Author/s:
Rubinstein, NJ;Turpin, A;Denniss, J;McKendrick, AM

Title:
Effects of criterion bias on perimetric sensitivity and response variability in glaucoma

Date:
2021-01-01

Citation:
Rubinstein, N. J., Turpin, A., Denniss, J. & McKendrick, A. M. (2021). Effects of criterion bias on perimetric sensitivity and response variability in glaucoma. Translational Vision Science and Technology, 10 (1), pp.1-12. https://doi.org/10.1167/tvst.10.1.18.

Persistent Link:
https://hdl.handle.net/11343/272791

tvst

Article

Effects of Criterion Bias on Perimetric Sensitivity and Response Variability in Glaucoma

Nikki J. Rubinstein[1,2], Andrew Turpin[2], Jonathan Denniss[3], and Allison M. McKendrick[1]

[1] Department of Optometry and Vision Sciences, The University of Melbourne, Australia
[2] School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia
[3] School of Optometry & Vision Science, University of Bradford, Bradford, UK

Correspondence: Allison M. McKendrick, Department of Optometry and Vision Sciences, The University of Melbourne, Parkville, 3010, Victoria, Australia. e-mail: allisonm@unimelb.edu.au

Received: January 22, 2020
Accepted: November 4, 2020
Published: January 8, 2021

Keywords: Perimetry; visual field; glaucoma; criterion; variability; bias

Citation: Rubinstein NJ, Turpin A, Denniss J, McKendrick AM. Effects of criterion bias on perimetric sensitivity and response variability in glaucoma. Trans Vis Sci Tech. 2021;10(1):18, https://doi.org/10.1167/tvst.10.1.18

Purpose: The purpose of this study was to isolate and quantify the effects of observer response criterion on perimetric sensitivity, response variability, and maximum response probability.

Methods: Twelve people with glaucoma were tested at three locations in the visual field (age = 47–77 years, mean deviation = −0.61 to −14.54 dB, test location Humphrey field analyzer [HFA] sensitivities = 1 to 30 dB). Frequency of seeing (FoS) curves were measured using a method of constant stimuli with two response paradigms: a "yes-no" paradigm similar to static automated perimetry and a criterion-free two interval forced choice (2IFC) paradigm. Comparison measures of sensitivity, maximum response probability, and response variability were derived from the fitted FoS curves.

Results: Sensitivity differences between the tasks varied widely (range = −11.3 dB to 21.6 dB) and did not correlate with visual field sensitivity nor whether the visual field location was in an area of steep sensitivity gradient within the visual field. Due to the wide variation in differences between the methods, there was no significant difference in mean sensitivity between the 2IFC task relative to the yes-no task, but a trend for higher sensitivity (mean = 1.9 dB, SD = 6.0 dB, $P = 0.11$). Response variability and maximum response probability did not differ between the tasks ($P > 0.99$ and 0.95, respectively).

Conclusions: Perimetric sensitivity estimates are demonstrably altered by observer response criterion but the effect varies widely and unpredictably, even within a single test. Response bias should be considered a factor in perimetric test variability and when comparing sensitivities to nonperimetric data.

Translational Relevance: The effect of response criterion on perimetric response variability varies widely and unpredictably, even within a single test.

Introduction

Functional effects of glaucoma on vision are commonly measured using static white-on-white automated perimetry (SAP). SAP sensitivity measurements are often used as an estimate of underlying true visual sensitivity in clinical science; for example, in relating perimetric estimates to structural measurements,[1–4] and in computer simulations of new perimetric procedures.[5–11]

In SAP, sensitivities are measured using a type of yes-no procedure in which the patient presses a button if they perceive a stimulus. In this procedure, there is an explicit "yes" button press) and an inferred "no" response, implied by the lack of a button press within a certain duration of the stimulus presentation. The patient's decision to press the button in this type of procedure is presumably guided by some internal model that the patient builds of what the stimulus looks like in terms of size, shape, location, and intensity. This model is often referred to within the psychophysical literature as the patient's "criterion", and this criterion can change both within and between tests, known as "criterion drift" or "criterion shift". This changing of internal criterion over time is a well-known form of cognitive bias that affects psychophysical measurements made by yes-no procedures.[12,13]

The yes-no task described above is used in clinical SAP tests in order to efficiently measure sensitivity at many spatial locations within a clinically acceptable test duration. In laboratory psychophysics, where more time is available, criterion bias can be reduced by the use of certain forced choice tasks. In a two-interval forced choice (2IFC) detection task, the observer is asked in which of two possible time intervals the stimulus appeared. In this case, the observer does not require an internal criterion for responding, as they simply report the interval in which they experienced the strongest sensation. As long as the two intervals are symmetrical in time and the stimulus occurs in either with equal probability on each trial, the task is assumed to be criterion free.

Although the existence of criterion bias in clinical SAP has long been recognized as a source of variability, it has not previously been quantified. Such bias will contribute to test-retest variability, observed learning effects, and the variability ubiquitously seen in structure-function studies. Quantifying such bias and identifying possible predictors of its direction and magnitude may be valuable in improving clinical tests and understanding of the relationships between them.

In this study, we explore and quantify the effects of response criterion in a yes-no SAP task by comparing thresholds, response variability, and maximum probability of seeing with those measured using a 2IFC task. Data were collected for a group of patients with glaucoma at a range of locations with different sensitivities, under conditions otherwise similar to those of clinical SAP tests.

Methods

Participants

Participants were recruited from a database of previous research participants. The study adhered to the tenets of the Declaration of Helsinki and was approved by the Human Research Ethics Committee of The University of Melbourne (HREC 1646955.2). Written informed consent was obtained from each participant.

Participants were required to meet the following inclusion criteria: an established clinical diagnosis of glaucoma, best-corrected visual acuity of 6/12 or better in the tested eye, no active ocular pathology, exclusive of changes associated with glaucoma, present on anterior and posterior segment biomicroscopic investigation (lens changes minimally affecting visual acuity and isolated retinal drusen normal for age were acceptable), areas of visual field loss with sensitivities < 20 dB on their most recent 24-2 SITA Standard visual field test (Humphrey Field Analyzer II; Carl Zeiss Meditec, Jena, Germany), and reliable visual field test results ($\leq$ 20% fixation losses and $\leq$ 15% false positive responses) on two SITA Standard 24-2 tests.

