Development and validation of a new glaucoma screening test using temporally modulated flicker

*Bruno R Fidalgo, Anish Jindal, Christopher W Tyler, Irene Ctori, John G Lawrenson**

Centre for Applied Vision Research, School of Health Sciences, City University of London, UK

***Corresponding author:** j.g.lawrenson@city.ac.uk

Keywords: Glaucoma; Screening; Perimetry; Flicker

Running head: Evaluation of a new flicker test for glaucoma

Acknowledgements: Supported by a grant from the College of Optometrists UK, who provided a PhD studentship to Anish Jindal.

Disclosure: The authors report no conflicts of interest and have no proprietary interest in any of the materials mentioned in this article.

Abstract

Purpose. Describing the psychometric characteristics and diagnostic accuracy of the Accelerator 4-Alternative Forced-Choice Flicker Test prototype (A4FTp) for detecting chronic open angle glaucoma (COAG)

Methods. A4FTp measures temporally modulated flicker thresholds in regions of the visual field with high susceptibility to glaucomatous loss. We initially evaluated its psychometric properties on 20 normals (aged 33.8 ± 8.5 years) who were tested multiple times over a period of 3m. All subjects underwent 4 repetitions for shorter (T8) and longer (T12) staircase termination criteria, to determine the most suitable threshold criterion. Four randomly selected subjects underwent a total of 10 repetitions to study test-retest repeatability and learning effects. To determine its diagnostic accuracy, one eye of 40 participants with COAG and 38 normal controls were tested with the A4FTp in comparison with the Frequency Doubling Technology (FDT) (C20-5 programme) and iVue Spectral Domain Optical Coherence Tomography (SD-OCT). Tests were conducted in a random order with results masked to the clinician conducting the reference ophthalmic examination. The accuracy of each test was determined by analysis of the area under the receiver operator characteristics curve (AROC).

Results. A4FTp flicker thresholds were stable, with standard deviations of only 0.52 decilog (dL) for T8, increasing to 1.32 dL for T12, and no significant flicker sensitivity threshold improvement over the 10 repeat runs. T8 was superior to T12 on several other measures, so it was used for the remaining comparisons. In terms of diagnostic accuracy, the mean AUROC for the three tests were; A4FTp (T8 criterion) (0.82, 95% confidence interval (0.73-0.92)), SD-OCT (any RNFL parameter p<1% level) (0.90 (0.83-0.97)) and FDT (one or more locations missed at p<5% level) (0.91 (0.82-0.96)). There was no statistical difference in AUROC between A4FTp and SD-OCT (p=0.18) or FDT (p=0.12). The A4FTp test duration averaged just over 2 minutes per eye, taking approximately one third of the time for completion of the HFA SITA 24-2 algorithm (conducted as part of the reference examination) and twice the time for the suprathreshold FDT.

Conclusion. Test accuracy for the A4FTp was comparable to those of the FDT and SD-OCT for the detection of COAG. Time taken to complete the A4FTp was relatively short and initial results are promising. With further refinement, the A4FTp could have a future role in glaucoma detection.

Introduction

Visual field testing remains one of the most important tools for identifying and monitoring vision loss in glaucoma.[1] Since the introduction of automated perimetry in the early 1970s, almost all developments in perimetry have focused on either improving the sensitivity of the tests to early visual field damage or increasing test efficiency and speed.[2, 3] However, the lack of portability of current perimeters means that the assessment of visual function remains difficult outside the normal clinical setting. The advent of touch screen and tablet-technology has created new opportunities for the development of a portable low cost glaucoma-screening test that could be used in a home setting or in remote underserved communities. An iPad-based threshold perimeter based on the Humphrey 24-2 test grid has recently been developed which shows good diagnostic performance for the detection of glaucoma.[4, 5]

The aim of this proof of concept study was to develop a simple, affordable and rapid glaucoma-screening test that could potentially be used to detect functional vision loss outside the normal clinical setting. We also compared the performance of the new test with commercially available glaucoma screening technologies that detect structural or functional glaucomatous damage. The new test incorporates a new psychophysical algorithm that measures flicker sensitivity thresholds in areas of the visual field that are susceptible to glaucomatous visual field loss. Temporal modulation flicker uses a stimulus that is matched in luminance to the background, the contrast of the stimulus is then modulated temporally at a fixed spatial frequency, and the amplitude of the flicker modulation needed to detect the stimulus is determined.[6] Sensitivity to temporal modulation flicker has been shown to be effective in separating normal subjects from glaucoma patients, suggesting that the stimulus is the method of choice for detecting glaucomatous damage using flicker perimetry.[7-10] The other advantage of flicker perimeter methods is that they are relatively unaffected by optical blur due to media opacification or uncorrected refractive error, which is particularly important for assessing visual function in the field.[11]

Another feature of the current test design is the use of extended stimulation areas at a small number of test locations. Previous studies have shown that the large numbers of stimulus locations used in current threshold perimeters are not always necessary to achieve high levels of sensitivity; good diagnostic performance can be achieved using relatively few test locations, confined to areas that are particularly prone to glaucomatous damage.[12-14] Moreover, one of the most common indices of glaucomatous loss is the Mean Deviation of the Humphrey Field Analyzer, which focuses on the average loss across large areas of the retina.[15] The philosophy of the stimulus design in the current test is therefore to use large stimuli that integrate information across extended retinal regions, with the further advantage of a potential reduction in test times.

We therefore determined that an appropriate paradigm for this purpose would be a four-alternative forced-choice temporal protocol (4AFTp), which is becoming increasingly popular in psychophysical investigations. It has two great advantages over 2AFC. One is that the false alarm rate is only 25%, giving greater discriminative power from each choice. The other is that all four selected locations are tested in the same state of the participant, so any practice or fatigue effects are equated across the four locations, giving a more accurate comparison of their relative sensitivities.

Materials and methods

This study was approved by the School of Health Sciences Research Ethics Committee, City, University of London and conducted per the tenets of Declaration of Helsinki. Written informed consent was obtained from all participants.

1. Development of the Accelerator 4-Alternative Forced-Choice Flicker Test prototype (A4FTp)

Stimulus Design

There is evidence to suggest that high sensitivity in glaucoma detection can be achieved with relatively few perimetric stimuli.[16-19] Whilst the informational value of stimulus locations in susceptible areas of the visual field may be high, the correlation between adjacent test locations limits the value of adding multiple stimuli in these areas The choice of stimulus location for this first iteration of the A4FTp was based on the study by Wang and Henson[14] who used optimised sub-sets of the standard 24-2 test pattern, based on the PPV of each test location that broadly corresponded to the typical patterns of glaucomatous visual field loss that reflect the course of retinal nerve fibers.[20, 21] Two large circular stimuli (11° diameter) were located in the temporal superior and inferior arcuate regions 9-21° from fixation, with the other two stimuli (11.7° diameter) spanning the horizontal meridian (14-26° from fixation) at the endpoints of the nasal arcuate sweep of the nerve fibre layer, the visual field location subject to the well-known 'nasal step' (Figure 1).

Stimuli were displayed at a viewing distance of 33 cm on a high refresh rate screen (120 Hz) in a uniform red field (610 nm), designed to minimise transmission losses in the optic media.[22]The target stimulus in each of four test fields was a burst of 30 Hz effectively sinusoidal flicker § with the same mean luminance as the background (19 cd/m^2), which was ramped on and off according to a 1 s raised cosine envelope in order to avoid onset and offset transients.

§ Since the harmonics of the 120 Hz sampled waveform at 60 Hz and above would be close to or beyond human temporal resolution, the waveform can be considered effectively sinusoidal

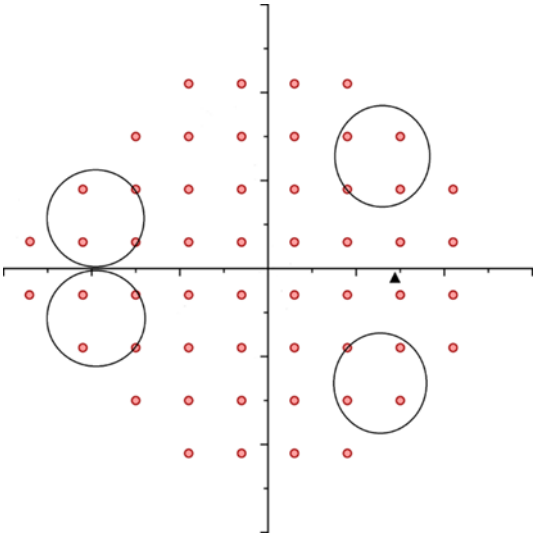


Figure 1. Spatial location of the A4FTp flicker regions for the right eye compared to the HFA 24-2 pattern. The black triangle corresponds to the location of the blind spot. The configuration was mirror-reversed for testing the left eye.

