectively, result from mean thresholds across all subjects.

Furthermore, as a Wilcoxon signed-rank test implicated, the differences in the FiveDot task between male and female on the one hand and between sequence 1 and sequence 2 on the other hand were significant (p = 0.0133) and highly significant (p = 0.001), respectively. Additionally, there was also highly significant difference in comparison of sequence 1 and sequence two in the Square configuration ($p = 2.77 \times 10^{-5}$).

3.2 Repeatability

The standard deviation (SD) and coefficient of variation (CV) were used to measure repeatability of the individual experimental conditions. SDs were highest in the FiveDot configuration (5.6 arcsec) and lowest in the Square task (1.0 arcsec). Landolt (2.8 arcsec), Vernier (4.3 arcsec) and Circle (2.8 arcsec) were lying in between. SD medians of Landolt, Square and Circle were in the same range, while SDs of Vernier and FiveDot stimulus were higher in comparison. The IQRs were highest in the Landolt and FiveDot configuration (2.7 and 2.3 arcsec, respectively) and in the same magnitude in the other three tasks: Vernier IQR = 0.9 arcsec, Square IQR = 1.0 arcsec, Circle IQR = 0.8 arcsec. Due to a Wilcoxon signed-rank test there was no significant difference in SDs, comparing Landolt and Vernier (p = 0.1212), Landolt and Square (p = 0.3705), Landolt and Circle (p = 0.7048) and Circle and Square (p = 0.2087). There was significant difference in comparison of Landolt and FiveDot (p = 0.0098; highly significant), Vernier and FiveDot (p = 0.0496; significant), Vernier and Square ($p = 1.14 \times 10^{-4}$; highly significant), Vernier and Circle ($p = 4.43 \times 10^{-4}$; highly significant), FiveDot and Square ($p = 1.95 \times 10^{-5}$; highly significant) and FiveDot and Circle (p = 0.0119; significant) (compare Figure 11a).

The CV median of Landolt tasks was lowest with 8.2%, whereas the CV in Vernier tasks was highest with 46.1%. The CVs of the other stimulus configurations were in the same range. The FiveDot median was at 18.9%, the Square median was at 22.5% and the Circle median was at 26.6%. The IQR in the Vernier task was largest and reached from 35.9% to 54.6% (width = 18.7%). IQRs of Landolt, FiveDot, Square and Circle configurations were 5.7%, 9.4%, 6.5% and 6.5% wide, respectively. Most differences in CVs were significant, except for the combinations Circle/Square and FiveDot/Square: Landolt/Vernier = highly significant (p = 1.42 x 10^{-6}); Landolt/FiveDot = highly significant (p = 2.61 x 10^{-4}); Landolt/Square = highly significant (p = 1.95 x 10^{-5}); Landolt/Circle = highly significant (p = 5.45 x 10^{-6}); Vernier/FiveDot = highly significant (p = 1.04 x 10^{-5}), Vernier/Square = highly significant (p = 6.46 x 10^{-5}) and Vernier/Circle = highly significant (p = 1.74 x 10^{-4}) (compare Figure 11b).