Data were collected from one eye each of 12 people with glaucoma (7 women; 8 right eyes; age range = 47–77 years). One additional participant was excluded from the study due to an inability to perform the 2IFC task described below. The range of mean deviations (MDs) on SITA Standard 24-2 examination was -0.61 to -14.54 dB (mean = -9.04 dB).

Equipment

Experimental software was run on a desktop computer (Optiplex 9010; Dell, Round Rock, TX, USA). White circular Goldmann size III (0.43 degrees diameter) luminance increment stimuli were displayed on the Octopus 900 (Haag-Streit Diagnostics, Bern, Switzerland); a projection-based bowl perimeter with a maximum stimulus luminance of 3183 cd/m^2 (10,000 asb) and background luminance of 10 cd/m^2 (31.4 asb). Four green dots presented centrally in the shape of a diamond were used to direct fixation during all experimental tasks. Responses were collected using a game controller (F310 Gamepad; Logitech, Lausanne, Switzerland).

Experimental Software

Experimental software was custom-written using the Open Perimetry Interface (OPI).[14] The OPI is a freely available R (http://www.r-project.org/, in the public domain)[15] package that can be used to control the Octopus 900 via the Eyesuite software available with the Octopus 900 (i8.2.0.0; Haag-Streit Diagnostics).

Test Procedure

A person's most recent Humphrey field analyzer (HFA) 24-2 SITA Standard result was used to identify potential study participants (collected between 3 and 16 months prior to testing). A second HFA 24-2 SITA Standard was performed on the first day of testing. The average of the two tests was used as an estimate of visual sensitivity at each location, which in turn was used to seed the collection of FoS curves, select test locations, and to select eyes.

For participants where both eyes were eligible, the eye containing the greatest number of locations with sensitivity estimates below 20 dB was chosen for testing. Three test locations, spanning at least

2 quadrants of the visual field, with average sensitivity estimates from the two SITA fields below 20 dB were chosen for each participant. If 3 locations did not meet this criterion, or a location failed the 0 dB test described below, additional locations were chosen, with an attempt to maximize spatial disparity between test locations. Increasing spatial disparity between locations reduces the incentive for participants to make eye movements away from the fixation target during testing, and divides spatial attention in a fashion that is more similar to clinical SAP. These three locations were determined separately for each participant and used for all experimental tasks. A histogram of the average SITA sensitivity of the test locations is shown in Figure 1B.

The nontested eye was covered by an opaque eye patch. Testing was performed with dim room lighting. Appropriate near spectacle correction for working distance was placed in the lens holder. Participants were instructed to look in the center of the green diamond at all times and eye position was monitored visually by the machine operator at regular intervals using the Octopus 900 display. Participants were reinstructed to maintain central fixation as necessary.

Testing was performed over 2 sessions separated by up to 3 weeks, of up to 1.5 hours duration each. Three different tasks (described in detail below and shown schematically in Fig. 1C) were performed: (1) 0 dB test; (2) FoS curves measured using a 2IFC task; and (3) FoS curves measured using a SAP-like yes-no task.

The longer test times of the 2IFC procedure (requiring 2 stimulus presentations per trial compared to one for the yes-no procedure) necessitated splitting the trials over 2 days to reduce fatigue. During the first test session, participants performed a SITA Standard 24-2 test, the 0 dB test and half the 2IFC procedure. During the second test session, participants performed the yes-no task followed by the remaining trials for the 2IFC procedure. For each test run, 1000 possible randomized stimulus presentation orders were precomputed, with the chosen test order being that which minimized the number of sequential test locations and intensity presentations.

The 0 dB Test

The 0 dB test was run in order to exclude from testing any locations where 0 dB is seen less than 50% of the time. Although the average SITA estimate from the 2 visual fields gives an indication of the degree of remaining functional vision, the high test-retest variability of SITA means that a location that returns 0 dB on one test, has a 90% retest interval from 0 to 24 dB for the next test.[16]

Thirty stimuli of 0 dB intensity were presented to each of the 3 test locations in a pseudo-random, interleaved spatial order. Participants were given 1500 ms to respond to the stimulus via a button press. If an observer detected fewer than 15 stimuli at any location, a new location was chosen and the test repeated.

FoS Curves Measured Using a Two-Interval Forced Choice Task

A method of constant stimuli (MOCS) was used to measure 2IFC FoS curves in which the effects of criterion bias are assumed to be greatly reduced. Stimulus intensity varied in seven steps, the spacing of which depended on the average sensitivity of the two SITA estimates (Table). The step intensities varied between individuals in order to ensure that at least two of the test intensities lay within two standard deviations of the mean of the FoS curve. This individualization was achieved by inspection of the responses during a training phase. Data collection commenced after each participant demonstrated understanding of the task and repeatable performance under observation (minimum number of practice trials was 63). Data from the training phase was not included in the data analysis. For one participant, the step-sizes required further adjustment after the first experimental run. No further adjustment to the step-size was made after this first run for any participants, and the same step intensities were used on the first and second day of testing.

Each trial consisted of two 200 ms test intervals, each preceded by an auditory cue, separated by an interstimulus interval of 750 ms (see Fig. 1C). The test stimulus was presented during either the first or second test interval at one of the test locations, chosen at random. Nothing was presented during the other test interval. Participants responded via a button press, indicating whether the stimulus appeared during the first or second test interval. If unsure, participants were required to make their best guess. Auditory feedback, indicating whether the response was correct, was provided after each trial in order to ensure the participant avoided potential inversion of the buttons. We assume that this auditory feedback does not affect the results in participants who understood the task correctly. The next trial commenced 800 to 1500 ms after the response was registered. This time interval was limited by the time taken for the Octopus 900's projector to move to the next location. This method is assumed to be criterion-free because the two possible stimulus presentations (interval 1 and 2) are symmetrical in time and equally probable.

