tvst

DOI: 10.1167/tvst.2.4.3

Article

Towards Patient-Tailored Perimetry: Automated Perimetry Can Be Improved by Seeding Procedures With Patient-Specific Structural Information

Jonathan Denniss[1,2] ✉, Allison M. McKendrick[1] ✉, and Andrew Turpin[2] ✉

[1] Optometry and Vision Sciences, The University of Melbourne, Australia
[2] Computing and Information Systems, The University of Melbourne, Australia

Correspondence: Andrew Turpin, Computing and Information Systems, The University of Melbourne, Melbourne, Victoria 3010, Australia; e-mail: aturpin@unimelb.edu.au

Received: 15 January 2013
Accepted: 22 April 2013
Published: 31 May 2013

Keywords: automated perimetry, visual field, perimetry, static perimetry, structure–function

Citation: 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, http://tvstjournal.org/doi/full/10.1167/tvst.2.4.3, doi:10.1167/tvst.2.4.3

Purpose: To explore the performance of patient-specific prior information, for example, from structural imaging, in improving perimetric procedures.

Methods: Computer simulation was used to determine the error distribution and presentation count for Structure–Zippy Estimation by Sequential Testing (ZEST), a Bayesian procedure with prior distribution centered on a threshold prediction from structure. Structure-ZEST (SZEST) was trialled for single locations with combinations of true and predicted thresholds between 1 to 35 dB, and compared with a standard procedure with variability similar to Swedish Interactive Thresholding Algorithm (SITA) (Full-Threshold, FT). Clinical tests of glaucomatous visual fields ($n = 163$, median mean deviation −1.8 dB, 90% range +2.1 to −22.6 dB) were also compared between techniques.

Results: For single locations, SZEST typically outperformed FT when structural predictions were within ± 9 dB of true sensitivity, depending on response errors. In damaged locations, mean absolute error was 0.5 to 1.8 dB lower, SD of threshold estimates was 1.2 to 1.5 dB lower, and 2 to 4 (29%–41%) fewer presentations were made for SZEST. Gains were smaller across whole visual fields (SZEST, mean absolute error: 0.5 to 1.2 dB lower, threshold estimate SD: 0.3 to 0.8 dB lower, 1 [17%] fewer presentation). The 90% retest limits of SZEST were median 1 to 3 dB narrower and more consistent (interquartile range 2–8 dB narrower) across the dynamic range than those for FT.

Conclusion: Seeding Bayesian perimetric procedures with structural measurements can reduce test variability of perimetry in glaucoma, despite imprecise structural predictions of threshold.

Translational Relevance: Structural data can reduce the variability of current perimetric techniques. A strong structure–function relationship is not necessary, however, structure must predict function within ±9 dB for gains to be realized.

Introduction

Standard automated perimetry (SAP) aims to measure contrast sensitivity to luminance increment stimuli at multiple locations across the visual field within a reasonable test time. Commonly, SAP is used for diagnosis and monitoring of visual field damage in patients with glaucoma where the goal is often to detect small changes in the visual field due to disease. Ideally, perimetric tests should make sensitivity estimates that are unbiased, repeatable, and robust to response errors. Laboratory test procedures exist that can achieve these goals, but they require many presentations per visual field location. In a clinical setting, the length of time a patient spends at a perimeter must be balanced against other factors such as clinic time and fatigue. For this reason, commercially available perimeters use adaptive threshold estimation strategies that aim to optimize the trade-off between bias/repeatability and test duration.[1] Sensitivity estimates from these procedures show significant test–retest variability, the magnitude of which is inversely related to sensitivity, becoming

more than half the measurement range of the perimeter in some cases,[2,3] hampering detection of change due to disease progression. A major goal of current research is to improve the precision of sensitivity estimates in SAP while maintaining clinically acceptable test times, generally less than 10 minutes per eye.

The adaptive procedures used in several current, commercially available perimeters (e.g., Swedish Interactive Thresholding Algorithm [SITA],[4] Zippy Estimation by Sequential Testing [ZEST][5]) are seeded with information based on population data rather than information from the specific test patient. One example of patient-specific prior information is the sensitivity estimates obtained from previous visual field tests. Seeding Bayesian procedures with the results of previous tests reduces test–retest variability in computer simulations of perimetric outcomes.[6] Sensitivities from previous tests are used in one commercially-available staircase procedure, German Adaptive Thresholding Estimation (GATE), which has similar variability characteristics to SITA, but reduces test duration.[7]

Since the conception of automated perimetry, there have been major advances in imaging devices that allow quantification of glaucomatous damage to many optic nerve head or retinal nerve fiber layer (RNFL) parameters. Numerous studies have explored relationships between these parameters and visual field damage in glaucoma.[8–11] Although reported relationships are not perfect, it is apparent that structural and functional measurements are not independent, and so measurements from imaging are another source of prior information that could be exploited in perimetric procedures.

The first aim of this study was to investigate, by computer simulation, how closely predictions of visual field sensitivity made from patient-specific prior information need to reflect underlying "true sensitivity" in order to reduce variability compared with current procedures when used to centre the prior distribution in a Bayesian procedure. Using this information, we then investigated what reductions in bias, test–retest variability and test duration could be expected using such a test compared with existing strategies. There are many possible ways in which prior information may be used to seed a SAP test; thus, an exhaustive study of all methods is not feasible. We concentrate on the possibility of using prior information from structural imaging, and provide quantitative results for a basic strategy that explores potential benefits and pitfalls of this approach. The results are broadly applicable to other prior information sources that relate to visual field sensitivity.

Methods

Computer Simulation of Patient Responses to SAP Stimuli

Computer simulation is the only method that allows the comparison of output thresholds of different SAP procedures when the underlying true threshold is known. Both procedure bias and variability are assessed by comparison of the output thresholds to the input ("true") thresholds, for a given set of response conditions. Test duration is estimated by the number of presentations required. To compare test procedures, we must first define a loss metric. For this, we chose mean absolute error (MAE, the mean absolute difference between output threshold and true threshold), as it penalises both bias and variability.[12]

The use of psychometric functions, or "frequency-of-seeing" curves, to model patient responses to psychophysical stimuli is well established. In this study, patient responses to stimuli presented at x decibels were simulated according to psychometric functions (Ψ) modelled by cumulative Gaussian functions with response errors incorporated according to the formula[13]:

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

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, t is the threshold, s is the SD of the cumulative Gaussian curve defining the spread of Ψ, and $G(x, t, s)$ is the value at x of a cumulative Gaussian distribution with mean t and SD s. As the slope of frequency-of-seeing curves for SAP stimuli flattens with decreasing sensitivity,[14,15] s in the above formula was varied with sensitivity according to a formula proposed by Henson et al.[15] that was based on empirical data for healthy subjects

$$s = \exp(-0.066 \times t + 2.81), \quad (2)$$

but capped at 6 dB.

