BJO Online First, published on August 16, 2017 as 10.1136/bjophthalmol-2017-310336

Comparing glaucomatous disc change using stereo disc viewing and the MatchedFlicker programme in glaucoma experts and trainees

Jamie L Schaefer,[1,2] Alissa M Meyer,[1] Cooper D Rodgers,[1] Nicole C Rosenberg,[1] Anthony J Leoncavallo,[1] Zachary L Lukowski,[1] Anthony B Greer,[1] Gina M Martorana,[1] Baiming Zou,[3] Jonathan J Shuster,[4] L Jay Katz,[5] Joel S Schuman,[6] Michael A Kass,[7] Mark B Sherwood[1]

[1]Department of Ophthalmology, University of Florida, Gainesville, Florida, USA
[2]Department of Ophthalmology, University at Buffalo, Ross Eye Institute, Buffalo, New York, USA
[3]Department of Biostatistics, University of Florida, Gainesville, Florida, USA
[4]Department of Health Outcomes and Policy, University of Florida, Gainesville, Florida, USA
[5]Glaucoma Service, Wills Eye Hospital, Thomas Jefferson University, Philadelphia, Pennsylvania, USA
[6]Department of Ophthalmology, New York University, New York City, New York, USA
[7]Department of Ophthalmology, Washington University School of Medicine, St. Louis, Missouri, USA

Correspondence to
Dr Mark B Sherwood, Box 100284 JHMHC, 1600 SW Archer Rd, Gainesville, FL 32610, USA; sherwood@ufl.edu

Received 20 February 2017
Revised 17 May 2017
Accepted 19 June 2017

ABSTRACT

Background/aims The objective of this study is to evaluate the accuracy and speed of trainees and experienced glaucoma specialists using the MatchedFlicker software against the manual examination of stereoscopic disc photographs for detecting glaucomatous optic disc change.

Methods Three experienced glaucoma specialists, two resident ophthalmologists and one glaucoma fellow from multiple institutions independently evaluated the same 140 image pairs from 100 glaucomatous/ocular hypertensive eyes using a handheld stereo viewer and the MatchedFlicker programme. Fifty had progression to glaucoma as determined by the Ocular Hypertension Treatment Study (OHTS) Optic Disc Reading Group and endpoint committee, and 50 more were negative controls for progression with photos taken a few minutes apart. Twenty photo pairs from each of the two groups were duplicated for reviewer variability analysis. The initial viewing method was randomised and then alternated for each group of 70 image pairs. Reviewer accuracy and evaluation time for each method were measured.

Results Evaluators averaged 8.6 s faster per image pair (26%) with the MatchedFlicker programme than with the stereo viewer (p=0.0007). Evaluators correctly identified more image pairs when using the MatchedFlicker software over the stereo viewer (p=0.0003). There was no significant difference between the expert and trainee group in speed or overall accuracy for either method. Experts were significantly more consistent than trainees with the duplicate image pairs (p=0.029). Trainees appeared more reluctant to designate eyes as showing glaucoma progression than experts.

Conclusions Both expert glaucoma specialists and ophthalmologists in various stages of training had greater accuracy and speed with the MatchedFlicker programme than with a handheld stereoscopic viewer.

To cite: Schaefer JL, Meyer AM, Rodgers CD, *et al*. *Br J Ophthalmol* Published Online First: [*please include* Day Month Year]. doi:10.1136/bjophthalmol-2017-310336

INTRODUCTION

An estimated 3 million people suffer from primary open angle glaucoma (POAG) in North America alone.[1] Meticulous examination of the optic nerve head is a crucial part of monitoring glaucoma progression. Often, the quality of the image scan or photograph limits the accuracy of this technique.

Clinicians must assess photos for signs of disease progression including disc haemorrhages, focal rim defects, optic nerve rim thinning and retinal nerve fibre layer (RNFL) abnormalities.[2] Optic disc stereoscopic photographic analysis is one of the standard methods for monitoring glaucomatous structural change in the optic disc and is deemed superior to written subjective documentation and drawings.[3]

In large randomised trials, the National Eye Institute (NEI) has utilised an optic disc reading centre as a primary means to detect glaucomatous progression.[4] In the Ocular Hypertension Treatment Study (OHTS), Kass *et al* reported that 55% of eyes that progressed to primary open angle glaucoma (POAG) endpoints were initially identified based on stereophotos alone, 35% by visual field changes and 10% based on both visual field and stereoscopic photographic changes simultaneously.[5]

Although disc photo evaluation is widely used throughout the world to assess structural progression, disc evaluation through examination of sequential simultaneous stereo photographs is not done routinely in practice, possibly due to the specialised equipment required to take a stereoscopic photo and the time-consuming nature of stereophoto evaluation. Additionally, stereophotos are not easily integrated into a patient's electronic medical record (EMR).