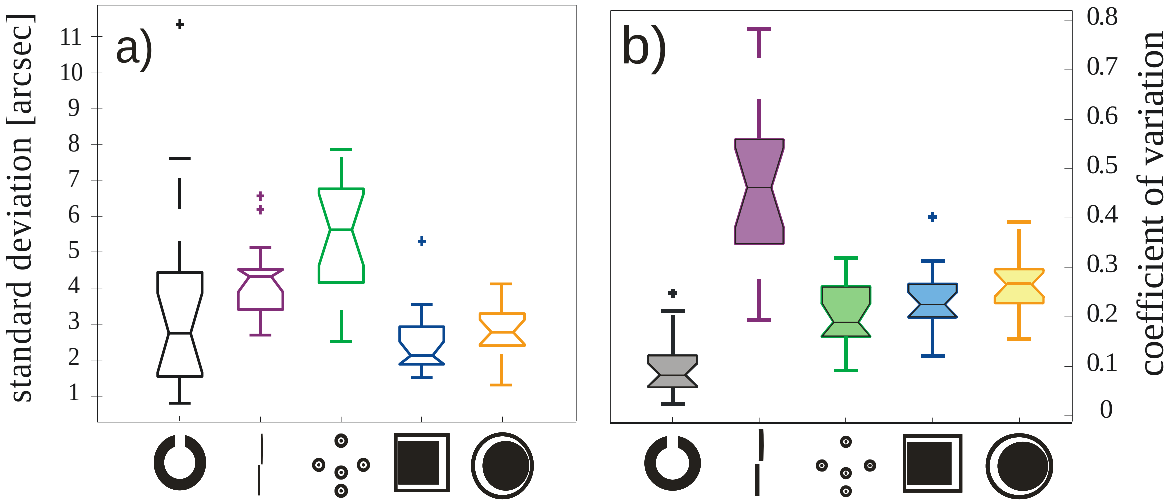

Figure 11: a) Boxplots of standard deviations across all subjects for each stimulus configuration; b) Boxplots of coefficients of variation across all subjects for each stimulus configuration.

All data of the different stimuli have been plotted in Bland-Altmann plots (compare Figure 12). An important information in BA-plots is given by the 95% mark (mean + SD x 1.96 and mean + SD x 1.96, respectively) which states the limit of agreement. In the Square configuration the 95% mark was at 6.1 and -6.1, respectively, which is equivalent with the lowest spread. In case of the Circle configuration the 95% mark was at 8.0 and -8.0, respectively and in the FiveDot configuration at 15.3 and -15.3, respectively. The Vernier stimulus lay somewhere in between with 10.4 and -10.4 representing the 95% mark, respectively. The spread of the data points did not show any special arrangement (compare Figure 12).