Three response error conditions were modelled for all simulations:

1. Reliable observers: False positive rate (fp) set to 3%, false negative rate (fn) set to 1%.
2. False positive (FP) observers: False positive rate (fp) set to 15%, false negative rate (fn) set to 3%.

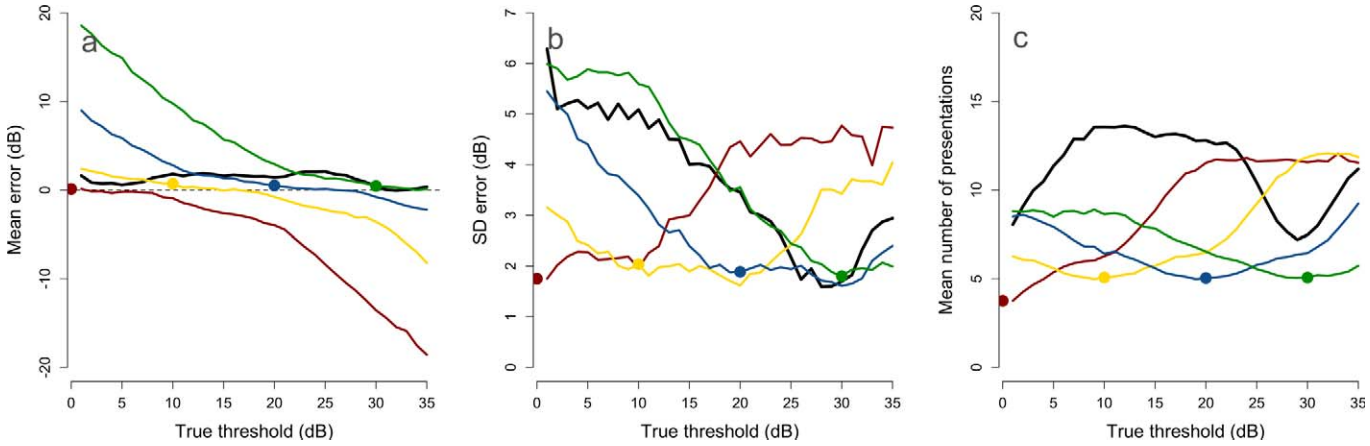


Figure 1. The effect of altering the prior probability mass function (*pmf*) of a ZEST procedure on the mean (a) and SD (b) error of its threshold estimates, and the mean number of presentations (c). The black lines show a standard ZEST procedure with a population-derived prior *pmf*. Colored lines show the ZEST procedure with Gaussian prior *pmfs* (SD 5 dB) as in SZEST, centered on 0 (red), 10 (yellow), 20 (blue), and 30 dB (green). Points on each colored line indicate the center of the prior *pmf*. Procedures were simulated 1000 times for each true threshold in the range of 1 to 35 dB.

3. False negative (FN) observers: False positive rate (*fp*) set to 3%, false negative rate (*fn*) set to 15%.

These conditions determine the rate of genuine response errors and are, therefore, not directly comparable to the estimated false response rates reported in clinical perimetry. While rates of genuine response errors in perimetry cannot be measured in real patients, we chose relatively high rates for FP and FN observers in order to test the robustness of procedures to response errors by unreliable patients.

Format of Experiments

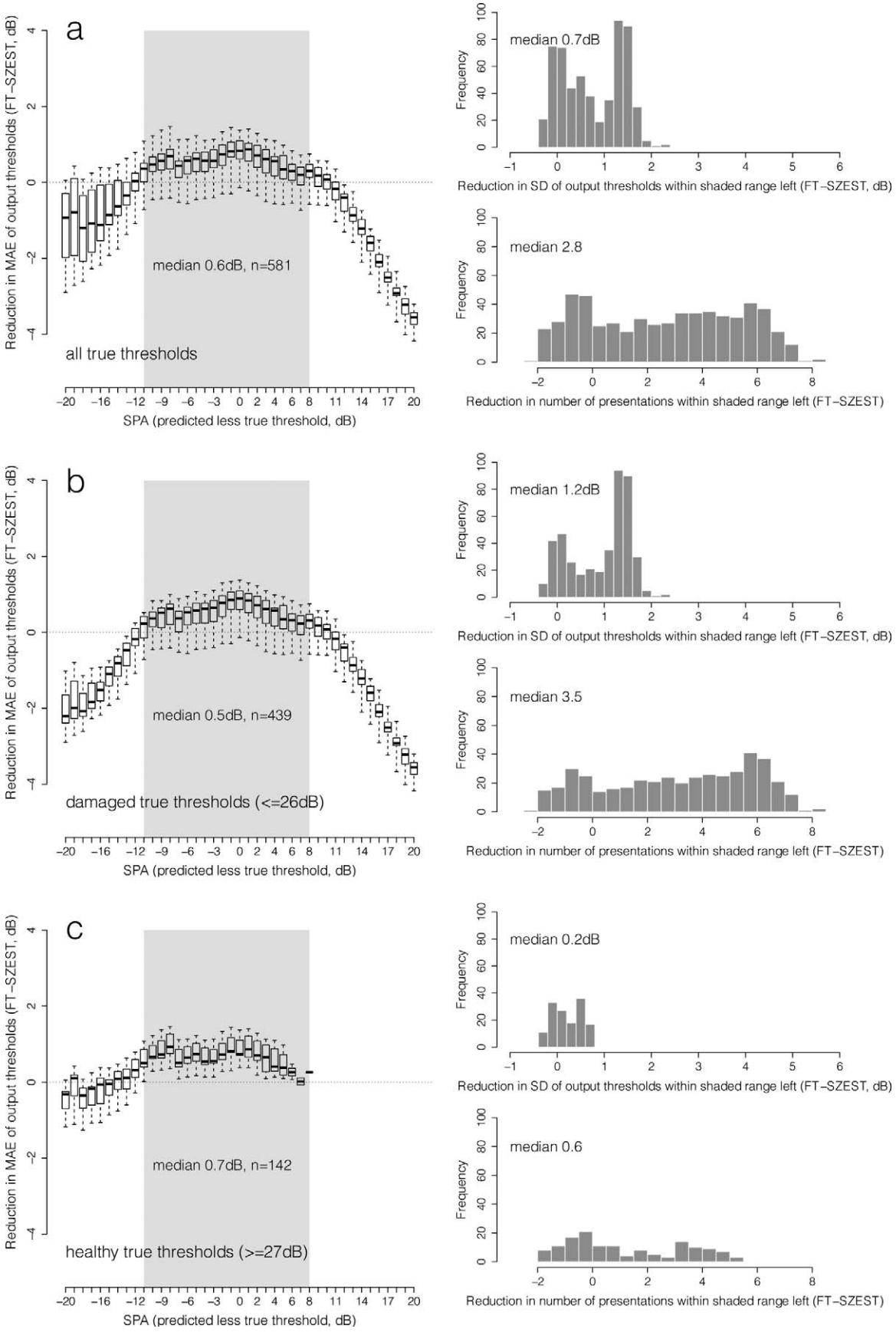
We compared two test procedures: Full Threshold (FT) and Structure-ZEST (SZEST), a procedure seeded by prior information from structural imaging. The procedures are described in full in the next section. Procedure behavior was first studied in detail for single locations to investigate how closely threshold predictions from hypothetical structural measures need to reflect true threshold for SZEST to outperform FT in terms of MAE (Experiment One). The parameters found in Experiment One were then used to compare the performance and presentation counts of the two procedures applied to a clinical database of 24-2 visual fields (Experiment Two).