Algorithm development

The screening algorithm uses a rapid four-alternative forced-choice staircase paradigm, with the results plotted in decilog (dL) units (or 1/10th of log10 base intervals, following the simplifying convention relative to dB units adopted by one of the authors.[6] Each one-up/two-down staircase in log modulation steps of 1 dL terminates when the standard deviation of the last *n* trials becomes less than 1 step (where *n* was either 8 steps (T8) or 12 steps (T12) in different conditions). Unlike conventional staircases, this straightforward algorithm ensures that the staircase has reached a stable asymptotic performance level of low variability.

The algorithm was implemented and tested on a Dell computer with specifications Inter core 2 Duo CPU E7500 @2.93 GHz and 4 GB RAM, displayed on a 144 Hz Asus VG248QE 24-inch 3D LED monitor running at 120 Hz, and used an Accmat™ USB wired 19 key number numeric keypad as the test input device. The computer was running on a 64 - Bit Windows 7 Enterprise, Service pack 1, Matlab R2014a and Psychtoolbox version 3.0.11.

Selection of optimal threshold criteria

An initial study was conducted to determine the optimal threshold criteria for the 4AFTp staircase. A randomly selected eye in a convenience sample of 20 healthy adults was tested with the flicker test four times on separate occasions during a period of two weeks for each criterion run length (T8 and T12). The starting level was set at 15 dL, which is half way between

the threshold of about 10 dL for healthy individuals and the maximum available modulation level of 20 dL. The overall mean modulation level for the 4 test locations was determined for each session and the time taken to perform the test was recorded. Inter-session differences were evaluated using Friedman's test and test–retest coefficients of repeatability were calculated for the flicker modulation levels.

To assess the inherent variability of the test for a typical participant, the inter-participant variability was isolated from the intra-participant variability by calculating the mean modulation level across participants for each test location and each threshold criterion and then normalizing the individual values to this group mean to determine the within-participant variability.

Assessment of individual reliability and learning effects

To assess learning effects over repeats, we evaluated the change in performance of the flicker algorithm for 4 volunteers who undertook the test 10 times over a period of 3 months. Learning effects were assessed by comparing the test duration and modulation levels of the first session with those of the other four sessions.

2. Diagnostic performance of the A4FTp flicker test in the detection of glaucoma

The diagnostic accuracy of the A4FTp flicker test was determined using a case-control design in which test results in series of patients with an established diagnosis of COAG were compared with those of a control group. Test performance was also compared with two commonly used tests for glaucoma case finding: The Frequency Doubling Technology Perimeter (FDT); and the iVue Spectral Domain Optical Coherence Tomography (SD-OCT). User acceptability of all screening tests was also determined. The study was designed and findings reported in accordance with the Standards for Reporting of Diagnostic Accuracy criteria.[23]

Study Participants

Participants eligible for inclusion were consecutive adults (≥40 years) with a clinical diagnosis of COAG and were recruited either from the university eye clinic or via a request for volunteers in the 'International Glaucoma Association' newsletter. The control participants were consecutive non-glaucomatous adults (≥40 years), who were recruited from the university eye clinic and local optometry practices.

Inclusion and Exclusion Criteria

There is no universally accepted reference standard for the diagnosis of glaucoma; however, optic disc and visual field damage are used to diagnose the presence of glaucoma.[24] The diagnosis of COAG was based on a reference standard ophthalmic examination and meeting the following diagnostic criteria: open anterior chamber angles; presence of glaucomatous optic neuropathy (indexed by localized absence of neuroretinal rim, cup-to-disc ratio 0.7, or interocular asymmetry in vertical cup-to-disc ratio 0.2 in similar sized discs and the presence of a concordant glaucomatous field defect using the 24-2 Swedish Interactive Thresholding Algorithm (SITA) on the Humphrey Visual Field Analyzer (HFA), based on criteria amended from Anderson.[25]

We excluded any participant who had a history of angle closure, significant ocular co-morbidity e.g. diabetic retinopathy, retinal vascular occlusions, peripheral retinal abnormalities, optic atrophy, visually significant cataract (indexed by the Lens Opacity Classification System III[26] (LOCIII) (N≥4.0, C≥2.0, P≥2.0) or a neurological field defect.

The inclusion criteria for the control participants included: freedom from ocular disease with normal appearance of the optic disc, normal fundus, intraocular pressure ≤21 mm Hg and full visual fields on the reference examination.

Test procedures for each index test

Figure 2 shows the flow of patients through the study. All participants underwent testing on both eyes with all three 'index tests', A4FTp flicker test, FDT and iVue OCT, performed in a random order by an experienced optometrist who was unaware of the participants' ocular status. This sequence was then followed by a reference standard ophthalmic examination by a clinician masked to the index test results. Participants were included in the analysis if they had interpretable results in both reference and index tests; those with uninterpretable results were excluded. Thresholds of abnormality for the index tests were based on cut-offs commonly reported in the literature and were pre-defined before data analysis.

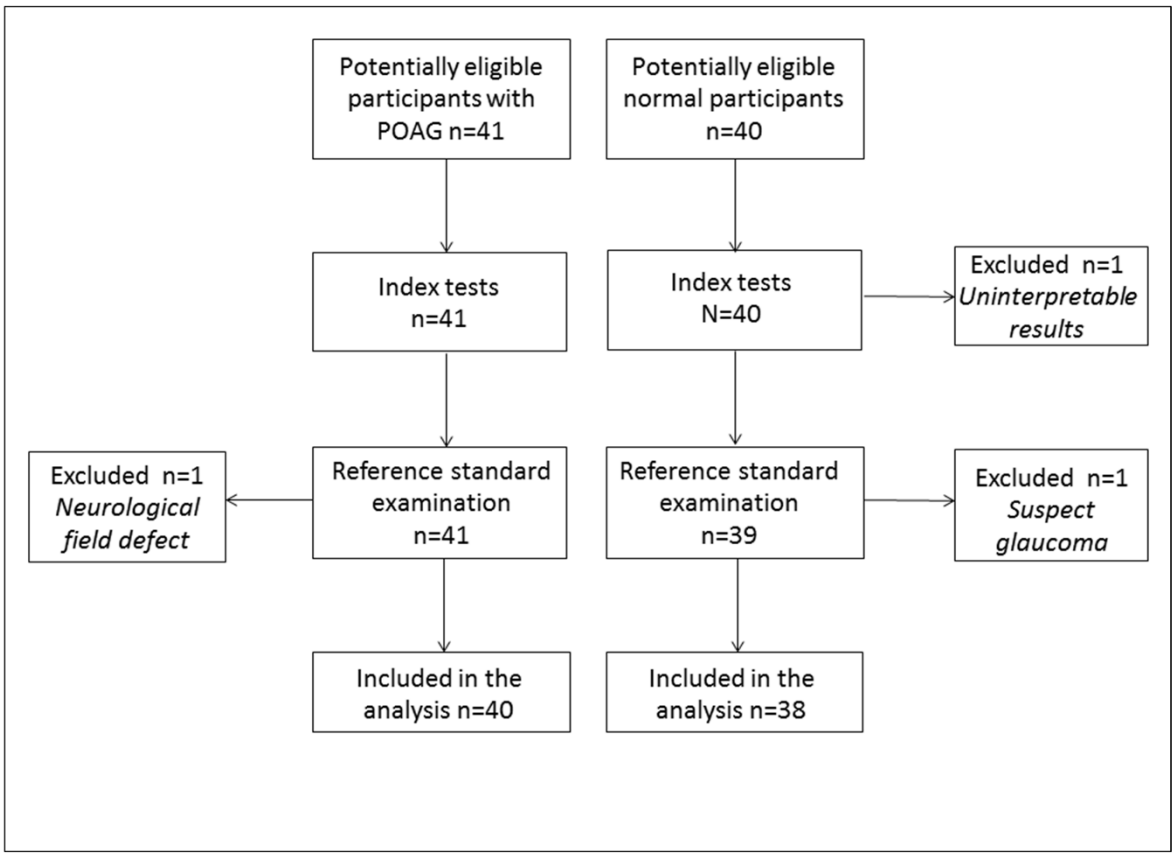


Figure 2. Study flow diagram

FDT

The first generation FDT (Carl Zeiss Meditec Inc,, software version 4.00.0, www.zeiss.com/meditec/int/products/ophthalmology-optometry.html) was used in the C20-5 supra-threshold mode. Contrast thresholds were determined at 17 locations within the central 20˚ of the visual field. Classification of an abnormal result was based on the following criteria: any location missed at the p<5% or any location missed at the p<1% significance level from

the manufacturers' normative database. If any of the indices were unreliable, which included false positives, false negative or fixation losses ≥33% or any point missed, the test was repeated.

