Published in final edited form as:

Am J Ophthalmol. 2016 July ; 167: 88–95. doi:10.1016/j.ajo.2016.03.031.

Comparing Glaucomatous Disc Change Using Stereo Disc Viewing and the MatchedFlicker® Software Program in Ophthalmologists in Training

Jamie L. Schaefer^{1,2}, Zachary L. Lukowski¹, Alissa M. Meyer¹, Anthony J. Leoncavallo¹, Anthony Greer¹, Gina M. Martorana¹, Baiming Zou¹, Jonathan J. Shuster¹, Ocular Hypertension Treatment Study Group³, and Mark B. Sherwood¹

¹University of Florida, 1600 SW Archer Road, Gainesville, FL 32610

²Ross Eye Institute, 1176 Main Street, Buffalo, NY 14209

³Washington University, 1 Brookings Drive, St. Louis, MO 63130

Abstract

Purpose—To compare the accuracy and speed of using the computerized MatchedFlicker® software program to evaluate glaucomatous optic disc change against the traditional gold-standard of manually examining stereoscopic disc photographs.

Design—A prospective evaluation of diagnostic technology.

Methods—Two resident ophthalmologists and one glaucoma fellow at the University of Florida independently evaluated 140 image-pairs from 100 glaucomatous/ocular hypertensive patient eyes using a handheld stereo-viewer and the MatchedFlicker® program. Fifty had progression to glaucoma as determined by the OHTS Optic Disc Reading Group and the OHTS end-point committee in the Ocular Hypertension Treatment Study (OHTS), and fifty more had photos taken a few minutes apart which were negative controls with no progression. Twenty photo-pairs from each group were duplicated to determine reviewer variability. Photos were examined in alternating blocks of 70 photo pairs for each method, with the starting viewing method randomized. Reviewer accuracy and time to review for each method were measured.

Results—Using the handheld stereo-viewer, the reviewers correctly identified progression or non-progression in 76.0% of the slide-pairs. Using the MatchedFlicker® software 87.6% were correctly identified ($p=0.011$). Evaluator speed averaged 34.1 seconds per image-pair with the stereo-viewer versus 24.9 seconds with the MatchedFlicker® program ($p=0.044$). Overall, Flicker was significantly more specific but less sensitive than stereo slides. Trainees appeared more

Corresponding Author: Dr. Mark B. Sherwood, 1600 SW Archer Rd. Suite M123D, Gainesville, FL 32610, Phone: (352) 283-0221, P.O. Box 100284, Fax: (352) 273-7402, sherwood@ufl.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Supplemental Material available at AJO.com

reluctant to identify glaucoma progression for slides than Flicker. For the two less experienced trainees Flicker was significantly more accurate.

Conclusion—The MatchedFlicker® software had a greater accuracy and was quicker to perform than using a handheld stereoscopic viewer.

Introduction

It is estimated that in the United States the incidence of those suffering from Open Angle Glaucoma (OAG) is over 2 million people.¹ Monitoring for glaucoma progression includes visual field and retinal nerve fiber analysis and careful examination of the optic nerve. All three methods have limitations, with quality of image scan or photograph being the primary problem with retinal nerve fiber layer (RNFL) optic coherence tomography (OCT) and disc evaluation respectively, and subjective factors related to patient alertness, their learning curve for the field testing, technician's experience, and small variations in head positioning also playing a role in visual field evaluation.

Accurate disc evaluation for focal rim defects, disc hemorrhages, optic nerve size, and nerve fiber layer abnormalities is important for identification of glaucomatous progression.² Photo documentation of the optic nerve head using stereo photos has been a traditional method for ophthalmologists to monitor glaucomatous progression because of its superiority to written subjective documentation and drawings.³ Large randomized National Eye Institute (NEI) trials have employed optic disc stereo photo analysis using a disc reading center as one of the primary methods to detect glaucoma progression.^{4–6}

Evidence suggests that structural damage can often be detected before the development of detectable visual field defects.⁷ In ocular hypertensive patients that were considered normal at baseline for automated visual field and disc morphology, Kass et al.⁸ in the Ocular Hypertension Treatment Study (OHTS) study (full study group listed in Appendix A: Supplemental Material at AJO.com) found that of 55% of eyes with primary open angle glaucoma (POAG) endpoints were initially identified on stereo photographs alone, 35% by visual field changes and 10% based on concurrent visual fields and stereo photograph changes.