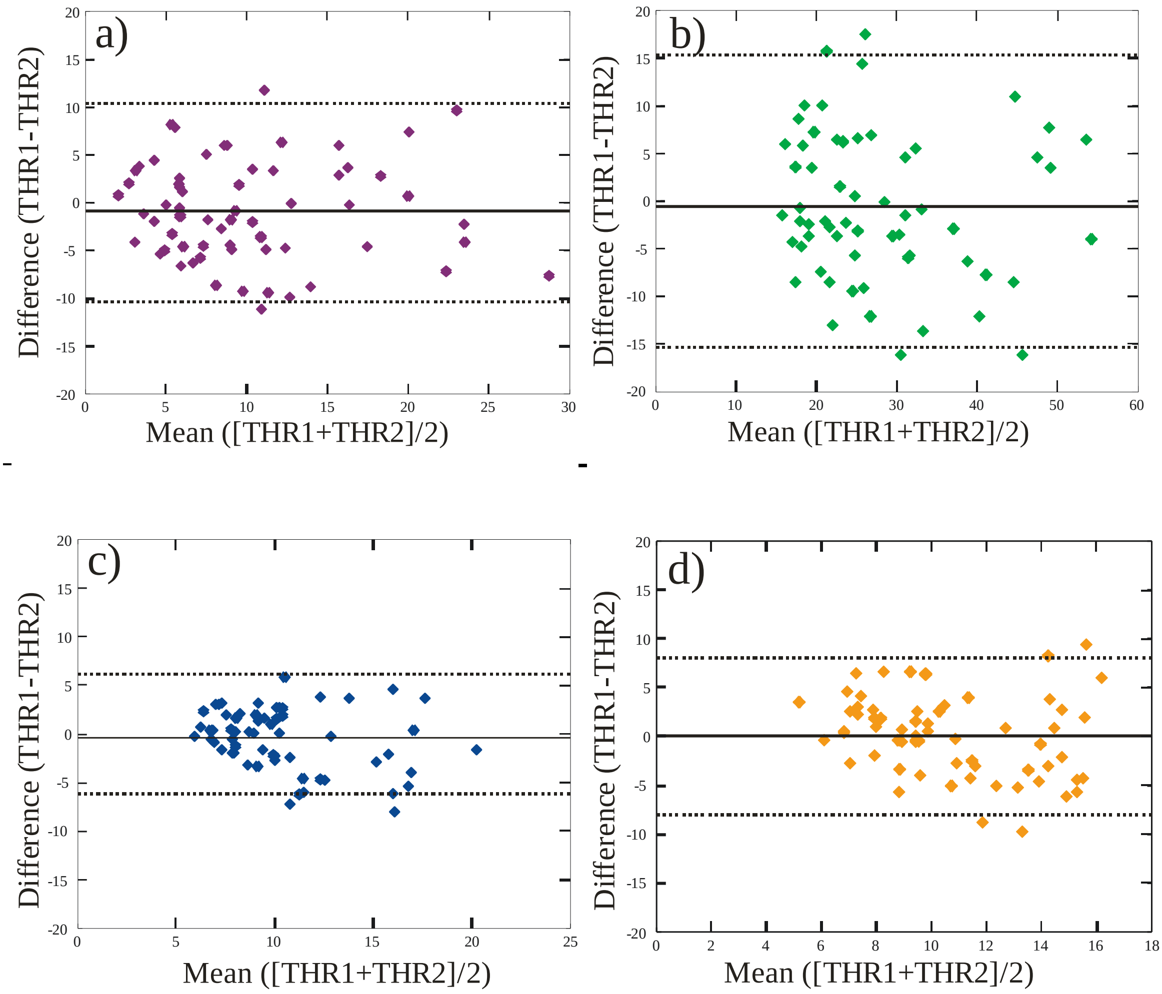

Figure 12: The continuous black line is the mean difference, while the upper and lower dotted lines are the 95% marks (mean + SD x 1.96 and mean - SD x 1.96, respectively) a) (purple) BA-plot Vernier, n = 63; b) (green) BA-plot FiveDot, n = 61; c) (blue) BA-plot Square, n = 62; d) (yellow) BA-plot Circle, n = 64

3.3 Subjective scores

The subjects were asked to evaluate the different stimuli with regard to subjective simplicity after they had run all tests. The scale reached from 1 to 6, with 1 representing ‘very easy’ and 6 representing ‘very hard’. They were not asked to guess how good they performed in the tests in particular, but should state how exhausting they found the tests were.

All tests were rated between 2.4 and 4.5 in the mean (compare Table 4). The test with lowest (easiest) mean rating was the Vernier test, whereas the test with highest (hardest) mean rating was the FiveDot configuration. Landolt, Square and Circle tests were rated almost the same (3.2, 3.3 and 3.2, respectively). There were no significant differences in comparison of female/male or sequence 1/sequence 2. Nevertheless, there was a difference in evaluation of the FiveDot stimulus as well in evaluation of men (4.0) and women (5.1) as in sequence 1 (5.0) and sequence 2 (4.0). The Landolt stimulus was rated 3.3 by females and 3.2 by males. The Vernier configuration was rated 2.3 and 2.5 by females and males, respectively. Subjects in sequence 1 rated the Vernier stimulus 2.1, subjects in sequence 2 2.6. Women rated the Square acuity test 3.6, whereas men rated the same stimulus 3.0. The same stimulus was rated 2.9 and 3.6 by sequence 1 and sequence 2, respectively. The Circle configuration was rated 3.1, 3.2, 3.0 and 3.3 by females, males, sequence 1 and sequence 2, respectively (compare Table 4). Personal communication showed that in former hyperacuity tests the Vernier task was rated harder than the Landolt task – but there was no written evaluation (Harmening 2010).

Table 4: Mean rating responses across all subjects and across sequence 1/sequence 2 and males/females. Note that there were no significant differences between the groups of male/female or sequence 1/ sequence 2.

	Landolt	**Vernier**	**FiveDot**	**Square**	**Circle**
Average all	3.2	2.4	4.5	3.3	3.2
Average female	3.3	2.3	5.1	3.6	3.1
Average male	3.2	2.5	4	3	3.2
Average sequence 1	-	2.1	5	2.9	3
Average sequence 2	-	2.6	4	3.6	3.3

Sequence 1: Landolt, Vernier, FiveDot, Circle, Square $p \leq 0.05$ *

Sequence 2: Landolt, Square, Circle, FiveDot, Vernier $p \leq 0.01$ **

In order to see if there was a correlation between obtained thresholds and subjective scores, rating responses of all subjects and mean thresholds of all subjects have been plotted (compare Figure 13). Generally, ratings of 1 or 6 were rare in comparison to ratings of 2, 3, 4, and 5 as there are n = 10 datapoints for rating 6 and 1, but n = 24, n = 40, n= 34 and n = 28 for ratings 2, 3, 4 and 5, respectively. The medians of the thresholds in tasks rated 1 to 4 were lower (between 8 and 15 arcsec) than the medians in thresholds rated 5 or 6 (approx. 26 to 27 arcsec).