This study is a continuation of our previous work comparing the accuracy and speed of the new MatchedFlicker (EyeIC Inc., Narberth, Pennsylvania, USA) technique to the traditional method of examining slides of stereoscopic disc photographs. In our original study, we compared the two techniques using non-expert trainee observers in various stages of their ophthalmology training.[6] For these trainees, the Flicker technique was both more accurate and time efficient when compared with stereo photos. It is also important to determine the efficacy of implementing the MatchedFlicker technique in practices manned by experienced clinicians where the use of stereoscopic photos is already a deeply ingrained skill and an established part of patient care, so we have now expanded our study to include highly experienced academic glaucoma specialists.

METHODS

Protocol approval was obtained prospectively from the University of Florida's Institutional Review

Board (UF IRB) for this experimental study, which adheres to the tenets of the Declaration of Helsinki. IRB approval was given for the participation of the photo evaluators. A full waiver of informed consent was granted for the use of de-identified patient optic nerve photographs from patient records and from the Bascom Palmer Eye Institute optic disc reading centre because no risk was posed to any patient with their use.

Three glaucoma specialists, two ophthalmology residents and one glaucoma fellow independently evaluated stereoscopic pairs of disc photographs of 100 eyes taken at two time points. As we described in our previous paper,[6] 50 optic nerve head (ONH) image pairs serving as positive controls were identified from 50 patients in the OHTS study that showed disc progression as determined by the OHTS Optic Disc Reading Group of the Bascom Palmer Eye Institute and confirmed as glaucomatous change by the OHTS endpoint committee.

Fifty other ONHS image pairs serving as negative controls were obtained from 50 additional patients at the University of Florida, using simultaneous stereo images of eyes taken just a few minutes apart with a Topcon TRC-50DX simultaneous stereo camera system (Topcon, Oakland, New Jersey, USA). Twenty of the 50 positive control pairs and 20 of the 50 negative control pairs were randomly duplicated to allow for the assessment of intraobserver variability in detecting progression. Thus, a total of 140 image pairs were analysed by each observer.

The slides from the OHTS study group and from the University of Florida were reformatted and rebound so that each had a consistent external appearance. Each observer used two different examination methods to judge the optic disc photos for glaucomatous progression: Pentax handheld stereo viewers (Pentax, Tokyo, Japan) (figure 1) and the computer-based Matched-Flicker programme (here on identified as 'Flicker') (figure 2).

Glaucomatous progression was defined by disc rim thinning (focal or diffuse), vessel movement related to increased cupping or detection of new or enlarged RNFL defects. All three trainee evaluators had some experience reviewing stereophotographs with a handheld stereo viewer, and all three expert evaluators were highly experienced in reviewing stereophotographs. One of the trainee evaluators and one of the experts had a very limited experience with the Flicker programme prior to the study, but the others were all naive to this programme. Prior to examining the study photographs, all six evaluators were given a practice session with 10 additional non-study photo pairs to familiarise them with both techniques.

Evaluators viewed four different blocks of 70 image pairs alternating the viewing method for each block. The initial method type (Flicker or stereophoto) was randomly selected prior to viewing. Observers were prompted to choose between 'progression' and 'non-progression' for each image pair. For all three expert and three trainee reviewers, the same study coordinator noted the evaluators' assessment for each pair and recorded the overall review time for the 140 image pairs for each method.

Statistical analysis

Statistical analysis starts with the assessment of reviewer accuracy comparisons. Accuracy was defined as the percentage of all designations that were correct out of the 100 slides. Specifically, we evaluated two different effects, that is, reviewing technique (Flicker vs slide) and reviewer expertise (expert vs trainee), on accuracy. Following the accuracy comparisons, we conducted the analysis for consistency. Consistency is defined as the percentage of all designations that matched each other in the reviewing of two duplicated images.

Figure 1 Pentax handheld stereo viewer.

