Vision Research 76 (2013) 1-10 doi:10.1016/j.visres.2012.09.019

Contents lists available at ScienceDirect

Vision Research

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

The slope of the psychometric function and non-stationarity of thresholds in spatiotemporal contrast vision

Stuart A. Wallis, Daniel H. Baker, Tim S. Meese & Mark A. Georgeson

School of Life and Health Sciences, Aston University, Birmingham B4 7ET, UK. Email: s.a.wallis2@aston.ac.uk

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form 6 July 2012
Available online 4 October 2012

Keywords:
Psychometric slope
Detection threshold
Psychometric function
Spatial frequency
Observer variability

ABSTRACT

The slope of the two-interval, forced-choice psychometric function (e.g. the Weibull parameter, β) provides valuable information about the relationship between contrast sensitivity and signal strength. However, little is known about how or whether β varies with stimulus parameters such as spatiotemporal frequency and stimulus size and shape. A second unresolved issue concerns the best way to estimate the slope of the psychometric function. For example, if an observer is non-stationary (e.g. their threshold drifts between experimental sessions), β will be underestimated if curve fitting is performed after collapsing the data across experimental sessions. We measured psychometric functions for 2 experienced observers for 14 different spatiotemporal configurations of pulsed or flickering grating patches and bars on each of 8 days. We found $\beta \approx 3$ to be fairly constant across almost all conditions, consistent with a fixed nonlinear contrast transducer and/or a constant level of intrinsic stimulus uncertainty (e.g. a square law transducer and a low level of intrinsic uncertainty). Our analysis showed that estimating a single β from results averaged over several experimental sessions was slightly more accurate than averaging multiple estimates from several experimental sessions. However, the small levels of non-stationarity (SD $\approx$ 0.8 dB) meant that the difference between the estimates was, in practice, negligible.

1. Introduction

Most studies of spatiotemporal contrast vision involve measuring the observers psychometric function: a measure of performance (d' or percent correct) as a function of contrast. This is usually done using a two-interval, forced-choice method (2IFC). The lateral position of the psychometric function is an indication of an observer's sensitivity to the stimulus and the contrast associated with a particularoften arbitraryperformance level (e.g. 75% correct) is sometimes referred to as a 'threshold' (though authors do not always wish to invoke the theoretical concept that this implies). Sometimes, the experimenter is also interested in how performance varies with signal strength. This involves measuring the slope of the psychometric function. When the results are plotted as d' against contrast, on log-log axes, then the psychometric function is approximately a straight line (e.g. Pelli, 1985) and the slope of the psychometric function is given by the gradient of this line (b). When the performance measure is 'percent correct', plotted against log(contrast), then the psychometric function is sigmoidal (S-shaped) in form and often fitted by a Weibull function, for which the slope is given by its β parameter (see results section for details). To fair approximations, $\beta = 1.3b$ (Tyler & Chen, 2000) or $\beta = 1.247b$ (Pelli, 1987). The slope parameter is of interest to experimenters because it can be used to estimate the form of the observers internal signal transducer (Nachmias & Sansbury, 1974) (e.g. linear vs. an accelerating square law), assuming no signal uncertainty (Foley & Legge, 1981; Lu & Dosher, 2008); the level of signal uncertainty (Lasley & Cohn, 1981), assuming a linear transducer (Georgeson *et al*, 2008; Pelli, 1985; Tyler & Chen, 2000); or some combination of the two (Meese & Summers, 2009). Note that if the contrast transducer (r) has the form $r = k.c^p$, where c is stimulus contrast and k is a constant, then in the absence of uncertainty, $b = p$.

The slope parameter is also of interest in contrast discrimination experiments, where very low pedestal levels produce steeper psychometric functions than higher pedestal levels (Bird *et al*, 2002; Meese *et al*, 2006). Similarly, contrast detection of target in noise can show a similar increase in slope as the spectral density of the noise increases (Legge *et al*, 1987; though see Baker and Meese, 2012). Changes in single interval psychometric slope have been used to inform models of decision-making (e.g. Wang, 2002), perceptual learning (e.g. Gold *et al*, 2010) and attention (e.g. Cameron *et al*, 2002). The slope of the psychometric function is also of interest in studies that measure a point of subjective equality and use the slope as a measure of discriminability, as is often done in work on cue combination (e.g. Ernst & Banks, 2002). However, to maintain focus, we restrict

ourselves here to the study of the 2IFC psychometric function for contrast detection (a form of the psychometric function whose lower asymptote is 50% correct).

1.1 Five unanswered questions about the slope of the 2IFC psychometric function

In spite of growing theoretical interest in the slope of the 2IFC psychometric function (e.g. Garca-Perez & Alcala-Quintana, 2007; Lu & Dosher, 2008; Meese *et al*, 2006; Meese & Summers, 2009; Petrov *et al*, 2006) few studies have provided a systematic empirical investigation of this parameter. The most obvious exception is a study by Mayer & Tyler (1986). Those authors measured thresholds and slopes (β) for 500 ms presentations of curved strips of grating for a wide range of sizes (4–48 grating cycles at 12 c/deg) and spatial frequencies (2–26 c/deg for 4 deg patches). Both of these manipulations were performed for gratings placed 3.5 deg into the periphery but only the spatial frequency manipulation was performed when the gratings were centred on the fovea. Mayer and Tyler reported some variation in β across their four observers but found no evidence for variation in β as functions of stimulus size or spatial frequency. On average, they found $\beta = 3.7$ for foveal viewing. Although broad in its scope, this study leaves several questions unanswered. In order of increasing priority these are:

1. Are similar results found using smoothly windowed stimuli such as Gabor patches (here we used log-Gabor stimuli) instead of hard-edged gratings? Although a fairly low-priority question, it is possible that performance in the Mayer & Tyler study was influenced by the high spatial frequency artefacts introduced by the hard-edged windowing of their stimuli.
2. Does the slope of the 2IFC psychometric function vary with stimulus size for foveal viewing? This has theoretical importance for understanding the processes of spatial summation (Tyler & Chen, 2000; see Summers & Meese, 2007 for a preliminary report). Some of the conditions in the present study bear on this issue.
3. Does the slope of the 2IFC psychometric function change when the number of cycles is reduced below 4 (the lower limit used by Mayer & Tyler (1986))? The preliminary cortical filtering stage probably involves receptive fields that respond to fewer than four grating cycles (Meese, 2010) whereas larger gratings are detected by either probability summation amongst multiple mechanisms (Robson & Graham, 1981) or higher-order mechanisms performing spatial pooling (Meese, 2010). An argument has been made for the slope of the psychometric function to be affected by probability summation (Wilson & Bergen, 1979; see also Mayer & Tyler, 1986) and it is plausible that the contrast response characteristic of higher-order pooling mechanisms might be different from that of their lower-order feeder units, as in the case of a cascade of accelerating contrast transducers (Meese & Baker, 2011; Sclar *et al*, 1990). Therefore, the slope of the psychometric function might be informative about the transition from a single (or few) mechanisms to many. More generally, localised stimulus patches containing few stimulus cycles have become the preferred contrast stimulus in vision science (e.g. see the ModelFest project: Watson & Ahumada, 2005) and a study of the slope of the psychometric function for these stimuli is long overdue.
4. Is the slope of the 2IFC psychometric function the same or different for light bars and dark bars? There is evidence from psychophysics that luminance increments and decrements can have different thresholds (e.g. Krauskopf, 1980; Short, 1966) and evidence from retinal anatomy and single-cell physiology that ON and OFF sub-systems in the retina are very distinct both structurally and functionally (e.g. Balasubramanian & Sterling, 2009; Burkhardt, 2011; Field & Chichilnisky, 2007). We asked whether such differences might be reflected in the threshold or slope of the psychometric function.
5. Is the slope of the 2IFC psychometric function the same or different in the two opposite 'speed' corners of spatiotemporal vision? It is thought that the high-speed[1] corner of spatiotemporal vision (high temporal frequency, low spatial frequency) is dominated by the magnocellular pathway and that the slow-speed corner of spatiotemporal vision (low temporal frequency, high spatial frequency) is dominated by the parvocellular pathway (Merigan *et al*, 1991; Merigan & Maunsell, 1990). The contrast responses of P-cells in the retina and lateral geniculate nucleus are far more linear than their M-cell counterparts, which first accelerate with contrast and then saturate (Croner & Kaplan, 1995; Shapley & Perry, 1986). Therefore, if psychophysical performance is determined by mechanisms with similar characteristics to the P- and M-streams, we should expect the slope of the psychometric function to increase with stimulus speed consistent with an increase in the underlying contrast response exponent (p; see above).

