

Characterisation of field loss based on microperimetry is predictive of face recognition difficulties

Thomas S A Wallis ^{*1,3,4}, Christopher Patrick Taylor ^{*1}, Jennifer Wallis ^{*2}, Mary Lou Jackson², and Peter J Bex^{†1}

¹Schepens Eye Research Institute, Harvard Medical School

²Massachusetts Eye and Ear Infirmary, Harvard Medical School

³School of Psychology, The University of Western Australia

⁴Present address: Universität Tübingen, Germany

October 30, 2013

Abstract

Purpose: To determine how visual field loss as assessed by microperimetry is correlated with deficits in face recognition.

Methods: Patients with impaired visual sensitivity in the central visual field caused by a variety of pathologies (N = 12, ages 26–70 years) and normally-sighted control subjects (CS group, N = 12, ages 20–68 years) performed a face recognition task for blurred and un-blurred

*These authors contributed equally

†Corresponding author: peter.bex@meei.harvard.edu

faces. For patients we assessed central visual field loss using microperimetry, fixation stability, Pelli-Robson contrast sensitivity, and letter acuity.

Results: Patients were divided into two groups by microperimetry: an LV group (N = 8) had impaired sensitivity at the anatomical fovea and / or poor fixation stability whereas an LV:F group (N = 4) was characterised by at least some residual foveal sensitivity but insensitivity in other retinal regions. The LV group performed worse than the other groups at all blur levels, whereas the performance of the LV:F group was not credibly different from the CS group. The performance of the CS and LV:F groups deteriorated as blur increased whereas the LV group showed consistently poor performance regardless of blur. Visual acuity and fixation stability were correlated with recognition performance.

Conclusions: Persons diagnosed with disease affecting the central visual field can recognise faces as well as persons with no visual disease provided that they have residual sensitivity in the anatomical fovea and show stable fixation patterns. Performance in this task is limited by the upper resolution of non-foveal vision or image blur, whichever is worse.

1 Introduction

A crucial function for independent living is recognising people around us. Facial features, such as the eyes, are the primary visual cues most humans use in discriminating people they know from those they do not (e.g., Sekuler et al., 2004). Reduced sensitivity in the central visual field (defined here as the central 20 degrees of vision) presents a set of challenges for patients to manage as part of their daily lives. For observers with central visual field loss face perception can be greatly impaired, and individuals who present for vision rehabilitation regularly report difficulty with face perception (Owsley et al., 2009; Watson et al., 1997). Since photoreceptor and ganglion cell density (Curcio et al., 2004; Curcio and Allen, 1990) as well as contrast sensitivity (Robson and Graham, 1981; Rovamo et al., 1978; Hilz and Cavonius, 1974; Kelly, 1984; Johnston, 1987) decrease as retinal eccentricity increases, it is no surprise that patients with sensitivity losses in the central visual field have difficulty perceiving faces (see Kwon

and Legge, 2011, for a quantification of the human CSF contribution to face recognition in the fovea and periphery).

Numerous studies have found impaired face recognition (Bullimore et al., 1991; Barnes et al., 2011; Glen et al., 2012, 2013; Peli et al., 1991; Tejeria et al., 2002; Yu and Chung, 2011) and emotional expression perception (Mei and Leat, 2009; Boucart et al., 2008) in individuals with central visual field loss. For example, Bullimore et al. (1991) determined a threshold viewing distance to reach a criterion level on tasks of identity and expression recognition. Patients with central field loss performed worse (required a nearer viewing distance) than normally-sighted controls, and were generally better at recognising the expression of a face than its identity. In addition they found that word reading acuity was better correlated with face recognition thresholds than letter chart acuity, grating acuity and contrast threshold for edge detection. In a face recognition task similar to the one we employ here, patients with AMD showed slower reaction times and poorer face recognition performance than age-matched controls (Barnes et al., 2011). Tejeria et al. (2002) correlated measures of visual function (letter acuity, reading, Pelli-Robson contrast sensitivity and colour vision) and self-reported difficulty in face recognition with performance on a familiar face identification and an odd-one-out expression discrimination task in 30 subjects with AMD. Letter and reading acuity were found to be correlated with performance on the face perception tasks, but there was little evidence for a relationship between self-reported difficulty in face perception and performance in either of the tasks. Using a “celebrity” / “not a celebrity” face recognition task, Peli et al. (1991) found that certain types of digital image enhancement improved recognition performance in some (16 of 38) patients with central vision loss (see also Peli et al., 1994). Patients with central field loss also demonstrated poorer performance on a famous face recognition task relative to controls in the study of Dulin et al. (2011). It is worth noting however that in face recognition tasks relying on famous faces, the contribution of cultural and general visual exposure cannot be separated from face identification performance per se.

Many of these studies have examined correlations between face recognition performance and standard metrics of functional vision such as acuity and contrast sensitivity and found

modest-to-strong relationships, depending on the type of stimulus used in assessment. However, the relationships between these quantities and the extent of central field loss are uncertain. To more precisely measure these relationships, in the present study we combine an objective measure of face recognition with clinical characterisation of retinal deficiencies using fundus-oriented microperimetry. This face identification task has been used previously (Gaspar et al., 2008) and shows many classical effects from the face perception literature: the inversion effect (Gaspar et al., 2008; Hussain et al., 2009b), the contrast-reversed face effect (Hussain et al., 2009a), the preferential use of horizontal orientation information (Pachai et al., 2012; Goffaux and Dakin, 2010; Dakin and Watt, 2009) and spatial frequency tuning of approximately 10 cy/face (Gaspar et al., 2008). In addition to letter acuity and contrast sensitivity, visual function was assessed with fundus-oriented microperimetry, which allows precise assessment of light sensitivity over a large (approximately 20 degree diameter) region of the retina. Perimetry also measures each patient's fixation stability and the locus of fixation, which may be foveal or non-foveal.

Individuals with foveal vision loss may use non-foveal areas of the retina for face perception (Schuchard, 2005). The peripheral retina has reduced sensitivity to high spatial frequencies compared to the fovea. We varied the amount of blur present in our face stimuli by selectively removing high spatial frequency content from the images. Individuals with foveal vision loss should be relatively unaffected by stimulus blur, if blur only affects the high spatial frequencies to which they are insensitive. We classified low vision participants into two groups on the basis of their sensitivity across the central visual field as well as the stability of their fixation in microperimetry. One group had impaired foveal vision and / or poor fixation stability (LV), whereas the other group had visual impairments excluding the fovea (LV:F). We hypothesised that the group with intact foveal function would show a consistent reduction in face recognition as a function of blur whereas the impaired fovea group would show poor performance for even relatively blur-free stimuli, consistent with the use of peripheral vision.

2 Methods

2.1 Participants

Patients with various pathologies affecting central visual field sensitivity were recruited from the Vision Rehabilitation Clinic at the Massachusetts Eye and Ear Infirmary. Twenty-one patients with low vision were initially enrolled in the study. Eight of these patients chose to withdraw after completing less than one full block of trials, in all cases citing frustration with task difficulty. This introduced a self-selection bias that may lead us to overestimate the performance of patients with low vision in our study. One additional patient was excluded from analyses because we did not have accompanying microperimetry data for that patient. These exclusions left a final group of 12 patients aged from 26 to 70 years (mean = 48.8 years). Three of the final group of patients completed less than the full three blocks due to time constraints (but did complete at least one block). These were P07 (1 block), P08 (2 blocks) and P09 (1 block). All patients were tested after initial consultation at the clinic with an ophthalmologist (MLJ), but before any vision rehabilitation training with an occupational therapist.

Twelve control subjects with normal retinal function were recruited, including two of the authors (C01 and C09). These subjects ranged between 20 and 68 years old (mean = 40.8). Demographic information for all subjects is provided in Table 1. All subjects provided informed consent; the procedures of the experiment were approved by the Institutional Review Board of the MEEI, and the experiment was conducted in accordance with the Declaration of Helsinki.

All participants were divided into one of three groups for data analysis. Control subjects (CS) were those with no known visual pathology. Patients with low vision were divided into two subgroups based on clinical evaluations and perimetry: those with at least partially functional foveal vision (LV:F) and those without (LV; see Figure 1). This classification was made for the eye with better acuity, or if acuity was equal in the two eyes, the eye with more stable fixation in microperimetry. The testing area for microperimetry was determined by initial fixation, so was not centred around the anatomical fovea for all patients. However, the testing array

always included the anatomical fovea. The anatomical foveal area was defined as the region of the retina located two vertical optic disk diameters to the temporal side of the optic disk (Jackson et al., 2013). Three conditions had to be met for a patient to be classified in the LV:F group. First, the perimetry array had to be centred around the anatomical fovea. Second, the patient had to report seeing at least one of the four perimetry targets in the centre of the array. Third, the patient had to show stable fixation, that is, more than 95% of fixations fell within two degrees of the anatomical fovea. Other patients with a low vision diagnosis failed to detect perimetry targets presented at the fovea and had unstable fixation. This group is labelled as LV. All patients were classified using these criteria by author JW. The classification was independently assessed by familiarising an occupational therapist of the Vision Rehabilitation Clinic (not an author) with the classification guidelines; this person made the same classifications as JW.

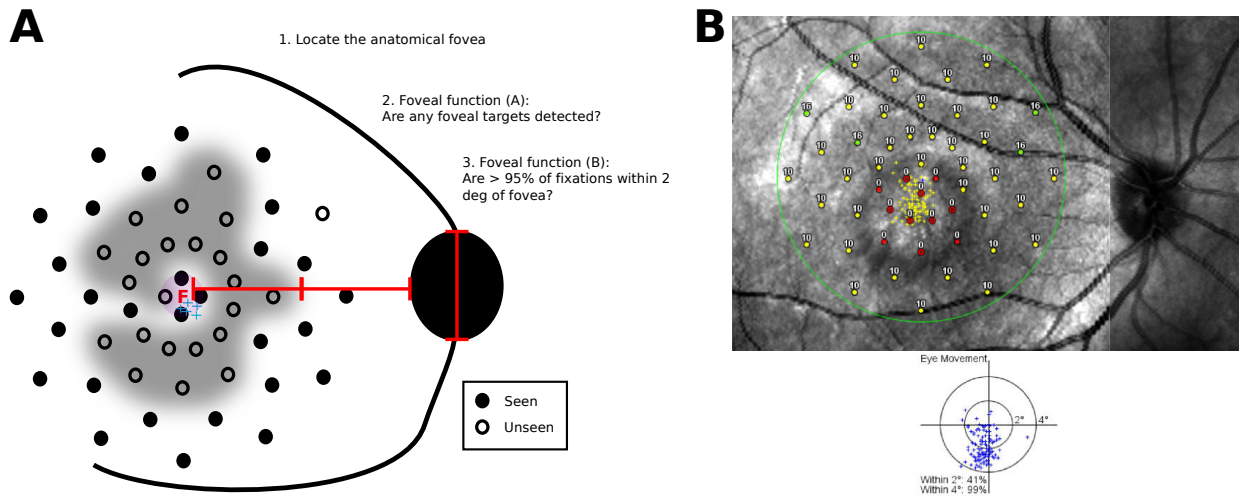


Figure 1: Classification of patients into groups based on microperimetry. (A): Schematic of classification. First, the location of the anatomical fovea (marked here by a red “F”) was estimated by measuring two vertical optic disk diameters (red line) to the temporal side of the optic disk. If any of the targets around the anatomical fovea were detected and > 95% of fixations (blue crosses) were within 2 degrees of the anatomical fovea in the eye with greater acuity, the patient was classified into the LV:F group. (B): Microperimetry output for a patient in the present study (P04) for the patient’s right eye (OD), tested with a modified threshold strategy. The foveal area is obscured by a scotoma. The patient attempted to fixate in the foveal region. Fixation stability is poor (lower polar plot, blue crosses) and no targets presented to the anatomical fovea were detected (red circles).