In the statistical analysis for assessing accuracy and consistency, the following logistic regression model was used to adjust for confounding factors:

$$logit\left(Pr\left(y_{ijk}=1\right)\right)=\alpha_0+\alpha_{rev}I\left(k=1\right)+\alpha_mI\left(i=1\right)$$

where y_{ijk} is the binary response for accuracy (1 for correct and 0 for incorrect) or consistency (1 for matched and 0 for unmatched) of subject j, reviewed by Reviewer k (1=expert and 0=non-expert) using method i (1=Flicker and 0=Slides). This produces an adjusted ORs showing the reviewer effect (parameter α_{rev}) and the method effect (parameter α_m) on accuracy or consistency. For parameter α_{rev}, ratios greater than 1.0 favour expert reviewers and ratios less than 1.0 favour trainees. For parameter α_m, ratios greater than 1.0 favour Flicker and ratios less than 1.0 favour the slide method.

Additionally, a two-sided paired t-test was used to compare the difference in mean time per evaluation between the two viewing methods. Mean time per evaluation for each practitioner was calculated by dividing the total time taken for an image block by the overall number of slides (140), yielding six paired observations.

True positive, true negative and true correct assessment rates between slides and Flicker were compared using individual experimental units: 50 units for true negatives, 50 for true positives and 100 total. The number of correct responses for both viewing methods in each experimental unit was scored from 0 to 6 for all evaluators and from 0 to 3 for the trainee and expert groups. All comparisons were conducted by paired (one-sample

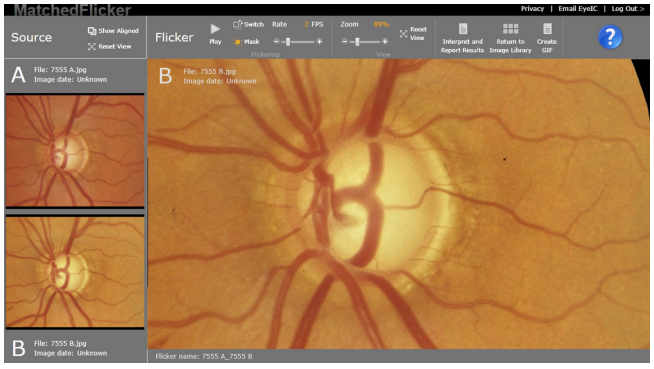

Figure 2 Images produced with the MatchedFlicker software programme.

Table 1 Comparison of mean time per image of individual reviewers between stereo slide viewing and the MatchedFlicker software programme

Method	Evaluator group	Mean seconds per image (SD), n=3	p Value* (expert vs trainee)	Mean seconds per image (SD), n=6	p Value* (slide vs Flicker)
Slide	Expert	32.01 (10.96)	0.78	33.024 (7.67)	0.0007
	Trainee	34.04 (4.90)			
Flicker	Expert	23.98 (8.47)	0.88	24.44 (6.46)	
	Trainee	24.90 (5.65)			

*p Value was assessed with a two-sided paired t-test.
A comparison of the time spent by evaluators using either the stereoscopic slide viewing technique or the Flicker programme to determine progression of the 140 image pairs.

Table 2b Comparison of accuracy between experts and trainees

Image group	Average % accuracy for experts	Average % accuracy for trainees	p Value*
Negative progression slides	88.7	96.7	0.0091
Positive progression slides	73.3	54.7	0.0018
Total slides	**81.0**	**75.7**	**0.1242**
Negative progression Flicker	95.3	88.0	0.0546
Positive progression Flicker	78.7	84.0	0.4045
Total Flicker	**87.0**	**86.0**	**0.7883**

*p Value was assessed with a two-sided paired t-test.
A comparison of the average per cent accuracy of the three experts and three trainees while using either the stereoscopic slide viewing technique or the Flicker programme to determine progression of the 50 negative progression image pairs, the 50 positive progression image pairs and the total 100 unduplicated image pairs.

matched) Z-tests. Comparisons included slide versus Flicker and within each method experts versus trainees.

An open source R statistical analysis package was used to analyse and create logistic regression models. Statistical Analysis System (SAS 9.4) software was used for comparing times and the various success rates. All p values are two-sided.