1.2 The issue of non-stationarity

There was one other important motivation for our study. The literature on sequential dependencies of observer responses (e.g. Howarth & Bulmer, 1956; Treisman & Williams, 1984) and perceptual learning (e.g. Gold *et al*, 2010) suggests that sensitivity can vary across repeated measures, implying that the observers 2IFC psychometric function is not stationary but slides along the contrast axis over time. Few studies have investigated this systematically, though there is some evidence for such variations from an early study using a now obsolete methodology (Hallet, 1969). If the psychometric function is non-stationary, this has potentially important implications for its measurement (Frund *et al*, 2011). When data are gathered from multiple experimental sessions (blocks), often spread over several days, there are two main ways in which investigators proceed. Data are either (i) collapsed across multiple sessions and a single fit performed to estimate threshold and slope (the 'pool-then-fit' method), or (ii) fitted separately for each session, and threshold and slope derived by averaging the multiple estimates (the 'fit-then-pool' method). The pool-then-fit method has the advantage of lessening the effects of binomial error inherent in the data because the fits are made to larger data sets. However, it has the disadvantage that the slope of the psychometric function will be underestimated if the observer is non-stationary, because it involves pooling multiple psychometric functions with different thresholds.

1.3 Aims and outcomes

To address the five questions posed above and the issue of non-stationarity, we measured the psychometric function for a large set of widely varying spatiotemporal stimuli and repeated this several times over several days. We analysed our results using both the pool-then-fit method and the fit-then-pool method. We found no systematic effect of stimulus type on the slope of the psychometric function (with only one exception) but did find low levels of non-stationarity. However, the amount of non-stationarity was so small that it had little impact on our estimates of pool-then-fit slopes, whereas the fit-then-pool slopes

[1] When we use the term 'speed' we refer to the scalar quantity given by dividing temporal frequency by spatial frequency. We do not imply that the stimulus is drifting.

were slightly over-estimated, due to undersampling. Thus, for well-practised observers at least, our conclusion is that the pool-then-fit method is slightly more accurate than the fit-then-pool method, but these small effects are unlikely to be of much practical concern.

2. Method

2.1 Equipment

Stimuli were displayed on a Nokia MultiGraph 445X CRT with a frame rate of 120 Hz using a CRS ViSaGe stimulus generator to render pseudo-14-bit greyscale resolution. The mean luminance of the central region (512×512 pixels; $10.7^\circ \times 10.7^\circ$) of the display was 60 cd/m^2. The surrounding region of the display was dark. Gamma correction was performed to ensure linearity over the full range of target contrasts. Observers sat in a dark room at a viewing distance of 91 cm with their head in a chin and headrest. The experiment was controlled by a PC.

2.2 Stimuli

There were 13 stimuli: 8 log-Gabor patches, 4 Gaussian bars and 1 Gaussian blob. Log-Gabors are similar in appearance to a conventional Gabor (a sinusoidal grating modulated by a 2D Gaussian), except that the carrier is not perfectly sinusoidal (Meese, 2010). Unlike conventional Gabor stimuli, they have the attractive property of containing no DC component for any carrier phase, including cosine phase. The set of log-Gabors was designed to span a range of sizes (full width of the envelope at half height: 163, 81.5, 40.8 and 10.2 min arc) whilst keeping the number of cycles constant (Figure 1, a-d), and to span a range of spatial frequencies (0.25, 0.5, 1 and 4 cycles/deg) whilst keeping the size constant (Figure 1, a and e-g). The pairings of b and e, c and f and d and g also allowed us to investigate the effects of varying stimulus size for a constant spatial frequency. There was also an elongated version of the smallest log-Gabor (Figure 1h). This was for direct comparison with the bar stimuli to test the possibility that a bar of a single polarity (dark or light) would result in less stimulus uncertainty (and hence a shallower psychometric function) than a bar containing both dark and light regions. Log Gabor stimuli were created in the Fourier domain. They were Cartesian separable (see Appendix C of Meese, 2010, for details) and were in positive cosine phase with stimulus centre (i.e. had a central light bar). The Gaussian bars (Figure 1i-l), were dark (Figure 1i, k) or light (Figure 1j, l) and wide (Figure 1i, j) or narrow (Figure 1k, l). The bar widths and lengths were matched to the central bars of the appropriate log-Gabor stimuli (e.g. Figure 1a). Contrast was defined as $\Delta L/Lb$, where Lb is the background luminance, and ΔL is the absolute difference between Lb and L_{max} (light bars or log-Gabors) or L_{min} (dark bars).

All of the stimuli above were temporally modulated by a positive 100ms pulse, which had an abrupt onset and offset. In two other 'temporal' conditions (designed to test the magno/parvo distinction described in the introduction), the temporal envelope was different from this. In a fast condition, a Gaussian blob (σ = 104.1 min arc, Figure 1m) was presented at 15Hz for 1 cycle of a temporal square-wave (4 frames light, then 4 frames dark). In the 'slow' condition, the contrast of the smallest log-Gabor (Figure 1d, n) was slowly ramped on and off by a Gaussian envelope whose full width at half-height was 400ms.

Groups of 4 fixation points (each 2 × 2 pixels) were used to avoid the masking of small targets by a single central fixation points (Summers & Meese, 2009). The fixation points were designed to reduce uncertainty by cueing the size and shape of each stimulus (Figure 1 and caption).

2.3 Procedure

We first estimated the approximate threshold for each stimulus using a staircase procedure. We then used the method of constant stimuli (MCS) with 6 contrast levels spaced at 2 dB intervals to determine the full psychometric function in each condition. Both procedures used a two-interval forced-choice (2IFC) technique, where one temporal interval contained the target and the other interval was blank. The onset of each interval was indicated by an auditory tone and the duration between the two intervals was 400 ms. Observers were required to select the interval containing the stimulus using one of two mouse buttons to indicate their response. Correctness of response was provided by auditory feedback, and the computer selected the order of the intervals randomly.