Table 1: Demographic and clinical information. The “instruction” column refers to those patients who received the pilot instruction manipulation. PR refers to log Pelli-Robson contrast sensitivity and VA refers to visual acuity. Visual acuity and contrast sensitivity were measured binocularly in control subjects, and where marked in patients.

Code	Age	Sex	Instruction	PR OD	PR OS	VA OD	VA OS	LogMAR VA better eye	Diagnosis
P01	52	M	No	N/A	1.5	N/A	20/150-200	0.88	Right prosthesis, left keratoplasty
P02	57	M	Yes	1.5 (binocular)		20/400	20/150-1	0.88	Congenital Nystagmus
P03	34	M	No	1.35	1.35	20/40	20/30	0.18	Retinitis Pigmentosa
P04	26	F	Yes	1.65 (binocular)		20/150	20/150	0.88	Stargards' Dystrophy
P05	30	M	Yes	1.05 (binocular)		1/12	1/16	1.08	Macular Malfunction
P06	57	F	No	1.2	0.45	20/30	20/200	0.18	Exposure Keratopathy/Neurotrophic Cornea OS
P07	70	F	No	1.05 (binocular)		20/100	20/150	0.70	Glaucoma
P08	32	M	No	1.65 (binocular)		20/150	20/150	0.88	Achromatopsia
P09	70	M	No	.45 (binocular)		20/400	20/200	1.00	Anterior ischemic optic neuropathy
P10	69	F	No	1.35 (binocular)		20/70	20/80	0.54	AMD
P11	58	F	No	1.05 (binocular)		20/70	20/100	0.54	Retinal degeneration
P12	31	F	No	1.65	1.65	20/25	20/20	0.00	Decreased vision post trauma
C01	29	M		42(1.95)				-0.13	
C02	20	F		42 (1.95)				-0.13	
C03	46	F		39 (1.8)				0.18	
C04	68	F		N/A				N/A	
C05	40	F		N/A				N/A	
C06	60	F		N/A				N/A	
C07	61	F		36 (1.65)				N/A	
C08	34	M		38 (1.8)				N/A	
C09	45	M		38 (1.8)				0.00	
C10	23	F		38 (1.8)				-0.13	
C11	34	M		42 (1.95)				-0.13	
C12	29	F		39 (1.8)				-0.13	

2.2 Apparatus & Stimuli

Face stimuli were presented on an iMac (late 2011 model) on a 27-inch LCD monitor with a spatial resolution of 2560 x 1440 pixels and a temporal resolution of 60 Hz using MATLAB and Psychtoolbox-3 (Kleiner et al., 2007; Pelli, 1997). The observers viewed the display under normal room illumination, and the mean luminance of the display was 250 cd/m². To increase the contrast resolution of the display, a bit-stealing routine was used to increase the number of grey-levels from 6-bits to 8.8 bits (Tyler, 1997). The stimuli were monochromatic photographs of faces presented at three levels of blur. Figure 2 depicts examples of representative face stimuli. Gaspar et al. (2008) provide a detailed description of how the faces were photographed and processed, thus only a brief description will be provided here ¹. The face set contains photographs of 10 individuals (five female), all Caucasian, with a mean age of twenty-four years; none of the models had facial hair, visible piercings, tattoos, or eye glasses. All the face images were cropped with an oval window with a width:height ratio of 1.5 to remove all clothing and hair. Each model was photographed from 5 view-points. To produce each view-point image the model was instructed to fixate on a marker positioned on the wall behind the camera. The face set used was made up of two left-, two right- and one front-facing view of each model for a total of 50 face images. To control for possible variations in detectability among the face images (Gold et al., 1999), we equated the Fourier amplitude spectra of the images by computing the average amplitude spectrum of the set of face images and then replacing the amplitude spectrum of each image with the average amplitude spectrum. As noted in the introduction, this stimulus set has been previously shown to demonstrate classic effects from the face perception literature. This indicates that our stimuli tap similar mechanisms to those used in more complex stimulus sets, despite using only a relatively small (10 individuals) and relatively unnatural (lacking features such as hair and skin pigment) set of faces.

To measure the effect of blur on recognition performance of both patients and controls, we tested face recognition performance at two levels of synthetic blur and a no-blur condition.

¹The face stimuli used in the experiment are available upon request

Blurred images were produced by applying a low-pass ideal filter (high frequency cutoff 32 or 16 cycles per image, ≈ 8 or ≈ 4 cycles per degree) to the Fourier amplitude spectrum of each image (see Figure 2). A zero-mean Gaussian pixel noise (R.M.S. contrast, $\sigma = 0.005$) was added to the test image on each trial to allow comparison with the ideal observer (see Supplementary Materials 7.1).

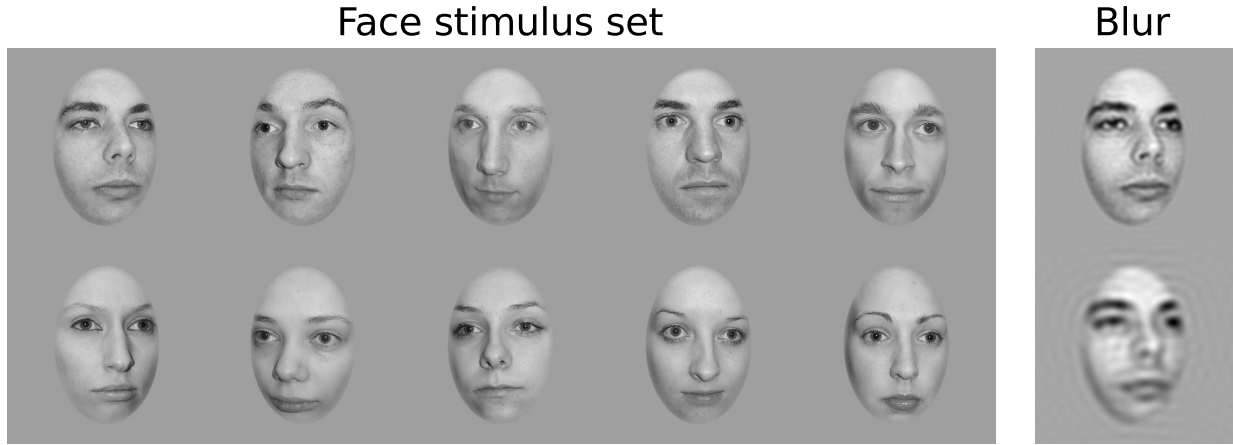


Figure 2: Stimuli used in the experiment. The left panel shows each stimulus face; each column shows one of the five possible viewpoints. The right panel shows an example of the blur kernels used. The top image depicts the 32 cycle per image cutoff, the bottom image depicts the 16 cy / image cutoff.

2.3 Perimetry

Microperimetry testing was conducted with an OCT/SLO microperimeter (OPTOS PLC). The extent of central field loss was assessed by having patients fixate a cross presented in the centre of the display, while a Goldmann III test stimulus (0.43 deg disk) was presented for 200 ms at each of 52 radially-organized locations within a 21 degree diameter circle centred around the initial fixation location. Seven patients (P03, P04, P06, P07, P08, P11 and P12) were tested with a modified threshold strategy, five patients (P01, P02, P05, P09 and P10) with a suprathreshold strategy. The testing regime was chosen depending on the clinical appropriateness for each patient, not as part of a research protocol. Suprathreshold testing entails the presentation of a target at an intensity that is anticipated to be seen (0 dB attenuation) and

the patient indicates when the target is detected by pressing a button. In modified stepwise threshold testing (using a 0-10-16 technique), a -10 dB point is presented on the first trial then lowered to -16 dB if seen or 0 dB if not seen. Testing positions that elicit no response suggest field loss. During testing, fixation is continuously monitored. Thus, a measure of fixation stability can be obtained (defined as fixations that remain within 2 degrees and 4 degrees of initial fixation). Unfortunately we were unable to save raw fixation coordinates due to computer malfunction, meaning that the more sensitive fixation measure of the Bivariate Contour Ellipse Area (BCEA; see Crossland et al., 2009) could not be calculated. However, since we observe credible correlations with our relatively blunt measure of fixation stability, the BCEA would likely only strengthen the relationships we observe.

All but one patient failed to report at least one microperimetry target entirely. The remaining patient (P06), tested in a modified threshold regime, detected all targets but required increased brightness relative to unimpaired vision (relative deficits). All other patients who were tested with a modified threshold strategy showed relative deficits in addition to their failures to report targets at the highest brightness (see for example P04 in Figure 1). All patients can therefore be considered to have some central vision loss (see Table 2).

2.4 Procedure

Participants completed a single interval 10 alternative forced choice face identification task. Observers freely viewed the display (binocular presentation, no restriction on eye movements) from approximately 57 cm and at this distance the test face images subtended 8.4 by 8.8 degrees (width by height) of visual angle. On each trial the observer saw a fixation point in the centre of the screen presented for 200 ms, followed by a test face image presented in the centre of the screen for 500 ms. The identity, view-point, and blur-level of the test face was randomly selected each trial. After the test image disappeared, 10 smaller images (6.1 by 6.4 degrees of visual angle) of each face were presented in a grid of 5 columns by 2 rows on the screen. These faces were of the 10 individuals from the face set and were presented in random order,

Table 2: Perimetry data for patients with low vision. The fixation percentages denote the percentage of all fixations made within 2 or 4 degrees of the centre of the testing grid. Any subject who showed both good fixation stability and reported seeing any of the four perimetry targets presented around the anatomical fovea was classified as having a functional fovea (LV:F). While P06 did not miss any perimetry targets at the highest intensity, extensive relative deficits were evident with lower intensity targets.