RESULTS

Evaluators were 8.6 s (26%) faster on average assessing images using the Flicker method (p=0.0007) (table 1). There was no significant difference in the mean time per assessment between the glaucoma expert and trainee groups for either the stereo photo or the Flicker method (p=0.78 and p=0.88, respectively).

Overall, the reviewers were more accurate using Flicker than using slides when analysing both the positive progression group (50 photos) and all 100 photo pairs (p<0.0001 and p=0.0003, respectively). For the negative progression control group (50 photos), there was no difference in accuracy using slides or Flicker (p=0.6) (table 2a).

When comparing experts and trainees using the slide method, trainees were significantly more accurate than experts for the negative progression group (p=0.009), but experts were more accurate for the positive progression group (p=0.002) (table 2b).

When comparing the accuracy of the three experts between the two viewing methods, experts were significantly more accurate using Flicker rather than slides for the 50 negative progression image pairs and for the 100 image pairs overall (p=0.039 and p=0.046). While not significant, the experts were also slightly more accurate (79% vs 73%) using Flicker rather than slides for the 50 positive progression image pairs (table 2c). For the three trainees, unlike the experts, the slides provided greater accuracy in the negative progression groups while for the positive progression and the overall slides the Flicker was more accurate (table 2b). However, trainees gave a non-progression designation on average to 71 of the 100 initial slides they reviewed (table 3).

Sensitivity is the percentage of positive diagnoses made that were correct, or 'true positive'. Specificity is the percentage of negative diagnoses that were correct, or 'true negative'. Trainees were slightly more sensitive using slides than Flicker (94% vs 88%), while experts were a little better using Flicker than slides (94% vs 87%). Overall, both groups of reviewers were somewhat more specific using Flicker than slides (85% vs 68% for trainees and 82% vs 77% for experts) (table 3).

When analysing combined Flicker and slide results, experts have a higher OR (1.25:1) of accurately determining progression compared with trainees, although this was not statistically significant (p=0.15). The Flicker method has a higher OR (1.78:1) of offering the correct diagnosis compared with the slide method after adjusting for reviewer effects, and this was statistically significant (p<0.001) (table 4).

For the 40 repeated image pairs, experts showed significantly higher OR (1.73:1) of achieving consistent diagnoses compared with trainees (p=0.029) (table 5a). Additionally, the slide method had a higher OR (1.20:1) of offering consistent reviews compared with the Flicker method overall, but this was not statistically significant (p=0.46). When comparing consistency between experts and trainees for each individual viewing method, experts were significantly more consistent for Flicker (p=0.005) but there was little difference for the slides (85.0% vs 84.2%) (table 5b).

Intrareviewer variability was 18% for Flicker and 15% for slides when combining the three trainee and three expert

Table 2a Comparison of accuracy between stereo disc viewing and the MatchedFlicker software programme

Image group	N	Average % accuracy while using slides	Average % accuracy while using Flicker	p Value*
Negative progression (50)	6	92.7	91.7	0.6267
Positive progression (50)	6	64.0	81.3	<0.0001
Total (100)	6	78.3	86.5	0.0003

*p Value was assessed with a two-sided paired t-test.
A comparison of the accuracy achieved by the six reviewers while using either the stereoscopic slide viewing technique or the MatchedFlicker software programme for determining progression of the 50 negative progression image pairs, the 50 positive progression image pairs and the total 100 unduplicated image pairs.

Table 2c Comparison of expert accuracy between stereo disc viewing and the MatchedFlicker software programme

Image group	Average % accuracy while using slides	Average % accuracy while using Flicker	p Value*
Negative progression (50)	88.7	95.3	0.039
Positive progression (50)	73.3	78.7	0.280
Total (100)	81.0	87.0	0.046

*p Value was found by fitting a logistic regression model.
A comparison of the accuracy achieved by the three expert reviewers in determining progression of the 50 negative progression image pairs, the 50 positive progression image pairs and the total 100 unduplicated image pairs while using either the stereoscopic slide viewing technique or the Flicker programme.