Test Procedures

The SZEST procedure, described in full below, represents a ZEST procedure with a prior probability mass function (*pmf*) centered on an estimate of threshold. This is in contrast to the ZEST procedures in the perimetry literature, where the *pmf* is typically derived from population data.[3,16] Figure 1 shows a comparison of threshold estimates and number of presentations made by SZEST with four Gaussian prior *pmf*s to the exact same procedure with a population-derived *pmf* (a standard ZEST procedure, the *pmf* used is shown in Turpin et al.,[3] Fig. 2). As would be expected, Figure 1 shows that using a prior *pmf* weighted close to the true threshold of an observer improves procedure performance compared with a generic prior *pmf*. However, using a prior *pmf* weighted away from true threshold can increase error and number of presentations above the population-based ZEST procedure. In this study, therefore, we investigate how good the prior *pmf* of SZEST must be in order to provide gains in clinical applications, where we consider SITA Standard to represent current clinical standards.

As the exact details of SITA Standard are not available, we evaluate SZEST against FT as a surrogate for SITA Standard. FT is a staircase procedure that has largely been superseded in clinical use by the SITA procedures.[4] The full details of the SITA algorithms are not publically available, so we use FT as a surrogate as it has similar test–retest characteristics to SITA Standard.[2,3,17–20] SITA Standard is on average 1 to 1.5 presentations faster than FT at each location, thus, we make sure that SZEST is also faster than FT to allow our indirect comparison between SZEST and SITA to be fair.

FT presents stimuli in 4 dB increments until a response reversal occurs, and in 2 dB increments


a
Reduction in MAE of output thresholds (FT−SZEST, dB)
median 0.6dB, n=581
all true thresholds
SPA (predicted less true threshold, dB)
median 0.7dB
Frequency
Reduction in SD of output thresholds within shaded range left (FT−SZEST, dB)
median 2.8
Reduction in number of presentations within shaded range left (FT−SZEST)
b
median 0.5dB, n=439
damaged true thresholds (<=26dB)
median 1.2dB
median 3.5
c
median 0.7dB, n=142
healthy true thresholds (>=27dB)
median 0.2dB
median 0.6

←

Figure 2. Box and whisker plots show median, interquartile range, and 90% range (whiskers: fifth to 95th percentile) of distribution of reduction in MAE (SZEST-FT matched by true threshold, dB) using SZEST across all combinations of structure-predicted and true threshold for each SPA (predicted less true threshold) given on the x-axis (from −20 to +20 dB) (reliable observers). Data are shown for all true thresholds simulated (*row a*), damaged true thresholds only (≤26 dB, *row b*) and healthy true thresholds only (≥27 dB, *row c*). Shaded area on boxplots represents the range of SPA where 25th percentile of MAE reduction was greater than 0 dB across all possible true thresholds. Histograms in each row show the distribution of reductions (SZEST-FT, dB) in SD of output threshold and presentation count for prediction accuracies within the shaded area.

thereafter. The output threshold is the intensity of the "last seen" stimulus after two response reversals. For primary locations of the 24-2 pattern (±9°, ±9° tested first), FT begins at 25 dB, so this is what we used as the first presentation for simulations of single locations (Experiment One). If a patient does not respond to the brightest stimulus (0 dB) twice then 0 dB is returned as the output threshold. In our FT implementation, if the output threshold was more than 4 dB away from the starting point, then a second staircase was started, identical to the first except starting from the initial estimate. This second estimate was taken as the output threshold, and presentations made in both staircases were counted.

For full-field simulations (Experiment Two), FT used a growth pattern, whereby after sensitivity was estimated at the four primary locations, thresholds at neighboring locations were estimated with the procedure starting from a decibel value equal to the neighboring location's estimate plus a correction for eccentricity. The growth pattern used in this study was identical to that illustrated in Turpin et al.,[3] figure 1.

SZEST is a modified ZEST procedure in which the prior *pmf* is a Gaussian distribution centered on a sensitivity prediction made by a structural measure. We chose a Bayesian procedure because Bayesian-like procedures are well established in perimetry and easily incorporate different sources of prior information or different desired outcomes. For example, the weighting applied to prior information can be adjusted according to the predictive power of the information and the consequences of erroneous predictions. Staircase procedures such as FT can also be modified to incorporate prior information by altering the start point of the staircase, but this has little effect on procedure bias when the starting guess is within ±10 dB of the true value.[21] Finally, Bayesian procedures have been shown in previous simulation studies to estimate threshold with less bias than FT.[3]

ZEST procedure performance is influenced by its prior *pmf*, likelihood function and termination criteria.[5,22] We were primarily interested in the effect of altering the prior *pmf*, so used the same likelihood function as in previous simulations (shown in Turpin et al.,[3] figure 2) whose slope is derived from frequency-of-seeing curves of healthy observers to perimetric stimuli.[23] We chose to use dynamic termination criteria as these reduce test variability.[22] We trialled many combinations of termination criteria and prior *pmf* SDs (data not shown), and settled on a prior *pmf* with SD 5 dB, and a termination criterion of posterior *pmf* SD less than 1.5 dB. This represented the best trade-off between accuracy and test duration in initial simulations where prior information perfectly predicted sensitivity. The procedure was implemented as a standard ZEST procedure.[3,6,22] Once the procedure terminated, the expectation of the final posterior *pmf* was rounded to the nearest integer decibel value and reported as the final threshold estimate. The simulated perimeter had a range of stimulus intensities from 0 to 40 dB, but *pmfs* were calculated over a range extended by 10 dB either side of this so that thresholds close to the range limits of the perimeter were achievable. For full-field simulations using SZEST each location was tested independently; no growth pattern was used.

Experiment One

Methods for Simulation of Single Locations

This experiment established how accurately prior information needs to reflect true threshold for SZEST to outperform FT; thus, we define structure–prediction accuracy (SPA) as

$$\text{SPA} = \text{structure predicted threshold} - \text{true threshold}. \quad (3)$$

If SPA is 0 dB, structure perfectly predicts true threshold; if negative, then structure underestimates true threshold; and if positive the structural value is an over estimate.