IVue SD-OCT

IVue SD-OCT (Optovue Inc., software version 3.2.0.42, www.optovue.com/) was used to scan the posterior segment of the eye, capturing data from the optic nerve head retinal nerve fibre layer (RNFL) and total retinal thickness. Scans were taken through un-dilated pupils under dark room conditions. Cut-offs for abnormality were for any RNFL parameter p<1% or p<5% from the manufacturers' normative database.

Reference test

All participants underwent a reference standard ophthalmic examination on the same day as the index tests by an experienced glaucoma-specialist optometrist (with training and accreditation within glaucoma clinics in the UK Hospital Eye Service), masked to the index tests results. The ocular examination comprised: Log MAR best correct visual acuity; refraction; intraocular pressure measurement (IOP) with a Goldmann Applanation Tonometer; slit-lamp biomicroscopy including grading of the crystalline lens with the LOCS III and van Herick assessment of limbal anterior chamber depth[27] (with potentially occludable angles examined using gonioscopy); dilated fundus examination and detailed disc assessment using indirect ophthalmoscopy. Visual fields were measured with the Humphrey Field Analyzer (HFA) (Carl Zeiss Meditec Inc., www.zeiss.com/meditec/int/products/ophthalmology-optometry.html) using the Swedish Interactive Thresholding Algorithm 24-2 standard pattern (model 720i software version 5.1.2). HFA field-testing was repeated for false positives >15%, false negatives or fixation losses >33%. Glaucomatous visual field loss was classified using Hodapp Parrish-Anderson criteria.[15]

Evaluation of user acceptability

Subjects were also asked to complete a questionnaire regarding the acceptability of the 3 index tests compared to the HFA at the end of reference examination. The questionnaire evaluated ease of use, test duration and test comfort using a 7-point Likert scale.

Sample Size Calculation

Sample size was calculated based on the precision around an anticipated sensitivity of 0.85 with a minimum acceptable precision of ±0.20 with 0.95 probability. This level required a sample of 33 participants with glaucoma.[28]

Statistical Analysis for Diagnostic Accuracy Study

All tests were performed on both eyes for comparison with the clinical assessment of individual eyes and the data from one eye was used in the analysis. In the case of participants with glaucoma, this was the eye with the greater visual field loss and the right eye for the controls; if the right eye in the control group was not eligible, the left eye was selected provided it met the inclusion criteria.

Receiver operator characteristics (ROC) curves were plotted for the ability of the index tests to discriminate glaucomatous from non-glaucomatous eyes. Differences in the area under the ROC curve (AUROC) for each test parameter at the 95% confidence interval (CI) was compared statistically using the DeLong method,[29] a p-value of 0.05 was considered as the threshold for significance. Sensitivity, specificity, positive and negative likelihood ratios were also calculated. Statistical analysis was performed using SPSS 23.0 software (www.ibm.com/SPSS_statistics) and MedCalc 17.4 (www.medcalc.org).

Results

Development of the Accelerator 4AFC flicker test

Determination of optimal threshold criteria

Twenty normal participants (14 males and 6 females) were included in the study, with mean age 33.8 years (SD± 8.5) and mean spherical refractive error -0.50 D ±2.41 (2). The analysis was based on 10 right eyes and 10 left eyes. Table 1 shows the mean inter-session modulation levels for all 4 locations and run durations for each threshold criterion (T8 vs T12). The intersession differences in modulation levels were not statistically significant for either the T8 or the T12 termination criterion. In terms of run duration, inter-session differences for the shorter T8 termination criterion were similar; however, there was a significant decrease in test duration for T12. Overall, the mean durations were nearly twice as long for the T12 than the T8 criterion.

Table 1: Inter-session modulation levels and test durations (means and 95% CI per eye)

	Session 1	Session 2	Session 3	Session 4	p-value
% modulation (T8)	9.22 ± 0.17	9.57 ± 0.27	9.17 ± 0.15	9.07 ± 0.18	0.11
% modulation (T12)	8.61 ± 0.43	8.69 ± 0.41	8.76 ± 0.40	8.90 ± 0.37	0.11
Duration (T8) (s)	81 ± 22	71 ± 9	76 ± 15	73 ±13	0.36
Duration (T12) (s)	166 ± 49	148 ± 40	148 ± 40	126 ± 31	**< 0.001**

The total number of responses in the T8 condition averaged 42, implying that the stable threshold values were obtained in just over 10 responses per visual field location. The corresponding number for the T12 condition averaged 76, or 19 responses per visual field location.

The differences in % modulation among the four chosen field locations with either the T8 or T12 criterion were not statistically significant (Figure 3). The average within-participant standard deviation (SD) across the four locations was 0.52 dL and 1.32 dL for T8 and T12 respectively. The corresponding values for the Coefficient of Repeatability are 1.44 dL and 3.65 dL.

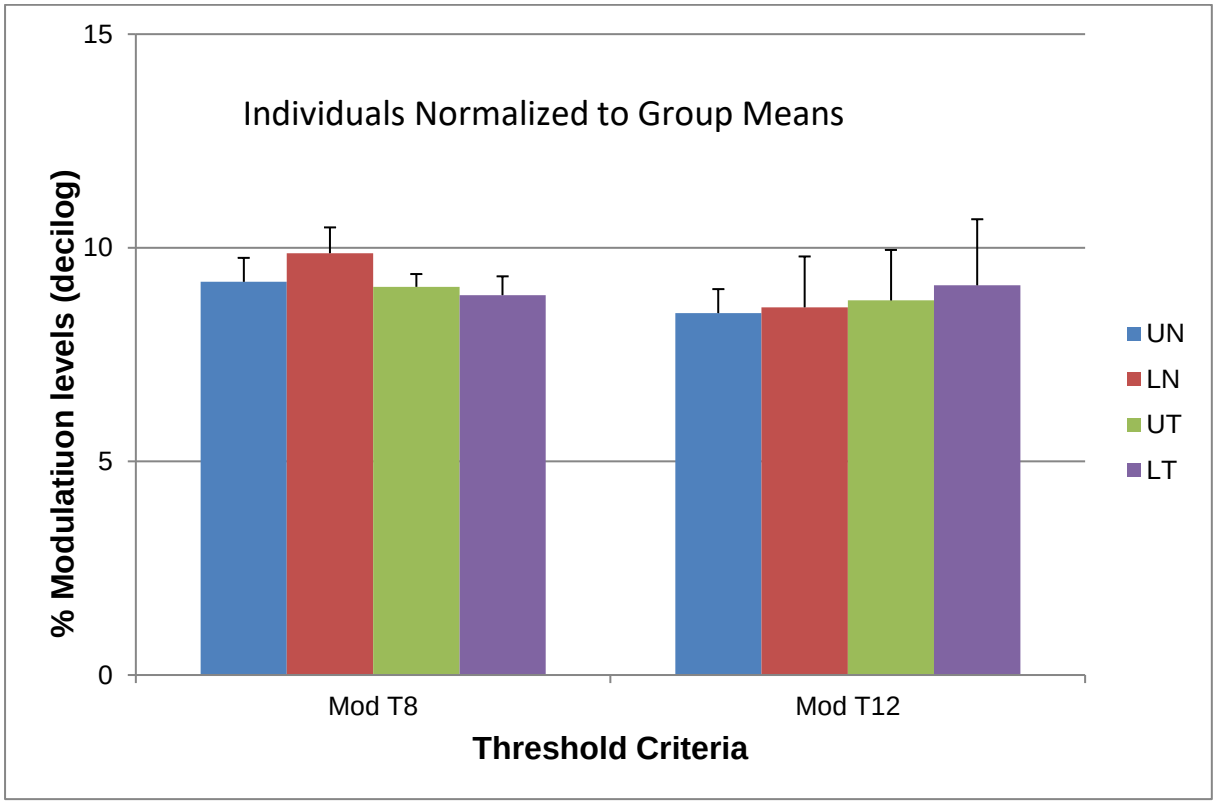
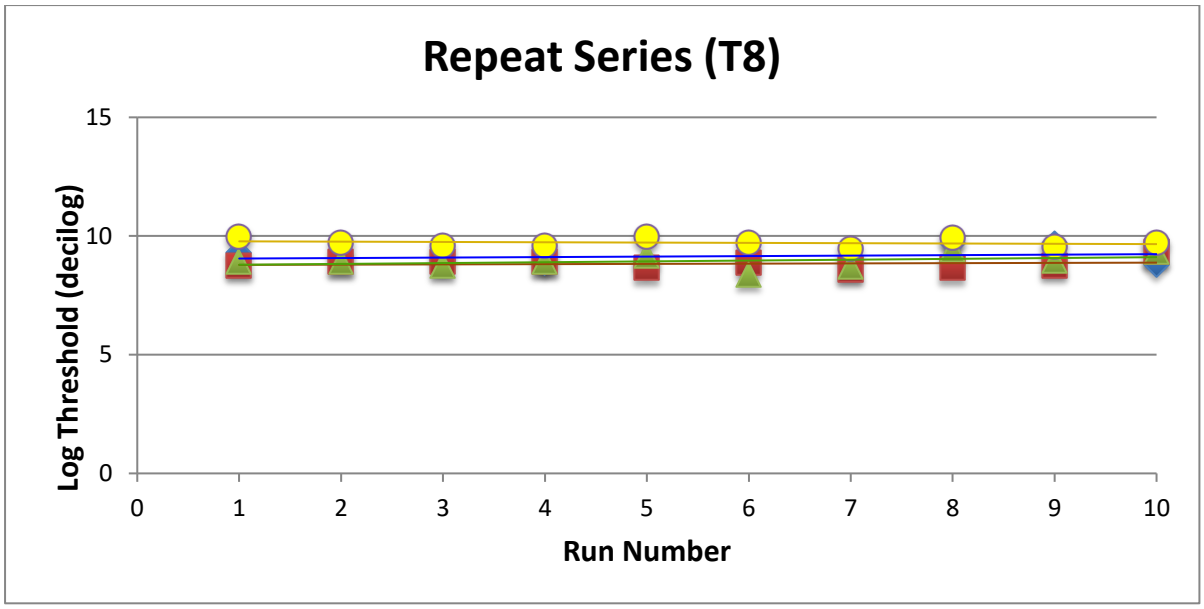