Table 3 Sensitivity and specificity of trainee and expert reviewers

Method	Reviewer group	n	Sensitivity (true positive)*	Specificity (true negative)†
Flicker	Trainee	3	126/144 (88%)	132/156 (85%)
	Expert	3	118/125 (94%)	143/175 (82%)
Slide	Trainee	3	82/87 (94%)	145/213 (68%)
	Expert	3	110/127 (87%)	133/173 (77%)

*Sensitivity is the percentage of positive diagnoses that were correct, or 'true positive'.
†Specificity the percentage of negative diagnoses that were correct, or 'true negative'.
An analysis of the three trainees and three experts' performance while using both the Flicker programme and the slide method in the three categories of sensitivity, specificity and accuracy. Sensitivity was determined by the percentage of positive diagnoses made that were correct, or 'true positive', specificity by the percentage of negative diagnoses made that were correct, or 'true negative'.

Table 5a Adjusted comparison of reviewing consistency between experts and trainees and between stereo disc viewing and the MatchedFlicker software programme

Parameter	OR	95% CI for OR	p Value*†
Reviewer effect†	1.731	(1.064 to 2.851)	0.029
Method effect‡	0.833	(0.512 to 1.351)	0.461

*p Value was found using a logistic regression model.
†Reviewer effect—an adjusted OR showing the reviewer's effect on consistency (expert:trainee), where an OR above one favours experts for accuracy.
‡Method effect—an adjusted OR showing the method's effect on consistency (flicker:slides), where an OR above one favours Flicker for accuracy.
A comparison of individual reviewer consistency achieved between the three experts and three trainees as well as between the stereoscopic slide viewing method and Flicker programme while evaluating the 40 duplicated image pairs after adjusting for individual reviewer effects.

reviewers for the 40 repeated slides. After combining the results of all six slide evaluators viewing 40 duplicated image pairs (240 duplicates in all among the six reviewers), there was a total of 37 cases where an incorrect diagnosis was made at both independent viewings of a duplicate, with these cases equally dispersed among the trainee and expert groups. When using Flicker, the same evaluators only had a combined 12 out of 240 of repeat incorrect designations, with nearly all of these being from the expert group (table 6).

To evaluate the inter-rater agreement, we calculated Cohen's kappa coefficient matrix for the stereophoto and MatchedFlicker viewing methods with the results summarised in table 7a and b, respectively. Table 7a indicates that only Expert 1 and Trainee 1, 2, 3; Expert 2 and Expert 3; Trainee 1 and Trainee 2, 3; Trainee 2 and Trainee 3 show relatively high agreement for the stereophoto viewing method. Examining table 7b, it reveals that Expert 1 and Expert 2, 3; Expert 2 and Expert 3; Trainee 1 and Trainee 3; Trainee 2 and Trainee 3 show relatively high agreement for the MatchedFlicker viewing method.

DISCUSSION

Detecting glaucomatous progression is a significant challenge facing clinicians today. Structural and functional parameters frequently do not show change at the same time.[5] Of patients in the OHTS who were determined to have reached an endpoint and progressed to glaucoma, 55% were sent to the endpoint committee based on disc change, 35% were sent based on data from the visual field reading centre and 10% showed change in both visual field and disc at a similar time.[5 7–9] In glaucoma, it is important to monitor both the anatomical and functional aspects of the eye regularly over time.

American Academy of Ophthalmology (AAO) Preferred Practice Pattern (PPP) guidelines recommend a range of follow-up optic nerve evaluation and visual field assessment from every month to a year in glaucoma patients (see the AAO PPP for POAG, table 6).[3] Optic disc examination is underused clinically. A 1996 study by Hertzog *et al* reported that 76.7% of patients did not have an optic nerve head drawing or photograph within 15 months of their most recent visit.[10] In a retrospective cohort study of 3623 diagnosed glaucoma patients and 1712 diagnosed suspects enrolled in a single insurance programme, Friedman *et al* noted that after a median of 440 days follow-up, only 13% of the suspects and 14% of those with diagnosed glaucoma had some kind of optic nerve head imaging, while 46% and 48%, respectively, had at least one billed visual field test.[11] Possible reasons for this paucity of optic nerve reviews and photography include the time-consuming nature of disc examination and the lack of stereoscopic photo camera equipment or disc digital image analysis systems for monitoring progression. Many clinicians rely on either old notes, hand drawings or non-stereoscopic photos of varying quality. According to the AAO, these techniques are less desirable alternatives.[12]

High-quality imaging is needed for accurate and early detection of glaucomatous progression. Defects in image illumination, centring and opacity may cause clinicians to miss the signs of progression that they otherwise would have noticed. The photos used in our study were relatively high-quality images from the OHTS disc reading centre and the University of Florida clinical archives. These photos did contain real-world abnormalities that challenge analysis including anatomical features such as tilted or ill-defined sloping disc rims commonly seen in clinical practice.