There were 26 trials (the first 12 conditions) or 20 trials (the final two 'temporal' conditions) for each contrast level, randomly interleaved from each of the six MCS levels. An additional 2 practice trials using the highest contrast level were included at the start of each session to indicate the target identity. Responses to these trials were ignored. One session was completed for each of the conditions in a random order. This process was repeated a

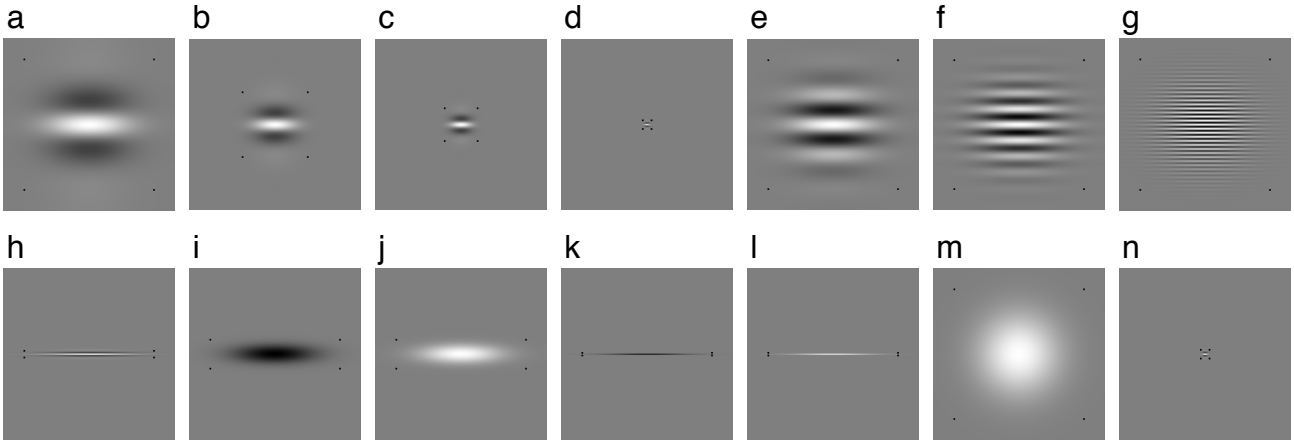

Figure 1: High contrast examples of the stimuli used in each of the 14 conditions. Fixation points are shown here at double size for clarity. They were placed symmetrically about the centre of the image and for the largest log-Gabors and the Gaussian blob (a, e-g, m) they were separated horizontally and vertically by 482 min arc. For the other 'circular' log-Gabors (b-d, n) the placement of the fixation points was scaled in proportion to the stimulus size. For the elongated log-Gabor and the Gaussian bars, the fixation points were separated horizontally by 482 min arc. The vertical separation was 30 min arc for the elongated log-Gabor (h), 110 min arc for the wide bars (i and j) and 10 min arc for the narrow bars (k and l).

further 11 times, with a different random order of conditions for each repetition. To ensure that the range of MCS contrast levels straddled each threshold, the psychometric functions were checked after each session and the range adjusted as appropriate. Thus, each final psychometric function was spread over 6-8 contrast levels, with up to 312 trials (12 × 26) per level. Data were collected from a total of 25,344 trials per observer.

Before data collection began, the following rejection and replacement criterion was set to lessen the impact of unreliable estimates of threshold. If the standard error of a threshold estimate from a single session was greater than 3 dB (estimated by probit analysis), the data for that condition were discarded and the condition was rerun. Only two out of 336 thresholds were rejected by this criterion. The standard error calculated by probit analysis was used only for assessing this rejection criterion.

2.4 Observers

Two psychophysically experienced observers performed the experiment. They were two of the authors (SAW and DHB). Both had normal uncorrected vision.

3. Results

3.1 Pool-then-fit method

For each observer, the data for each condition were pooled across sessions and fitted, using a maximum likelihood method (Wichmann & Hill, 2001), with a Weibull function defined as

$$W = \left(1 - \exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right)\right) \tag{1}$$

where α is the threshold, β is the slope and x is contrast in percent. This was scaled for 2IFC proportion correct by,

$$p = 0.5\,(W\,(1-\lambda)) + 0.5 \tag{2}$$

where λ is the lapse rate and controls the upper asymptote of the psychometric function. The proportion of lapses was free to vary between 0% and 1%, but constrained to a common value across all 14 conditions for each observer. The fitting was implemented in *Matlab* by using *Palamedes* functions (Kingdom & Prins, 2010) and the resulting lapse rate was 0.008 for each observer.

3.2 Fit-then-pool method

We also explored a second method of combining data across sessions. A Weibull function was fitted to the data from each session (168 fits per observer, 120 or 156 trials per fit), with the proportion of lapses fixed at 0.008 for each of the two observers (determined from the pool-then-fit method). This provided 12 estimates of threshold and slope for each condition, which were averaged (geometric mean) across the 12 repetitions. The geometric mean was used rather than the arithmetic mean because the distribution of slopes was not normal in linear units (Lilliefors test: $k = 0.131$, $p<0.001$) but was closer to normal when transformed to log units (Lilliefors test: $k = 0.048$, $p = 0.063$), as illustrated in Figure 2a, b.

Figure 3 shows the best and worst fit (defined as lowest and highest deviance) of the 168 (14 conditions × 12 repetitions) fitted Weibull functions to each observers data. Only 19 of the 336 fits produced a p(deviance) less than 0.05. This represents 5.65% of the fits, and is close to the expected value of 5%, suggesting that the Weibull function provides an acceptable fit to this set of data.

The top two panels of Figure 4 show the thresholds from the fits to each of the 168 psychometric functions for each observer. As expected, thresholds varied across the different targets. This is of little interest here other than to note that the variation in sensitivity with spatial frequency is consistent with the shape of a typical contrast sensitivity function when the size of the stimulus is fixed (Campbell & Robson, 1968) and that the peak of this function shifts substantially to lower spatial frequencies when the number of cycles is fixed (a, b, c & d; Watson & Ahumada, 2005).

Averages for each condition are shown by the red and black horizontal lines for the pool-then-fit and the fit-then-pool methods, respectively. The error bars show 95% confidence intervals. The superposition of the red and black lines in the top two rows of Figure 4 confirms that estimates of threshold were very similar for the two methods of analysis.

The slopes of the psychometric functions are shown in the bottom two panels of Figure 4. These were similar across the 14 stimulus conditions and where they did vary, this was not consistent across observers (discussed further below). For each observer, the slopes (β) were always slightly shallower for the pool-then-fit method (geometric mean: $\beta = 2.78$, 95% conf: 2.32-3.35) than for the fit-then-pool method (geometric mean: $\beta = 3.16$, 95% conf: 2.52-3.96). This is consistent with a small level of drift (non-stationarity) of the observers thresholds over time. (We describe Monte Carlo simulations that attempt to quantify this non-stationarity in the next section.)

A two-way ANOVA was performed on the rank transformed[2] slopes of the psychometric functions (Conover & Iman, 1981), which revealed a significant effect of condition ($F_{13,308} = 2.29, p = 0.007$) and observer ($F_{1,308} = 7.6, p = 0.006$) but no significant interaction ($F_{13,308} = 0.98, p = 0.469$). Post-hoc analysis, with Bonferroni correction, revealed that the significant effect of condition arose only from a difference between the 'elongated' condition (steep slopes in Figure 4, target h) and the 'slow' condition (shallow slopes in Figure 4, target n). All other pairwise comparisons between conditions were not significant, including the three pairwise comparisons where the spatial frequency was constant but the stimulus size varied (b and e; c and f; d and g).

To improve the strength of our conclusions, we performed a second analysis that used Akaike's Information Criteria (AIC; Akiake, 1974) to compare the fit of 16 competing models to the pool-then-fit data. (We used the pool-then-fit results because our analysis below indicates that they provide a slightly better estimate of the underlying 'true' psychometric slopes than the fit-then-pool estimates; see ahead to Figure 5). One model allowed all 28 slopes (14 per observer) to be free. The second model constrained the slopes to a common value for the 14 conditions of each observer, but allowed SAW's slope to differ from DHB's slope. The remaining 14 models each allowed the slopes for one condition to be free, but the other 13 conditions to be constrained to a common value for each observer. In all models, thresholds were free to vary across conditions and observers. For each model, the likelihood of the data given the model was computed, and this value (L) was used to calculate the AIC, given by

$$AIC = -2\,\ln(L) + 2k \tag{3}$$

where k is the number of free parameters in the model.