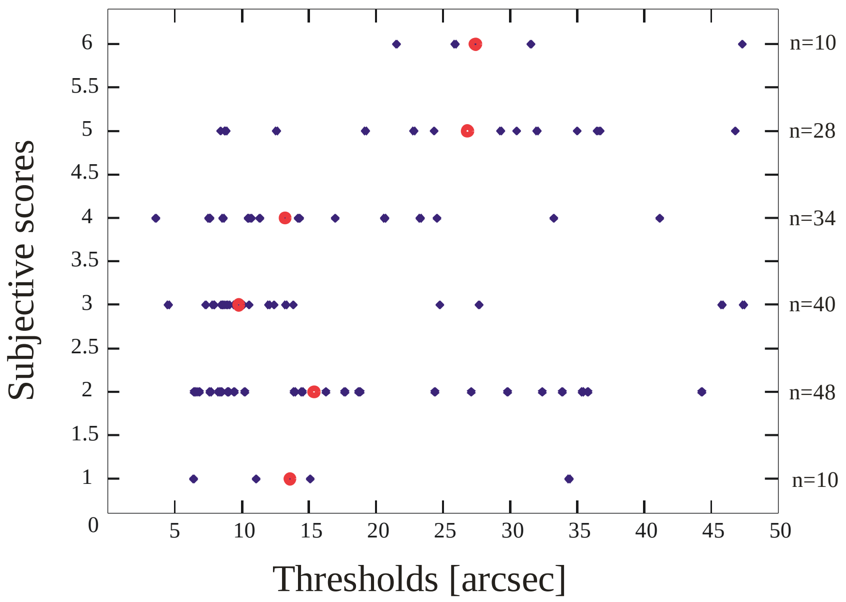

Figure 13: Correlation between subjective scores and actual thresholds; Red dots state medians, blue dots state mean thresholds of different subjects.

4. Discussion

The aim of the thesis was to find at least one stimulus that provides hyperacute thresholds, but with better repeatability as the common Vernier stimulus and effortless performance. The outcome of the experiments is satisfying, because I indeed found two stimuli that are qualified to replace the usual Vernier stimulus.

4.1 Acuity

In my experiments it was confirmed that Landolt acuity/2-dot resolving power is in the range of 1 arcmin, whereas results in hyperacuity/Vernier tasks are in the magnitude of a few arcsec (Westheimer, 1987). My subjects had above average visual acuity of 34.7 arcsec, which may have been due to low average age and beneficial viewing conditions. The results in my new stimuli regarding acuity were not exactly as I expected. The Circle and Square stimulus worked quite well, as thresholds in these tasks were in a hyperacute range. Contrary, the FiveDot stimulus did not meet our expectations regarding acuity. Due to the fact that the design was developed from the 3-dot Vernier, which is known to provide thresholds very similar to the Vernier stimulus (Beck & Schwartz, 1979), we expected this stimulus to be in a hyperacute range as well. One explanation for worse performance in this task might be that the perceptual task was somehow different. In Circle and Square stimulus the subjects had more 'references' in the form of the outer square/circle to compare the inner square/circle with. In case of the Square they had four couples of edges to evaluate the shifting of the inner square – similar in case of the Circle, where they had plenty of points at the edges that could be analyzed. In the FiveDot configuration the task was insofar different that the subjects imagined a cross (formed of the four outer dots) and tried to figure out if the inner dot was shifted up, down, left or right in relation to this cross. So there was much less reference than in the two stimulus configurations mentioned before. According to the Wilcoxon signed-rank test the difference between the FiveDot task and the Vernier task was highly significant ($p = 1.11 \times 10^{-34}$), as one would expect considering the outcome of these two stimulus configurations. Unexpectedly, the outcome of the Circle test showed a significant difference in comparison to the Vernier stimulus ($p = 0.0309$), as well, while the Square outcome was not significantly different from the Vernier outcome ($p = 0.0748$).