Code	2 deg fixation (%)	4 deg fixation (%)	Perimetry targets missed (%)	Analysis group
P01	57	94	2	LV
P02	9	37	46	LV
P03	100	100	67	LV:F
P04	41	99	13	LV
P05	42	100	44	LV
P06	100	100	0	LV:F
P07	18	100	42	LV
P08	41	80	2	LV
P09	76	100	83	LV
P10	100	100	6	LV:F
P11	72	100	4	LV
P12	100	100	2	LV:F

at the same viewpoint as the test image, but without blur. The observer used the arrow keys on the keyboard to move a response selection rectangle over the response faces, then pressed the space-bar to enter their response. If the observer was unable to use the keyboard, they pointed at a face and the experimenter keyed in their response. The ten response-selection faces remained onscreen until a response was entered. If the observer's response was correct then a clearly audible high-pitched tone was played; if the observer's response was incorrect a tone of lower pitch was played. Prior to the experimental trials, each observer completed 5 practice trials to introduce them to the task.

Each observer completed 3 blocks of 75 trials, except one control subject (C02), who completed 3 blocks of 150 trials, and three patients (P07, P08 and P09) who completed only one, two and one blocks respectively. The three levels of blur were each tested 25 times per block, in a random order for each subject.

Three of the patients with low vision (P02, P04 and P05) received an instruction manipulation designed to improve their guidance of eye movements to important regions of the faces. After completing the first block of trials, these subjects were instructed to "focus on the eyes,

nose and mouth” by the experimenter, who asked the subject to practice by looking at the experimenter’s face. This manipulation was designed to ensure that patients were placing regions of functional retina over facial features important for identification as far as possible. However, a pilot analysis showed no evidence that the instruction improved performance beyond a simple learning effect. All of these patients reported that they were already able to place their gaze to perceive important facial features, making it unlikely that this instruction manipulation would have any further effect. We therefore group together patients who received this instruction manipulation with patients who did not in further analyses.

2.5 Data Analysis

To analyse this data we used a multi-level logistic Generalised Linear Model (GLM; see Supplementary Material [7.2](#) for details). Multilevel models are extensions of traditional regression models that treat regression parameters themselves as conditional on higher-order structure in the data (Gelman, 2006; Gelman and Hill, 2007). Our data has hierarchical structure in that some parameters (such as a given subject’s performance level) can be considered to arise from distributions of mean performance within a group (such as “control” or “low vision”), which themselves are conditional upon an overlying population distribution of mean performance level (the distribution for humans). Intuitively, this is similar to running three regressions: one on the data for each subject to find the parameter weights at the subject level, a second on the weights from the first regression to estimate the group parameters, and a third on the group estimates to find the population-level parameters. Here, however, all of these parameters are estimated simultaneously rather than in a stepwise fashion as implied by this example.

A multilevel Bayesian model like the one we use is a more general form of the more common multilevel generalised linear mixed model, which has been recently promoted within the psychophysics (Moscatelli et al., 2012; Knoblauch and Maloney, 2012) and clinical vision literature (Cheung et al., 2008) since it allows the analysis of data on the population level. A mixed model is “mixed” because it contains both fixed effects (in which a parameter is not

allowed to vary at some level of the model) and random effects (for which the uncertainty in the parameter estimates is considered throughout all levels of the model). One way to consider a fixed effect is that it is a special case of a random effect, in which the variance parameter is set to be zero (i.e. there is no uncertainty around the point estimate of the parameter; see Gelman and Hill, 2007, p.245). In the framework we employ, all parameters can be thought of as random effects (Gelman and Hill, 2007). This treatment has the desirable characteristic that uncertainty at all levels of the model is considered at all others (rather than reducing the model to a point estimate at some level), and that this uncertainty can be propagated into predictions. For our data set, this is particularly useful since not all subjects completed the same number of trials. Therefore our approach takes the individual subject uncertainties into account when we make inferences at the group level.

We base statistical inferences on the full Bayesian posterior distribution of credible parameter values rather than on null hypothesis significance testing. Desirable properties of this approach are discussed briefly in the Discussion. Since the posterior distributions for the models we use are not derivable analytically, we numerically estimate proportional distributions using Markov Chain Monte Carlo (MCMC) sampling implemented in the Stan package (Stan Development Team, 2013a; Hoffman and Gelman, *ress*). Results were analysed using the R statistical computing environment (R Development Core Team, 2011) using additional packages (Wickham, 2009, 2011; Xie, 2013). The estimate of the posterior distribution for the primary analysis is based on 4 independent chains of 100,000 samples using the NUTS II sampling method (Hoffman and Gelman, *ress*), with half of those samples being warmup (adapting the step size of the sampler), and saving every 100th sample to reduce both autocorrelation in the chains and data file size, resulting in 2,000 final samples. Convergence was assessed by the \hat{R} value (the ratio of between-to-within chain variance; see Gelman and Rubin, 1992) as well as by inspection of traceplots. Code to perform these analyses and raw data can be found on the first author's Github page (https://github.com/tomwallis/microperimetry_faces).

We were primarily interested in how performance differed between the three groups as a

function of the blur level. In addition, since there could have been a learning effect and not all participants completed the same number of blocks, we included a covariate for block. The trial-by-trial (correct or incorrect) data for each subject was modelled as a linear function of the log high frequency cutoff of the blur filter and the block:

$$\eta = a + b \cdot \log_{10}(\text{HFC}) + c \cdot \text{block} \quad (1)$$

This yielded three free parameters for each subject: intercept a , blur slope b and block slope (linear learning effect) c ². In addition, to make the regression coefficients easier to interpret (see Gelman and Hill, 2007, p.55), we centred each input predictor by subtracting the mean (or 2 in the case of block). Therefore the intercept parameter represents performance when blur was at its mean in the second block. We additionally examined a model (not shown here) that contained an interaction term between block and blur slope. The interaction term was not credibly different from zero so we excluded it for simplicity. The linear predictor η is then passed through the inverse logit function to give the expected value for a Bernoulli trial, with an additional lower bound of 0.1 (reflecting chance performance).

For the two conditions in which we imposed physical stimulus blur, the high frequency cutoffs correspond to 32 cycles per image (less blurry) and 16 cy / image (more blurry). However, for the un-blurred condition no frequencies were filtered. To express the un-blurred condition on the same scale, we assume the high frequency cutoff to be the Nyquist limit of the stimulus, which given that our images were 372 pixels square, the highest frequency present in our stimuli was 186 cy / image, ≈ 46 cy / deg.

Each coefficient from each subject is hypothesised to arise from a distribution for that subject's group. A population hyperparameter is also placed over all groups, causing group differences to be pulled towards the population mean and making our inferences on group differences more conservative where parameter estimates are uncertain. The full specification

²Including age as an additional covariate did not substantively change our conclusions, and additional covariates such as visual acuity cannot be included in this model since they were reported as corrected to normal for control subjects, in which case there is no variation in acuity among controls.

of the model is provided in the Supplementary Materials (7.2), along with a discussion of the tolerance of our conclusions to the prior distributions chosen. All groups are given the same mean (of zero) and variance (broad) in their regression parameters *a priori*, so it is not the case that our use of priors trivially presumes the conclusion we report.

3 Results

3.1 Recognition performance

Figure 3 shows recognition performance for each subject for each blur level and block, with the corresponding fits of the multi-level model at the subject level. Each MCMC sample post-warmup represents a plausible value of the posterior probability over the full parameter space. We computed a distribution of expected proportion correct for a continuum of blur values from the subject-level parameters a , b and c for each of the post-warmup samples of the MCMC chains after pooling the chains. The solid curves in Figure 3 show the mean of this distribution for each subject, and the shaded regions are the values containing 95% of the distribution's mass: the highest-density interval (HDI). Unlike a frequentist confidence interval, the 95% HDI of a posterior distribution allows us to say that, given the model and the data, there is a 95% chance that the true parameter value lies in this interval. These are 95% *credible* intervals (as distinct from *confidence* intervals, Kruschke, 2011b). As expected, performance increased as a function of the logarithm of high frequency cutoff (i.e., decreased as a function of blur level) for most subjects. The model provides reasonable fits to the data at the individual subject level.

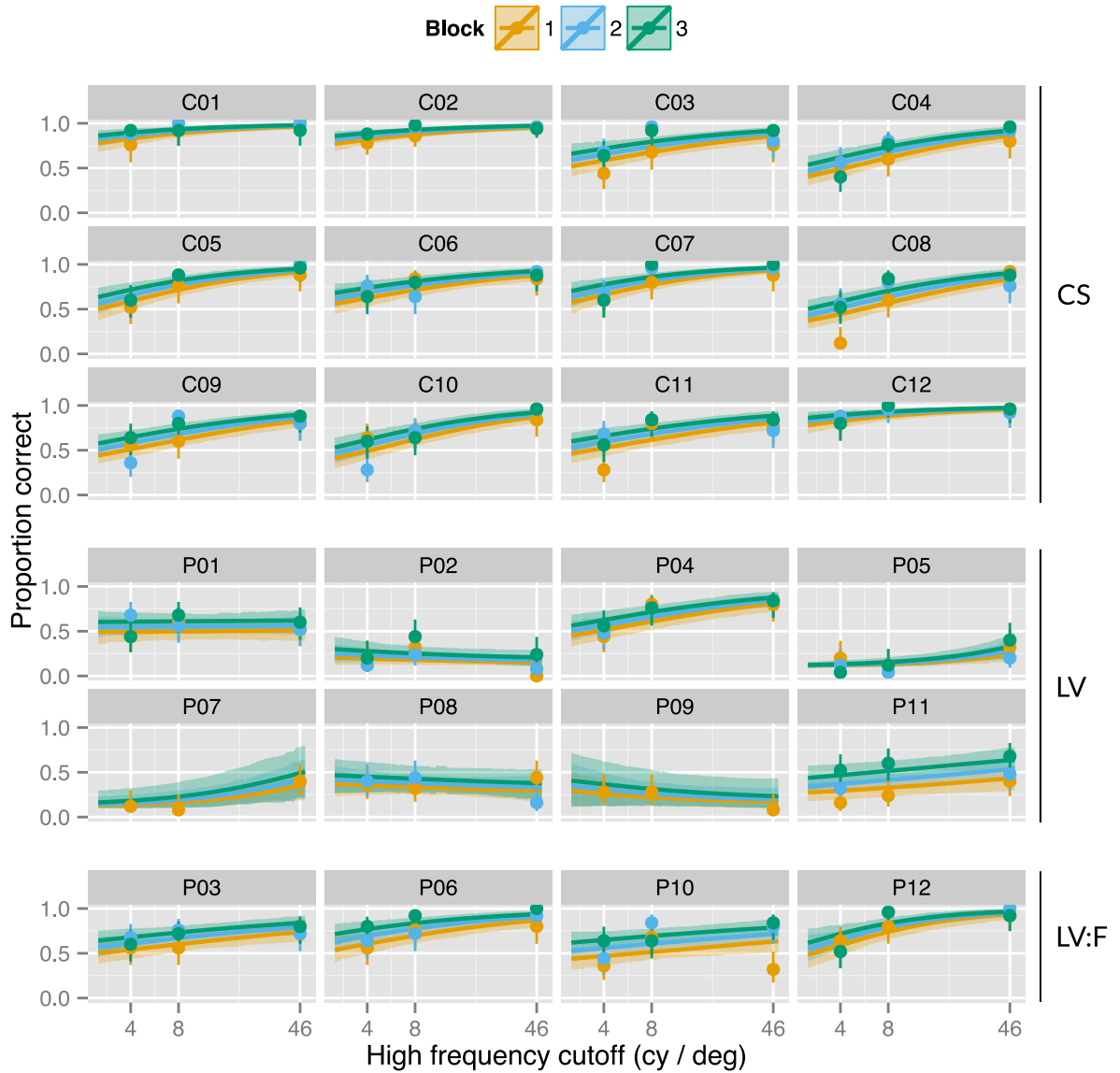


Figure 3: Face recognition performance as a function of high frequency cutoff (blur level) and block for each subject (note logarithmic x-axis). A lower cutoff value corresponds to a more blurred stimulus. Here the subject code displayed in the top margin corresponds to Table 1 and the side text shows the analysis group for that subject (CS = control subjects, LV = low vision, LV:F = low vision with foveal sensitivity). Points show the average performance of the subject; error bars on points depict 95% beta distribution confidence intervals. Solid lines represent the predictions of the Bayesian multilevel model at the subject level and shaded regions represent 95% credible intervals for these predictions.