To determine the range of SPA for which SZEST outperforms FT, 1000 patients for each integer's true threshold from 1 to 35 dB were simulated with FT and SZEST with all possible SPA values (predicted threshold ranging from 0 to 35 dB). Simulations were carried out under each of the three response error conditions (reliable, FP, FN).

This simulation provided us with the MAE for both FT and SZEST for each SPA and for all true thresholds. We defined the range of SPA across which SZEST outperforms FT for a particular response error condition to be the range of SPA where SZEST MAE was less than FT MAE for at least 75% of true thresholds. This gives bounds on how closely structural measurements need to predict visual field sensitivity for SZEST to outperform FT. Since the clinical consequences of MAE are difficult to interpret, we also report the differences in SD of output thresholds (variability) and number of presentations per location for each response error condition, within the range of prediction accuracies where SZEST outperforms FT. Although not reported for all simulations, examples of the magnitude and direction of bias in threshold estimates made by SZEST can be seen in the colored lines in Figure 1A, where the structure predictions are assumed to be 0, 10, 20, and 30 dB. As would be expected, bias increases as true threshold becomes more distant from the center of the prior *pmf*, and the direction of the bias is towards the center of the prior *pmf*. Hypothesis testing was performed with two-tailed Wilcoxon Rank-sum tests.

Results of Simulation of Single Locations

As our simulations of the FT procedure were identical to those reported previously, its error and presentation count distributions were similar to those previously reported.[6]

Figure 2 shows the distribution of reduction in MAE (FT-SZEST) across all possible combinations of structure-predicted and true threshold for each SPA given on the x-axis (reliable observers); Figures 3 and 4 show the same for FP and FN observers, respectively. For this analysis, the MAE for SZEST was subtracted from the MAE for FT with equivalent true threshold. Data are shown for all true thresholds simulated (row a), damaged true thresholds only ($\leq$26 dB, row b) and healthy true thresholds only ($\geq$27 dB, row c). Although somewhat arbitrary, this cutoff was chosen to show the impact of only using SZEST on locations already known to have visual field damage. The grey shaded areas on the boxplots represent the prediction accuracies across which MAE was less for SZEST than FT in at least 75% of instances across all possible true thresholds; that is, SZEST outperformed FT according to the criterion detailed in "Methods for Simulation of Single Locations" earlier. The histograms in each row show the distribution of reductions (SZEST-FT) in variability and presentation count for when SZEST outperformed FT (prediction accuracies within the shaded area). When structural predictions were sufficiently accurate for SZEST to outperform FT, output thresholds from SZEST were significantly less variable than those from FT ($P < 0.001$, two-tailed Wilcoxon Rank-sum test) and SZEST made significantly fewer presentations than FT ($P < 0.001$, two-tailed Wilcoxon Rank-sum test) across all response conditions. Reductions in variability and presentation count were generally greater for damaged true thresholds than healthy true thresholds.

Figures 2 to 4 also show the penalty of using SZEST when the structural measure is a poor predictor of visual field sensitivity. When SPA is worse than approximately $\pm$12 dB the MAE of SZEST begins to exceed that of FT on average, and MAE and presentation count increases further with reduced prediction accuracy (Figs. 2–4).

Given that the MAE of FT varies with true threshold, it is feasible that the range of SPA across which SZEST outperforms FT should also vary with true threshold. This is explored in Figure 5, where the bars show this range for true thresholds in 5 dB bins. Figure 5 shows that SZEST rarely outperforms FT when true threshold is 21 to 25 dB, the area where FT performs best.[3,6] When there were few response errors (black bars in Fig. 5) SZEST generally outperformed FT when SPA was within approximately $\pm$10 to 12 dB. The position of the bars for true thresholds less than or equal to 5 dB and greater than or equal to 31 dB reflects the reduced range of possible SPA in these situations due to floor/ceiling effects. The general skew of the red and blue bars in Figure 5 reflects the effect of response errors on the performance of both procedures.

Experiment Two

Methods for Simulation of Clinical Visual Field Tests

To compare SZEST to FT on clinical visual fields we used a dataset of 163 glaucomatous visual fields as input ("true") thresholds (FT algorithm, 24-2 pat-