The Cross stimulus had already been sorted out after pilot testing (Figure 4d). Results in this kind of test were that bad that I did not want to test the stimulus on our subjects. I wanted to keep the time for testing procedure as limited as possible. The Cross stimulus was even worse than the FiveDot stimulus. In this task, the relative position of the cross should be judged. Maybe the problem in this task was that there were no direct references. One could try to connect the four outer dots, so that they would either form an ‘X’ or a square. In both cases there are no direct references that could be used to compare the cross with as there were in Circle and Square configurations.

Maybe performance could generally be improved if the perfect relationships in size and spacing could be determined (Sullivan, Oatley, & Sutherland, 1972). I only tried to find optimal size and spacing in some pilot testing – thus, the size and spacing I considered to be optimal was not confirmed statistically.

Comparing performance in men/women and sequence1/sequence 2 showed that there were no remarkable differences in the Landolt task, Vernier task and Circle task. In contrast to that there was a highly significant difference in average thresholds of sequence 1 and sequence 2 (but not in male and female) in the Square task. Performance in sequence 2 was 3.2 arcsec better than in sequence 1. In the FiveDot task there were remarkable differences as well in comparison of male and female (significant) as in sequence 1 and sequence 2 (highly significant). In this case males were 3.0 arcsec better than women and subjects in sequence 2 had a 7.3 arcsec improved performance in comparison to subjects in sequence 1.

The subjects that participated in the runs of sequence 1 were tested on the first two days of my experimental runs, whereas the subjects that participated in the runs of sequence 2 were all tested on the third day of my experimental runs. Thus, maybe weather or other outside influences had an effect on the subjects’ concentration and/or motivation. Another explanation for improved performance in sequence 2 in the FiveDot task may be that there was a training effect as it was generally observed in hyperacuity tasks (Poggio, Fahle, & Edelman, 1992). In sequence 1 the FiveDot stimulus was on position three, while in sequence 2 it was on position four. The difference between these two positions was quite small so that one would not consider this difference to play such a major role, but on the other hand a combination of training effect and outside influences could have been responsible for this remarkable difference in thresholds. The same explanation could be used for the difference in sequence 1 and sequence 2 looking at the Square stimulus. The difference in the FiveDot stimulus between male and female is harder to explain and not as remarkable as in comparison of the two sequences. Some of the male participants explained that they did not find the FiveDot stimulus very hard (see 3.3 Evaluation), because they imagined a crosshair and then tried to judge whether the dot in the middle was shifted up, down, left or right. Female participants found this stimulus more exhausting. To get an appropriate explanation for this remarkable difference in male and female performance further studies would be necessary.

Concerning the spread of the thresholds in the different tasks the boxplots in Figure 10 should be observed. The IQR in the FiveDot task was largest with 13.2 arcsec. As one could assume looking at Figure 9, the IQR confirmed the very high scatter in the FiveDot task. The IQR in Landolt and Vernier task, 9.8 and 9.2 arcsec, respectively, were in the same range. The spread was lower in comparison to the FiveDot task, but higher than in Circle and Square stimulus, where IQRs were 4.5 and 5.2 arcsec, respectively. Obviously, the spread was independent of response possibilities, as Landolt, Vernier and the other three stimuli have different numbers of response possibilities. The spread also seemed to be independent of the magnitude of the stimulus, comparing the Vernier task and Circle and Square task. At the moment there is no good explanation what the scatter in these tasks depended on.

4.2 Repeatability

In order to achieve better test repeatability we created stimulus configurations with a 4-AFC paradigm, instead of the usual 2-AFC Vernier stimulus. The 4-AFC configuration is considered to be the most efficient in naive observers (Jäkel & Wichmann, 2006). However, our expectations to the new stimuli have not exactly been met, since our aim was to get CVs lower than 10% with our new stimuli.