It is unclear as to why disc evaluation through examination of sequential simultaneous stereo photographs is not done routinely in clinical practice. Possible explanations include the need for a special simultaneous-stereoscopic camera and stereo-viewer (although it is possible to obtain less-standardized stereoscopic photos with a non-simultaneous camera), and the difficulty and time-consuming nature of the stereo photo evaluation. With the increased use of electronic medical records (EMR), it is important to have a disc imaging technique which is both accurate at detecting glaucoma progression and is also well adapted to the new electronic format.

This study was designed to compare the accuracy and speed of the new MatchedFlicker® (EyeIC Inc., Narberth, PA), technique using a computer screen with the traditional gold-standard method of examining slides of stereoscopic disc photographs in a group of non-expert observers who were in various stages of ophthalmology training.

Methods

All research in this experimental study was performed in accordance with the University of Florida's Institutional Review Board (IRB) with protocol approval prior to initiation of the study. Approval was given for the collection of optic nerve photographs from patient records and from the Bascom Palmer Eye Institute optic disc reading center, as well as for the participation of the photo evaluators. A full waiver of informed consent was obtained for the use of de-identified patient optic nerve photographs, because no risk was posed to any patient with their use. This study adhered to the tenets of the Declaration of Helsinki and all federal and state laws.

From February to May of 2013 two ophthalmology residents (PGY-2 and PGY-3) and one glaucoma fellow (PGY-5) at the University of Florida independently evaluated stereoscopic pairs of disc photographs of 100 eyes taken at two time points. Fifty eyes were identified from patients in the OHTS study that showed glaucoma progression. This progression was previously determined by the OHTS Optic Disc Reading Group of the Bascom Palmer Eye Institute and confirmed as glaucomatous change by the OHTS end-point committee. The patient's comparison photos were taken between 1 and 5 years apart, depending upon when the disc progression was first noted by the study committee. Another group of 50 eyes were obtained that were known to show no progression since the photos were taken at the University of Florida using a Topcon TRC-50DX simultaneous stereo camera system (Topcon, Oakland, NJ) just a few minutes apart. Duplicate photographs of twenty of the fifty eyes that showed progression and 20 of the 50 eyes with no progression were randomly selected to allow for assessment of intra- and inter-observer variability in detecting progression. Thus, a total of 140 image pairs (280 photos) were examined by each observer.

Two different examination methods were used by each observer for judging the optic disc photos for glaucomatous progression; handheld stereo-viewers and the computerized MatchedFlicker® program (identified further in the study as "Flicker"). Using two Pentax handheld stereo viewers (Pentax, Tokyo, Japan) (Figure 1), 35 mm simultaneous-stereoscopic color slides of optic disc photographs were reviewed with the slide from the first timepoint in one stereoviewer and the slide from the second timepoint in the other stereoviewer. No poor quality photos were used, but photos were of naturally slightly varying quality to reflect what a practitioner would encounter in real-world clinical practice.

The evaluators reviewed the same sets of photographs using the Flicker software. For this computerized analysis, stereoscopic images are not required. The right image of the stereo pair from each of the 35mm stereo-photographs was digitized using a Nikon Super Coolscan 5000 (Tokyo, Japan) and uploaded into the Flicker alignment and presentation software. In four eyes the right image was blurred and the left image of the stereo pair was selected instead. The right image at the second time point was also digitized and uploaded into the Flicker software. The Flicker program took these two digitized images, registered that they were photos of the same eye taken at two time points, and aligned the object (the optic disc) in the two images so that they were superimposed (Figure 2). The program then "flickered" rapidly between the two images, which simulated a sensation of movement if any structural alterations had occurred. In this way, any features that underwent a potentially glaucomatous

structural change between the photos would be detected as movement. A 16 inch Samsung Galaxy laptop computer screen was used for the Flicker presentations.