Figure 3. Decilog modulation levels over the four visual field locations, normalized to the group mean, for each threshold criterion. These error bars thus represent the average individual

variability at the four visual field locations. (UN – Upper Nasal, LN – Lower Nasal, UT- Upper Temporal and LT – Lower Temporal)

Assessment of threshold stability and learning effects

To study threshold stability, we evaluated the test-retest repeatability of the flicker algorithm by repeating the test 10 times over a 4-month period in four control observers to determine stability of the estimates. The average threshold values for the T8 and T12 are plotted in Figure 4A and B. For the T8 condition, all four slopes are less than 0.5 dL over the 10 runs, with small differences in the average sensitivity across observers and an average standard deviation (SD) of 1.66 dL. For the T12 condition, the data are generally more variable, with an average SD of 1.75 dL and somewhat lower threshold values, but there are again no significant trends for either improvement with practice.



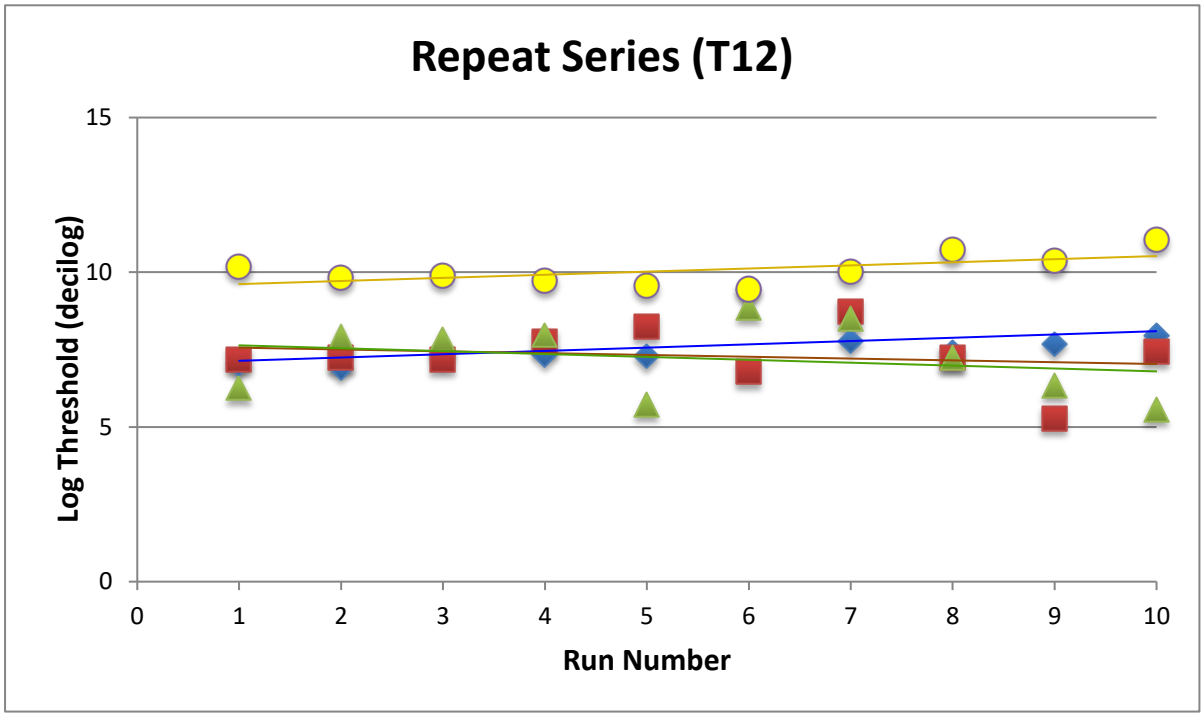


Figure 4. Trend analysis over 10 runs of the thresholds for each field location averaged for four control observers. A: Data for the criterion run length of 8 (T8 condition). B: Data for the criterion run length of 12 (T12 condition). None of the slopes differed significantly from zero ($p<0.05$).

Diagnostic performance of the A4FTp flicker test in the detection of glaucoma

Participant characteristics

Eighty-one participants were assessed for initial eligibility and invited to participate. Three participants were excluded from the analysis: 2 control participants (one had uninterpretable test results on all visual function tests and another had suspected glaucoma) and 1 glaucoma participant (who had a bilateral neurological visual field defect). In total, 38 controls and 40 participants diagnosed with glaucoma were included in the final analysis. The glaucoma group contained more pseudophakic test eyes (n=22, 55%) compared to the controls (n=2, 5.3%). There was no statistically significant difference between the two groups in visual acuity, refractive error or gender. Most participants were of Caucasian origin and the average ages for the control and glaucoma groups were 61.6 years (95% CI 58.1-65.0) and 71.9 years (68.8-74.9) respectively.

Glaucoma cases were further categorised according to glaucoma severity using the criteria from Hodapp et al[15] as: early (n=13, 32.5%), moderate (n=14, 35%) and advanced glaucoma

(n=13, 32.5%). A summary of the demographic and clinical data for the controls and glaucoma participants is provided in Table 2**;** t-tests were used to compare the differences between the two groups, apart from gender, where a chi-squared test was used.

Table 2. Demographical and summary clinical data for all participants and each group.

	Overall	Control	Glaucoma	P
No. Participants	78	38	40	
Age Years (Mean ± SD) (years)	66.8±11.2	61.6±10.6	71.9±9.4	<0.001
Female No. (%)	46 (59%)	22 (57.9%)	24 (60%)	0.85
Ethnicity Caucasian No. (%)	65 (83.3%)	28 (73.7%)	37 (92.5%)	
Asian Indian No. (%)	12 (15.4%)	10 (26.3%)	2 (5%)	
African Origin No. (%)	1 (1.3%)	0 (0%)	1 (2.5%)	
Visual Acuity (Log) (Mean ± SD)	0.07±0.15	0.04±0.17	0.09±0.12	0.092
IOP (mm Hg) (Mean ±SD)	16.4±4.42	17.5±2.5	15.3±5.5	0.024
Refractive Error (DS) (Mean ± SD)	-0.50±3.20	-0.26±3.59	-0.72±2.81	0.53
Refractive Error (DC) (Mean ± SD)	-0.87±0.84	-0.72±0.71	-1.01±0.93	0.13
Humphrey SAP SITA 24-2 Threshold MD (Mean ± SD) (dB)	-5.75±7.41	-0.71±1.55	-10.53±7.61	<0.001

DS: Dioptric Sphere; DC: Dioptric Cylinder; MD: Mean Deviation; PSD: Pattern Standard Deviation; SAP; Standard automated perimetry.