Each subject's three free parameters are conditional on the parameters of their group, which

are themselves conditional on the population-level parameters. Therefore, the data from one subject inform the fits to the data of another subject, more so for subjects in the same group. Statistical inferences at the group level can be made by comparing the group-level parameter distributions. Of principal interest here is the group-level distribution for performance as a function of blur.

Predictions at the group-level are shown in Figure 4, and were derived from the MCMC samples as in the subject-level data. It can be seen that the control subjects (CS) had higher performance than the patients with low vision (LV) using diseased retina or with unstable fixation. In addition, those patients with low vision (LV:F) with remaining foveal sensitivity are similar to control subjects in their performance.

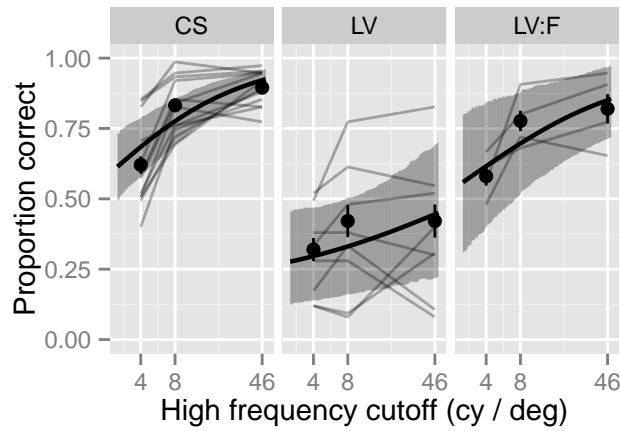


Figure 4: Face recognition performance as a function of high frequency cutoff (blur level) for each group (note log unit x-axis). A lower cutoff value corresponds to a more blurred stimulus. Points show each group mean, and error bars depict ± 1 s.e.m between subjects. Solid black lines represent the predictions of the Bayesian multilevel model at the group level and shaded regions represent 95% credible intervals on these predictions. Faint grey lines connect the mean proportion correct for each subject (data replotted from Figure 3, averaged across block).

The group differences apparent in Figure 4 can be formally compared by calculating the difference score between parameter estimates in each MCMC sample to create a distribution of difference scores. That is, we calculate the value of the linear predictor for each MCMC sample (Equation 1), at each block and blur level, then apply the inverse logit to convert these

values into predicted probability correct, then take a difference score between the groups for each sample. We then base our inferences about whether two groups have credibly different performance on the distribution of difference scores between those groups, by asking whether 0 (i.e., no difference) is within the range of credible values (Kruschke, 2011b). Note that no correction is necessary for making multiple comparisons since all possible comparisons are encapsulated by the full posterior distribution (Kruschke, 2012), which remains unchanged whether we compare two parameters or many. The posterior distributions for the group-level parameters are provided in the Supplementary Materials (Figure 9).

At all blur levels tested, and as early as the first 25 trials³, the performance of the CS group was credibly higher than that of the LV group (Figure 5). In addition, the LV group was credibly lower than the LV:F group. No credible differences were evident between the CS and LV:F group.

³ In this model, the values for parameters a , b and c are estimated from all the data. While strictly this means that the parameters are not independent of block, this is not important for our statement that group differences emerge as early as block one. The same group differences are evident after fitting a two parameter (a and b) model to only the data from block one.

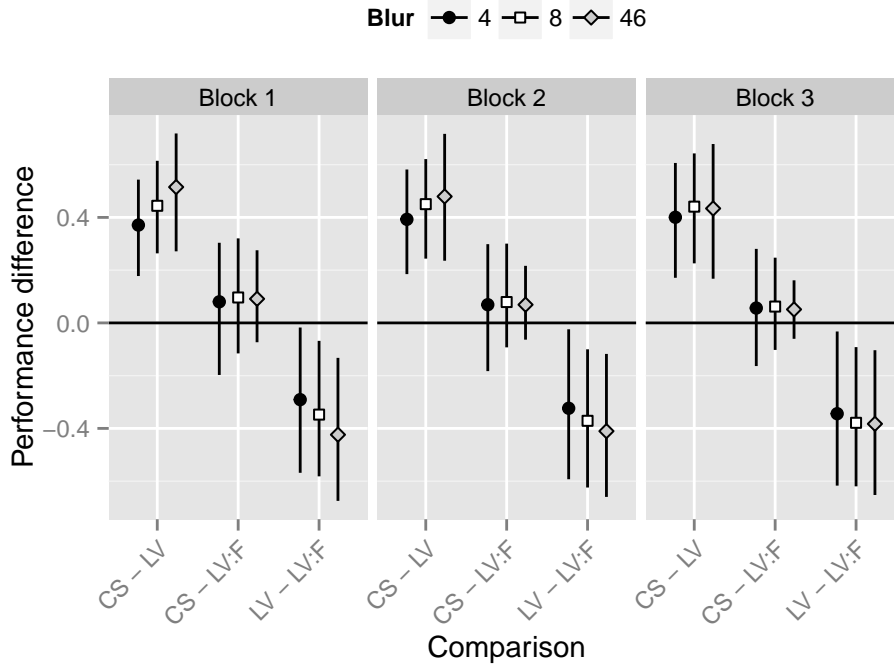


Figure 5: Mean and HDIs of pairwise comparisons between groups. These values are calculated from the posterior distributions of difference scores in predicted proportion correct. Differences are shown for each blur level at each block of trials. At all blur levels and as early as block 1, the CS and LV:F groups performed better than the LV group.

We can also examine both the absolute slope parameters for blur and block, and differences between groups for these slope parameters. Slopes credibly greater than zero signify a linear relationship between (the logit of) performance and the variable of interest (log high frequency cutoff for b , block for c ; see Equation 1). The specific value of the slope parameter depends on the scaling of the predictor, since the slope multiplies the value of the input predictor. If the value of the parameter is zero this indicates no linear relationship (i.e. the logit of performance does not change linearly as a function of the predictor).

For the slope as a function of the log high frequency cutoff (parameter b), performance for the CS and the LV:F groups increased as high frequency cutoff increased (CS mean = 1.66, 95% HDI = [1.26, 2.03], LV:F = 1.34, [0.26, 2.46]). That is, as the amount of blur increased performance got worse. However, the slope for the LV group was not credibly

different from zero (0.83, -0.34, 1.89). This indicates that the performance of the LV group was not credibly affected by changes in blur whereas the other groups showed effects of blur such that performance improved as blur decreased (i.e. as high frequency cutoff increased). However, these comparisons should be treated with caution since there were also no credible differences between the group slopes (see Supplementary Materials, Section 7.3), indicating that there is large uncertainty for these comparisons. While all groups showed a credible learning effect (parameter c was credibly greater than zero, see Supplementary Figure 9), there were no credible differences between group learning rates (Section 7.3).

3.2 Correlations between visual function measures

	TM	FS	CS	VA	PC
TM	1	-0.11 [-0.48,0.29]	-0.26 [-0.62,0.13]	0.23 [-0.17,0.6]	-0.3 [-0.67,0.1]
FS	–	1	-0.14 [-0.5,0.23]	-0.37 [-0.71,0.03]	0.48 [0.1,0.78]
CS	–	–	1	-0.09 [-0.46,0.28]	0.35 [-0.04,0.68]
VA	–	–	–	1	-0.47 [-0.81,-0.09]
PC	–	–	–	–	1

Table 3: Rank-order correlations (Spearman’s ρ) between visual function measures for patients with low vision estimated using a Bayesian framework. TM denotes the proportion of targets missed in perimetry, FS denotes fixation stability (percentage of fixations within 2 degrees), CS denotes Pelli-Robson contrast sensitivity, VA denotes visual acuity, and PC denotes proportion correct on the face identification task. The posterior mean rank order correlation coefficient is given in each cell, with 95% credible intervals in square brackets.

Scatterplots comparing clinical measures of visual function and face identification performance are shown in Figure 6 for the patients with low vision. We estimated the posterior distribution over the correlation coefficients between variables using MCMC as in the main analysis; these are shown in Table 3. Since these calculations are based on only 12 data points, we urge caution in drawing strong conclusions from the correlations. To encourage conservative inferences, we place a weakly-informative prior distribution over the correlation matrix with greater mass on the unit diagonal. That is, we *a priori* assign higher probability to correlations around zero (see Supplementary Materials, Section 7.2.1), but with high uncertainty.

Using the conservative rank order Spearman's ρ , fixation stability (FS) and visual acuity (VA) were credibly correlated with face recognition performance (i.e. the credible intervals on these comparisons did not overlap zero). More stable fixation and better acuity were associated with better performance. Credible intervals for all other correlation coefficients overlapped zero. Similar results were found using a bootstrapping approach and traditional calculations of correlation coefficients, with the exception that fixation stability and visual acuity were significantly correlated with each other. We believe our more conservative approach here is warranted given the small number of data points.