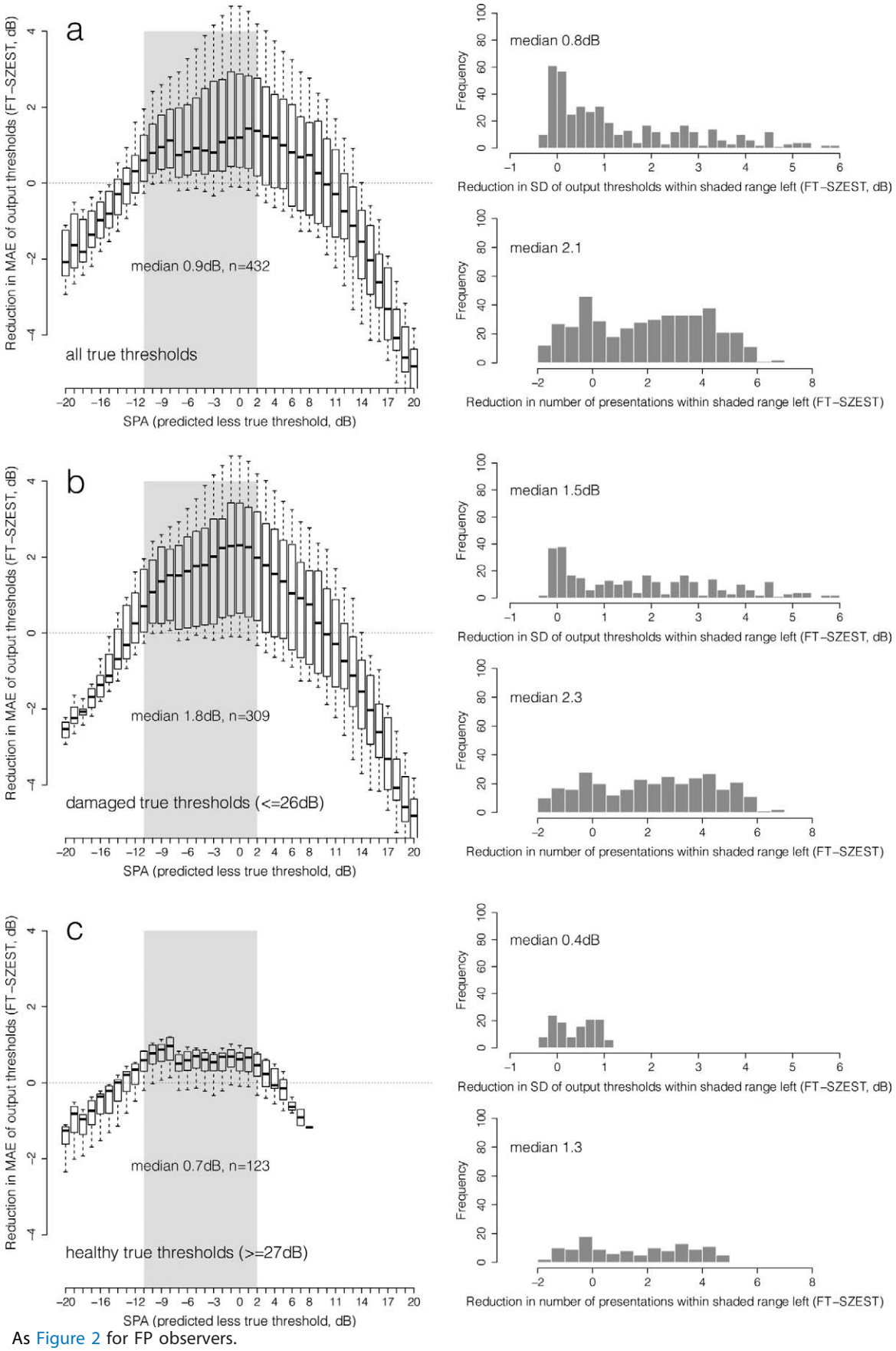


Figure 3. As Figure 2 for FP observers.

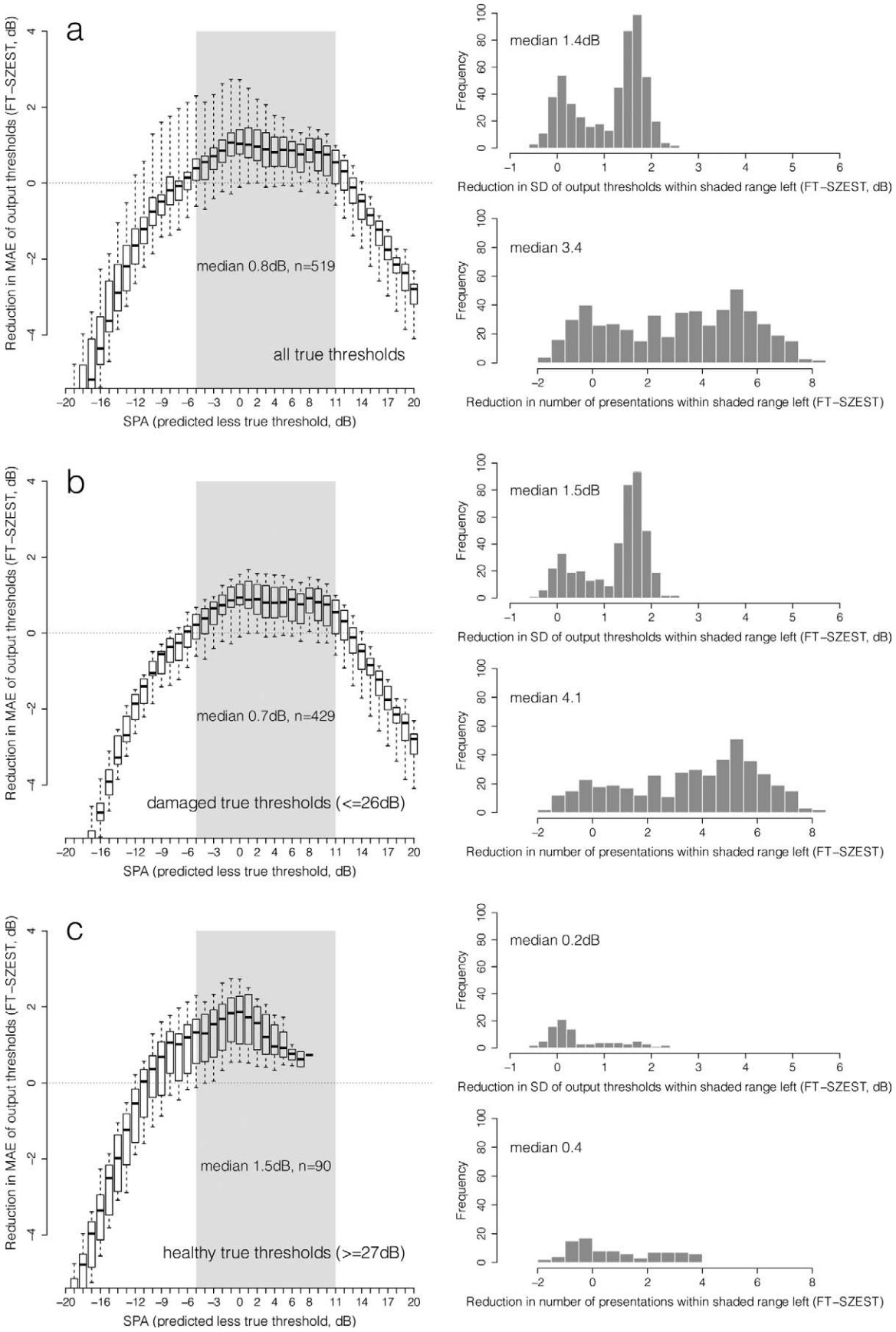


Figure 4. As Figure 2 for FN observers.

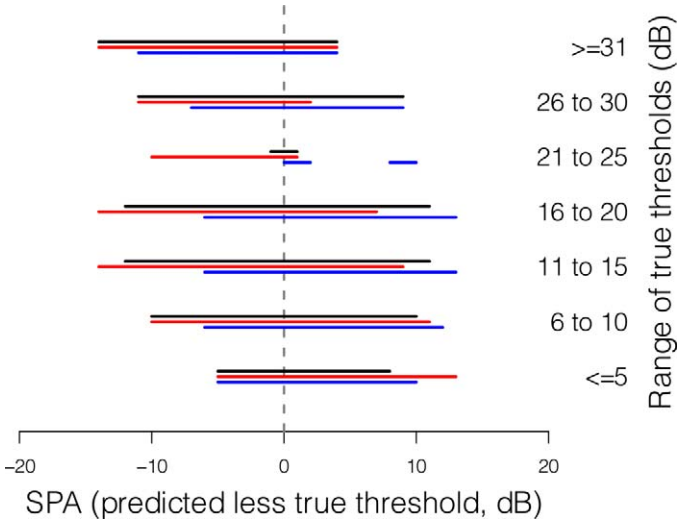


Figure 5. Bars show the range of SPA (predicted less true threshold) across which SZEST outperforms FT in 75% of instances for all possible true thresholds within the range given to the right of each bar. Data are shown for reliable observers (black bars), FP observers (red bars) and FN observers (blue bars).

tern). The fields were collected for a previous study where written informed consent in accordance with the tenets of the Declaration of Helsinki was obtained from all participants for the data to be stored in a de-identified database for further research. Patients mean age was 61 years, SD 13 years, and had a wide range of visual field defects (median mean deviation [MD] −1.8 dB, 90% range +2.1 to −22.6 dB) after age correction to 45 years (correction factor 1 dB per decade). Blind spot locations were excluded from analysis.