Generally, SD values were very similar to SDs in former experiments (Garcia-Suarez, Barrett, & Pacey, 2004) in the Vernier task. In the Landolt configuration the SD median was 2.8 arcsec, while the IQR was 2.7 arcsec wide. A similar IQR (2.3 arcsec), but higher SD median (5.6 arcsec) was found in the FiveDot configuration. The IQRs of the Vernier, Square and Circle stimulus, and hence the spread, were very similar (0.9, 1.0 and 0.8 arcsec, respectively). The SD medians in Landolt, Square and Circle were in the same magnitude, as well. Again, there was no obvious relation between spread and absolute SD values or response possibilities. There may have been a connection between absolute thresholds and the SD spread, as Landolt and FiveDot SD boxplots suggest. In both tasks thresholds were remarkably higher than in the Vernier task – and so was the spread of the SD. Maybe there were more widely scattered SDs in tasks with relatively high thresholds, in general.

The most important information on repeatability were given by the CVs. One problem in evaluating CVs was that the CV automatically decreased, if the threshold value increased, while the SD remained the same. This suggests that it is best to compare CVs of stimuli that were in the same range of thresholds. As it was my main aim to improve hyperacuity tasks concerning repeatability and I indeed found two stimuli that are in a hyperacute range it was especially necessary to compare CVs of these two new stimuli and the usual hyperacuity stimulus: The CV median in the usual Vernier task was at 46.1%, whereas the CV medians of Square and Circle were at 22.5% and 26.6%, respectively. Since these three tasks were not very different in mean thresholds (Vernier 10.3 arcsec, Square and Circle 10.6 arcsec, each) these CVs were very well comparable. As a conclusion, repeatability in the Square task was best, immediately followed by the Circle configuration. Repeatability in these two tasks was a factor 1.9 better than in the usual Vernier task. The CV in Landolt tasks was even lower than in Circle and Square stimulus, e.g.. An explanation for lower CVs in these tasks could be the number of response possibilities. In the Vernier task there were only two response possibilities (left or right; 2-AFC), whereas in the Circle and Square configurations 4-AFC paradigms were used. In the Landolt task the CV median was 8.2% - that is a factor 3 better than in the Circle and Square configuration. This factor may have been a product of eight response possibilities (only 12.5% guessing rate) on the one hand and high thresholds on the other hand. Probably the repeatability in this task would still have been better if lower thresholds were achieved, because of the 8-AFC paradigm. Thus, another approach to further improve repeatability in hyperacuity tasks would be to create an 8-AFC stimulus. We had two prototypes of such a stimulus (the Circle with double response possibilities and a similar octagon stimulus), but these stimuli were not appropriate, because it was not possible to discriminate between the eight directions in these configurations. Contrary to the presumption that lower CVs could be due to lower guessing rates, we found that the number of response possibilities had no influence on the test outcome, as it has been pointed out in the last part of ‘4.1 Acuity’.

Higher CVs in hyperacuity tasks could also have been due to ‘*neuronal noise*’. If a stimulus is presented with very low intensities it will be difficult for the subject to discriminate between the neuronal noise and the stimulus. The neuronal noise is a product of spontaneous activity of neurons. The firing rate of these neurons has to be exceeded by a defined amount to detect a stimulus. If the stimulus intensity is low the stimulus ‘drowns’ in the spontaneous activity of the neurons.

4.3 Subjective scores

The evaluation was one important part of the stimulus design. We wanted to make sure that the stimulus would be feasible in clinical and scientific daily routine. Thus, one important goal was to create a stimulus configuration that facilitates effortless testing procedure for participating subjects.

In general, subjects reacted rather different on the whole testing procedure. Some found it very exhausting and too longsome, some thought it was fun, were honestly interested in their performance and did not complain at any time. The evaluation had to be taken in account very carefully. Probably, those subjects who found the entire testing procedure rather exhausting evaluated the stimuli ‘harder’ in general. Contrariwise, equivalently in subjects who had fun doing those tests. Probably these two ‘types’ of subjects and intermediates cancelled each other.

Regarding evaluation, the newly created tests did not perform as well as we hoped. The Vernier stimulus was rated easier than Circle and Square stimulus. On the other hand the Landolt-C was rated nearly the same as Circle and Square stimulus so that one could still assume performance in this tests to be rather effortless, as Landolt-C tests are known to be done quite easily.