Diagnostic performance of the A4FTp

The best performing parameter for the A4FTp (as determined by the greatest AUROC) was the mean threshold from all four tested locations. The A4FTp achieved a sensitivity and specificity of 83% and 74% respectively, with an AUROC of 0.82 (95% CI 0.73-0.92). Figure 5 shows a histogram of the mean log flicker thresholds for the control and glaucoma subgroups with their 95% confidence intervals. Mean log threshold increased with increasing disease severity, although there was overlap in the distribution for control participants and those with early glaucoma. Consequently, whilst the A4FTp test identified 93% and 100% of moderate and severe glaucoma respectively, it identified half of those diagnosed with early glaucoma. Although the sensitivity of the test could potentially be increased by lowering the log threshold, this would lead to an unacceptable reduction in specificity (see Figure 6, which shows a sensitivity/specificity plot for different thresholds for the A4FTp).

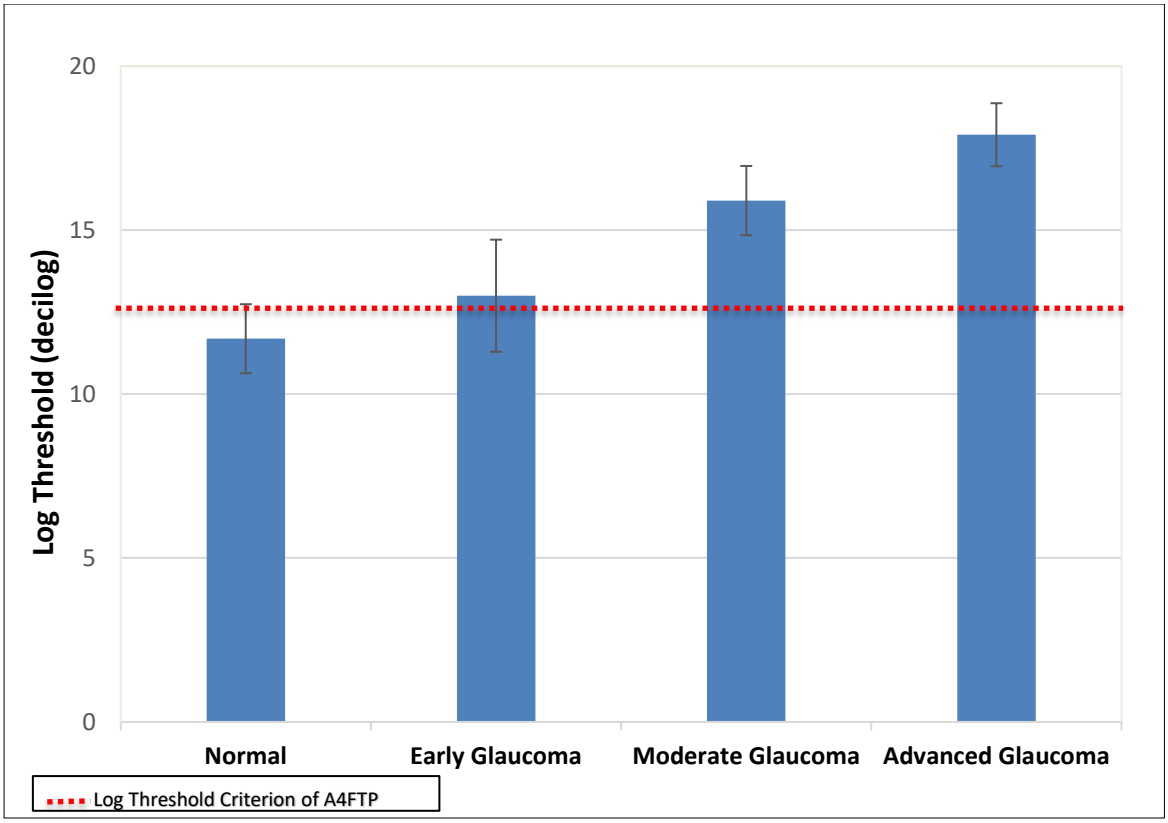


Figure 5. Histogram of the mean log thresholds of control and glaucoma subgroups.

.

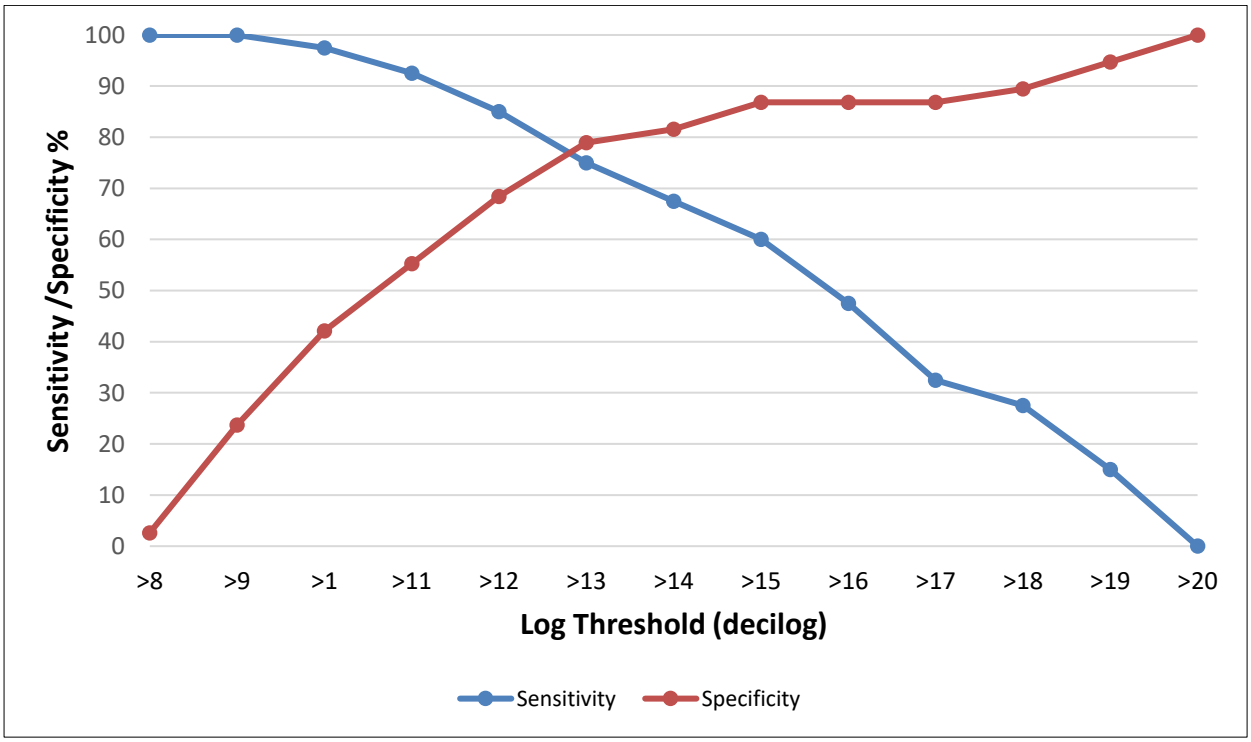


Figure 6. Sensitivity and specificity plot for A4FTp

Comparison of the A4FTp with other index tests

Table 3 displays the diagnostic performance of all the index tests, the optimal threshold for the FDT was: any location missed at p<5% level, and for the SD-OCT: any quadrant at the p<1% level in the RNFL. The AUROC for the FDT and SD-OCT, respectively were 0.91 (0.82-0.96) and 0.90 (0.83-0.967); sensitivities and specificities were (90%, 92%) and (88%, 84%), respectively.

Table 3. Diagnostic performance of the index tests.

	Index Test Parameter	Best Cut-off Value (dL)	Sensitivity (%)	Specificity (%)	PLR	NLR	AUC at 95% Confidence Interval
A4FTp	Mean of 4 Locations	>12.3	82.5	73.7	3.1	0.2	0.82 (0.73-0.92)
	Superior hemifield	>10.8	95.0	55.3	2.1	0.1	0.79 (0.69-0.88)
	Inferior hemifield	>13.1	75.0	73.7	2.9	0.3	0.78 (0.67-0.87)
	Nasal hemifield	>13.3	82.5	73.7	3.1	0.2	0.80 (0.70-0.88)
	Temporal hemifield	>10.5	95.0	57.9	2.3	0.1	0.79 (0.68-0.87)
	Superior nasal quadrant	>10.4	92.5	50.0	1.9	0.2	0.77 (0.66-0.86)
	Superior temporal quadrant	>10.9	77.5	68.4	2.5	0.3	0.76 (0.65-0.85)
	Inferior nasal quadrant	>14.1	67.5	73.7	2.6	0.4	0.74 (0.63-0.84)
	Inferior temporal quadrant	>10.3	97.5	47.4	1.9	0.1	0.75 (0.64-0.84)
FDT	Any point missed at p<5% level	>0	90.0	92.1	11.4	0.1	0.91 (0.82- 0.96)
	Any point missed at p<1% level	>0	82.5	97.4	31.7	0.2	0.90 (0.81-0.96)
iVue SD-OCT RNFL	Any quadrant at p<1%	-	87.5	84.2	5.5	0.1	0.90 (0.83-0.97)
	Any quadrant at p<5%	-	95	68.4	3.0	0.1	0.82 (0.74-0.90)

PLR: Positive Likelihood ratio; NLR: Negative Likelihood Ratio.