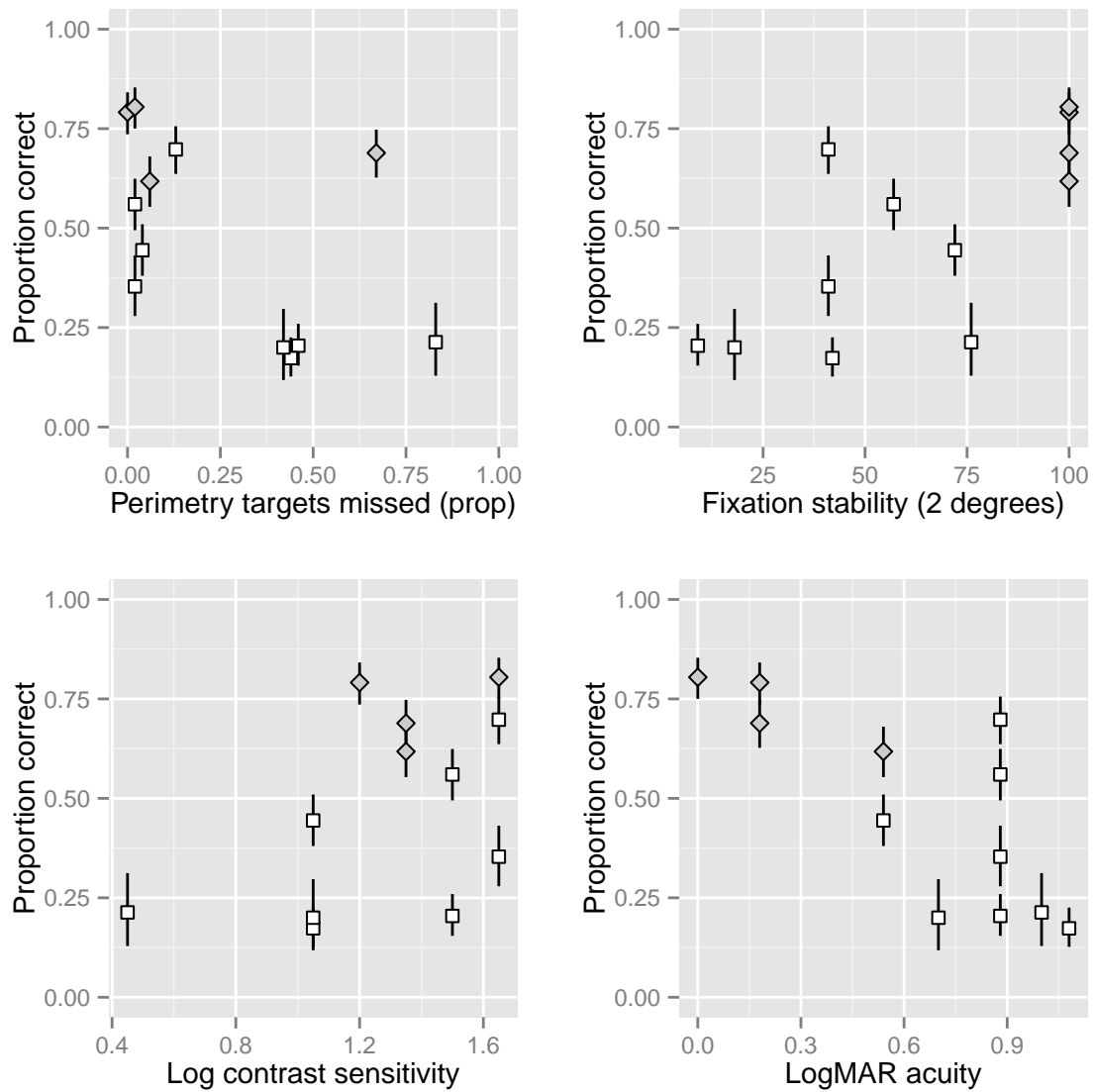


Figure 6: Scatterplots of face identification performance (proportion correct) in low vision participants as a function of scotoma level, fixation stability within two degrees, contrast sensitivity and letter acuity. The LV group is shown as white squares, the LV:F group as grey diamonds. Points show the mean performance for each subject and error bars show beta distribution 95% confidence intervals. Similar patterns were found when performance was broken down for each blur level.

4 Discussion

Many patients with pathology affecting the fovea are forced to use non-foveal retina to accomplish visual tasks. We examined how the extent of central field loss affected face identification. The LV group in our study contained individuals who failed to perceive perimetry targets centred over the anatomical fovea at maximum intensity (i.e., had dense central scotomas), demonstrated poor fixation stability, or presented a combination of these deficits (see Table 2). This group performed worse in the face identification task compared with the other groups (see Figure 4 and Figure 5). Patients classified into the LV:F group saw at least one of the perimetry targets centred over the anatomical fovea and showed stable fixation. Unlike the LV group, the LV:F group performed similarly to control subjects despite having either dense or relative deficits in their central visual field (see Table 2). The data also suggest that a patient's letter acuity and fixation stability during perimetry are correlated with difficulties in face recognition. For this task, functional macular retina seems to be the primary determinant of performance among a group of patients with varied degrees of paracentral field loss.

We also examined how stimulus blur affected face identification in these groups. We expected that blur would only affect performance in participants who could resolve high spatial frequencies and thus visual function would be limited by the lower bound of blur in the stimulus or visual resolution. Face identification performance steadily decreased as blur increased for normally-sighted and patients with low vision with (at least partially) functional foveal vision: the slope parameters for these groups were credibly different from zero. Patients with low vision and scotomas including their fovea (the LV group) were relatively unaffected by stimulus blur, performing worse than control subjects and LV:F subjects at every blur level tested and as early as the first block of trials (see Figure 5). As expected, the LV group showed poor performance at all levels of blur, whereas the group with intact foveal function (LV:F) showed a consistent reduction in performance as a function of blur, similar to controls. While we did not track the gaze position of our participants during the face recognition task, the result is consistent with LV subjects using peripheral vision to perform the task. The limiting factor

on the performance of the LV group is likely to be the poor resolution of non-foveal vision rather than stimulus blur at the levels we imposed.

To provide an illustration of the level of impairment demonstrated by LV patients in the context of our experiment, we calculate how much the stimulus would need to be blurred for control subjects to produce the same performance as the LV group with no stimulus blur. For an unblurred stimulus, patients in the LV group have an average performance level of approximately 46% (see Figure 4). According to the model, the control group would need a blur cutoff of ≈ 0.8 cy / degree to reach a performance of 46% correct on average. A face stimulus filtered with this blur level is depicted in Figure 7. Within the context of stimulus blur, the amount of information LV patients can extract from a face is clearly limited. Of course, many factors influence performance as stimuli move away from the foveal retina, such as reduced contrast sensitivity (Kelly, 1984) and crowding (Bouma, 1970). We do not claim that the reduced performance of the LV group here is caused by intrinsic blur alone, nor does this demonstration mean that stimuli appear blurred to LV patients, owing to sharpness overconstancy (Galvin et al., 1997). We provide this illustration to give the reader some intuition as to the level of impairment shown by these patients in comparison to normally-sighted individuals.

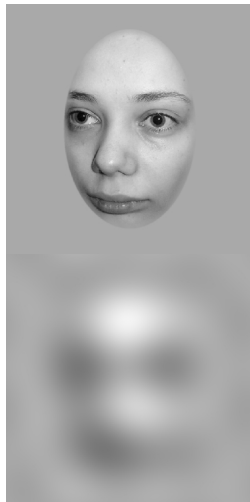


Figure 7: An unblurred face (top) and the same face blurred to the level our model predicts would produce performance in control subjects equivalent to the LV group with no stimulus blur (see Figure 4). The lower image gives some intuition as to the level of impairment demonstrated by LV patients in our task.

Some patients spontaneously used a preferred retinal locus (PRL) during perimetry. We cannot say whether these patients also used a PRL to perform face recognition because gaze position was not monitored in this task. Since these patients showed no sensitivity in the fovea they were included in the LV group for analysis. In addition, as faces in the recognition task were viewed binocularly and perimetry was monocular, any conclusions we could draw about the influence of a monocular PRL in our task are limited. This may be improved in future studies by using binocular perimetry (currently in development; see Timberlake, et al., IOVS 2013; 54: ARVO E-Abstract 2182).

Visual acuity and fixation stability correlated with face recognition performance. More stable fixation and better letter acuity were associated with higher task performance. Using our Bayesian estimation of correlation coefficients, we also found no credible association between visual acuity and fixation stability. However, a bootstrapping estimation procedure showed this association to be significant. This contrasts with previous studies (von Noorden and Mackensen, 1962; Timberlake et al., 1986; White and Bedell, 1990; Crossland et al., 2004) that reported no association between visual acuity and fixation stability. The difference can be attributed to our intact foveal function group, who showed stable fixation and relatively good acuity, driving up the strength of this association relative to studies whose patient populations did not include patients with functional foveal vision. Indeed, any study such as ours that includes diverse groups of patients spanning a wide range of acuities and fixation stabilities is likely to find a stronger correlation than a study within a single group of patients. A scatterplot of fixation stability versus visual acuity (see Supplementary Materials, Figure 11) shows two clusters, one for each of the patient groups, with little evidence of correlations within each group (though note that the LV:F group was selected on the basis of steady fixation). Of course, these correlational analyses must all be treated cautiously due to our small patient sample size.

It is possible that fixation stability is a more useful correlate of face recognition performance than acuity. Acuity was more variable than fixation stability in patients with (at least partial) foveal vision (see Figure 6): subjects with foveal vision all showed stable fixation (100% within 2 degrees), however, their visual acuity ranged from 20/20 to 20/70. Fixation

stability is important for visual performance in a range of tasks; for example, Crossland et al. (2004) showed that unstable fixation is associated with slower reading speeds. Dulin et al. (2011) found that patients with retinitis pigmentosa (in which the macular is spared) and patients with Stargardt’s dystrophy (in which the macular is diseased) performed equivalently poorly in a famous face recognition task despite showing markedly different visual acuity. Potentially, fixation stability is a more useful indicator of face recognition problems than acuity, despite both being associated with performance in our study.

Recent work (Seiple et al., 2013, 2011; Seiple, 2005; Glen et al., 2013) suggests that more general eye movement behaviour (not just fixation stability) is crucial for visual tasks in patients with visual field loss. For example, AMD patients with an established PRL show a different fixation pattern than controls when looking at a face; patients placed fewer fixations on the internal features (e.g., eyes) of a target face than controls (Seiple et al., 2013). Eye movement control training can improve performance on visual tasks such as reading (Seiple et al., 2011), and potentially also improve face perception. However, in many real-world situations people would have the opportunity to view a face for longer than in our study, so fixation stability may be less important for face perception tasks when faces are presented for extended durations. Conversely, fast face recognition may be required to comprehend television or films, particularly those with multiple and rapid scene cuts.

The main finding of our study can be visually assessed from Figure 4: the LV group performs worse than the other groups at all blur levels. What does our modelling add to this? First, it allows us to quantify the size of the differences between groups, and the performance changes as a function of blur and block. Second, analysing the data in a multilevel framework means that our inferences for groups with fewer participants are likely to be more robust, since these estimates will be pulled towards the population mean (“shrinkage”) to the extent that the data suggest. Third and most importantly, estimating the posterior probabilities over the parameter space using Bayesian inference has a number of desirable properties over traditional null-hypothesis significance testing (NHST). For the present application three pertinent advantages of the Bayesian approach over NHST are that inferences over the Bayesian posterior

do not require corrections for making multiple comparisons, that inferences do not depend on the experimenters' intentions for stopping data collection, and that Bayesian analyses coherently handle unbalanced designs by propagating uncertainty through all levels of the model. Interested readers are referred to the following papers (Gelman, 2006; Gelman et al., 2012; Kruschke, 2010a,b, 2012, 2011a; Wagenmakers, 2007; Wetzels et al., 2009, 2011) and textbooks (Gelman and Hill, 2007; Kruschke, 2011b) for recent introductions to and discussions of this type of inference, which is gaining traction in many scientific fields.