For SZEST, a set of structure-predicted thresholds was required to seed the prior *pmfs*. We constructed a hypothetical situation where structural measurements predict threshold with accuracy similar to the range across which SZEST outperformed FT in Experiment One. For simplicity, a single SPA range of ±9 dB was chosen as a reasonable approximation of the range across which SZEST performs well (Fig. 5). Gaussian noise (SD 3 dB) was added to the true thresholds to generate predicted thresholds that were within approximately ±9 dB of true threshold. Any predicted thresholds below 0 dB were set to 0 dB to constrain predictions to the measurement range of the simulated perimeter.

Each visual field in the dataset was then tested 1000 times with each simulated procedure, for each response error condition. Pointwise average MAE, SD of output thresholds, and number of presentations were compared for each procedure using two-tailed Welch's *t*-tests. The 90% retest limits were compared for both procedures at each input ("true") threshold and for each response error condition using two-tailed Wilcoxon Rank-Sum tests.

Results of Simulation of Clinical Visual Field Tests

Pointwise average MAE, output threshold variability (SD) and numbers of presentations for SZEST and FT under each response error condition are shown in Table 1 for the 163 glaucomatous visual fields simulated. For all response error conditions both MAE and variability were less for SZEST than FT ($P < 0.001$, two-tailed Welch's *t*-tests). On average, SZEST made approximately one fewer presentation per location than FT ($P < 0.001$, two-tailed Welch's t-test).

Figure 6 shows the retest distribution of SZEST according to true threshold and response error condition. The 90% retest limits of FT on the same data are shown for comparison (solid lines), and dashed lines show the range where output threshold is within ± 9 dB of input threshold as a reference for the range within which structural predictions of threshold fell. Table 2 gives the median and interquartile range for 90% retest limit width across the dynamic range for both procedures across the three response conditions. Narrower interquartile ranges indicate less variable 90% retest limits across the dynamic

Table 1. Average Pointwise Properties of Output Thresholds and Presentation Counts for SZEST and FT for the 163 Clinical Glaucomatous Visual Fields for All Response Error Conditions

	Reliable Observers		Typical FP Observers		Typical FN Observers	
	SZEST	FT	SZEST	FT	SZEST	FT
Mean absolute error (dB)	1.4	2.0	1.8	2.3	1.7	2.9
SD output thresholds (dB)	1.8	2.1	2.2	2.8	2.1	2.9
Number of presentations	5.2	6.3	5.3	6.4	5.2	6.4

Values for all measures across all response error conditions were significantly lower for SZEST than FT ($P < 0.001$, two-tailed Welch's *t*-test).

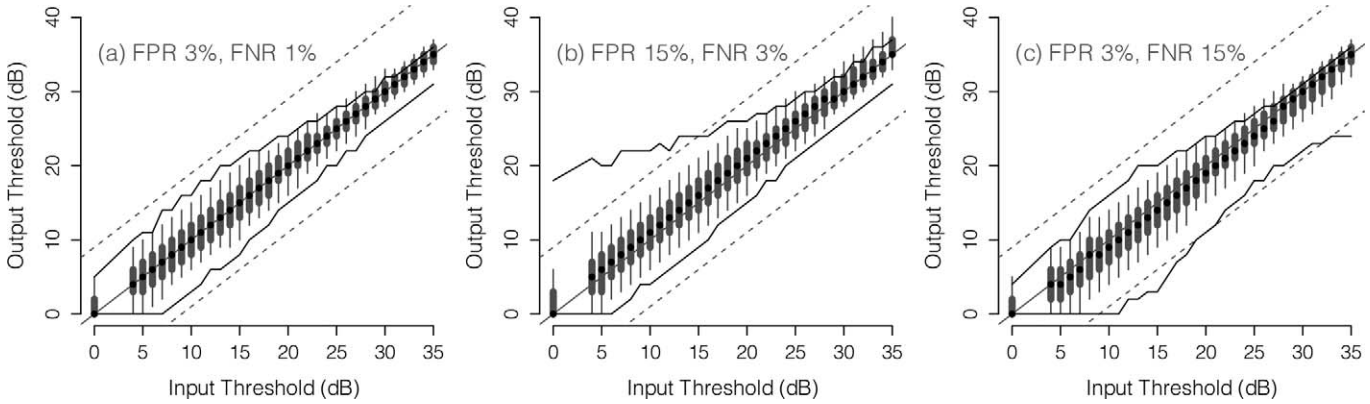


Figure 6. Box and whisker plots show median, interquartile range, and 90% range (whiskers: fifth to 95th percentile) of output thresholds for each input threshold in the clinical dataset. Data are shown only for input thresholds that occurred at least 20 times in the dataset. The solid lines show the 90% range (fifth and 95th percentiles) of output thresholds from the FT procedure in the same data. Dashed lines show the range within which output threshold is within ±9 dB of input threshold as a reference for the range within which structural predictions of threshold fell. Data are shown for reliable observers (a), FP observers (b) and FN observers (c).

range. SZEST had narrower median 90% retest limits than FT across all response error conditions ($P < 0.001$, two-tailed Wilcoxon Rank-sum tests), particularly at reduced true thresholds. The retest limits of SZEST were also more consistent in width across both the dynamic measurement range of the simulated perimeter and the different response error conditions than those for FT (Table 2). Given the decreasing slope of the psychometric function used to model patient responses at reduced thresholds, these results suggest that, under the conditions simulated, output thresholds estimated using SZEST may be less affected by the flattening of frequency-of-seeing curves with decreasing sensitivity[15] than those estimated by FT.

Discussion

SZEST represents a simple approach to using patient-specific prior information in perimetric procedures. SZEST yields improvements over FT in measurement error, variability, and number of presentations when prior data predict visual field sensitivity within approximately ±9 dB, depending on response error rates and true threshold (Fig. 5). This paper does not address the myriad of possible ways in which prior information may be incorporated into possible perimetric procedures. Rather, we have concentrated on a basic approach that has been shown to work well for perimetry in previous studies and that is similar to those used in commercially-available perimeters. This "first-pass" approach demonstrates some potential benefits of seeding perimetric procedures with patient-specific prior data, which are likely to extend to various sources of prior information.