Table 4 Adjusted comparison of reviewing accuracy between experts and trainees and between stereo disc viewing and the MatchedFlicker software programme

Parameter	OR	95% CI for OR	p Value†
Reviewer effect†	1.248	(0.925 to 1.687)	0.148
Method effect‡	1.774	(1.311 to 2.412)	<0.001

A comparison of accuracy achieved between the three experts and three trainees as well as between the stereoscopic slide viewing method and Flicker programme while evaluating the 100 unduplicated image pairs after adjusting for individual reviewer effects.
*p Value was found using a logistic regression model.
†Reviewer effect—an adjusted OR showing the reviewer's effect on accuracy (expert:trainee), where an OR above one favours experts for accuracy.
‡Method effect—an adjusted OR showing the method's effect on accuracy (flicker:slides), where an OR above one favours Flicker for accuracy

Table 5b A comparison of consistency between experts and trainees

Method	Reviewer group	N	Consistency	p Value*
Flicker	Trainee	3	90/120 (75.0%)	0.005
	Expert	3	107/120 (89.2%)	
Slide	Trainee	3	101/120 (84.2%)	0.858
	Expert	3	102/120 (85.0%)	

*p Value was assessed with a t-sided t-test.
A comparison of the three trainees and three experts' consistency while using both the Flicker programme and slide method. Consistency was determined by the percentage of the 40 repeated image pairs (20 positive and 20 negative) that received identical diagnoses both times they were reviewed.

Table 6 Intrareviewer variability of the experts and trainees

Method	Reviewer group	N	Negative progression (20)	Positive progression (20)	Total consistency
Flicker	Trainee	3	47/60	43/60 (1)*	90/120 (1)*
	Expert	3	57/60 (1)*	50/60 (10)*	107/120 (11)*
Slide	Trainee	3	59/60	42/60 (18)*	101/120 (18)*
	Expert	3	50/60 (3)*	52/60 (16)*	102/120 (19)*

*Parentheses, the number of times there was agreement on the incorrect diagnosis. An analysis of evaluation variability of the individual reviewers for the 40 image pairs that were repeated randomly throughout the trials, for the Flicker programme and for the stereoscopic slide-viewing technique. The total agreeing responses for the three trainees and three experts are shown in the table.

In clinic, the images required to perform Flicker analysis are more easily integrated into EMR because a non-stereo image can be used, which does not require special camera equipment to produce nor stereoscopic viewing equipment to create three-dimensional depth for the analysis. Unlike stereophoto analysis, Flicker image examination is designed to be conducted through a standard computer system.

Radcliffe and coworkers compared stereodisc photographs with the alternating Flicker programme several years ago for detecting glaucoma progression and showed similar agreement between perimetric and disc progression using either technique.[13] Combining use of alternating Flicker and stereoscopic photography examination was felt to potentially enhance a clinician's ability to detect progressive glaucomatous nerve damage.[14] In a retrospective cohort study, an acceptable interobserver agreement using Flicker chronoscopy was later demonstrated between two glaucoma fellowship trained ophthalmologists, masked to temporal sequence.[15] Using the MatchedFlicker programme, Radcliffe and colleagues have further shown that retinal blood vessel positional shifts are significantly associated with neuroretinal rim loss and disc haemorrhages.[16]

In our previous study, we found that Flicker was more accurate, specific and time efficient than stereophoto examination for ophthalmologists in training.[6] Our current study also includes data from academic glaucoma experts from three centres to examine the efficacy of implementing Flicker into clinical practice where stereophoto disc review is already an integral part of patient care.

Due to increasing patient volume, time efficiency has become an increasingly important concern to physicians in all specialties. For both expert and trainee groups, Flicker was significantly faster to perform than the slide analysis. On average, evaluators were 8.6 s (26%) faster using Flicker than slides. There was no significant difference in mean evaluation time between expert and trainee groups using either technique (table 1).