We found that the lowest value of AIC (46527.16, indicating the 'best' model) was provided by the model that constrained the slope to a common value for 13 conditions and allowed it to be free for the 'elongated' condition (see Figure 4h). The 2nd lowest value of AIC (46531.82) was provided by the model that constrained the slope to a common value for 13 conditions and allowed it to be free for the slow' condition (see Figure 4n). These findings are consistent with the conclusion from the ANOVA post hoc test, above, that a statistically significant difference exists

[2]We initially examined the distributions of the log-transformed slopes for normality, but the high outlier in DHB's slope for the 'slow' condition (Figure 4n) meant that normality could not be assumed.

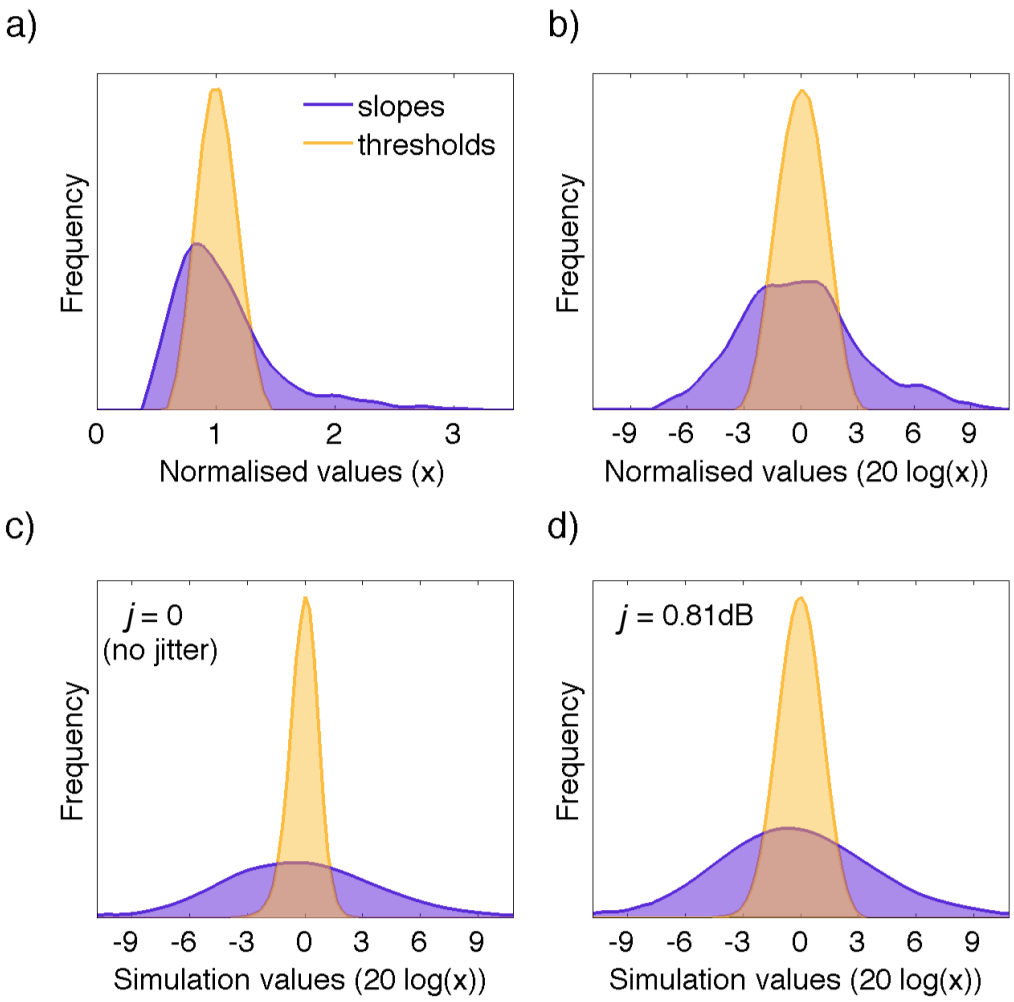

Figure 2: Figure 2. a) Distributions of normalized Weibull slopes and thresholds pooled across two observers, shown on linear axes (x refers to threshold (α) or slope (β)). Each histogram is based on 336 individual psychometric functions (14 conditions × 12 sessions × 2 observers) and for each observer the results for each condition were normalized by their mean. b) The same results as in (a), but with a logarithmic abscissa. Distribution of slopes has positive skew in (a) but is approximately normal on log axes (b). c) Distributions of Weibull slopes and thresholds from the Monte Carlo simulations for a stationary observer (see text for details). d) As c, but for a non-stationary observer with threshold jitter, $j = 0.81$ dB. In all four panels, the frequency axis is normalized to the peak of the threshold function.

between these two conditions.

Although the 2nd best model produced an AIC value that was 4.66 higher than the best model, the remaining models produced AIC values that were 5.86 to 13.87 higher. Given this variability in AIC values, it can be difficult to intuit the magnitude of a difference in AIC that represents a substantial difference in the abilities of two models to fit the data, over and above data 'noise'. In order to resolve this problem, we calculated the Akaike weights (Wagenmakers & Farrell, 2004), which can be interpreted as the probability that a given model is the best model. The set of weights is given by

$$w = \exp \frac{-\Delta\alpha}{2} \tag{4}$$

followed by

$$w = \frac{w}{\Sigma\,(w)} \tag{5}$$

where $\Delta\alpha$ is the vector of AIC values minus the minimum AIC value. For the 16 models considered here, the weights of the best and 2nd best models were 0.80 and 0.08 respectively. Thus it can be concluded that the best model does represent a substantial improvement over its competitors.

We also computed the set of AICs and weights for each observers data separately. For DHB, the pattern of results was similar to both observers combined, described above. But for SAW, there were 4 models that had low AIC values, and they produced weights from 0.14 to 0.22, suggesting no clear 'winner'. Thus, the identification of the best-fitting model appears to be driven primarily by DHB's data and not SAW's, and this is consistent with his steeper slopes for the 'elongated' condition, apparent in Figure 4h.

Close inspection of Figure 4 suggests that for SAW (but not DHB), there was a small decrease in slope as spatial frequency increased when the number of cycles was fixed (targets a-d, fourth row of Figure 4), but this spatial frequency effect was not replicated when the target size was fixed (targets a, e, f and g) and its cause (if real) remains unclear. Neither observer showed a consistent variation of slope with bar polarity (targets i-l). For SAW the slope of the psychometric function was slightly steeper for the 'fast' condition than for the 'slow' condition but there was little evidence of this difference for DHB, and the ANOVA post hoc test above demonstrated that this difference was not statistically significant. Nevertheless, to check this more thoroughly, we gathered data from an additional two observers (RJS and ASB) for these two conditions. They were both experienced psychophysical observers and were naive to the aims of the study. Independent t-tests were performed on the log-transformed slopes for these