Table 4 shows pairwise comparisons of the AUROC curves between the index tests. There was no statistical significances between the A4FTp and FDT p<1% level (p=0.15) or FDT p<5% level (p=0.12). Comparison with respect to the A4FTp and the SD-OCT parameters; SD-OCT RNFL any quadrant at the p<1% level (p=0.18) and p<5% level (p=0.91) also showed no statistically significant differences.

Table 4. Pairwise comparison of ROC curves of the index tests.

Test comparisons	Difference between the areas	95% Confidence Interval	P
A4FTp vs FDT p<1% any point missed	0.08	-0.03-0.18	0.15
A4FTp vs FDT p<5% any point missed	0.09	-0.02-0.02	0.12
A4FTp vs SD-OCT p<1%	0.07	-0.04-0.18	0.18
A4FTp vs SD-OCT p<5%	0.01	-0.11-0.12	0.91

Figure 7 shows a Venn diagram for the best performing criteria of each index test in identifying the glaucoma participants. The A4FTp detected slightly fewer glaucoma cases (n=33, 83%) than the FDT (n=36, 90%) or SD-OCT (n=35, 88%). Two cases (5%) were missed by all three index tests. The diagram shows that a screening strategy that combines a structural test (SD-OCT) with a functional test (FDT or A4FTp) increases the likelihood of detecting the disease.

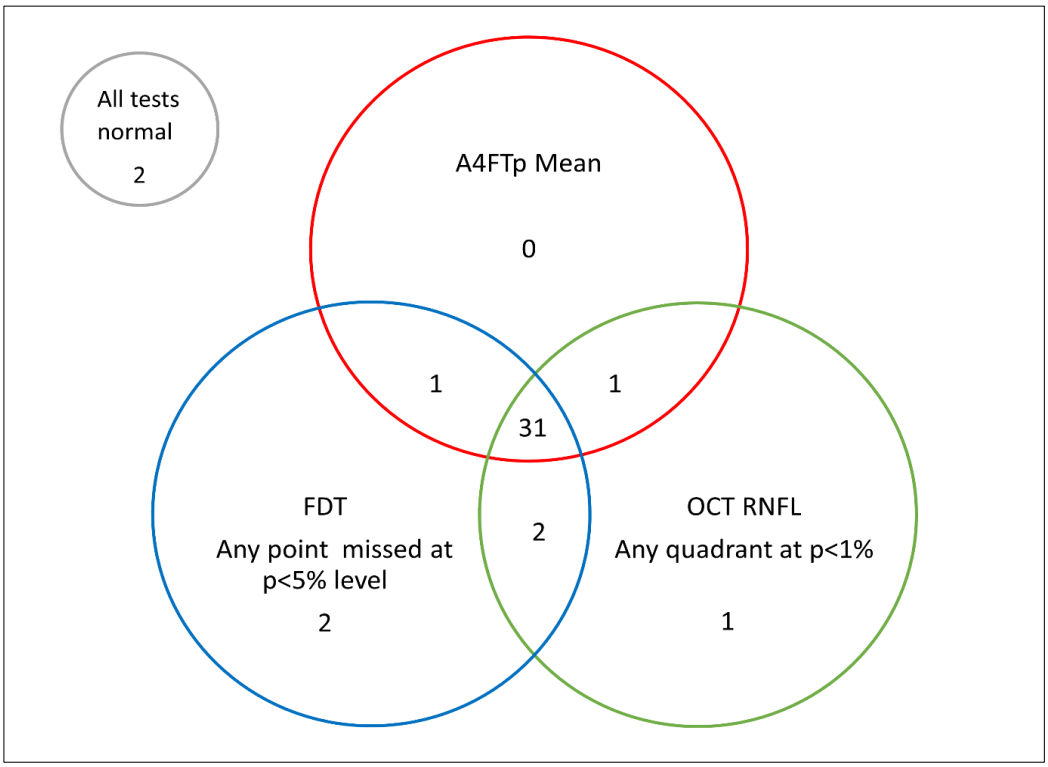


Figure 7. Venn diagram of best performing parameter from the index tests in identifying the glaucoma cases alone or combined with the other tests. The numbers within the circles represent the number of participants identified by each test. In two participants all tests were normal.

Test duration

Table 5 and Figure 8 depicts shows the time taken to perform the functional tests (including the HFA 24-2, which was undertaken as part of the reference examination). The mean time taken for all participants in the tests were FDT C20-5 (61.6±34.8 seconds), A4FTp (141.8±85.6 seconds) and HFA (364.0±79.3 seconds), which were significantly different between all three tests ($p<0.001$). There were also statistically significant differences between the control and glaucoma groups for the HFA ($p<0.001$), FDT C20-5 ($p<0.001$) and A4FTp ($p<0.008$).

Table 5. Time taken to perform each functional test.

	Total	Control	Glaucoma	P
HFVA SITA 24-2 Threshold time taken (sec) Mean ± SD)	364.0±79.3	306.5±29.4	418.6±73.0	<0.001
FDT time (Supra-threshold C20-5) time taken (sec) Mean ± SD	61.6±34.8	36.5±8.9	85.5±33.4	<0.001
A4Tp Threshold time taken (sec) Mean ± SD)	141.8±85.6	115.7±44.5	166.5±106.3	0.008

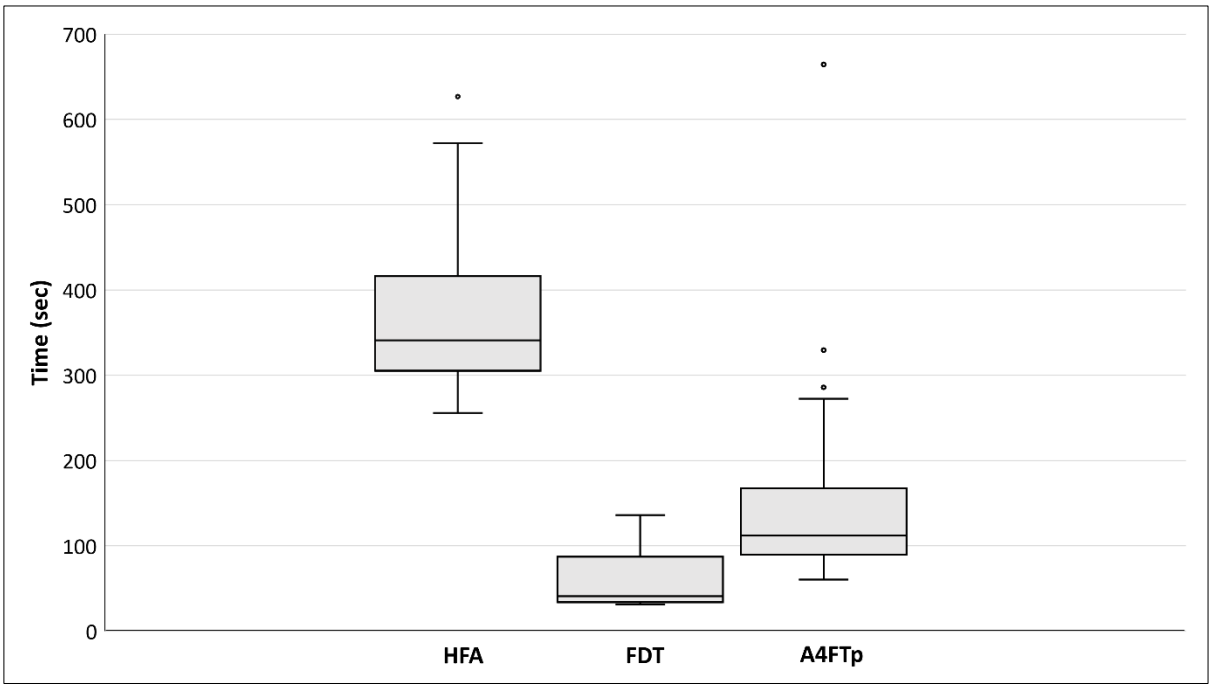


Figure 8. Box and Whiskers plot of time taken to perform each functional test.