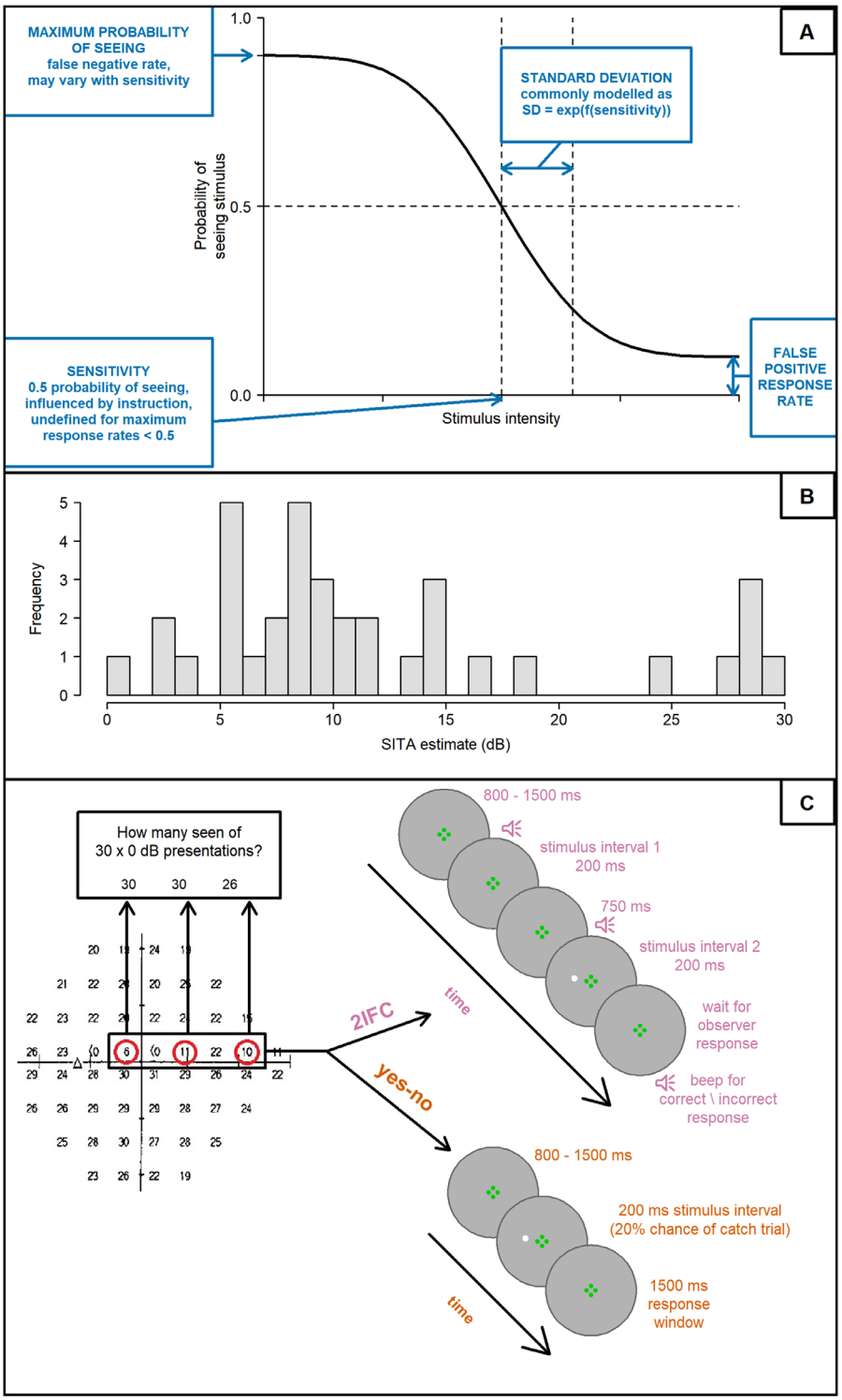

Figure 1. Procedural schematic. Panel (**A**) shows an example of a frequency of seeing curve for a yes-no task. Raw data for each location were fit with a cumulative Gaussian curve for each task. Panel (**B**) shows the distribution of average SITA estimates for all 36 test locations from 12 participants. Panel (**C**) shows the experimental procedure. Three test locations were chosen, as shown on the SITA field result in Panel **C** (participant 2). Participants were required to detect at least 15 of 30 × 0 dB stimuli at each location. Participants then performed a 2IFC MOCS procedure (*pink*) and a yes-no MOCS procedure (*orange*), using step sizes described in the Table. Each test stimulus was presented a total of 30 times for each procedure.

Table. Initial MOCS Step Sizes for Different Average SITA Standard Sensitivity Values, Where S is the Average Threshold Over Two SITA Fields

S, dB	MOCS steps, dB						
0–15	0	4	8	12	16	20	28
16–19	S − 15	S − 8	S − 4	S	S + 4	S + 8	S + 12
20–29	S − 10	S − 4	S − 2	S	S + 2	S + 4	S + 8
30–40	S − 10	S − 2	S − 1	S	S + 1	S + 2	S + 5

Where average SITA sensitivity was 0 to 15dB, initial MOCS steps were always as shown in the first row. For average SITA sensitivities greater than 15 dB, the initial MOCS steps depended on the average SITA sensitivity as shown.

Each intensity value was tested 30 times, resulting in 630 test trials (3 locations × 7 intensity steps × 30 repeats). Testing was broken up into 10 runs, each of 63 trials. One of the 12 participants completed only 15 repeats for each location and intensity (participant 7) due to logistic constraints associated with the long test session time. Between each test run, participants were asked if they wanted to take a short break. If so, participants moved away from the chin-rest to stretch, and then were repositioned according to standard clinical procedures. Room lighting was not altered during these short rest periods and brief readaptation to the perimetry bowl occurred during the realignment and reinstruction by the perimetrist.

FoS Curves Measured Using a SAP-Like Yes-No Task

A MOCS procedure was used with the same step sizes and response windows as used for the 2IFC task (see the Table). Each trial consisted of a single stimulus presentation at a test location and stimulus intensity chosen at random (see Fig. 1C). Observers responded via a button press each time a stimulus was detected. No auditory cues were provided for this task in order to mimic clinical perimetry. Participants were instructed using neutral instructions, similar to those described in Kutzko et al.[17] Unlike the 2IFC method described above, this method is subject to observer criterion effects, as described in the Introduction.

False positive catch trials were presented throughout the test. Each trial had a 20% probability of being a catch trial. During these trials, no stimulus was presented such that any response was recorded as a false positive.

Each intensity value was tested 30 times, resulting in 630 test trials (3 locations × 7 intensity steps × 30 repeats), broken up into 5 runs, each of 126 trials plus catch trials. All 12 participants completed this task.