The only difference in evaluation between men and women was found in FiveDot tests. Men found the tests easier than women and also showed better performance. Some of the male participants said that they thought it was easy, because they imagined a crosshair in order to figure out if the inner dot was either shifted up, down, left or right. Unfortunately, there is no good explanation why men found the FiveDot stimulus easier than women and had better results in this test, as well. However, difference between males and females in this task was not statistically significant due to a Wilcoxon signed-rank test. Evaluation in sequence 1 and sequence 2 in the FiveDot stimulus was very similar to evaluation in men and women, but again there was no statistically proved significant difference.

Concerning Square and Circle stimulus, the subjects reported that they had very strong afterimages when they did these tests. They found it exhausting trying to ignore/to get rid of these afterimages and felt disturbed by them. They did not report such strong afterimages in the Vernier test. This may be an explanation of the fact that the Vernier stimulus was rated easier than the Circle and Square stimulus.

There seemed to be a connection between the subjective scores and mean thresholds of the subjects, as tasks rated 1 to 4 had lower thresholds than tasks rated 5 or 6. If a specific task was exhausting for a subject (as ratings of 5 or 6 suggest) concentration and/or motivation could have decreased. Consequently, performance got worse, resulting in higher thresholds. Another explanation may be that some subjects did not rate how effortless they could do the test, but tried to guess how good or bad their performance was. However, this is very unlikely, because I explained the sense of this evaluation and what they should evaluate several times.

Conclusion

As a conclusion the Circle and the Square stimulus seem to be qualified to replace the usual Vernier stimulus in clinical tests. Both stimuli provided hyperacute thresholds that were very much the same as in Vernier tests. Additionally, we had only 35 trials per run and CVs were considerably lower than in usual hyperacuity tasks. That was the most important criterion in order to create a test with better repeatability. The new test stimuli have not been evaluated as good as the Vernier stimulus, but nevertheless were considered not to be extraordinary hard.

References

Abbud, C. M., & Cruz, A. A. (2002). Variability of Vernier acuity measurements in untrained subjects of different ages. *Brazilian Journal Of Medical And Biological Research, 35*, 223-227.

Bach, M. (1996). The Freiburg Visual Acuity test-automatic measurement of visual acuity. *Optometry & Vision Science*, *73*(1), 49. Retrieved from http://journals.lww.com/optvissci/Abstract/1996/01000/The_Freiburg_Visual_Acuity_Test_Automatic.8.aspx.

Bach, M. (2007). The Freiburg Visual Acuity Test-variability unchanged by post-hoc re-analysis. *Graefe's archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv für klinische und experimentelle Ophthalmologie*, *245*(7), 965-71. doi: 10.1007/s00417-006-0474-4.

Banton, T., & Levi, D. M. (1991). Binocular summation in vernier acuity. *Journal of the Optical Society of America A*, *8*(4), 673–680. OSA. Retrieved from http://www.opticsinfobase.org/oe/ViewMedia.cfm?id=4041&seq=0.

Beck, J., & Schwartz, T. (1979). Vernier acuity with dot test objects. *Vision research*, *19*(3), 313-9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/442553.

Campbell, F. W., & Green, D. G. (1965). Optical and retinal factors affecting visual resolution. *J. Physiol.*, *181*, 576-593.

Carney, T., & Klein, S. A. (1997). Resolution acuity is better than vernier acuity. *Vision Research*, *37*(5), 525-539. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0042698997858786.

Carney, T., Amnon Silverstein, D., & Klein, S. A. (1995). Vernier acuity during image rotation and translation: visual performance limits. *Vision research*, *35*(14), 1951–1964. Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/004269899400288W.

Enoch, J. M., Essock, E. A., & Williams, R. A. (1984). Relating vernier acuity and Snellen acuity in specific clinical populations. *Documenta Ophthalmologica*, *58*(1), 71–77. Springer. Retrieved from http://www.springerlink.com/index/RL05H2VPP3223T95.pdf.

Essock, E. A., Williams, R. A., Enoch, J. M., & Raphael, S. (1984). The effects of image degradation by cataract on vernier acuity. *Investigative ophthalmology & visual science*, *25*(9), 1043-50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6469488.

Fahle, M. (1991). Parallel perception of vernier offsets, curvature, and chevrons in humans. *Vision research*, *31*(12), 2149–2184. Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/004269899190170A.