Table 7a Kappa statistics between reviewers using the stereophoto viewing method

	Expert 1	Expert 2	Expert 3	Trainee 1	Trainee 2	Trainee 3
Expert 1		−0.005	0.061	0.480	0.563	0.324
Expert 2	−0.005		0.202	−0.113	−0.058	−0.136
Expert 3	0.061	0.202		0.066	0.061	−0.056
Trainee 1	0.480	−0.113	0.066		0.480	0.307
Trainee 2	0.563	−0.058	0.061	0.480		0.424
Trainee 3	0.324	−0.136	−0.056	0.307	0.424	

Table 7b Kappa statistics between reviewers using the Matchedflicker viewing method

	Expert 1	Expert 2	Expert 3	Trainee 1	Trainee 2	Trainee 3
Expert 1		0.586	0.276	−0.087	0.033	−0.006
Expert 2	0.586		0.361	−0.051	−0.076	−0.143
Expert 3	0.276	0.361		−0.041	0.020	−0.041
Trainee 1	−0.087	−0.051	−0.041		−0.006	0.469
Trainee 2	0.033	−0.076	0.020	−0.006		0.235
Trainee 3	−0.006	−0.143	−0.041	0.469	0.235	

Using the slide method, trainees showed a higher percentage accuracy for negative progression image pairs while experts showed higher accuracy for the positive progression (table 2b). Part of this difference may be due to the fact that evaluators using stereo slides assigned a disproportionally large number of the image pairs as negative for glaucomatous progression, especially among the trainees (71% of the trainee group designations vs 58% for the experts), while using Flicker the positively and negatively designated groups were more evenly balanced (45% vs 55% overall) (table 3). For Flicker analysis, there was no significant difference in accuracy between experts and trainees (table 2b). Of note, the experts will have had many more years of experience judging stereo slides for progression, especially compared with the PGY-2 and PGY-3 trainees, but with regard to using the Flicker programme the experts and trainees were equally naive.

Using a logistical regression model analysis, the OR of a correct response was significantly higher using Flicker than with slides ($p<0.001$). The effect of reviewer experience was not significant, although there was a trend for the experts to do better (table 4). When consistency of response is compared for the 40 double reviewed slides, the experts were significantly better than the trainees ($p=0.029$), while there was no difference in consistency between Flicker and slide (table 5a). When the improved consistency for experts compared with trainees was further analysed, it was the Flicker technique that was significantly better while there was no difference between groups for the slides in consistency (table 5b).

When individual reviewer variability using the 40 repeated slides was analysed, the experts did better with Flicker compared with trainees when examining the number of single incorrect responses (ie, the same slide called positive for progression at one review and negative for progression at the second review) but had a higher incidence of providing an incorrect designation at both instances that a repeated photo pair was analysed. For slide review, there was no difference between trainees and experts in either single or double erroneous assignments (table 6).

This study confirms the results in our previous publication demonstrating that the MatchedFlicker software is more accurate and time efficient than stereophoto optic disc evaluation. Additionally, this follow-up study has effectively tested the value of MatchedFlicker for experienced ophthalmologists by including data from academic glaucoma specialists. Our results indicate that both well-practiced and less experienced physicians show improvement in accuracy and speed using Flicker. The implementation of an optic disc evaluation technique that utilises easier to acquire mono-photos and which is both quicker to perform and possibly more accurate than the traditional method could make it more likely that ophthalmologists would track optic disc structure in glaucoma patients in conjunction

with visual fields and RNFL OCT regardless of their experience level. For future long-term clinical glaucoma studies, Flicker may augment the ability of an optic disc reading centre to determine progression. With the potential to be easily integrated into EMR, MatchedFlicker shows promise to enhance the ability of both glaucoma experts and non-experts in the detection of glaucomatous progression from optic nerve examination.

Acknowledgements We gratefully acknowledge Mae Gordon and the OHTS for providing the photo pairs in the positive control group.

Contributors All the authors have provided substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work, as described for each author below. JLS developed methodology to be used in experiment, managed execution of research project, provided the study materials and conducted experiments and performed data collection. AMM managed execution of research project and coordinated data analysis. CDR and NCR coordinated data analysis. AJL, ZLL, ABG, LJK and JS conducted experiments and performed data collection. GMM developed methodology to be used in experiment, provided the study materials and analyzed the data. BZ and JJS analyzed the data. MAK provided the study materials. MBS conceptualised the research idea, analyzed data, developed methodology to be used in experiment, managed execution of research project, provided the study materials, conducted experiments and performed data collection. All the authors have participated in drafting the work or revising it criticallyfor important intellectual content, as described for each author below. JLS, AMM,GMM, BZ, JJS and MBS were involved in writing the initial draft and preparing data tables and criticalreview and commentary for editing. CDR, NCR, AJL, ZLL, ABG, LJK, JS and MAK did the criticalreview and commentary for editing. All the authors have given final approval of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding This study was supported in part by an unrestricted grant from Research to Prevent Blindness (New York City, New York, USA), to the Department of Ophthalmology, University of Florida.