Analysis

All analyses were performed using the open source statistical programming language R (http://www.r-project.org/, in the public domain)[15] in the RStudio environment.[18] FoS curves were constructed for each of the three tested locations by using a maximum likelihood estimation method to fit the following function:

$$\Psi(x,t) = fp + (1 - fp - fn) \times [1 - G(x,t,s)]$$

where *fp* is the false positive rate defining the lower asymptote of Ψ, *fn* is the false negative rate defining the upper asymptote of Ψ, and $G(x,t,s)$ is the value at x of a cumulative Gaussian function with mean t and standard deviation s. The mean (t), standard deviation (s), and upper asymptote (*fn*) were free parameters in the fitting procedure. The lower asymptote (*fp*) was set at 0.5 for the 2IFC task as per the recommendations of Wichmann and Hill[19] (Fig. 2; Fig. 3), whereas for the yes-no task, *fp* was set at the false positive response rate, estimated as the proportion of false positive catch trials for which a response was detected (i.e. the lower asymptote is set to 0 if no catch trial responses are detected). The difference in lower asymptote position for the 2 tasks is inherent to the nature of the tasks: choosing between 2 alternatives results in a chance level of 50% of picking correctly, whereas for a yes-no task an observer may not detect any stimuli, resulting in a possible 0% detection rate. The FoS curve shown in Figure 1A is for the yes-no task.

FoS curves were used to quantify sensitivity (intensity value corresponding to 0.5 seen for yes-no task and 0.75 correct for 2IFC task), response variability (standard deviation of the fit), and maximum response rate (upper asymptote of the FoS curve). In the absence of criterion effects in the yes-no task, sensitivity and response variability quantified this way are mathematically equivalent between yes-no and 2IFC FoS curves despite the difference in scaling of the FoS curve. Consequently, we assume that within-participant, within-location differences in sensitivity and response variability between the two tasks result from criterion bias in the yes-no task. Because an unseen stimulus has a 50% probability of a correct response in the 2IFC task, false negative rates (*fn*) are expected to differ by a factor of 2 between the 2 methods, such that maximum response rates (MR = 1-*fn*) are expected to be related by the function $1\text{-}MR_{yes\text{-}no} = 1\text{–}2MR_{2IFC}$. Note that maximum

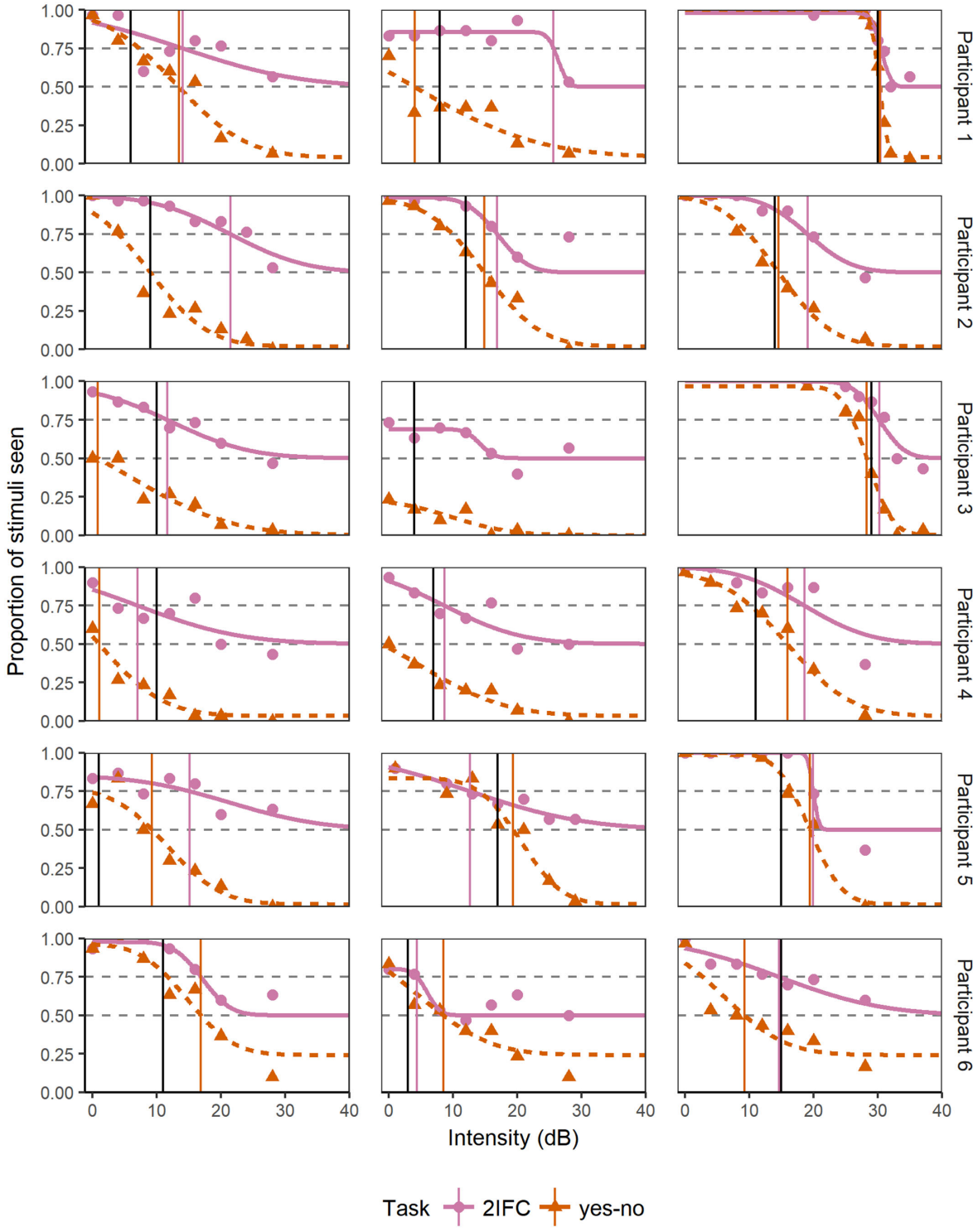

Figure 2. FoS curve fits for participants 1 to 6. Each row comprises the curve fits from three locations. *Pink*: 2IFC. *Orange*: yes-no. *Points*: raw data. *Thick lines*: FoS curve fits. *Vertical lines*: sensitivity values of curve fits (*pink* and *orange*) and average HFA sensitivity (*black*).

response rates were the upper asymptote of the fitted functions and were not constrained to the intensity range of the perimeter, so some maximum response rates were inferred from stimulus intensities below 0 dB.