As many in the field of vision rehabilitation have stated (e.g., Trauzettel-Klosinski, 2010), to offer appropriate vision rehabilitation practitioners must understand the impact of each patient's vision loss on various visual functions. Our patient sample included patients with very different pathologies and visual function (see Table 1). While a variety of pathologies impair face recognition (see also Dulin et al., 2011), our results suggest the key determinant of face recognition performance was whether the fovea was preserved or impaired. Patients in the LV:F group recognised as many faces as control subjects despite significant field loss. While our experiment measured face recognition performance in an objective paradigm with high sensitivity (10 AFC), the strength of the conclusions we can draw are limited by our small sample of patients. An independent replication of our result in a larger sample of patients would strengthen the conclusion of this study: that patients with preserved foveal vision (but with other visual deficits) can recognise faces as well as someone with healthy vision.

5 Conclusions

We examined the effects of image blur on face identification in patients with low vision. Patients with at least partially preserved foveal vision and stable fixation performed similarly to normally-sighted control observers and face identification steadily decreased as stimulus blur increased. Patients with no foveal vision and unstable, eccentric fixation were severely impaired in the task at all levels of image blur. The results suggest that face recognition performance is limited by the highest spatial frequencies that are available to the observer and this

may be predicted by visual acuity and fixation stability, as assessed in the clinic.

6 Acknowledgements

Designed the experiment: CPT, JW, TSAW. Programmed the experiment: CPT. Collected the data: JW. Analysed the data: TSAW, CPT. Wrote the paper: TSAW, CPT, JW, MLJ, PB. The authors thank Dr William H. Seiple for comments and suggestions and Ms Joy Facella-Ervolini for assistance with data collection. TSAW was supported by a fellowship from the National Health and Medical Research Council of Australia (634560). PJB was supported by NIH grants EY019281 and EY018664.

7 Supplementary materials

7.1 Ideal observer

Performance in our task can be compared to the performance of an *ideal observer* that performs the task optimally given the information available in the stimulus. The optimal decision rule for this task, assuming Gaussian contrast noise, is to compute the 2D (pixel wise) cross-correlation between each possible response face and the test face, then to select the response with the maximum cross correlation (Gold et al., 1999; Tjan et al., 1995). We estimated the performance of the ideal observer as a function of the signal-to-noise ratio (SNR; for Gaussian pixel noise added to the stimulus) by computer simulation (see Figure 8). For the stimulus conditions shown to observers (dashed vertical line), the ideal observer never made mistakes. That is, humans perform the task sub-optimally.

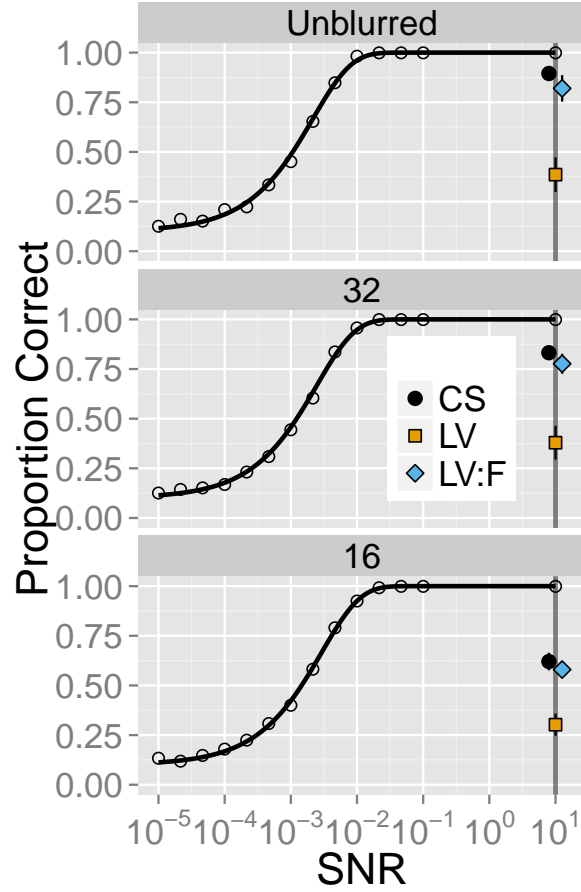


Figure 8: Face recognition performance for the ideal observer as a function of the ratio of signal contrast to Gaussian luminance noise contrast (SNR), for the unblurred stimulus and high frequency cutoffs of 32 and 16 cy / image. The performance of the ideal observer (open circles) was evaluated via Monte Carlo simulation. Test faces were presented with the same three blur levels as shown to human observers, at the same white noise contrast, and at 13 additional white noise contrasts in log-spaced steps. The ideal observer provided 1000 simulated trials at each combination of blur and noise contrast for a total of 14,000 trials per blur level. Solid lines represent the fits of a Weibull GLM (Knoblauch and Maloney, 2012). The grey vertical line marks the SNR shown to subjects. Average performance in each blur condition for human subjects are shown as filled points (x position is offset to aid readability). The error bar on these points shows the s.e.m between subjects in each group.

The ratio of ideal-to-human performance (“absolute efficiency”) is usually used to quantify this sub-optimality. We cannot compute this measure here since we did not measure full psychometric functions for human performance as a function of the SNR; we arbitrarily selected

an SNR to present to the human observers. That is, in Figure 8, the horizontal location of the human data is arbitrary and the slope of human performance as a function of contrast noise level is unknown.

However, some sense of the discrepancy between human and ideal can be gained from the ideal observer’s psychometric functions in Figure 8, as the ideal observer requires ≈ 2 log units more noise contrast than humans to even begin to make mistakes, and around 4 log units to reach equivalent performance to humans. This is consistent with previous studies that have found that face recognition efficiency is extremely poor even in normally-sighted observers (Gold et al., 1999; Mäkelä et al., 2001; Melmoth et al., 2000; Näsänen, 1999). In contrast, relatively good efficiencies have been shown for other tasks, such as simple detection (Kersten, 1987; Taylor et al., 2009) and for extracting information about a data set from graphs (Legge et al., 1989). This comparison suggests that human face recognition in tasks such as ours must be limited mainly by constraints inherent to the visual system rather than by a lack of information in the stimulus.

By measuring the full psychometric functions for face recognition in patients with low vision as a function of SNR, future studies could determine the theoretical limit on performance improvements expected from image modification techniques (such as contrast enhancement) for a particular task. If humans are primarily limited by internal noise rather than by noise in the stimulus, it is possible that no amount of stimulus enhancement could effect performance improvement. Potentially, knowing the limits of internal noise could help direct low vision research to effectively target visual tasks in which improvements could be expected given particular stimulus changes or training.

7.2 Multilevel model and Bayesian inference

We fit the data with a multilevel model with three levels of parameters: subject, group and population⁴. The trial-by-trial binary correct or incorrect response data y for each subject i in

⁴R code for reproducing this analysis is available from the first author’s Github page: https://github.com/tomwallis/microperimetry_faces

group j were assumed to arise from a Bernoulli distribution with expected value p :

$$y_{ij} \sim \text{Bernoulli}(p_{ij}) \quad (2)$$

where the \sim symbol denotes “is distributed as”. The expected value p was modelled as a logistic function with a lower asymptote γ , fixed at the chance performance level of 0.1:

$$p_{ij} = \gamma + (1 - \gamma) \times \text{logit}^{-1}(\eta_{ij}) \quad (3)$$

where the inverse logit function is $1/(1 + \exp(-\eta))$. The linear predictor η_{ij} for subject i was given by a linear function of the intercept a plus a slope coefficient of log high frequency cutoff b plus a slope with block c (to capture any linear effect of learning):

$$\eta_{ij} = a_{ij} + b_{ij} \log_{10}(\text{HFC}) + c_{ij} \text{block} \quad (4)$$

The three coefficients from each subject were assumed to arise from a group-level distribution:

$$a_{ij} \sim \text{Normal}(\mu_j, \sigma_j) \quad (5)$$

Since the same structure is used for all coefficients we simplify the notation here to only refer to one, but the model contained parameters μ_j and σ_j for all coefficients. The group level mean and standard deviations are themselves conditional upon population-level distributions:

$$\begin{aligned} \mu_j &\sim \text{Normal}(\mu_\mu, \sigma_\mu) \\ \sigma_j &\sim \text{Gamma}(\alpha_\sigma, \beta_\sigma) \end{aligned} \quad (6)$$

Here we use the subscripted letter to denote the group-level parameter that is conditional on the higher-level parameter. Gamma denotes a gamma distribution with shape α_σ and scale β_σ . A gamma distribution is used for the standard deviation parameter σ_j since standard deviation cannot be negative and the gamma distribution is only defined for non-negative values.

To estimate the posterior distribution over model parameters using Bayes’ rule, it is necessary to specify prior distributions on the model parameters. In this respect our entire model structure can be considered a “prior” (see Gelman and Hill, 2007): we believe it is logical to structure the model with individual subjects nested in groups, and groups nested in populations. Having defined this model structure in which low level parameters are conditional on higher-level parameters, we now must specify priors for the top-level hyperparameters (i.e. at the population level). The credible values for the lower levels of the model are then calculated based on the top-level priors and the data. We take the approach of using vague prior distributions (i.e. with large variance) reflecting a relatively agnostic stance on the model parameters. Since these distributions reflect large uncertainty about the true parameter values, they are quickly overwhelmed by the data causing them to have little influence on the posterior distribution (Kruschke, 2012). The model results presented in the paper are fit with the following priors on the population level:

$$\begin{aligned}
\mu_\mu &\sim \text{Normal}(0, 3) \\
\sigma_\mu &\sim \text{Uniform}(0, 10) \\
m_\sigma &\sim \text{Uniform}(0, 10) \\
sd_\sigma &\sim \text{Uniform}(0, 10)
\end{aligned} \tag{7}$$

We specify priors on the mean m_σ and standard deviation sd_σ of the Gamma distribution for σ_j since it is more intuitive to place priors over the mean and deviation than over shape and scale (see for example Kruschke (2011b, , Figure 9.8, page 170). These parameters are then reparameterised into shape and scale for the gamma distribution as:

$$\begin{aligned}
\alpha_\sigma &= m_\sigma^2 / sd_\sigma^2 \\
\beta_\sigma &= m_\sigma / sd_\sigma^2
\end{aligned} \tag{8}$$

To summarise the model using the blur slope coefficient b as an example, the parameter b for subject i in group j is hypothesised to arise from a group normal distribution with mean μ_j and standard deviation σ_j . The mean of each group’s blur slope is conditional upon a single

population normal distribution with mean μ_μ and standard deviation σ_μ , and the standard deviation of each group's blur slope is conditional upon a single population gamma distribution with shape α_σ and scale β_σ . These population-level distributions are given weakly-informative priors. For MCMC sampling the model was parameterised slightly differently, with each model coefficient described as a shifted-and-scaled distribution of errors to increase sampling efficiency (Stan Development Team, 2013b). For more details than can be described here, please see the analysis code on the first author's Github page.