User acceptability

Figure 9 shows the Likert scores responses from the tests from an acceptability survey; there was a higher proportion of participants who found the HFA uncomfortable, too long and difficult to perform, compared to the other index tests. The A4FTp, FDT and SD-OCT had a similar proportion of participants (>90%) rating the tests as not uncomfortable or not too long. However, more participants found the A4FTp difficult to perform in its current configuration.

There were no adverse events occurring when performing any of the index tests.

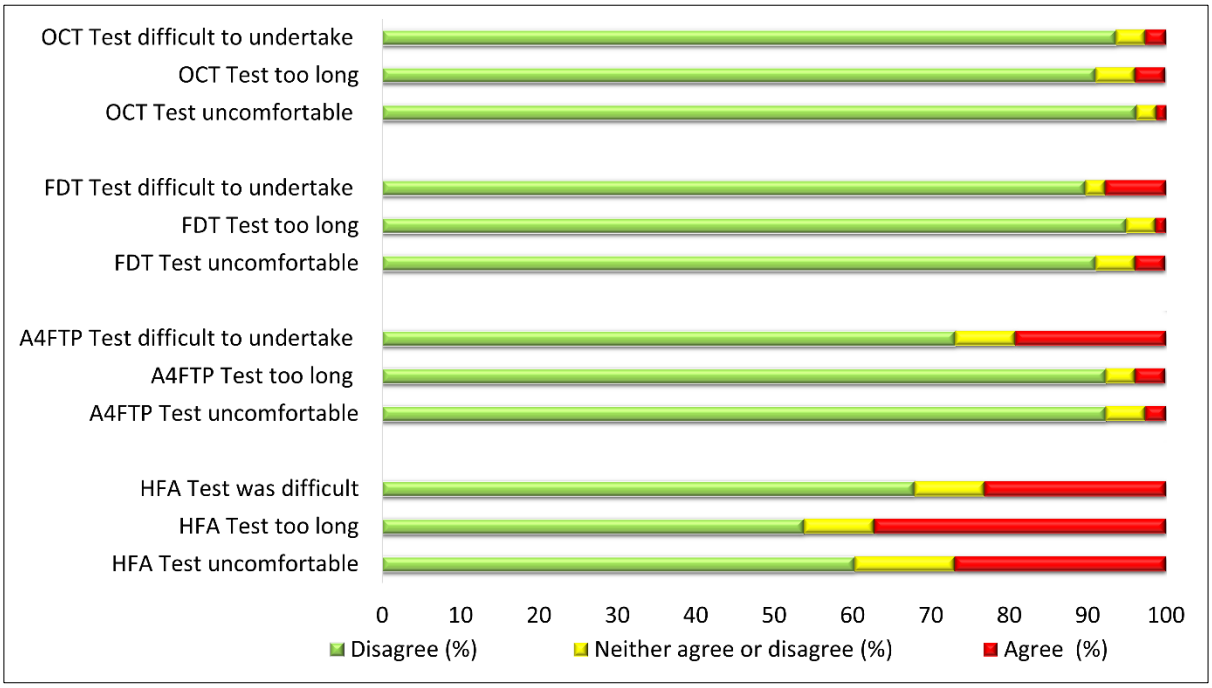


Figure 9. Likert responses from user acceptability survey from all participants.

Discussion

Glaucoma is the leading cause of irreversible blindness worldwide. A recent systematic review estimated that glaucoma was responsible for 6.6% of total global blindness and 2.2% of all moderate and severe visual impairment.[30] The lack of symptoms in its early stages, coupled with inequitable access to healthcare in many countries, contributes to high rates of undiagnosed glaucoma and late clinical presentation. The main aim of the study was therefore to evaluate the diagnostic performance of a simple screening test that could be potentially used in conjunction with other clinical tests to detect glaucoma in a home setting or in underserved populations.

The A4FTp measures flicker thresholds at four fixed locations strategically placed to detect the arcuate or nasal step defects that are commonly seen in glaucoma. The flicker rate was set at 30Hz, the value found to be diagnostic of glaucoma in previous studies.[7, 31] Flicker sensitivity has been shown to be resilient to optical blur[11] and the large stimulus size minimised the impact of fixation losses. Furthermore, presenting the stimuli in a red field reduces optical scatter due to opacities in the ocular media, including cataract.

Psychophysical tests that are used for clinical purposes should have a minimum set of psychophysical attributes before they can be introduced for clinical testing, including an acceptable within and between test variation. Within-test variability is caused by several aspects: neural noise, decision criteria, and thresholding strategy.[32, 33] Between-test variability has been attributed to ocular and neural sensitivity fluctuations.[34] The variability that occurs during test and retest is quantified by the degree of scatter between measurements taken at different test sessions.[34-36] Considering first the stability over time, we found that there was no significant drift in the estimated values for either threshold criterion (T8 or T12) in normals for either the short- or long-term test series (although the duration for T12 did significantly decrease over time). Thus, the A4FTp staircase was equally stable under all test conditions. However, there was a notable difference between the criteria in the variation across individuals, which had the remarkably low standard deviation of 0.52 dL for T8 compared with 1.32 dL for T12. These values may be compared with the standard deviation values of approximately 1.7 dL reported for the population variation of 2AFC flicker threshold estimates in a previous study.[37] Thus, while the standard deviation for the 12-point staircase criterion is close to the previously reported range, the 8-point criterion markedly reduces the standard deviation to about one third of this level, implying that the physiological strain of the longer runs overcame the statistical advantage of the increased number of samples. Since variance is defined as the square of the standard deviation, this result further implies that about 90% of

the variance in the previous methods, and in the 12-point criterion staircase, was due to methodological variations, with no more than 10% of the variance attributable to inherent population variability. Moreover, there is no significant gain in reliability for the extra time spent to reach the 12-point criterion, and, though slightly lower on average, the threshold values themselves are not significantly different from those for the 8-point criterion. Thus, we have to conclude from the 20-normal participant results that, of the two approaches evaluated, the 8-point criterion staircase is a strikingly effective approach to flicker threshold measurement, with no evident disadvantages and a clear time advantage over the 12-point criterion staircase.

The most important characteristic of a screening test is its ability to differentiate between normal and diseased individuals. To assess the diagnostic accuracy of the A4FTp, we measured flicker thresholds in 40 subjects with COAG and 38 normal controls. Based on a comparison of areas under the ROC curve, the overall performance of the A4FTp was similar to the FDT (C20-5 algorithm) and the SD-OCT (RNFL thickness outside normal limits). The best performing criterion for the A4FTp was the mean threshold of all 4 stimulus locations. Using this criterion, the test identified 33 out of the 40 glaucoma cases in our sample (83%). All cases of advanced glaucoma were detected, all but one of the moderate glaucoma's and half of the early glaucoma's. The mean time taken for all participants to perform the A4FTp was within the clinically recognised '3-minute period of vigilance'; this is advantageous in terms of response variability, as beyond this time it has been noted that the response variability increases.[38] User acceptability of the A4FTp was positive, with similar acceptability questionnaire scores to the FDT and SD-OCT, in terms of comfort and participant's opinion on the duration of the test.

The 'optimal' threshold criterion for the A4FTp was based on an equal weighting for sensitivity and specificity. Sensitivity and specificity are inversely proportional and therefore lowering the threshold to maximise sensitivity leads to a corresponding reduction in specificity. A further issue with population screening for glaucoma is the relatively low prevalence of the disease. Based on a glaucoma prevalence of 5% in high-risk populations[30] and a test with 90% sensitivity and 90% specificity, only one in three persons screening positive will have the disease. Currently, no single test for glaucoma has been shown to have a sufficiently high sensitivity and specificity for the detection of glaucoma in the general population.[39] It is therefore likely that future screening strategies for glaucoma will employ combinations of tests and will target those most likely to be affected by vision loss in their lifetime,[40] putting a premium on the rapidity of the tests selected for inclusion. Combining structural and functional testing can be used to improve the sensitivity or specificity for glaucoma detection, depending on whether the priority is to maximise true positives or minimise false positives.[41] For

example, if structural and functional screening tests are used and disease positives are defined as those who test positive by either test, there will be a net increase in sensitivity; conversely, a strategy where disease negatives are defined as those who test negative on both tests will maximise specificity.

This study has a number of strengths: the design, analysis and reporting of the diagnostic accuracy study complied with the principles of the Standards for Reporting of Diagnostic Accuracy (STARD). The performance of the index tests was compared in a representative sample of participants with the target condition, with a range of disease severities. The reference standard and index tests were conducted on the same day as the index tests by an experienced clinician masked to the index test results. The reference standard for COAG was based on a comprehensive ophthalmic examination typical of that conducted in a hospital glaucoma unit. However, since both structural and functional tests were evaluated for the case definition, this required both established visual field loss and glaucomatous disc damage to be present. Although the use of a case-control design may have artificially overestimated the performance of the A4FTp, given the comparative nature of the study, the same degree of bias would apply to all tests. Similarly, although the age of the control and glaucoma groups differed by approximately a decade, the relative effect of an age-related loss of flicker sensitivity would have applied equally to the performance of the FDT.