Competing interests None declared.

Patient consent A full waiver of informed consent was granted for the use of de-identified patient optic nerve photographs from patient records and from the Bascom Palmer Eye Institute optic disc reading center because no risk was posed to any patient with their use.

Ethics approval The University of Florida's Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Tham YC, Li X, Wong TY, *et al*. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology* 2014;121:2081–90.
2 Mackenzie PJ, Mikelberg FS. Evaluating optic nerve damage: pearls and pitfalls. *Open Ophthalmol J* 2009;3:54–8.
3 Prum BE, Rosenberg LF, Gedde SJ, *et al*. Primary open-angle glaucoma preferred practice pattern guidelines. *Ophthalmology* 2016;123:P41–P111.
4 Gordon MO, Kk S. The ocular hypertension treatment study. *Arch Ophthalmol* 1999;117:573.
5 Kass MA, *et al*. The ocular hypertension treatment study. *Arch Ophthal* 2002;120:701.
6 Schaefer JL, Lukowski ZL, Meyer AM, *et al*. Comparing glaucomatous disc change using stereo disc viewing and the MatchedFlicker software program in ophthalmologists-in-training. *Am J Ophthalmol* 2016;167:88–95.
7 Breusegem C, Fieuws S, Stalmans I, *et al*. Agreement and accuracy of non-expert ophthalmologists in assessing glaucomatous changes in serial stereo optic disc photographs. *Ophthalmology* 2011;118:742–6.
8 Reus NJ, Lemij HG, Garway-Heath DF, *et al*. Clinical assessment of stereoscopic optic disc photographs for glaucoma: the European Optic Disc Assessment Trial. *Ophthalmology* 2010;117:717–23.
9 Xin D, Greenstein VC, Ritch R, *et al*. A comparison of functional and structural measures for identifying progression of glaucoma. *Invest Ophthalmol Vis Sci* 2011;52:519–26.
10 Hertzog LH, Albrecht KG, LaBree L, *et al*. Glaucoma care and conformance with preferred practice patterns. examination of the private, community-based ophthalmologist. *Ophthalmology* 1996;103:1009–13.
11 Friedman DS, Nordstrom B, Mozaffari E, *et al*. Glaucoma management among individuals enrolled in a single comprehensive insurance plan. *Ophthalmology* 2005;112:1500–4.
12 Shaffer RN, Ridgway WL, Brown R, *et al*. The use of diagrams to record changes in glaucomatous disks. *Am J Ophthalmol* 1975;80(3 Pt 1):460–4.
13 Radcliffe NM, Sehi M, Wallace IB, *et al*. Comparison of stereo disc photographs and alternation flicker using a novel matching technology for detecting glaucoma progression. *Ophthalmic Surg Lasers Imaging* 2010;41:629–34.
14 Syed ZA, Radcliffe NM, De Moraes CG, *et al*. Detection of progressive glaucomatous optic neuropathy using automated alternation flicker with stereophotography. *Arch Ophthalmol* 2011;129:512–2.
15 Chee RI, Silva FQ, Ehrlich JR, *et al*. Agreement of flicker chronoscopy for structural glaucomatous progression detection and factors associated with progression. *Am J Ophthalmol* 2013;155:983–90.
16 Radcliffe NM, Smith SD, Syed ZA, *et al*. Retinal blood vessel positional shifts and glaucoma progression. *Ophthalmology* 2014;121:842–8.

Comparing glaucomatous disc change using stereo disc viewing and the MatchedFlicker programme in glaucoma experts and trainees

Jamie L Schaefer, Alissa M Meyer, Cooper D Rodgers, Nicole C Rosenberg, Anthony J Leoncavallo, Zachary L Lukowski, Anthony B Greer, Gina M Martorana, Baiming Zou, Jonathan J Shuster, L Jay Katz, Joel S Schuman, Michael A Kass and Mark B Sherwood

Br J Ophthalmol published online August 16, 2017

Notes