Garcia-Suarez, L., Barrett, B., & Pacey, I. (2004). A comparison of the effects of ageing upon vernier and bisection acuity. *Vision research*, *44*(10), 1039–1045. Elsevier. doi: 10.1016/j.visres.2003.11.018.

Glaser, J. S. (1999). *Neuro-ophthalmology* (p. 667). Lippincott Williams & Wilkins. Retrieved from http://books.google.com/books?id=eVU2CODGj98C&pgis=1.

Jäkel, F., & Wichmann, F. a. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of vision*, *6*(11), 1307-22. doi: 10.1167/6.11.13.

Klein, S. a., & Levi, D. M. (1985). Hyperacuity thresholds of 1 sec: theoretical predictions and empirical validation. *Journal of the Optical Society of America. A, Optics and image science*, *2*(7), 1170-90. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4020514.

Lee, B. B., Wehrhahn, C., Westheimer, G., & Kremers, J. (1995). The spatial precision of macaque ganglion cell responses in relation to vernier acuity of human observer. *Vision Research, 35*, 2743-2758.

Mckee, S. P., & Westheimer, G. W. (1978). Improvement in vernier acuity with practice. *Perception and Psychophysics, 24*, 258-262.

Mckee, S. P., Welch, L., Taylor, D. G., & Bowne, S. F. (1990). Finding the commom bond: Stereoacuity and the other hyperacuities. *Methods, 30*, 879-891.

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science (New York, N.Y.), 256*(5059), 1018-21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1589770.

Snowden, R., Thompson, P., & Troscianko, T. (2006). *Basic Vision: An Introduction to Visual Perception* (p. 408). OUP Oxford. Retrieved from http://www.amazon.co.uk/Basic-Vision-Introduction-Visual-Perception/dp/0199286701.

Sullivan, G., Oatley, K., & Sutherland, N. (1972). Vernier acuity as affected by target lenght and separation. *Percept. Psychophys., 12*, 438–444.

Wehrhahn, C., & Westheimer, G. (1990). How vernier acuity depends on contrast. *Experimental Brain Research, 80*, 618-620.

Westheimer, G. (1987). Visual acuity and hyperacuity: resolution, localization, form. *American journal of optometry and physiological optics, 64*(8), 567-74. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8994586.

Westheimer, G., & McKee, S. P. (1975). Visual acuity in the presence of retinal-image motion. *Journal of the Optical Society of America*, *65*(7), 847-50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1142031.

Williams, R. A., Enoch, J. M., & Essock, E. A. (1984). The resistance of selected hyperacuity configurations to retinal image degradation. *Investigative ophthalmology & visual science, 25*(4), 389-99. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6706503.

Declaration

I warrant that the thesis is my original work and that I have not received outside assistance. Only the sources cited have been used in this thesis. Parts that are direct quotes or paraphrases are identified as such.

Aachen, 21.09.2010

(Anna-Marina van der Meer)

Acknowledgements

At first, special thanks go to Dr. rer. nat. Wolf M. Harmening who awakened my interest in this topic and supervised this whole thesis from the very beginning to the very end. He always supported me, came up with the right ideas and was really patient with me, when I came up with new programming, MATLAB or any other problems.

Moreover, I owe thanks to my two reviewers Univ.-Prof. Dr. rer. nat Hermann Wagner and Univ.-Prof. Dr. med. Peter Walter.
Prof. Wagner was the one who sparked my interest in animal physiology and neurobiology and therefore is responsible for that I decided to write my bachelor thesis at the Institute of Biology II.
Prof. Walter from the Department of ophthalmology in Aachen complied with reviewing my thesis in the very short term for what I'm really thankful.

Furthermore, I would like to thank my roommates and roomneighbors at the institute for participating in pilot testing and a real good time.

Last, but not least, very special thanks go to my subjects who voluntarily participated in my experiments. These are: Tobias Reiß, Jens Oltmanns, Michael Hennig, Marlene Leucker, Florian Hischen, Saskia Bock, Freya Sautner, Niklas Domdei, Konrad Ricke, David Mondrzyk, Eike Poelders, Kolja Mertens, Vera Päfgen, Stefan Hallmann, Nadja Wenter and Christina Hölscher.