Comparisons were performed using linear mixed models, accounting for within-subject effects. Models of the form x ~ 1 + (1|participant) in which x denotes the parameter of interest, 1 denotes the intercept representing the fixed effect of mean paired

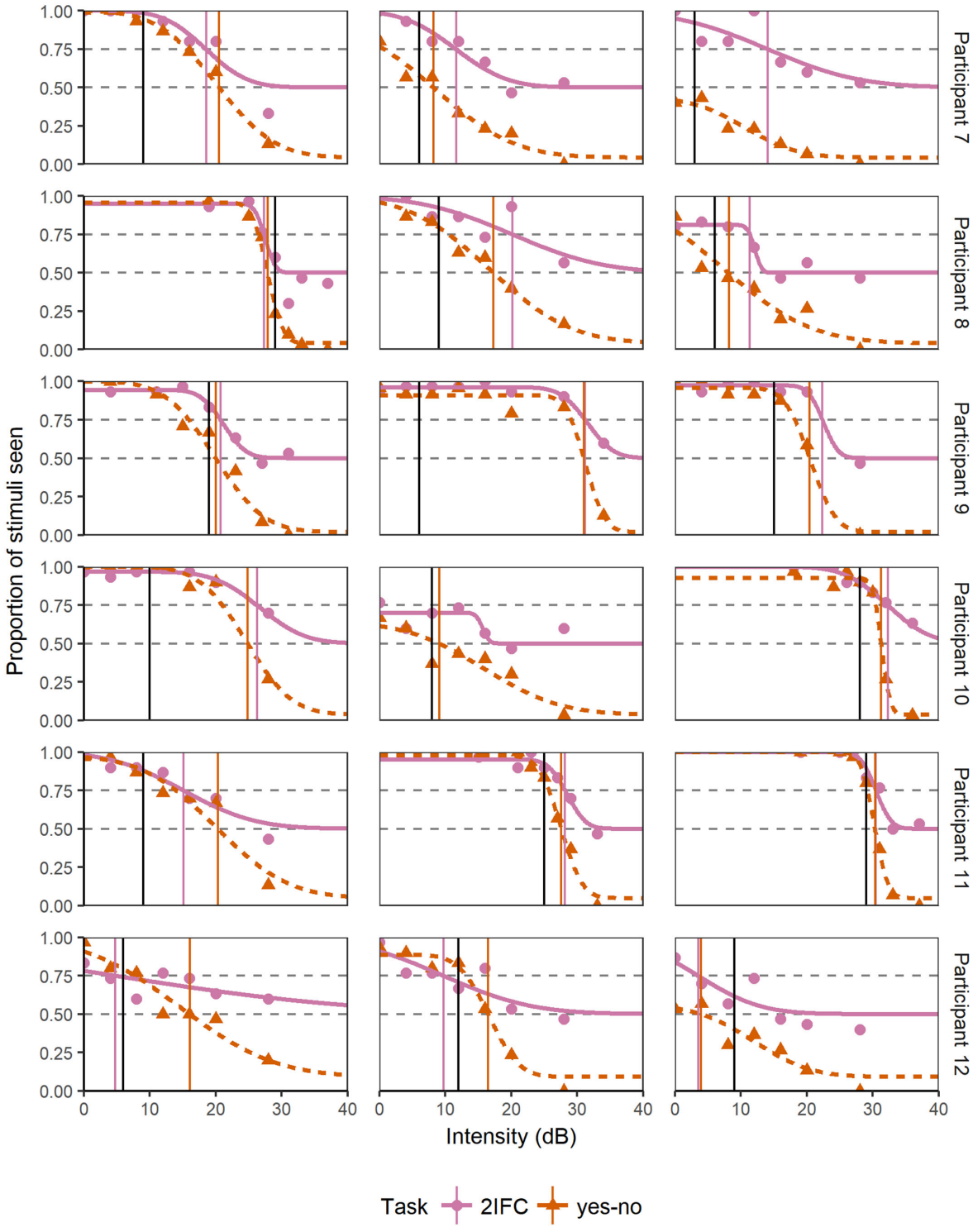

Figure 3. FoS curve fits for participants 7 to 12. Formatting is the same as for Figure 2.

difference between the 2IFC and yes-no tasks, and (1|participant) represents random effects of participant were compared with null models without the fixed effect. Residuals were checked for all models and found to have approximately Gaussian distributions. Models were compared by χ^2 likelihood ratio test, with $P < 0.017$ being considered statistically significant after accounting for familywise error rate