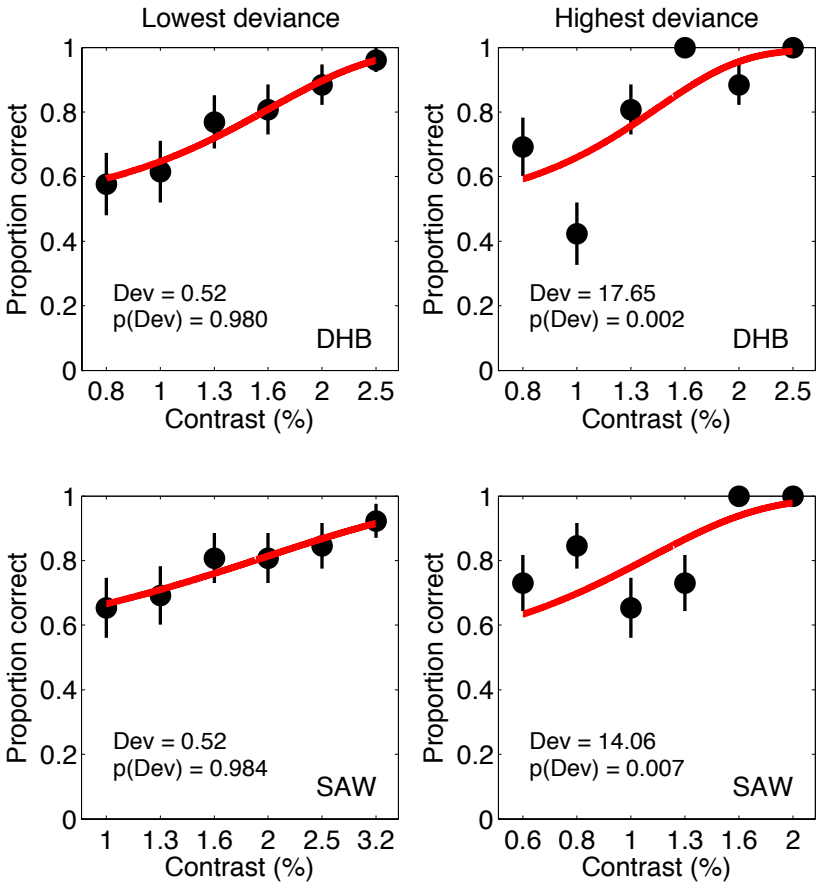

Figure 3: Figure 3. Example data and fitted Weibull functions. Rows show data from each observer, and columns show the best (left) and worst (right) fits (defined as lowest and highest deviance) of the 168 Weibull functions to the results. Error bars show the expected binomial error (±1se) given by $se = \sqrt{(p\,(1-p))\,/n}$, where p is the proportion correct, and n is the number of trials at a given contrast level.

two conditions for observers SAW, RJS and ASB, and a Mann-Whitney U test was performed on DHB's slopes (because of the outlier in his 'slow' condition). This revealed that there was no significant difference between the 'fast and 'slow' conditions for any of the four observers, even without correcting for the use of multiple tests (SAW: t(22)=1.62, p=0.12; RJS: t(22)=1.03, p=0.31; ASB: t(22)=0.12, p=0.90, DHB: U=173, p=0.71).

4. Monte Carlo Simulations

Our results show that estimations of the slope of the psychometric function depend to a small extent on the way in which the results are pooled. This is to be expected if the location of the observers threshold fluctuates a little over sessions. How might we estimate the magnitude of this non-stationarity? It would be overestimated by the standard deviation of the distribution of threshold estimates from different sessions because even for a stationary observer, this would be non-zero owing to binomial errors in the data. To tackle this problem, we performed Monte Carlo simulations for various levels of simulated non-stationarity to estimate the level needed to account for the differences in the estimates of the empirical slopes from the two methods of analysis. Details of the simulations were as follows.

We ran 1400 Monte Carlo simulations of every condition of the experiment, each of which had the same number of contrast levels, trials per datum and repetitions as were used in the experiment. In keeping with the empirical results, the simulated proportion of lapses was fixed at 0.008. In other words, on every simulated trial there was a 0.8% probability that no signal event was simulated on that trial (equivalent to the observer missing the entire trial), in which case there was a 50% probability that the trial was recorded as correct. For a range of generative slopes, we estimated the slope of the psychometric function using the pool-then-fit and fit-then-pool methods at various levels of non-stationarity, j. This was the standard deviation of normally-distributed jitter, in logarithmic (dB) units, applied to the 'true' generative threshold between simulated experimental sessions. Thus, we assumed that the observer was stationary within a session but non-stationary between sessions (i.e. across different days).

When the observer was stationary ($j = 0$, Figure 5a), the simulated pool-then-fit slopes (red line) were very close to the generative slope (grey dashed line). The simulated fit-then-pool slopes (black line) were slightly steeper than the generative slopes, by an amount that increased with the generative slope. For a typical experimental slope of $\beta = 3$, the simulated fit-then-pool slope was $\beta = 3.3$. This small overestimation is an inherent consequence of undersampling (Wichmann & Hill, 2001) by this method (in experiment and simulations) and can be completely overcome in the simulations by increasing the number of simulated trials (effectively, this is shown by the red line in Figure 5a).

Non-stationarity ($j > 0$) can have no effect on the estimate of slope using the fit-then-pool method, since the overall estimate of slope derives from those measured within each session where we have assumed (in our simulations) that the observer is stationary (see next section). This is confirmed in Figure 5 by the black line, which is identical for the three levels of non-stationarity considered here ($j = 0$, $j = 0.75$ dB and $j = 0.98$dB; different panels). However, as the simulated observer became

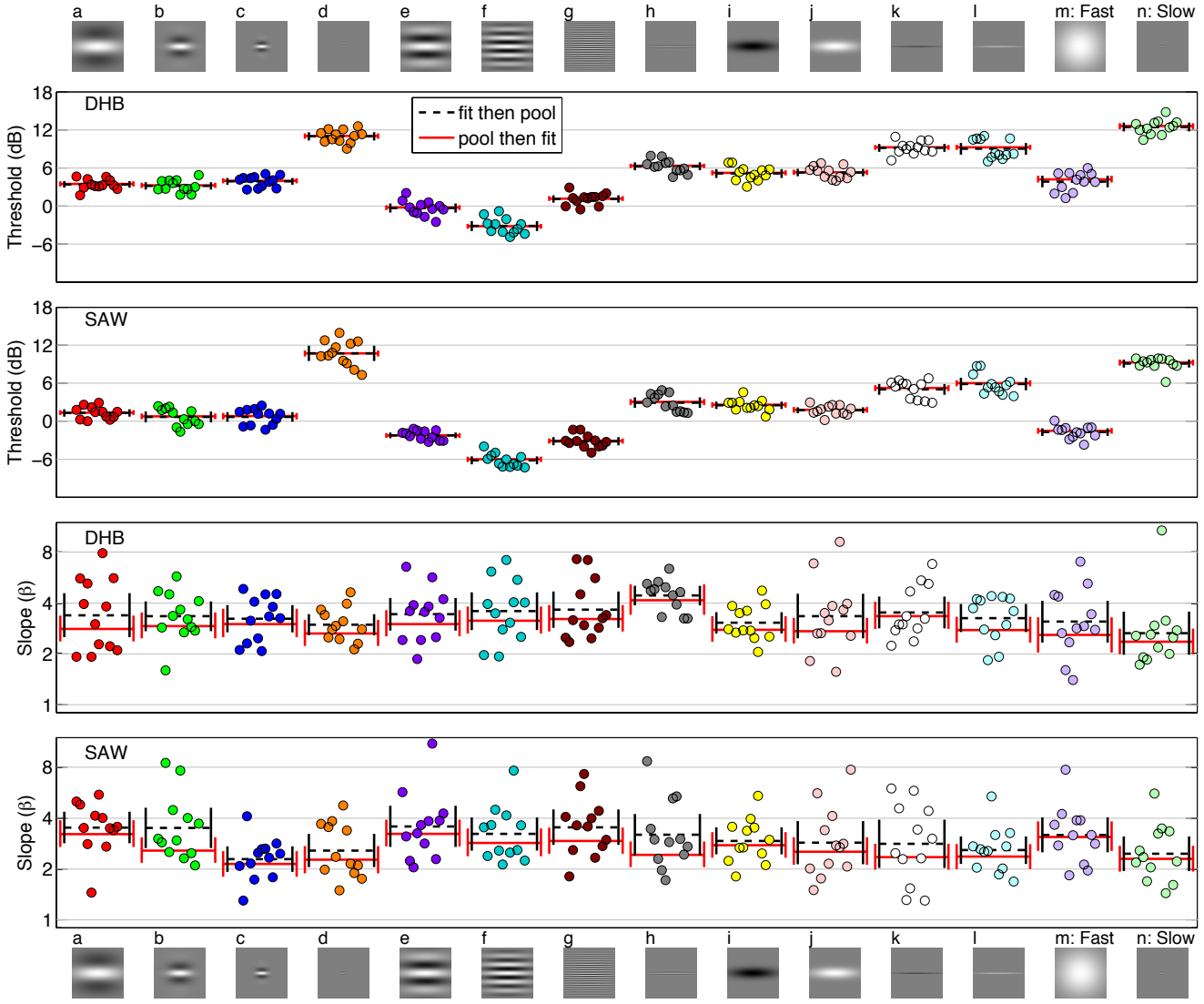

Figure 4: Figure 4: Thresholds and slopes for each observer and condition. The red horizontal lines show the results of fitting to the pooled data (the pool-then-fit method) and the black dashed horizontal lines show the geometric mean across the 12 repetitions of each condition (the fit-then-pool method). Error bars are 95% confidence intervals, obtained from bootstrapping (pool-then-fit) or 2.2 × se (fit-then-pool).