Using the two different methodologies the evaluators assessed the optic nerve head (ONH) images for evidence of progression including disc rim thinning (focal or diffuse), vessel movement related to increased cupping, or detection of new or enlarged RNFL defects.

Prior to examining the study photographs, each evaluator was given a practice session with 10 additional photo pairs to familiarize them with each technique. For two of the examiners this practice with the Flicker technique was their only exposure to the program, as they had not used it previously. All three had some experience reviewing disc stereo-photographs with a handheld stereo-viewer.

The method of examining the disc pairs was divided into 2 alternate blocks of 70 paired images, using one method for the entirety of that block. The method that was used first was randomized. All 280 images were reviewed separately by each of the three observers in a single session. A forced decision on progression or non-progression was required for each image, even if the observer was not certain. A study coordinator recorded the evaluator's assessment of disc progression for each photo and recorded the time taken for the overall disc determination using each method. The time taken for each evaluator to complete the set of photographs using the two different techniques was divided by the number of photo pairs viewed to give a mean 'time-per-assessment' measurement.

Statistical Analysis

The difference in accuracy between the Flicker and slides method was initially analyzed with a 2-sided paired t-test comparing each reviewer's percent accuracy for both the Flicker and slides method. A more precise analysis in which the image pairs were scored for both Flicker and for slides and the paired differences in score compared is shown in Appendix B (Supplemental Material at AJO.com). Each subject slide or image was scored 100% (3/3 reviewers correct), 66.7% (2/3 correct), 33.3% (1/3 correct), or 0% (0/3 correct) and the p-value assessed by a 2-sided paired t-test for all 100 image pairs.

The difference in the average evaluation time per assessment between the two methods was also analyzed with a 2-sided paired t-test comparing the total time of each reviewer divided by all 140 slides in the block for each method.

Information on each reviewer's sensitivity, specificity, and overall accuracy was compared between the two methods using the following logistic regression model:

$$\text{logit} [\text{Pr} (y_{ijk}=1)] = \text{Log} [\text{odds} (y_{ijk}=1:y_{ijk}=0)] = \alpha_0 + \alpha_{trt} I(i=1) + \alpha_1 I(k=2) + \alpha_2 I(k=3)$$

Where y_{ijk} is the binary response (i.e. 1 for correct and 0 for incorrect) of subject j , reviewed by Reviewer k ($=1, 2$, or 3) using method i ($=1$ for Flicker and $=0$ for slides).

Reviewer 1 (PGY-5) is used as the reference. $I(x=\text{value})$ is the indicator function (i.e. $=1$ if $x=\text{value}$ and 0 otherwise). This produces for each of the groups (a, b, and c) an adjusted odds

ratio for Flicker to slide, with values above 1.0 favoring Flicker and values below 1.0 favoring slides.

Sensitivity was determined by the percentage of progression-positive diagnoses that were correctly attributed to truly glaucoma positive, i.e. “true positive” images. Similarly, specificity was determined by the percentage of progression-negative diagnoses that were correct (truly negative). The logistic regression model compares the probability of a reviewer’s response being correct between the two methods, Flicker and slides, while controlling for reviewer variability. The model was used to compare reviewer accuracy in three different sets of images: (a) the full image set of 100 unique image pairs, (b) the set of 50 true positives, and (c) the set of 50 true negatives.

McNemar’s chi-squared was used to assess discordant pairs in reviewer responses for the same images (inter-reviewer variability), and between the two responses given by reviewers for the same image using the two different viewing methods.

The logistic regression models were analyzed using the open source R statistical analysis package, while McNemar’s test was conducted in SAS, using PROC UNIVARIATE, which provides exact P-values. All tests were two-sided and statistical significance was qualified by whether or not $P < 0.05$.

Results

For all three reviewers, the number of correct responses using Flicker was significantly higher ($P = 0.011$) than that achieved while using the stereo-slide viewer (Table 1 and Appendix B).