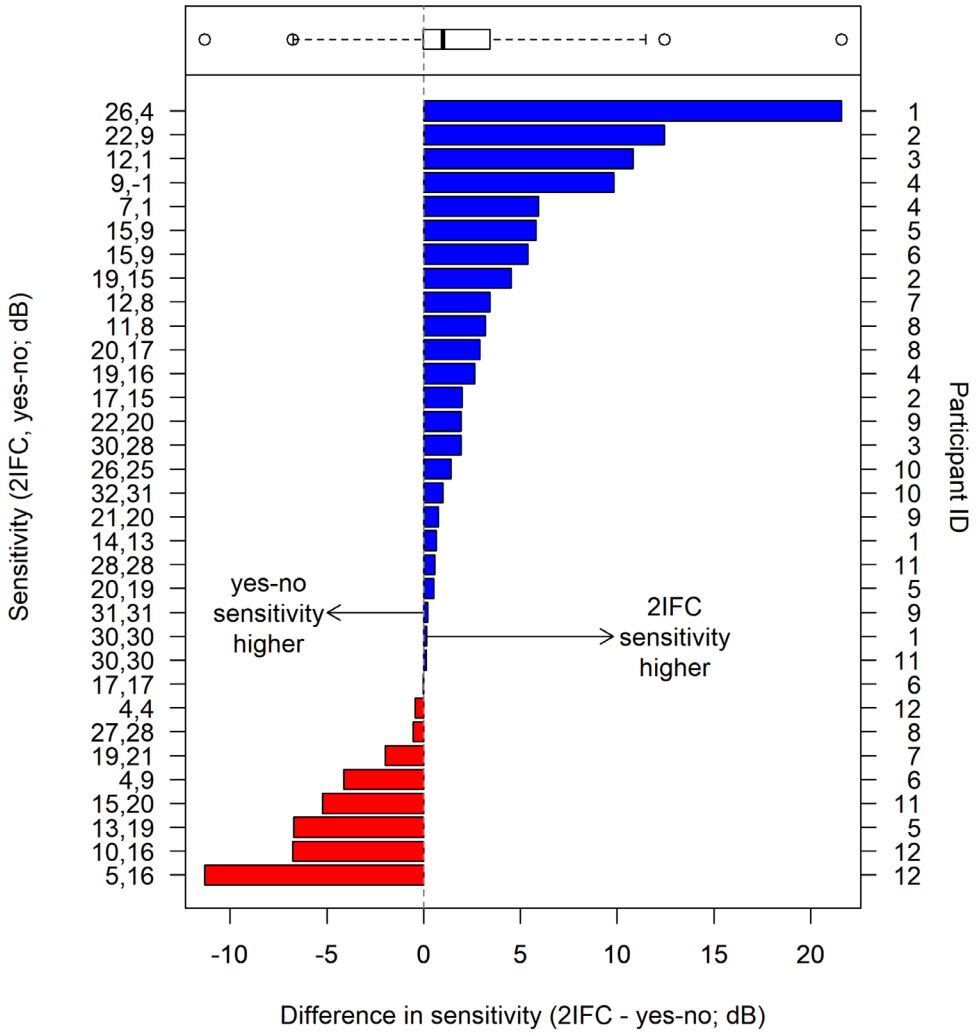

Figure 4. Difference in sensitivity between the 2IFC and yes-no tasks for each location. *Blue* = 2IFC sensitivity greater than yes-no. *Red* = yes-no sensitivity greater than 2IFC. The individual sensitivity values are given on the left axis, rounded to the nearest integer (2IFC, yes-no). The right axis gives the participant number, which corresponds to the participant numbers in Figures 2 and 3. The sensitivity difference data is represented as a boxplot in the *top panel*: *thick line*: median; box: 25th and 75th percentiles; whiskers: 5th and 95th percentiles.

by Bonferroni correction. This approach is analogous to paired *t*-tests while accounting for within-subject effects.

Results

The raw data, along with FoS curve fits, for each of the 3 tested locations for each of the 12 participants are shown in Figures 2 and 3. Two yes-no FoS curves (participants 3 and 7) and two 2IFC FoS curves (participants 3 and 10) had sensitivity estimates that could not be calculated, due to the maximum probability of seeing falling below 0.5 and 0.75, respectively.

The difference in sensitivity between the 2IFC and yes-no tasks is shown in Figure 4. On average, FoS curve sensitivities were higher for the criterion-free 2IFC method but this difference was not statistically significant (mean difference = 1.9 dB, SD = 6.0 dB, $P = 0.11$). Note there is a large range of differences in sensitivity between the 2IFC and yes-no task (−11.3 dB to 21.6 dB). There was no significant relationship between this difference in sensitivity measure and stimulus eccentricity (calculated as the Euclidean distance of the location from the fovea: rho = −0.06, $P = 0.74$).

In order to explore whether the difference in sensitivity between the 2IFC and yes-no task is related to the severity of visual field damage, the left-hand panel of Figure 5 plots this difference against HFA

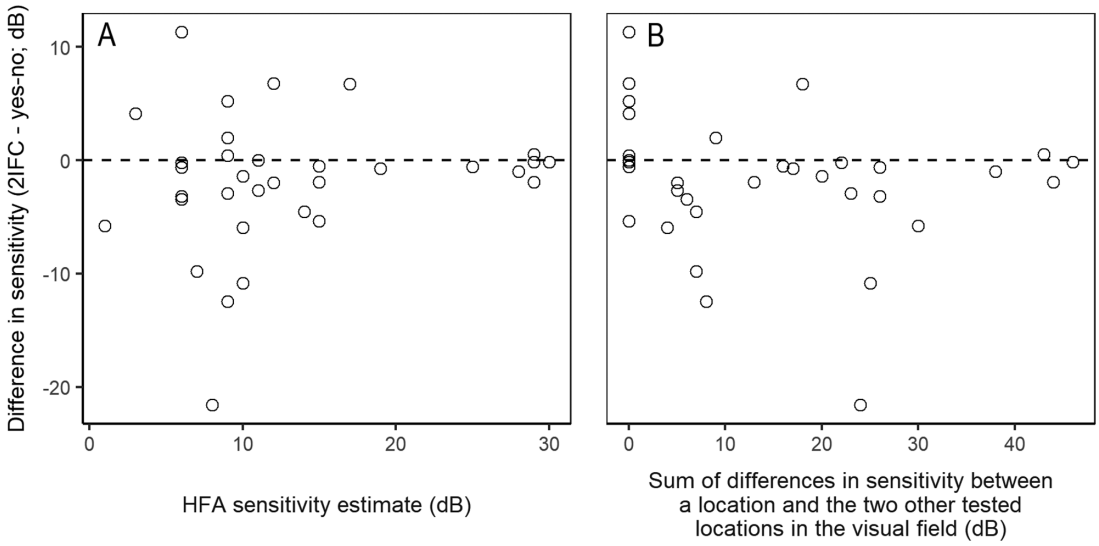

Figure 5. For each tested location, the relative difference in sensitivity between the two methods (2IFC and yes-no) is plotted relative to the HFA sensitivity estimate for the specific location (panel **A**) and the sum of the absolute differences in HFA sensitivity for the specific location and the two other locations tested in the visual field (panel **B**).

sensitivity for each location. These measures were not correlated (Pearson $R^2 = 0.019$, $P = 0.44$). We also explored whether this difference in sensitivity between the 2IFC and yes-no task was related to the magnitude of the sensitivity difference between test locations within an observer (i.e. was the 2IFC result further from the yes-no result when low and high sensitivity locations were used in the same test compared to when only locations with low sensitivity were tested?). The right-hand panel of Figure 5 plots the difference in sensitivity between the two test methods against the sum of differences in HFA sensitivity between each location and the two other locations tested in the visual field. A high sum of differences indicates that a location's HFA sensitivity was very different from the HFA sensitivity of the other two locations tested in the visual field. However, these measures were also not correlated (Pearson $R^2 = 0.037$, $P = 0.28$).

After correction for the expected difference in maximum response rate between the 2IFC and yes-no tasks (see Analysis section), there was no difference in maximum response rates between the two tasks (mean difference = 0%, standard deviation = 0.2%, $P = 0.95$).