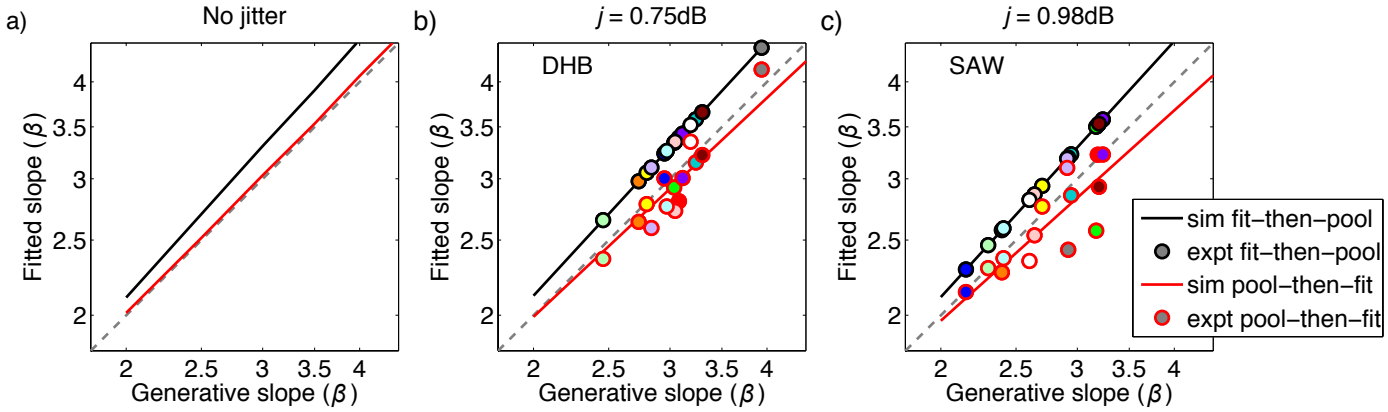

Figure 5: Figure 5: Simulated Weibull slopes (solid lines) for: a) a simulated stationary observer ($j = 0$); b) observer DHB with simulations using $j = 0.75$ dB and c) observer SAW with simulations using $j = 0.98$ dB. Symbols show experimental slopes derived from the pool-then-fit estimates (red-edged symbols) and the fit-then-pool estimates (black-edged symbols). They have the same y-axis values as the red and black markers in Figure 4. The horizontal position of the pair of symbols (same fill colour) from each experimental condition is fixed by forcing the fit-then-pool estimates (black-edged symbols) to fall on their simulated values (black line).

less stationary (j increased), the simulated pool-then-fit slopes decreased, underestimating the true generative slopes. This occurred because the generative slope was 'blurred' by the non-stationarity across sessions.

The symbols in Figure 5 (b and c) are the empirical slopes estimated using the pool-then-fit (red-edged symbols) and fit-

then-pool methods (black-edged symbols). These were replotted from Figure 4, as follows. We associated the fitted slope parameters from the fit-then-pool results with the fitted slopes from the simulations. By forcing these points to fall on the black line this set their horizontal position and solved the inverse problem of estimating the true slope of the empirical psychometric function from the slopes generated by the fits to the original data. By association, this also set the horizontal positions of the pool-then-fit estimates (i.e. there is one red edged and one black edged estimate in each pair of points with identical positions along the x-axis). The level of non-stationarity, j, was then varied in increments of 0.01 dB to find the minimal (summed square) error between the logarithm of the simulated and experimental pool-then-fit slopes (the red lines in Figure 5b and c). The optimal values were $j = 0.75$ dB (DHB) and $j = 0.98$ dB (SAW).

4.1 The overall variability of thresholds is predicted by combining two sources

The variability in threshold estimates from undersampling the psychometric function (in the fit-then-pool method) was estimated from the simulation of the stationary observer. Figure 2c shows this distribution, which has a standard deviation of 0.76 dB. Our estimates of the non-stationarity for each observer (expressed as standard deviations) are given by j in the previous section (e.g. Figure 5b and c). Summing the variances of these two sources of variability (undersampling and non-stationarity) for each observer gives us very good predictions (SD = 1.08 dB for DHB and SD = 1.24dB for SAW) for the overall variability in our estimates of threshold across experimental sessions (SD = 1.08 dB for DHB and SD = 1.18 dB for SAW).

Although the variability of thresholds is well predicted by combining the two sources above, we cannot rule out the possibility that observers were non-stationary within individual experimental sessions as well, even though each of these lasted only 4-5 minutes. Relevant factors might include drifting of attention, learning or adaptation effects (either to the target waveform or the background luminance). However it was not practical to examine these possibilities because dividing the results from each session into smaller parts reduced the number of trials to an extent that made the estimates in the slopes of the psychometric functions too unreliable.

5. Discussion

We gathered extensive data (42,048 trials, 336 psychometric functions) to address the issue of observer non-stationarity, and to answer five questions about the slope of the psychometric function set out in the introduction. The answers to those questions are as follows.