To understand why these priors represent vague values we must consider the scale of the model space. We tested three levels of high frequency cutoff (186, 32, and 16), and the model is fit to the logarithm of these (2.27, 1.50 and 1.20). The value of p (Equation 3) can range from 0.1 to 1. The value of η (given by Equation 4) is unbounded (i.e. from negative to positive infinity), but for practical purposes resides in the range of -2.5 to 2.5 (which give expected values of 0.106 and 0.994). Therefore, for any given pairing of intercept a , blur slope b and block slope c , the value of Equation 4 becomes relatively meaningless outside of the $+/- 2.5$ range, so the values of these parameters must produce an η around this range. For example, the prior on μ_μ for all parameters has a mean of zero, together corresponding to an η of 0 and an expected performance level of 0.55 (including γ , the lower bound correction for chance performance). The standard deviation of 3 spreads most of the prior certainty over a range of -6 to +6 (i.e. 2 standard deviations around the mean of zero), which can be seen to over-span the useful range of the parameter space and therefore represents a relatively flat prior centred on zero.

This prior can be compared to the group-level posteriors in Figure 9. The posterior distributions for all groups fall within the range of -0.5 to 2, and the central tendency for three of the groups lie between values of 0.5 and 1. That is, the data has caused the posterior densities to become more certain (less variable) than the weak prior and also to shift away from the prior mean of zero for three of the groups, to the extent that the data suggest that the true mean is different. The uniform prior over variance parameter σ_μ in the range of 0 to 10 indicates that we are uncertain about the true variance between the groups on intercept a and slope b .

Group variance hyperpriors m_σ and sd_σ are given uniform priors over the range of 0 to 10. This selection of priors allows the variance between individuals in the group to take on a wide range, from individuals being similar to individuals being dissimilar.

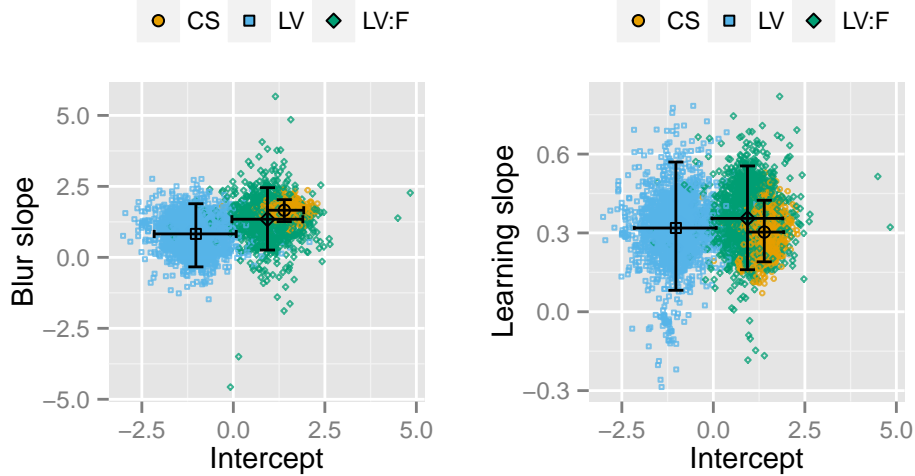


Figure 9: Samples from the posterior distribution for intercept, slope of high frequency cutoff (blur), and slope of block (learning) parameters at the group level. Samples from the final MCMC chain are shown as coloured points, with mean and 95% credible intervals (HDIs) shown as black points and error bars.

Multilevel models have the desirable property of creating “shrinkage” between individual subject estimates. Since subjects are assumed to be drawn from a common population, estimates of their effects “share strength” which has the effect of reducing the influence of outliers on inference from the model fits (Gelman, 2006). Estimates of individual subject effects therefore lie between the completely pooled population estimates (fitting a single model to all data across subjects, or equivalently fixing σ_μ to be zero) and the within-subjects effects (fitting independent models to each subject). The degree of shrinkage is dictated by the data and the prior, since we estimate σ_μ from data: if subjects or groups show highly variable performance, the population-level parameter estimates will be very uncertain and there will be little shrinkage. If many subjects show similar estimates except for one, the estimate of that subject’s true mean will be strongly pulled towards the (otherwise highly certain) population estimate. This

reduces the influence of outliers on inferences made on the population scores in a principled way (see Gelman et al., 2004; Kruschke, 2012, 2011b, for further examples).

Like any choice of statistical model, the hierarchy we hypothesise here along with the specific prior distributions we use can be regarded as “priors” (Gelman et al., 2004). How robust are our conclusions to the modelling choices we make? We have tested several alternative models and discuss one here. We examined a similar model to the one reported in the paper, but without the third (population) level hyperparameters. That is, the fits to each group are completely independent, and no shrinkage will occur between group estimates. The results of this fitting showed greater uncertainty for the group-level parameters compared to the 3-level model as expected, but the key results of the paper did not change substantively. Since we wish to draw inferences about human performances on our task, we believe the use of the 3-level model in the paper is justified. The top-level hyperparameter helps the data constrain credible values for model parameters in groups with less data, as well as drawing all groups towards the grand mean where less data are available, theoretically producing more robust predictions. We believe this choice should be acceptable to a skeptical reader, since it is far from trivially presuming our result by specifying it in our prior: the prior assumes that all the groups are the same, with large uncertainty; the group differences we observe in the posterior are purely driven by the data.

7.2.1 Bayesian correlation coefficients

To compute Bayesian correlation coefficients we estimate the posterior over the correlation matrix between the five predictor variables. We first compute the rank orders of the data matrix, then normalise these ranks (z-score) to compute the Spearman rank order correlation coefficient. These data are considered to arise from a multivariate normal distribution with a mean of zero and a covariance matrix that is estimated from the data and the prior. To encourage conservative inference, we place weak *a priori* credibility on correlations of zero using the following correlation matrix prior distribution (see Stan Development Team, 2013b,

p. 315):

$$M(\Sigma|\alpha) \propto \det(\Sigma)^{\alpha-1} \quad (9)$$

where α is a shape parameter and Σ is a positive definite symmetric matrix with unit diagonal (i.e. a correlation matrix). When α is one, the prior has uniform density (i.e. all correlation matrices are equally plausible). As α approaches infinity, the prior becomes the identity matrix (i.e. all correlations are zero). We set α to be 2, which is a weak prior that represents the belief that zero correlations are more plausible than strong correlations, but with high uncertainty.

Estimating the correlation coefficients in this way did not produce drastically different conclusions from traditional statistical estimation. In a bootstrapping regime we found that visual acuity and fixation stability were significantly associated with performance (Spearman's ρ), consistent with the conclusions of our Bayesian approach. In addition, the association between fixation stability and acuity was also significant, where it did overlap zero in our method. We believe our method is the more conservative approach, since bootstrapping estimates can be unstable with such few data.

7.3 No differences evident for blur slope and block slope

There were no credible differences between the groups in either blur slope or in learning over blocks (see Figure 10).

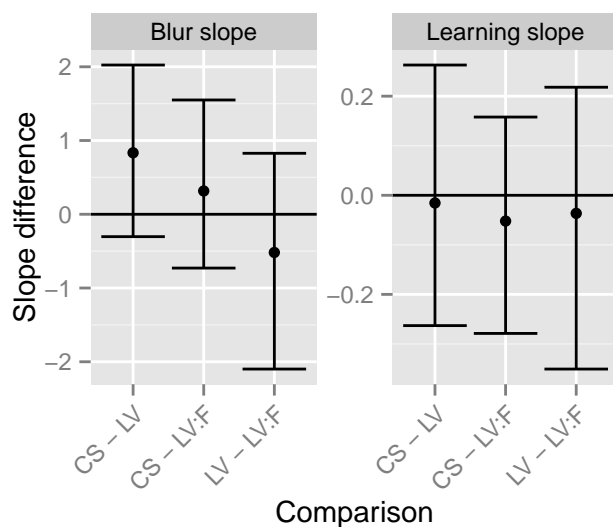


Figure 10: Mean and HDIs of the distributions of difference scores between group level parameters for blur slope and learning slope. No credible differences are observed (all distributions overlap zero).

7.4 Correlation between acuity and fixation stability

It is possible that the correlation between acuity and fixation stability we observe is driven by the use of diverse patient populations, and that within each group there is little evidence for a correlation.

Below is a scatterplot of these data. It is extremely difficult to draw conclusions about this due to the limited group size, however it is entirely possible that little correlation exists within the LV group. Note that LV:F patients were selected partly on the basis of their good fixation stability, and so the lack of relationship here is by design.

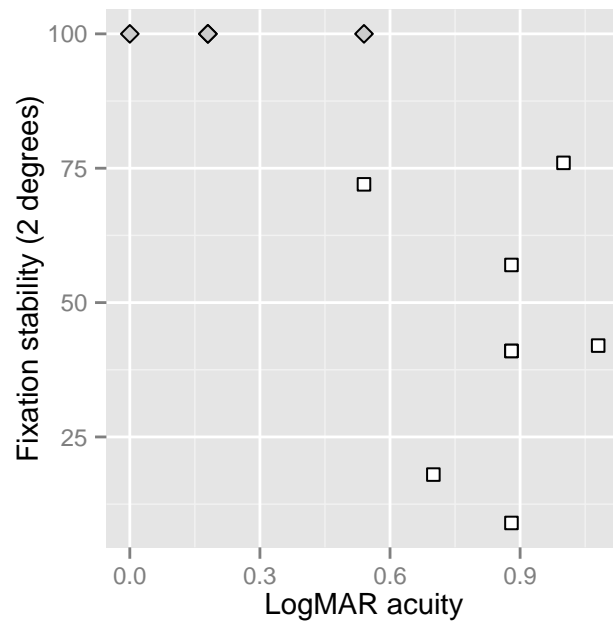


Figure 11: Scatterplot of fixation stability versus letter acuity. The LV group is shown as white squares, the LV:F group as grey diamonds.