The observed time with Flicker was 9.2 seconds (27%) faster per assessment on average than those using conventional stereo slides ($P = 0.04$) (Table 2).

Specificity is the percentage of negative diagnoses that are truly negative, and sensitivity is the percentage of those diagnosed with true glaucoma progression. As a whole, reviewers were more specific and more accurate with Flicker than with slides (Tables 3a and 3b). Reviewers were also less sensitive with Flicker than with slides. While using Flicker, the reviewers assigned both positive and negative responses at almost an equal rate, identifying 49–54 of the 100 image pairs as negative (Table 3a). However, reviewers identified 66–79 of the 100 image pairs they viewed as negative for glaucomatous progression while using the slides (Table 3b).

According to Table 4, reviewers were significantly more accurate overall ($P = 0.001$) and more likely to correctly diagnose the group of 50 image pairs depicting glaucomatous progression ($P < 0.001$) when using the Flicker method. However, the group of 50 image pairs depicting no glaucomatous progression were significantly more likely to be designated correctly by reviewers using the slides than those using Flicker ($P = 0.008$) (Table 4). This difference from Tables 3a and 3b is driven, as mentioned above, by the relative reluctance of the reviewers to call a positive progression with slide viewing, so that although they correctly called nearly all of the true negatives, there was also a group of true positives that

were designated as negative by the reviewers, thus decreasing the percentage of negatives that were actually correctly called.

Reviewers 2 and 3 (the PGY-2 and PGY-3, respectively) had significantly more instances where they could correctly decide whether there was glaucomatous progression between a set of image pairs while using Flickr, but were unable to decide correctly while using the slides for the same image pair ($P=0.043$ and $P=0.026$ for reviewer 1 and reviewer 2, respectively) (Appendix C: Supplemental Material at [AJO.com](http://ajophth.com)). While not significant, reviewer 1 (the PGY-5) also had more instances of correctly deciding whether or not an image pair showed progression or not when using Flickr.

When analyzing the discordant pairs between any two reviewers while using Flickr, there were no cases of one reviewer having significantly more correct decisions on glaucomatous progress in image pairs than his cohort (Table 5). However, it should be noted that when analyzing the cases of discordant responses while using slides, reviewer 1 and reviewer 2 (PGY-5 and PGY-2) differed significantly in the number of correct decisions on glaucomatous progress, with reviewer 1 making a significantly greater amount of correct decisions ($P=0.012$).

Intra-observer agreement was generally high between the two techniques when analyzing the forty duplicates, with a mean of 30 and 33.67 of the 40 repeated images (75% and 84%) being given the same designation with the use of Flickr and Slides respectively (Table 6). While using the slides, there were 18 cases out of 42 agreements where the wrong decision on the presence of glaucomatous progression was made twice. In all 18 cases, the reviewers determined that there was no progression when there actually was. While using Flickr, there was only 1 case of 43 agreements where the wrong determination of glaucomatous progression was chosen twice. The error made was the same as with the slides: the reviewer saw no progression between the two image pairs, when progression had actually taken place.

Discussion

A significant problem in the management of glaucoma is the difficulty in detecting progression so that adequate therapy can be instituted. Xin *et al*⁹ have noted the importance of following both structural and functional parameters in combination when monitoring for progression, because frequently these do not change at the same time. Disc changes commonly precede visual field loss in early glaucoma.^{8, 10, 11}

Examination of the optic disc is an acquired skill and is under-performed. The recommended AAO PPP guidelines for follow up of glaucoma and evaluation of optic nerve and visual field assessment are from 3 months to 2 years at a maximum.¹² Hertzog and co-workers¹³ found that only 23.3% of patients had an optic nerve head drawing or photograph within 15 months of the most recent visit. In addition, the authors found that 37.8% of patient charts had neither an optic nerve head drawing nor a photograph documented after the initial visit. Friedman *et al*¹⁴ noted in a retrospective cohort study of 1712 diagnosed suspects and 3623 diagnosed glaucoma patients enrolled in a single comprehensive insurance program that

after a median follow up of 440 days, while 46% had a billed visual field testing only 13% had some sort of optic nerve head imaging.