1. The use of smoothly windowed stimuli here produced similar results to Mayer & Tyler (1986), in that pooled psychometric slopes remained fairly constant ($\beta = 2.78 \pm 0.07$) across all conditions. Moreover, our slopes were of similar magnitudes to Mayer & Tyler's observers (3.24 ± 0.39 for DD, 5.08 ± 0.72 for MM and 2.45 ± 0.24 for JB in that study). This suggests that the high spatial frequency artefacts introduced by the hard-edged windowing of their stimuli had little or no effect on the psychometric slopes.
2. The 2-way ANOVA of the 12 rank-transformed measures of slope from each of the 14 conditions for observers SAW and DHB revealed no significant difference between the stimuli that had the same spatial frequency but varied in size (Figure 1, stimuli b and e; c and f; d and g). Thus, we found no evidence that the slope of the psychometric function depends on stimulus size for foveal viewing.
3. The similarity of the slopes of the psychometric functions for stimuli with few cycles (Figure 4, targets a-d) compared to those with many cycles (Figure 4, targets e-g) shows that the slope of the psychometric function does not change when the number of cycles is reduced below four. This extends Mayer and Tyler's (1986) conclusions to grating patches containing small numbers of cycles, including single bars.
4. The empirical slopes and thresholds were very similar for light bars and dark bars (Figure 4, targets i-l). Many, but not all, previous studies have reported consistently lower detection thresholds for decrements than increments, by about 2 dB (0.1 log unit) (e.g. Krauskopf, 1980; Patel & Jones, 1968; Short, 1966). In these earlier studies such a small difference might be attributed to a criterion shift (though it would have to be a surprisingly consistent one). The light/dark asymmetry in thresholds was more prominent at low background luminances where threshold contrasts were higher (Patel & Jones, 1968; Short, 1966). Indeed, contrast level may be the key factor, because when the retinal response to luminance is nonlinear and compressive, the response gain for increments is lower than decrements. The difference may be trivial for small luminance changes (low contrasts) but very significant at high contrasts (see Kingdom & Whittle, 1996; McIlhagga & Peterson, 2006, for a full discussion). Consistent with this argument, in a forced-choice study similar to ours, Legge & Kersten (1983) reported that thresholds for dark bars were on average just 0.04 log units (0.8 dB) lower than for light bars (their table 2). Thus our data reinforce the conclusion that at photopic luminances, and with forced-choice methods, light and dark bars are almost equally detectable. Our data show that the psychometric slopes are also equal. This implies that nonlinearity and/or uncertainty in the response to contrast are the same for localized increments and decrements. Light-dark asymmetries arise only at much higher contrasts.
5. We compared the slopes of the psychometric functions from the two opposite 'speed' corners of the spatiotemporal frequency domain. For SAW, the slope was slightly steeper for the 'fast' stimulus than the 'slow' stimulus (Figure 4, targets m and n), as we had anticipated (see Introduction), but this difference was not significant and not at all apparent in the results from the other three observers. Thus, we were unable to find any evidence for a difference in the slopes of the psychometric functions for 'fast' and 'slow' (flickering) stimuli. This implies that any differences (e.g different exponents) in the early contrast responses to these stimuli are irrelevant at the point of the decision variable. Assuming that our stimuli were successful in differentially tapping the M- and P-pathways, then one interpretation (following Birdsall's theorem and Lasley & Cohn, 1981) is that performance limiting noise is injected after the nonlinearity that distinguishes the M- and P- pathways, but also after subsequent (e.g. cortical) nonlinearities (response exponents or uncertainty) that control and equate the slope of the psychometric function across the various stimulus conditions. With this arrangement, the distinction between the nonlinearities of the M- and P-pathways would be lost in the performance data. Put another way, our results imply that observers did not tap the direct outputs of pure M- and P-pathways here because that would have produced differing slopes in the psychometric functions.

5.1 Polarity uncertainty

Wallis & Georgeson (2007) examined detection performance for Gaussian bars, and found that the slope of the psychometric function was slightly steeper when there was uncertainty about the polarity of the target (the bar could be light or dark on

each trial) compared with when its polarity was known. We wondered whether a similar increase in slope might occur for a stimulus containing adjacent light and dark thin bars, where small fixation errors might induce uncertainty about polarity, when compared with the slope for a thin bar of known polarity. The ANOVA revealed no significant difference between the 'elongated' log-Gabor condition (Figure 1h) and the thin light or dark bars (Figure 1k, l). Thus, polarity uncertainty (if present) appears to have little or no impact on the slope of the psychometric function for these stimuli.

5.2 The slope of the psychometric function is invariant with stimulus condition

In general, the slopes of the psychometric functions showed no consistent departure from the mean across any of the conditions, apart from the difference between the 'elongated' log-Gabor and 'slow' conditions described above (and the reason for that difference is unclear). This general uniformity suggests a common form of nonlinear contrast transducer, or constant intrinsic stimulus uncertainty, or a fixed contribution from the combination of both factors. For example, using Monte Carlo simulations, it can be shown that if the transducer is a square-law ($p = 2$) and the observer monitors about two or three times as many mechanisms as are useful, then these effects will combine to predict our average $\beta = 2.78$ to 3.16 (from our two different methods of analysis).

5.3 Pool-then-fit slopes vs fit-then-pool slopes

One of the main aims of this study was to discover how best to combine data sets across multiple experimental sessions. Figure 4 shows that there was a small difference in the empirical estimates of the slopes of the psychometric functions (0.38 β units) when they were derived by fitting to each of 12 sessions and then pooling (fit-then-pool method), compared with a single fit to data pooled across the 12 sessions (pool-then-fit method). In every condition, the slope from the fit-then-pool method was slightly steeper than that from the pool-then-fit method. Which of the two methods is most appropriate for estimating the true slope of the psychometric function? Using the same curve-fitting methods and the same number of trials as typically used in psychophysical experiments, the Monte Carlo simulations in Figure 5a indicate that part of the difference between the two estimates derives from the undersampling of the psychometric function in the session-by-session basis of the fit-then-pool method and that this causes the slope of the psychometric function to be slightly overestimated (i.e. to be slightly too steep) (see also Wichmann & Hill, 2001). This is quantified in Figure 5a, which shows that when the threshold was completely stationary the fit-then-pool slopes (y) are given by[3]

$$log_{10}(y) = 1.09 log_{10}(x) - 0.004 \quad (6)$$

where x = generative β. This error could be reduced by substantially increasing the number of simulated trials for each psychometric function either by increasing the number of trials per contrast level or the number of contrast levels, or a combination of the two. However, this is usually impractical in experimental situations where time constraints can be important. Can estimation be improved by using the pool-then-fit method where the large number of trials in the pooled psychometric functions mitigate the effects of undersampling in the fit-then-pool method? Unfortunately, Figure 5 shows that this method also comes with a cost, since the blurring of the psychometric function caused by non-stationarity causes its slope to be underestimated (i.e. to be slightly too shallow). For one of our observers (DHB) the blurring was fairly minor, leading to only small errors in the estimate (Figure 5b). For the other observer (SAW) the blurring was a little more severe, causing the magnitude of the errors to approach those inherent in the fit-then-pool method, but in the opposite direction (Figure 5c). Nonetheless, it might be argued that this is the preferred method for estimating the slope of the psychometric function because the solid red lines lie closer to the generative slopes (dashed grey lines) than do the black lines. In fact, our initial concern that the pool-then-fit method would be unduly compromised by the non-stationarity of the psychometric function was not borne out because the level of non-stationarity for our experienced observers was so small. However, as the level of non-stationarity (j) became more severe in the simulations, the method eventually underperformed the pool-then-fit method (not shown). This could be a problem in circumstances where the levels of non-stationarity are greater than those estimated here. These situations could include results from less experienced observers or results from observers spread over a longer period of time. In these cases, a better estimate of the slope of the psychometric function might be achieved by taking the average of the pool-then-fit and the fit-then-pool estimates.

The slope from the fit-then-pool method was slightly steeper than that from the pool-then-fit method in all 14 conditions for both observers. Although this pattern looks systematic, the magnitude of this effect is very small. Nevertheless, it raises the interesting possibility that a future study addressing the sole question of non-stationarity would further delineate between these two pooling rules.

5.4 Conclusion

For practised observers, non-stationarity of the psychometric function is of little practical concern, meaning that reasonable estimates can be achieved using either the fit-then-pool method or the pool-then-fit method. Computational models for contrast detection can be simplified by assuming a stationary observer and a slope of the psychometric function that is common across stimuli that vary in area, number of spatial cycles, spatial frequency, contrast polarity and 'speed'.

5.5 Acknowledgments

This work was supported by BBSRC grant BB/H00159X/1 and EPSRC grant EP/H000038/1 awarded to Mark Georgeson and Tim Meese. We would like to thank two anonymous reviewers and Ingo Frund for their comments on an earlier draft of this paper.

6 References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6), 716-723. doi: 10.1109/tac.1974.1100705

Baker, D.H. & Meese, T.S. (2012). Zero-dimensional noise: the best mask you never saw. *Journal of Vision*, 12(10), art. 20, 1-12. http://dx.doi.org/10.1167/12.10.20.

Balasubramanian, V. & Sterling, P. (2009). Receptive fields and functional architecture in the retina. *Journal of Physiology-London*, 587(12), 2753-2767.

Bird, C. M., Henning, G. B. & Wichmann, F. A. (2002). Contrast discrimination with sinusoidal gratings of different spatial frequency. *Journal of the Optical Society of America A*, 19(7), 1267-1273.