No difference was found for the standard deviation of the FoS curves between the 2IFC and yes-no tasks (mean difference [2IFC - yes-no] = 0.0 dB, standard deviation = 5.66 dB, $P > 0.99$). To confirm that the FoS curves measured in this study were consistent with previous studies measuring yes-no FoS curves, the relationship between sensitivity and standard deviation was compared to the data from three previous yes-no FoS curve studies: Henson et al.,[20] Chauhan et al.,[21] and Wall et al.,[22] as shown in Figure 6. The distribution of the data in the current study is different to the distribution of the data in the above-mentioned studies when compared using the Kernel density based global two-sample comparison test (t = 0.0085, z = 1.88, $P = 0.03$). However, this difference is likely caused by

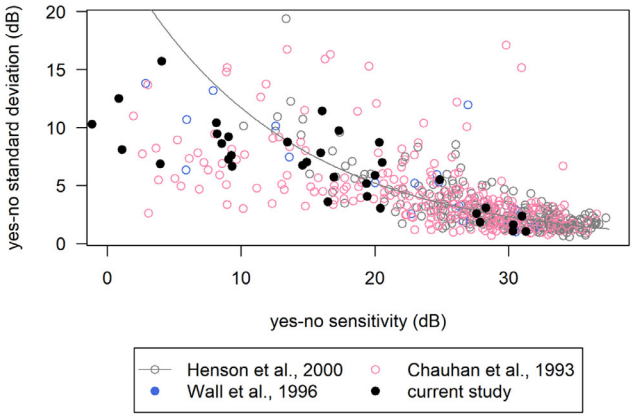

Figure 6. FoS curve standard deviations plotted against sensitivity for the yes-no task overlaid on the data from Henson et al.,[20] Chauhan et al.,[21] and Wall et al.[22] Data were extracted directly from the graphs presented in the aforementioned papers for normal, suspect, and diseased data sets. Interquartile ranges were converted to standard deviations by dividing by 1.349. The exponential curve fit derived by Henson et al.[20] is shown by the *grey line*.

the relatively large numbers of locations with sensitivities greater than 20 dB in the comparison distributions relative to the data in the current study. When this difference is taken into account by only looking at the data points with sensitivity values less than or equal to 20 dB, the two distributions are no longer different ($t = 0.0022$, $z = 0.172$, $P = 0.43$).

Discussion

The results of the current experiment showed that when criterion bias is reduced through the use of forced choice methodology, visual field sensitivity measurements are not predictably different on average than when measured using the yes-no methods used in clinical perimetry. Although the average effect of criterion bias was relatively small (mean = 1.9 dB, standard deviation = 6.0 dB, $P = 0.11$), it formed part of a broad distribution of criterion effects (see Fig. 4) that were not consistent even within a single observer. For example, participant 5 had one location whose sensitivity increased for the 2IFC relative to the yes-no task, one location that was similar for the 2 tasks, and one location that reduced the 2IFC relative to the yes-no task (see Fig. 2). This is despite the test being performed with all three stimulus locations interleaved within the test procedure. There was no relationship between the difference in sensitivity for the two tasks and defect depth, nor with relative asymmetry in sensitivity between that location and the other two tested locations (see Fig. 5). The differences in sensitivity between the two tasks were not accompanied by differences in response variability or maximum response probability, which were both similar between the two tasks.

These experiments were reasonably demanding on the participants. The 2IFC trials were split over 2 visits on separate days in an attempt to reduce the influence of fatigue for this test procedure. Although participants were provided with practice trials to familiarize themselves with the requirements of the task prior to the first session, it is possible that splitting the trials across sessions may have resulted in learning improving performance across the 2 days. To determine whether learning had occurred, FoS curves were fit to the data from each of the two sessions separately. Although no difference was found for the spread ($P = 0.32$) or maximum response probability ($P = 0.66$) of the FoS curves, sensitivity values increased on average from the first to the second day ($P = 0.008$, median [5th to 95th percentile] improvement = 2.1dB [−6.0 dB to 13.5 dB]). It should be noted that the psychometric function estimates based on only 15 trials per level (from a single visit), in people with expected shallow functions are noisy. Indeed, we chose to test over two visits in order to build up sufficient data to improve confidence in the estimates. Nevertheless, our between-session estimate of test-retest sensitivity difference for the 2IFC task is markedly less than the difference between the 2IFC task and the yes-no task (ranging from −11.3dB to 21.6dB).

Due to the requirement for a large number of trials to be collected per visual field location, we only collected data from three spatial locations. While we endeavored to have these as spatially separate as possible, the required division of spatial attention in this task is not the same as for a standard visual field test. Furthermore, the limited number of locations creates a reasonable likelihood of sequential stimuli in the same location. It is possible that the probability of seeing a particular location-intensity combination may differ for sequential versus nonsequential presentations. We randomized the time window between trials to avoid stimuli in the same location appearing more rapidly in sequence than for more distant locations (due to the time taken for the mechanical movement of the projector in the O900). We also attempted to minimize the number of sequential stimuli by precomputing the stimulus order to minimize sequential presentations. The actual percentage of sequential locations tested varied from 13 to 21% between individuals. Hence, for the 30 presentations of any given location-intensity pair in the 2IFC, there were therefore on average 4 to 6 occurrences that were sequential pairs, and 24 to 26 occurrences that were not sequential. However, it is also important to note that these sequential presentations were rarely at the same intensity level, so they may have been both "seen", both "unseen", or one "seen" and one "unseen" regardless of the sequential nature. Unfortunately, this amount of data does not allow robust determination of whether sequential pairs resulted in a different probability of seeing compared with nonsequential occurrences.

Our data did not reveal any obvious covariates to assist in explaining the large variation in the differences between thresholds measured with the forced choice and yes-no procedure. Even within the same individual, there were marked differences between locations, and these were not explained by obvious candidates, such as visual field sensitivity or differences in sensitivity between tested locations. Clearly our sample size is low, hence it is possible that with many more participants, a multivariate analysis might reveal that some proportion of the variance can be explained by some additional parameter. However, from our data here, it seems unlikely that there is a relationship of

sufficient strength to be clinically meaningful, that is, strong enough to be used to incorporate corrections for criterion to SAP sensitivity estimates.

One possible explanation for differences in response criterion shown at different locations of the visual field of the same patient is altered perceptual expectations due to long-standing visual loss. For example, a person may be more conservative in responding to stimuli within a region of known visual field loss, compared to a relatively normal region. Further research could investigate the effects of patient's awareness or longevity of visual field loss on measured sensitivities and criterion bias.

The exponential relationship between sensitivity and variability shown by Henson et al.[20] (shown in Fig. 6) is commonly used in computer simulation experiments with a maximum standard deviation cutoff of 6 dB:[5–7,10,11] standard deviation = exp(−0.081 × "sensitivity" +3.27). The data from this study provides further evidence for the utility of using a cutoff value when using the Henson equation to simulate FoS curves for varying sensitivity levels, as this more closely mirrors empirical data for low sensitivities than the original formula.