Whilst the relative performance of the three index tests was similar based on AUROC analysis, the specificity for the A4FTp was the lowest of all three tests. Three participants in the control group had log threshold values for the A4FTp well outside the 95% confidence interval. If these ‘outliers’ were removed from the A4FTp, the overall specificity would increase from 74% to 80%. With further refinement and optimisation, the data suggest that the A4FTp could have a role in combination with other tests for glaucoma screening. The test has a number of advantages including: its easy administration, ready interpretation, relatively short testing time and robustness to the effects of media opacity or refractive error. Following the success of this proof of concept study, we are currently in the process of optimising the A4Ftp for a touch screen tablet display to increase its portability.

References

1. Jampel HD, Singh K, Lin SC, Chen TC, Francis BA, Hodapp E, et al. Assessment of visual function in glaucoma: a report by the American Academy of Ophthalmology. Ophthalmology. 2011;118(5):986-1002.
2. McKendrick AM. Recent developments in perimetry: test stimuli and procedures. Clin Exp Optom. 2005;88(2):73-80.
3. Wu Z & Medeiros FA. Recent developments in visual field testing for glaucoma. Curr Opin Ophthalmol. 2018;29(2):141-146.
4. Vingrys AJ, Healey JK, Liew S, Saharinen V, Tran M, Wu W, et al. Validation of a Tablet as a Tangent Perimeter. Transl Vis Sci Technol. 2016;5(4):3.
5. Schulz AM, Graham EC, You Y, Klistorner A & Graham SL. Performance of iPad-based threshold perimetry in glaucoma and controls. Clin Exp Ophthalmol. 2018;46(4):346-355.
6. Tyler CW. Analysis of normal flicker sensitivity and its variability in the visuogram test. Invest Ophthalmol Vis Sci. 1991;32(9):2552-2560.
7. Tyler CW. Specific deficits of flicker sensitivity in glaucoma and ocular hypertension. Invest Ophthalmol Vis Sci. 1981;20(2):204-212.
8. Lachenmayr BJ & Drance SM. The selective effects of elevated intraocular pressure on temporal resolution. Ger J Ophthalmol. 1992;1(1):26-31.
9. Horn FK, Jonas JB, Korth M, Junemann A & Grundler A. The full-field flicker test in early diagnosis of chronic open-angle glaucoma. Am J Ophthalmol. 1997;123(3):313-319.
10. Yoshiyama KK & Johnson CA. Which method of flicker perimetry is most effective for detection of glaucomatous visual field loss? Invest Ophthalmol Vis Sci. 1997;38(11):2270-2277.
11. Lachenmayr BJ & Gleissner M. Flicker perimetry resists retinal image degradation. Invest Ophthalmol Vis Sci. 1992;33(13):3539-3542.
12. Westcott MC, Fitzke FW, Crabb DP & Hitchings RA. Characteristics of frequency-of-seeing curves for a motion stimulus in glaucoma eyes, glaucoma suspect eyes, and normal eyes. Vision Res. 1999;39(3):631-639.
13. Bosworth CF, Sample PA & Weinreb RN. Motion perception thresholds in areas of glaucomatous visual field loss. Vision Res. 1997;37(3):355-364.
14. Wang Y & Henson DB. Diagnostic performance of visual field test using subsets of the 24-2 test pattern for early glaucomatous field loss. Invest Ophthalmol Vis Sci. 2013;54(1):756-761.
15. Hodapp E, Parrish RK & Anderson DR. Clinical decisions in glaucoma. St Louis: Mosby; 1993.
16. Henson DB, Chauhan BC & Hobley A. Screening for glaucomatous visual field defects: the relationship between sensitivity, specificity and the number of test locations. Ophthalmic Physiol Opt. 1988;8(2):123-127.
17. Sugimoto K, Schotzau A, Bergamin O & Zulauf M. Optimizing distribution and number of test locations in perimetry. Graefes Arch Clin Exp Ophthalmol. 1998;236(2):103-108.
18. Krakau CE. Visual field testing with reduced sets of test points. A computerized analysis. Doc Ophthalmol. 1989;73(1):71-80.
19. Gonzalez de la Rosa M, Reyes JA & Gonzalez Sierra MA. Rapid assessment of the visual field in glaucoma using an analysis based on multiple correlations. Graefes Arch Clin Exp Ophthalmol. 1990;228(5):387-391.
20. Keltner JL, Johnson CA, Cello KE, Edwards MA, Bandermann SE, Kass MA, et al. Classification of visual field abnormalities in the ocular hypertension treatment study. Arch Ophthalmol. 2003;121(5):643-650.
21. Nicholas SP & Werner EB. Location of early glaucomatous visual field defects. Can J Ophthalmol. 1980;15(3):131-133.
22. Wyszecki G & Styles WS. The Eye. Color science: concepts and methods, quantitative data and formulae. 2nd Edition ed. New York: Wiley; 1982.

23. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138(1):W1-12.
24. Michelessi M, Lucenteforte E, Oddone F, Brazzelli M, Parravano M, Franchi S, et al. Optic nerve head and fibre layer imaging for diagnosing glaucoma. Cochrane Database Syst Rev. 2015(11):CD008803.
25. Anderson D & Patella V. Interpretation of a single field. Automated static perimetry. . 2nd Edition ed. St Louis: Mosby; 1999.
26. Chylack LT, Jr., Wolfe JK, Singer DM, Leske MC, Bullimore MA, Bailey IL, et al. The Lens Opacities Classification System III. The Longitudinal Study of Cataract Study Group. Arch Ophthalmol. 1993;111(6):831-836.
27. Van Herick W, Shaffer RN & Schwartz A. Estimation of width of angle of anterior chamber. Incidence and significance of the narrow angle. Am J Ophthalmol. 1969;68(4):626-629.
28. Flahault A, Cadilhac M & Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol. 2005;58(8):859-862.
29. DeLong ER, DeLong DM & Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-845.
30. Kapetanakis VV, Chan MP, Foster PJ, Cook DG, Owen CG & Rudnicka AR. Global variations and time trends in the prevalence of primary open angle glaucoma (POAG): a systematic review and meta-analysis. Br J Ophthalmol. 2016;100(1):86-93.
31. Tyler CW, Hardage L & Stamper RL. The temporal visuogram in ocular hypertension and its progression to glaucoma. J Glaucoma. 1994;3 Suppl 1:S65-72.
32. Johnson CA, Adams CW & Lewis RA. Fatigue effects in automated perimetry. Appl Opt. 1988;27(6):1030-1037.
33. Gonzalez de la Rosa M, Gonzalez-Hernandez M, Garcia Feijoo J, Morales J & Azuara-Blanco A. Diagnostic accuracy and reproducibility of tendency oriented perimetry in glaucoma. Eur J Ophthalmol. 2006;16(2):259-267.
34. Spry PG & Johnson CA. Identification of progressive glaucomatous visual field loss. Surv Ophthalmol. 2002;47(2):158-173.
35. Spry PG, Henson DB, Sparrow JM & North RV. Quantitative comparison of static perimetric strategies in early glaucoma: test-retest variability. J Glaucoma. 2000;9(3):247-253.
36. Chauhan BC & Johnson CA. Test-retest variability of frequency-doubling perimetry and conventional perimetry in glaucoma patients and normal subjects. Invest Ophthalmol Vis Sci. 1999;40(3):648-656.
37. Heijl A & Bengtsson B. The effect of perimetric experience in patients with glaucoma. Arch Ophthalmol. 1996;114(1):19-22.
38. Henson DB & Emuh T. Monitoring vigilance during perimetry by using pupillography. Invest Ophthalmol Vis Sci. 2010;51(7):3540-3543.
39. Mowatt G, Burr JM, Cook JA, Siddiqui MA, Ramsay C, Fraser C, et al. Screening tests for detecting open-angle glaucoma: systematic review and meta-analysis. Invest Ophthalmol Vis Sci. 2008;49(12):5373-5385.
40. Friedman DS. Issues in screening for glaucoma. Ophthalmic Epidemiol. 2007;14(3):101-102.
41. Shah NN, Bowd C, Medeiros FA, Weinreb RN, Sample PA, Hoffmann EM, et al. Combining structural and functional testing for detection of glaucoma. Ophthalmology. 2006;113(9):1593-1602.