In Experiment One, we compared SZEST with a 4-2 staircase procedure beginning at 25 dB at single locations. SZEST yielded improvements that were greatest at damaged locations (Figs. 2–4). The aim of Experiment One was to establish parameters for Experiment Two, where full clinical visual fields were considered. In Experiment Two, FT used a growth

Table 2. Median and Interquartile Range for 90% Retest Limit Widths of SZEST and FT Across the Range of Input ("True") Thresholds Simulated in Experiment Two (All Response Conditions)

	Reliable Observers		Typical FP Observers		Typical FN Observers	
	SZEST	FT	SZEST	FT	SZEST	FT
Median width of 90% retest limits (dB)	8	9	9	12	9	11
Interquartile range of 90% retest limits (width) (dB)	6–9 (3)	6–12 (6)	7–10 (3)	7–18 (11)	7–10 (3)	10–15 (5)

Across all response conditions median 90% retest limits were narrower for SZEST than FT ($P < 0.001$, two-tailed Wilcoxon Rank-sum tests).

pattern such that 48 of the 52 staircases were seeded with prior information from neighboring locations. SZEST still made improvements over FT. While these improvements over FT were statistically significant, it is important to consider whether gains are clinically meaningful. One criterion is whether improvements would allow progression to be detected at an earlier clinic visit. Previous simulations have shown that this likely requires a reduction in variability of 20% to 40% at damaged locations.[24] Experiment Two demonstrated this magnitude of reduction only for observers making high rates of response errors. Smaller reductions in test–retest variability may still be beneficial, though, particularly in clinical trial situations where different approaches may be taken to progression detection, such as the "wait and see" approach recently suggested by Crabb and Garway-Heath.[25]

As a baseline for comparison, we have used the FT procedure to represent current clinical procedures as the error characteristics of FT are the same as for SITA Standard. It is well established, though, that SITA Standard makes 1 to 1.5 fewer presentations per location than FT.[18,19] Further reductions in test times are achieved by dynamically altering response window duration during the test, and estimating false responses without catch trials,[4] but these are not relevant for our simulations. SZEST made on average about one presentation fewer per location than FT, making it similar to SITA Standard. However, since SITA Standard has similar test–retest characteristics to FT,[2,3,17–20] SZEST is likely to have less bias and variability than SITA Standard. Procedures with patient-specific priors based on structural data (SZEST) or previous visual field data (retest minimizing uncertainty [REMU])[6] demonstrate more consistent retest variability across the dynamic range, and across different variability conditions (Fig. 6, Table 2), potentially leading to an improved ability to detect progression at reduced sensitivities or in the presence of high rates of response errors.

This study has also demonstrated a major potential pitfall of incorporating patient-specific prior information into perimetric procedures. When prior information was a poor predictor of true threshold ($|SPA| \gg 9$dB in our simulations) SZEST was slow and showed large amounts of bias towards the prior prediction resulting in inaccurate threshold estimates. Consequently, whether there is any benefit in using *structural* prior information hinges on the availability of structural measurements that are sufficiently closely related to visual field threshold. The Hood–Kardon model[26] relating structural and functional measurements assumes an underlying 1:1 linear relationship between retinal ganglion cell axon loss and visual field sensitivity loss in glaucoma. Hence, a 3 dB loss of visual field sensitivity (a doubling in stimulus luminance at threshold) equates to a 50% loss of axons according to their model. Similarly, a 10 dB loss of sensitivity equates to a 90% loss of axons and, therefore, more than half of the dynamic range of commercially available perimeters is determined by less than 10% of retinal ganglion cells. Given the resolution and measurement variability of current imaging devices, the variability in clinical visual field threshold measurements and the population variance in both parameters, it is not surprising that in clinical data RNFL neuronal component thickness below 10% of mean normal (>90% loss of retinal ganglion cell axons) can only predict that visual field sensitivity will be reduced by around 10 dB or more compared with mean normal.[27] Predictions from this model would, on average, only allow SZEST to perform well in early to moderate glaucoma, though it is difficult to draw firm conclusions from these data as we require knowledge of how structural measures relate to *true* threshold, not clinically measured threshold. A recent study has highlighted that the between–subject variation seen as scatter in current structure-function relationships is still compatible with a close underlying structure-function relationship when measurement variability in both structure and function are taken into account, suggesting the relationship between structural measurements and underlying true threshold may be strong.[28]

Histological studies of retinal ganglion cell density and visual field sensitivity have indicated that underlying structure may relate to function in a way that could be exploited for perimetry. Harwerth and Quigley[29] made predictions of ganglion cell density from visual field sensitivity in glaucoma patients. Post mortem, they made histological counts of ganglion cell density and found that their predictions were accurate to within ±8 dB, with the residuals approximately normally distributed within this range. There was a lack of deep visual field defects in their sample, so it is unknown whether their predictions maintain the same level of accuracy at lower sensitivities. Another approach to predicting visual field sensitivity from structural measures is the Bayesian radial basis model described by Zhu et al.[30] This model predicts threshold sufficiently accurately for SZEST to perform well until visual field sensitivity falls below approximately 10 dB. Since this model was developed and tested on scanning laser

polarimetry data, it might be possible to obtain improved predictions from newer, higher-resolution devices.

One potential barrier to the use of SZEST is the spatial resolution with which we can relate structural measures (e.g., peripapillary RNFL thickness or neuroretinal rim area) to locations of the visual field. Recent studies have increased the resolution of structure-function maps averaged across study populations,[31,32] but these do not adjust for individual variability. We have recently developed an anatomically-customizable model for mapping visual field locations to the optic nerve head, which we hope will aid development of more individualized perimetric tests for glaucoma.[33]

Bayesian procedures such as SZEST are readily adaptable to different prior information. SZEST uses the structural prediction to seed the Gaussian prior *pmf* and then continues from that point as in a standard ZEST algorithm. Other Bayesian procedures could be used in place of ZEST, or different parameters used within SZEST. For example, if structural sensitivity predictions were found to be nonnormally distributed about the true sensitivity then a prior *pmf* more closely related to that distribution could be used. The SD of the *pmf* may also be altered to reflect the predictive power of a given source of prior information. Our study provides a framework for the evaluation of such modifications in the future.

Seeding any visual field test with data from a structural measure somewhat increases the dependence of the functional measurement on the structural measurement; hence, current analyses that treat these two variables as independent sources of evidence would need to be altered. Further study is required to evaluate whether improving functional measures through an approach like SZEST provides overall benefit in the diagnosis and monitoring of glaucoma over the post hoc combination of existing functional and structural measures.

In conclusion, seeding a Bayesian perimetric procedure (SZEST) with patient-specific prior information yields reduced test–retest variability and number of presentations compared to FT, if the prior information predicts sensitivity within approximately ±9 dB. Improvements were greatest for damaged visual field locations and in the presence of higher rates of response errors. Further prospective study is required to assess and refine the performance of SZEST-like procedures when seeded specifically by measurements from current imaging devices.

Acknowledgments

The authors thank Chris A. Johnson (University of Iowa, IA) for providing the clinical dataset used in the simulations, and the reviewers for helpful comments.