References

- Barnes, C. S., De L'Aune, W., and Schuchard, R. A. (2011). A test of face discrimination ability in aging and vision loss. *Optometry and Vision Science*, 88(2):188–199.
- Boucart, M., Dinon, J.-F., Despretz, P., Desmettre, T., Hladiuk, K., and Oliva, A. (2008). Recognition of facial emotion in low vision: A flexible usage of facial features. *Visual Neuroscience*, 25(04).
- Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, 226(5241):177–178.
- Bullimore, M. A., Bailey, I. L., and Wacker, R. T. (1991). Face recognition in age-related maculopathy. *Investigative Ophthalmology & Visual Science*, 32(7):2020–2029.
- Cheung, S. H., Kallie, C. S., Legge, G. E., and Cheong, A. M. Y. (2008). Nonlinear Mixed-Effects Modeling of MNREAD Data. *Investigative Ophthalmology & Visual Science*, 49(2):828–835.
- Crossland, M. D., Culham, L. E., and Rubin, G. S. (2004). Fixation stability and reading speed in patients with newly developed macular disease. *Ophthalmic and Physiological Optics*, 24(4):327–333.
- Crossland, M. D., Dunbar, H. M. P., and Rubin, G. S. (2009). Fixation stability measurement using the MP1 microperimeter. *Retina*, 29(5):651–656.
- Curcio, C. A. and Allen, K. A. (1990). Topography of ganglion cells in human retina. *The Journal of Comparative Neurology*, 300(1):5–25.
- Curcio, C. A., Sloan, K. R., Kalina, R. E., and Hendrickson, A. E. (2004). Human photoreceptor topography. *The Journal of Comparative Neurology*, 292(4):497–523.
- Dakin, S. C. and Watt, R. J. (2009). Biological "bar codes" in human faces. *Journal of Vision*, 9(4).

- Dulin, D., Cavezian, C., Serrière, C., Bachoud-Levi, A.-C., Bartolomeo, P., and Chokron, S. (2011). Colour, face, and visuospatial imagery abilities in low-vision individuals with visual field deficits. *Quarterly Journal of Experimental Psychology*, 64(10):1955–1970.
- Galvin, S. J., O’Shea, R. P., Squire, A. M., and Govan, D. G. (1997). Sharpness overconstancy in peripheral vision. *Vision Research*, 37(15):2035–2039.
- Gaspar, C., Sekuler, A. B., and Bennett, P. J. (2008). Spatial frequency tuning of upright and inverted face identification. *Vision Research*, 48(28):2817–2826.
- Gelman, A. (2006). Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics*, 48(3):432–435.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.
- Gelman, A. and Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge Univ Press, New York, NY.
- Gelman, A., Hill, J., and Yajima, M. (2012). Why We (Usually) Don’t Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211.
- Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.
- Glen, F. C., Crabb, D. P., Smith, N. D., Burton, R., and Garway-Heath, D. F. (2012). Do Patients with Glaucoma Have Difficulty Recognizing Faces? *Investigative Ophthalmology & Visual Science*, 53(7):3629–3637.
- Glen, F. C., Smith, N. D., and Crabb, D. P. (2013). Saccadic eye movements and face recognition performance in patients with central glaucomatous visual field defects. *Vision Research*, pages 1–10.

- Goffaux, V. and Dakin, S. C. (2010). Horizontal information drives the specific signatures of face processing. *Frontiers in Perception Science*, 1:1–14.
- Gold, J. M., Bennett, P. J., and Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research*, 39(21):3537–3560.
- Hilz, R. and Cavonius, C. R. (1974). Functional organization of the peripheral retina: sensitivity to periodic stimuli. *Vision Research*, 14(12):1333–1337.
- Hoffman, M. D. and Gelman, A. (in press). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*.
- Hussain, Z., Sekuler, A. B., and Bennett, P. J. (2009a). Contrast-reversal abolishes perceptual learning. *Journal of Vision*, 9(4).
- Hussain, Z., Sekuler, A. B., and Bennett, P. J. (2009b). Perceptual learning modifies inversion effects for faces and textures. *Vision Research*, 49(18):2273–2284.
- Jackson, M., Crossland, M. D., Wallis, J., and Seiple, W. (2013). Microperimetry in Rehabilitation. In Mastropasqua, L., Vingolo, E. M., Midena, E., Jackson, M., and Mark, i. S., editors, *Handbook of Microperimetry in Visual Rehabilitation*. Fabiano.
- Johnston, A. (1987). Spatial scaling of central and peripheral contrast-sensitivity functions. *Journal of the Optical Society of America A*, 4(8):1583–1593.
- Kelly, D. H. (1984). Retinal inhomogeneity. I. Spatiotemporal contrast sensitivity. *Journal of the Optical Society of America A*, 1(1):107–113.
- Kersten, D. (1987). Statistical efficiency for the detection of visual noise. *Vision Research*, 27(6):1029–1040.
- Kleiner, M., Brainard, D., and Pelli, D. G. (2007). What’s new in Psychtoolbox-3? *Perception*, 36(ECVP Abstract Supplement).

- Knoblauch, K. and Maloney, L. T. (2012). *Modeling Psychophysical Data in R*. Springer, New York.
- Kruschke, J. K. (2010a). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5):658–676.
- Kruschke, J. K. (2010b). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14(7):293–300.
- Kruschke, J. K. (2011a). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, 6(3):299–312.
- Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Academic Press / Elsevier.
- Kruschke, J. K. (2012). Bayesian Estimation Supersedes the t Test. *Journal of Experimental Psychology: General*.
- Kwon, M. and Legge, G. E. (2011). Spatial-frequency cutoff requirements for pattern recognition in central and peripheral vision. *Vision Research*, 51(18):1995–2007.
- Legge, G. E., Gu, Y. C., and Luebker, A. (1989). Efficiency of graphical perception. *Perception & Psychophysics*, 46(4):365–374.
- Mäkelä, P., Näsänen, R., Rovamo, J., and Melmoth, D. (2001). Identification of facial images in peripheral vision. *Vision Research*, 41(5):599–610.
- Mei, M. and Leat, S. J. (2009). Quantitative Assessment of Perceived Visibility Enhancement with Image Processing for Single Face Images: A Preliminary Study. *Investigative Ophthalmology & Visual Science*, 50(9):4502–4508.
- Melmoth, D. R., Kukkonen, H. T., Mäkelä, P. K., and Rovamo, J. M. (2000). The effect of contrast and size scaling on face perception in foveal and extrafoveal vision. *Investigative Ophthalmology & Visual Science*, 41(9):2811–2819.

- Moscatelli, A., Mezzetti, M., and Lacquaniti, F. (2012). Modeling psychophysical data at the population-level: The generalized linear mixed model. *Journal of Vision*, pages 1–17.
- Näsänen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research*, 39(23):3824–3833.
- Owsley, C., McGwin, G., Lee, P. P., Wasserman, N., and Searcey, K. (2009). Characteristics of low-vision rehabilitation services in the United States. *Archives of Ophthalmology*, 127(5):681–689.
- Pachai, M. V., Sekuler, A. B., and Bennett, P. J. (2012). Practice with inverted faces selectively increases the use of horizontal information. *Journal of Vision*, 12(9):975–975.
- Peli, E., Goldstein, R. B., Young, G. M., Trempe, C. L., and Buzney, S. M. (1991). Image enhancement for the visually impaired. Simulations and experimental results. *Investigative Ophthalmology & Visual Science*, 32(8):2337–2350.
- Peli, E., Lee, E., Trempe, C. L., and Buzney, S. (1994). Image enhancement for the visually impaired: the effects of enhancement on face recognition. *Journal of the Optical Society of America A*, 11(7):1929–1939.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10:437–442.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Robson, J. G. and Graham, N. (1981). Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research*, 21(3):409–418.
- Rovamo, J., Virsu, V., and Näsänen, R. (1978). Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature*, 271(5640):54–56.

- Schuchard, R. A. (2005). Preferred retinal loci and macular scotoma characteristics in patients with age-related macular degeneration. *Canadian Journal of Ophthalmology*, 40:303–312.
- Seiple, W. (2005). Eye-Movement Training for Reading in Patients with Age-Related Macular Degeneration. *Investigative Ophthalmology & Visual Science*, 46(8):2886–2896.
- Seiple, W., Grant, P., and Szlyk, J. P. (2011). Reading Rehabilitation of Individuals with AMD: Relative Effectiveness of Training Approaches. *Investigative Ophthalmology & Visual Science*, 52(6):2938–2944.
- Seiple, W., Rosen, R. B., and Garcia, P. M. T. (2013). Abnormal fixation in individuals with age-related macular degeneration when viewing an image of a face. *Optometry and Vision Science*, 90(1):45–56.
- Sekuler, A. B., Gaspar, C. M., Gold, J. M., and Bennett, P. J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology*, 14(5):391–396.
- Stan Development Team (2013a). Stan: A C++ Library for Probability and Sampling, Version 2.0.1. Technical report.
- Stan Development Team (2013b). *Stan Modeling Language User’s Guide and Reference Manual, Version 2.0.1*.
- Taylor, C. P., Bennett, P. J., and Sekuler, A. B. (2009). Spatial frequency summation in visual noise. *Journal of the Optical Society of America A*, 26(11):B84–93.
- Tejeria, L., Harper, R. A., Artes, P. H., and Dickinson, C. M. (2002). Face recognition in age related macular degeneration: perceived disability, measured disability, and performance with a bioptic device. *British Journal of Ophthalmology*, 86(9):1019–1026.
- Timberlake, G. T., Mainster, M. A., Peli, E., Augliere, R. A., Essock, E. A., and Arend, L. E. (1986). Reading with a macular scotoma. I. Retinal location of scotoma and fixation area. *Investigative Ophthalmology & Visual Science*, 27(7):1137–1147.

- Tjan, B. S., Braje, W. L., Legge, G. E., and Kersten, D. (1995). Human efficiency for recognizing 3-D objects in luminance noise. *Vision Research*, 35(21):3053–3069.
- Trauzettel-Klosinski, S. (2010). Rehabilitation for Visual Disorders. *Journal of Neuro-Ophthalmology*, 30(1):73–84.
- Tyler, C. W. (1997). Colour bit-stealing to enhance the luminance resolution of digital displays on a single pixel basis. *Spatial Vision*, 10(4):369–377.
- von Noorden, G. K. and Mackensen, G. (1962). Phenomenology of eccentric fixation. *American Journal of Ophthalmology*, 53:642–660.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5):779–804.
- Watson, G. R., De l’Aune, W., Stelmack, J., Maino, J., and Long, S. (1997). National survey of the impact of low vision device use among veterans. *Optometry and Vision Science*, 74(5):249–259.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., and Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, 6(3):291–298.
- Wetzels, R., Raaijmakers, J. G. W., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, 16(4):752–760.
- White, J. M. and Bedell, H. E. (1990). The oculomotor reference in humans with bilateral macular disease. *Investigative Ophthalmology & Visual Science*, 31(6):1149–1161.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer New York.
- Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40(1):1–29.

- Xie, Y. (2013). knitr: A comprehensive tool for reproducible research in R. In Stodden, V., Leisch, F., and Peng, R. D., editors, *Implementing Reproducible Computational Research*. Chapman & Hall/CRC.
- Yu, D. and Chung, S. T. L. (2011). Critical orientation for face identification in central vision loss. *Optometry and Vision Science*, 88(6):724–732.