The range of possible stimulus intensities used in the current experiment was limited by the hardware used, resulting in a maximum luminance (0 dB) of 3183 cd/m^2. Some FoS curves could not be completely described within this intensity range, requiring extrapolation of results for stimulus intensities greater than 0 dB. An additional limitation of the hardware was that fixation stabilization was not possible at the time of data collection. When test stimuli are presented near the edge of scotomata, small fixational eye movements can affect measurements of sensitivity.[23–25] To minimize the incentive to make eye movements, test locations were spread spatially across at least two quadrants of the visual field. Participants' eyes were monitored visually by the examiner during testing and participants were reinstructed as necessary to maintain fixation. Even in the absence of explicit eye movements, variations in visuo-spatial attention are likely to be present in these experiments relative to standard perimetry because it becomes obvious to participants that stimuli are only appearing in a small number of fixed spatial locations. Repeating this experiment with an increased number of locations could help minimize any change in visuo-spatial awareness. However, the benefits of doing this need to be carefully weighed against increased test times and associated patient fatigue.

In summary, the yes-no task used in most clinical SAP is subject to criterion bias. On average, this results in a small reduction in sensitivity, although the size of the effect varies considerably and unpredictably, including within a single test of a single patient. As such, an individual's response criterion is one of the factors contributing to the variability of perimetric sensitivity measurements and also to the scatter in observed structure-function relationships.

Acknowledgments

Supported by ARC LP130100055; ARC LP150100815 (AT and AMM), College of Optometrists Research Fellowship (JD).

Disclosure: **N.J. Rubinstein**, None; **A. Turpin,** Heidelberg Engineering GmBH (F), Haag-Streit AG (F), CentreVue SpA (C); **J. Denniss**, None; **A.M. McKendrick,** Heidelberg Engineering GmBH (F), Haag-Streit AG (F), CentreVue SpA (C)

References

1. Harwerth RS, Wheat JL, Fredette MJ, Anderson DR. Linking structure and function in glaucoma. *Prog Retin Eye Res*. 2010;29:249–271.
2. Medeiros F, Lisboa R, Weinreb RN, Girkin CA, Liebmann JM, Zangwill LM. A combined index of structure and function for staging glaucomatous damage. *Arch Ophthalmol*. 2012;130:1107–1116.
3. Price DA, Swanson WH, Horner DG. Using perimetric data to estimate ganglion cell loss for detecting progression of glaucoma: a comparison of models. *Ophthalmic Physiol Opt*. 2017;37:409–419.
4. Hood DC, Kardon RH. A framework for comparing structural and functional measures of glaucomatous damage. *Prog Retin Eye Res*. 2007;26:688–710.
5. Ganeshrao SB, McKendrick AM, Denniss J, Turpin A. A perimetric test procedure that uses structural information. *Optom Vis Sci*. 2015;92:70–82.
6. Rubinstein NJ, McKendrick AM, Turpin A. Incorporating spatial models in visual field test procedures. *Trans Vis Sci Tech*. 2016;5:7.
7. Turpin A, Jankovic D, McKendrick AM. Retesting visual fields: utilizing prior information to decrease test-retest variability in glaucoma. *Invest Ophthalmol Vis Sci*. 2007;48:1627–1634.
8. Turpin A, McKendrick AM, Johnson CA, Vingrys AJ. Development of efficient threshold strategies for frequency doubling technology perimetry using

computer simulation. *Invest Ophthalmol Vis Sci*. 2002;43:322–331.
9. Wild D, Kucur SS, Sznitman R. Spatial entropy pursuit for fast and accurate perimetry testing. *Invest Ophthalmol Vis Sci*. 2017;58:3414–3424.
10. Denniss J, McKendrick AM, Turpin A. Towards patient-tailored perimetry: automated perimetry can be improved by seeding procedures with patient-specific structural information. *Trans Vis Sci Tech*. 2013;2(4):3.
11. Chong LX, McKendrick AM, Ganeshrao SB, Turpin A. Customized, automated stimulus location choice for assessment of visual field defects. *Invest Ophthalmol Vis Sci*. 2014;55:3265–3274.
12. Kaernbach C. Adaptive threshold estimation with unforced-choice tasks. *Percept Psychophys*. 2001;63:1377–1388.
13. Klein SA. Measuring, estimating, and understanding the psychometric function: a commentary. *Percept Psychophys*. 2001;63:1421–1455.
14. Turpin A, Artes PH, McKendrick AM. The Open Perimetry Interface: An enabling tool for clinical visual psychophysics. *J Vision*. 2012;12(11):22.
15. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org/; 2017.
16. Artes PH, Iwase A, Ohno Y, Kitazawa Y, Chauhan BC. Properties of perimetric threshold estimates from Full Threshold, SITA Standard, and SITA Fast strategies. *Invest Ophthalmol Vis Sci*. 2002;43:2654–2659.
17. Kutzko KE, Brito CF, Wall M. Effect of instructions on conventional automated perimetry. *Invest Ophthalmol Vis Sci*. 2000;41:2006–2013.
18. RStudio Team. *RStudio: Integrated Development for R. RStudio*, PBC, Boston, MA; 2020, http://www.rstudio.com/.
19. Wichmann FA, Hill NJ. The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Percept Psychophys*. 2001;63:1314–1329.
20. Henson DB, Chaudry S, Artes PH, Faragher EB, Ansons A. Response variability in the visual field: comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes. *Invest Ophthalmol Vis Sci*. 2000;41:417–421.
21. Chauhan BC, Tompkins JD, LeBlanc RP, McCormick TA. Characteristics of frequency-of-seeing curves in normal subjects, patients with suspected glaucoma, and patients with glaucoma. *Invest Ophthalmol Vis Sci*. 1993;34:3534–3540.
22. Wall M, Maw RJ, Stanek KE, Chauhan BC. The psychometric function and reaction times of automated perimetry in normal and abnormal areas of the visual field in patients with glaucoma. *Invest Ophthalmol Vis Sci*. 1996;37:878–885.
23. Demirel S, Vingrys AJ. Eye movements during perimetry and the effect that fixational instability has on perimetric outcomes. *J Glaucoma*. 1994;3:28–35.
24. Maddess T. The influence of sampling errors on test-retest variability in perimetry. *Invest Ophthalmol Vis Sci*. 2011;52:1014–1022.
25. Vingrys AJ, Demirel S. The effect of fixational loss on perimetric thresholds and reliability. *Perimetry Update 1992/93, Proceedings of the Xth International Perimetric Society Meeting 1992; Kyoto, Japan* 1992; Kugler Publications, Amsterdam/New York.