Supported by the Australian Research Council Linkage Project LP100100250 (with Heidelberg Engineering GmbH, Germany) and the Australian Research Council Future Fellowships FT0990930 (to AMM) and FT0991326 (to AT).

Disclosure: J. Denniss, Heidelberg Engineering GmbH (F); **A.M. McKendrick,** Heidelberg Engineering GmbH (F); **A. Turpin,** Heidelberg Engineering GmbH (F)

References

1. McKendrick AM. Recent developments in perimetry: test stimuli and procedures. *Clin Exp Optom*. 2005;88:73–80.
2. Artes PH, Iwase A, Ohno Y, Kitazawa Y, Chauhan BC. Properties of perimetric threshold estimated from Full Threshold, SITA Standard, and SITA Fast strategies. *Invest Ophthalmol Vis Sci*. 2002;43:2654–2659.
3. Turpin A, McKendrick AM, Johnson CA, Vingrys AJ. Properties of perimetric threshold estimates from full threshold, ZEST, and SITA-like strategies, as determined by computer simulation. *Invest Ophthalmol Vis Sci*. 2003;43:4787–4795.
4. Bengtsson B, Olsson J, Heijl A, Rootzen H. A new generation of algorithms for computerized threshold perimetry, SITA. *Acta Ophthalmol Scand*. 1997;75:368–375.
5. King-Smith PE, Grigsby SS, Vingrys AJ, Benes SC, Supowit A. Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. *Vision Res*. 1994;34:885–912.
6. 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.
7. Schiefer U, Pascual JP, Edmunds B, et al. Comparison of the new perimetric GATE strategy with conventional full-threshold and SITA standard strategies. *Invest Ophthalmol Vis Sci*. 2009;50:488–494.
8. Bowd C, Zangwill LM, Madeiros FA, et al. Structure-function relationships using confocal scanning laser ophthalmoscopy, optical coher-

ence tomography and scanning laser polarimetry. *Invest Ophthalmol Vis Sci*. 2006;47:2889–2895.
9. Hood DC, Anderson SC, Wall M, Kardon RH. Structure versus function in glaucoma: an application of a linear model. *Invest Ophthalmol Vis Sci*. 2007;48:3662–3668.
10. Denniss J, Schiessl I, Nourrit V, Fenerty CH, Gautam R, Henson DB. Relationships between visual field sensitivity and spectral absorption properties of the neuroretinal rim in glaucoma by multispectral imaging. *Invest Ophthalmol Vis Sci*. 2011;52:8732–8738.
11. Malik R, Swanson WH, Garway-Heath DF. 'Structure-function relationship' in glaucoma: past thinking and current concepts. *Clin Experiment Ophthalmol*. 2012;40:369–380.
12. Kelareva E, Mewing J, Turpin A, Wirth A. Adaptive psychophysical procedures, loss functions, and entropy. *Atten Percept Psychophys*. 2010;72:2003–2012.
13. Treutwein B. Adaptive psychophysical procedures. *Vision Res*. 1995;35:2503–2522.
14. 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.
15. 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.
16. Vingrys AJ, Pianta MJ. A new look at threshold estimation algorithms for automated static perimetry. *Optom Vis Sci*. 1999;76:588–595.
17. Wild JM, Pacey IE, Hancock SA, Cunliffe IA. Between-algorithm, between-individual, differences in normal perimetric sensitivity: full threshold, FASTPAC, and SITA. *Invest Ophthalmol Vis Sci*. 1999;40:1152–1161.
18. Bengtsson B, Heijl A, Olsson J. Evaluation of a new threshold visual field strategy, SITA, in normal subjects. *Acta Ophthalmol Scand*. 1998;76:165–169.
19. Bengtsson B, Heijl A. Evaluation of a new perimetric strategy, SITA, in patients with manifest and suspect glaucoma. *Acta Ophthalmol Scand*. 1998;76:368–375.
20. Shirato S, Inoue R, Fukushima K, Suzuki Y. Clinical evaluation of SITA: a new family of perimetric testing strategies. *Graefes Archive Clin Exp Ophthalmol*. 1999;237:29–34.
21. Johnson CA, Chauhan BC, Shapiro LR. Properties of staircase procedures for estimating thresholds in automated perimetry. *Invest Ophthalmol Vis Sci*. 1992;33:2966–2974.
22. McKendrick AM, Turpin A. Advantages of terminating Zippy Estimation by Sequential Testing (ZEST) with dynamic criteria for white-on-white perimetry. *Optom Vis Sci*. 2005;82:981–987.
23. Spry PGD, Johnson CA, McKendrick AM, Turpin A. Variability components of standard automated perimetry and frequency-doubling perimetry. *Invest Ophthalmol Vis Sci*. 2001;42:1404–1410.
24. Turpin A, McKendrick AM. What reduction in standard automated perimetry variability would improve the detection of visual field progression? *Invest Ophthalmol Vis Sci*. 2011;52:3237–3245.
25. Crabb DP, Garway-Heath DF. Intervals between visual field tests when monitoring the glaucomatous patient: wait-and-see approach. *Invest Ophthalmol Vis Sci*. 2012;53:2770–2776.
26. Hood DC, Kardon RH. A framework for comparing structural and functional measures of glaucomatous damage. *Prog Retin Eye Res*. 2007;26:688–710.
27. Hood DC, Anderson SC, Wall M, Raza AS, Kardon RH. A test of a linear model of glaucomatous structure-function loss reveals sources of variability in retinal nerve fiber and visual field measurements. *Invest Ophthalmol Vis Sci*. 2009;50:4254–4266.
28. Gardiner SK, Johnson CA, Demirel S. The effect of test variability on the structure-function relationship in early glaucoma. *Graefes Arch Clin Exp Ophthalmol*. 2012;250:1851–1861.
29. Harwerth RS, Quigley HA. Visual field defects and retinal ganglion cell losses in patients with glaucoma. *Arch Ophthalmol*. 2006;124:853–859.
30. Zhu H, Crabb DP, Schlottmann PG, et al. Predicting visual function from the measurements of retinal nerve fiber layer structure. *Invest Ophthalmol Vis Sci*. 2010;51:5657–5666.
31. Jansonius NM, Nevalainen J, Selig B, et al. A mathematical description of nerve fiber bundle trajectories and their variability in the human retina. *Vision Res*. 2009;49:2157–2163.
32. Turpin A, Sampson GP, McKendrick AM. Combining ganglion cell topology and data of patients with glaucoma to determine a structure-function map. *Invest Ophthalmol Vis Sci*. 2009;50:3249–3256.
33. Denniss J, McKendrick AM, Turpin A. An anatomically customizable computational model relating the visual field to the optic nerve head in individual eyes. *Invest Ophthalmol Vis Sci*. 2012;53:6981–6990.