Burkhardt, D.A. (2011). Contrast processing by ON and OFF bipolar cells. *Visual Neuroscience*, 28(1), 69-75.

Cameron, E., Tai, J.C. & Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision Research*, 42(8), 949-967. doi: 10.1016/s0042-6989(02)00039-1.

[3]The values of 1.09 and 0.004 were obtained from a least squares fit (employing Matlab's fminsearch function) between the fit-then-pool slopes and a linear function (on logarithmic axes).

Campbell, F.W. & Robson, J.G. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology*, 197(3), 551-556.

Conover, W.J. & Iman, R.L. (1981). Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics. *The American Statistician*, 35(3), 124-129.

Croner, L.J. & Kaplan, E. (1995). Receptive fields of P-ganglion and M-ganglion cells across the primate retina. *Vision Research*, 35(1), 7-24.

Ernst, M.O. & Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870), 429-433.

Field, G.D. & Chichilnisky, E.J. (2007). Information Processing in the Primate Retina: Circuitry and Coding. *Annual Review of Neuroscience*, 30(1), 1-30.

Foley, J.M. & Legge, G.E. (1981). Contrast detection and near-threshold discrimination in human vision. *Vision Research*, 21(7), 1041-1053.

Frund, I., Haenel, N.V. & Wichmann, F.A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, 11(6), art. 16, 1-18, doi: 10.1167/11.6.16.

Garca-Perez, M.A. & Alcala-Quintana, R. (2007). The transducer model for contrast detection and discrimination: formal relations, implications, and an empirical test. *Spatial Vision*, 20, 5-43.

Georgeson, M.A. Yates, T.A., & Schofield, A.J. (2008). Discriminating depth in corrugated stereo surfaces: Facilitation by a pedestal is explained by removal of uncertainty. *Vision Research*, 48(21), 2321-2328.

Gold, J.I., Law, C., Connolly, P. & Bennur, S. (2010). Relationships Between the Threshold and Slope of Psychometric and Neurometric Functions During Perceptual Learning: Implications for Neuronal Pooling. *Journal of Neurophysiology*, 103(1), 140-154. doi: 10.1152/jn.00744.2009.

Hallett, P. E. (1969). Variations in Visual Threshold Measurement. *Journal of Physiology-London*, 202(2), 403-419.

Howarth, C.I. & Bulmer, M.G. (1956). Non-random sequences in visual threshold experiments. *Quarterly Journal of Experimental Psychology*, 8(4), 163-171.

Kingdom, F. & Prins, N. (2010). *Psychophysics: a practical introduction*. London: Academic Press.

Kingdom, F.A.A. & Whittle, P. (1996). Contrast discrimination at high contrasts reveals the influence of local light adaptation on contrast processing. *Vision Research*, 36(6), 817-829.

Krauskopf, J. (1980). Discrimination and Detection of Changes in Luminance. *Vision Research*, 20(8), 671-677.

Lasley, D.J. & Cohn, T.E. (1981). Why luminance discrimination may be better than detection. *Vision Research*, 21(2), 273-278.

Legge, G.E., Kersten, D. & Burgess, A.E. (1987). Contrast discrimination in noise. *Journal of the Optical Society of America A*, 4(2), 391-404.

Legge, G.E. & Kersten, D. (1983). Light and dark bars - contrast discrimination. *Vision Research*, 23(5), 473-483.

Lu, Z.L. & Dosher, B.A. (2008). Characterizing observers using external noise and observer models: Assessing internal representations with external noise. *Psychological Review*, 115(1), 44-82.

Mayer, M.J. & Tyler, C.W. (1986). Invariance of the slope of the psychometric function with spatial summation. *Journal of the Optical Society of America A*, 3(8), 1166-1172.

McIlhagga, W. & Peterson, R. (2006). Sinusoid = light bar plus dark bar? *Vision Research*, 46(12), 1934-1945.

Meese, T.S. (2010). Spatially extensive summation of contrast energy is revealed by contrast detection of micro-pattern textures. *Journal of Vision*, 10 (8), art. 14, 1-21, doi: 10.1167/10.8.14.

Meese, T.S. & Baker, D.H. (2011). Contrast summation across eyes and space is revealed along the entire dipper function by a "Swiss cheese" stimulus. *Journal of Vision*, 11(1), art. 23, 1-23, doi: 10.1167/11.1.23.

Meese, T.S., Georgeson, M.A. & Baker, D.H. (2006). Binocular contrast vision at and above threshold. *Journal of Vision*, 6(11), art. 7, 1224-1243, doi: 10.1167/6.11.7.

Meese, T.S. & Summers, R.J. (2009). Neuronal convergence in early contrast vision: Binocular summation is followed by response nonlinearity and area summation. *Journal of Vision*, 9(4), art. 7, 1-16, doi: 10.1167/9.4.7.

Merigan, W.H., Katz, L.M. & Maunsell, J.H.R. (1991). The Effects of Parvocellular Lateral Geniculate Lesions on the Acuity and Contrast Sensitivity of Macaque Monkeys. *Journal of Neuroscience*, 11(4), 994-1001.

Merigan, W.H. & Maunsell, J.H.R. (1990). Macaque vision after magnocellular lateral geniculate lesions. *Visual Neuroscience*, 5(4), 347-352.

Nachmias, J. & Sansbury, R.V. (1974). Grating contrast - discrimination may be better than detection. *Vision Research*, 14(10), 1039-1042.

Patel, A.S. & Jones, R.W. (1968). Increment and Decrement Visual Thresholds. *Journal of the Optical Society of America*, 58(5), 696-699.

Pelli, D.G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A*, 2(9), 1508-1532.

Pelli, D.G. (1987). On the relation between summation and facilitation. *Vision Research,* 27(1), 119-123.

Petrov, Y., Verghese, P. & McKee, S.P. (2006). Collinear facilitation is largely uncertainty reduction. *Journal of Vision*, 6(2), 170-178, doi: 10.1167/6.2.8.

Robson, J.G. & Graham, N. (1981). Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research*, 21(3), 409-418.

Sclar, G., Maunsell, J.H.R. & Lennie, P. (1990). Coding of image contrast in central visual pathways of the macaque monkey. *Vision Research*, 30(1), 1-10.

Shapley, R. & Perry, V.H. (1986). Cat and monkey retinal ganglion-cells and their visual functional roles. *Trends in Neurosciences*, 9(5), 229-235.

Short, A.D. (1966). Decremental and incremental visual thresholds. *The Journal of Physiology*, 185(3), 646-654.

Summers, R.J. & Meese, T.S. (2007). Area summation is linear but the contrast transducer is nonlinear: Models of summation and uncertainty and evidence from the psychometric function. *Perception*, 36(5) suppl, 38.

Summers, R.J. & Meese, T.S. (2009). The influence of fixation points on contrast detection and discrimination of patches of grating: Masking and facilitation. *Vision Research*, 49(14), 1894-1900.

Treisman, M. & Williams, T.C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, 91(1), 68-111.

Tyler, C.W. & Chen, C.C. (2000). Signal detection theory in the 2AFC paradigm: attention, channel uncertainty and probability summation. *Vision Research*, 40(22), 3121-3144.

Wallis, S.A. & Georgeson, M.A. (2007). Seeing light vs dark lines: Psychophysical performance is based on separate channels, limited by noise and uncertainty. *Perception*, 37(2), 315.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1), 192-196.

Wang, X. (2002). Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron*, 36(5), 955-968. doi: 10.1016/s0896-6273(02)01092-9.

Watson, A.B. & Ahumada, A.J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, 5(9), 717-740, doi: 10.1167/5.9.6.

Wichmann, F.A. & Hill, N.J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8), 1293-1313.

Wilson, H.R. & Bergen, J.R. (1979). A four mechanism model for threshold spatial vision. *Vision Research*, 19(1), 19-32.