It can be a time-consuming process to accurately and thoroughly examine each disc every year. In addition, many clinicians do not have stereoscopic disc photos or computerized image analysis systems for examining discs for progression and have to rely on either old notes describing the nerve or hand-drawings or non-stereoscopic photographs of varying quality. The AAO preferred practice pattern suggests that these are less desirable alternatives.¹⁵

With 243 ophthalmologist participants, Reus et al¹⁰ found the accuracy for diagnosing glaucomatous discs by reviewing stereoscopic optic disc photographs using stereo-viewers at best 80.5%. A different study by Breusegem et al¹¹ looked at the inter-observer agreement and accuracy of glaucoma experts and non-expert ophthalmologists. They found that the inter-observer agreement of non-expert ophthalmologists was significantly lower than that of the experts.

As with any imaging modality correctly interpreting the image is dependent on the quality; whether it's well centered, well focused, well illuminated, without motion artifact, and without medial opacities. Poor quality images reduce the ability to detect subtle changes and therefore subtle glaucomatous changes may be overlooked.¹⁶ Our study included good quality photographs from the OHTS Disc Reading Center and from the University of Florida clinical archives, while possessing real-life characteristics such as ill-defined sloping of some disc rims or disc-tilting that would be regularly encountered by physicians in a clinical setting.

To the best of our knowledge, this is the first study to compare two different methods for accuracy of detecting glaucomatous disc progression over time using photographs with a known status in regard to progression. The negative controls in which simultaneous stereo-photos were taken within a few minutes of each other can reliably be stated to show no progression. It is more difficult to be certain that subtle disc progression has definitely occurred, but by utilizing the results of a well-recognized NEI study disc reading center, our study aimed to get as close as is possible to being certain that there is real progression. Indeed, this system of having a disc reading center review stereoscopic disc photographs is the current gold standard for detecting disc progression for NIH sponsored glaucoma clinical trials. We considered using slides of non-progressing discs from the OHTS study disc-reading center as our negative control in view of the concern regarding increased lighting and other photographic differences when multiple clinics take photographs over longer time periods. However, we felt that there was some risk that discs that were graded in OHTS as not showing progression by the disc reading center might in fact have a minute number of subtle signs of progression, thus introducing a type II error (false negative). Although there are pros and cons of each negative control, on balance, we felt that using slides taken a few minutes apart and that were certain to have no progression would be better.

A recent study by Radcliffe *et al*¹⁶ viewed stereo photo image sets from patients with glaucoma using an automated alternation flicker (the Flicker software program) and found

retinal blood vessel shifts in 26.4% of the longitudinally followed glaucomatous eyes. In addition, they found a statistically significant association between blood vessel movement and increased rate of visual field progression and neuroretinal rim loss.

Our study compares the new Flicker software program viewed on a standard computer screen with the traditional technique of examining a 35 mm slide in a hand-held stereo viewer, for detecting glaucomatous optic disc structural progression. Our 3 observers included 2 resident ophthalmologists at different stages of their training and a glaucoma fellow at the University of Florida. The primary end-point was the accurate identification of disc progression over time. A higher percentage of correct answers was found using the Flicker computerized program for all three reviewers ($P=0.01$) (Table 1, Appendix B). There were no statistically significant differences in accuracy between the trainees with different levels of experience using Flicker, but when using slides the resident with the least experience examining stereo photographs was significantly less accurate than the glaucoma fellow (Table 5). This may reflect the fact that all 3 observers had a similar, limited experience using the Flicker program with just the same 10 training photo pairs.

As mentioned previously in the statistical analysis section, specificity is the percentage of negative designations that are truly negative, and sensitivity is the proportion of positive diagnoses that are truly positive (e.g. have true glaucomatous progression). In our opinion, both are important, but perhaps having high specificity is of particular importance so that patients who are progressing are detected and can receive treatment. According to tables 3a and 3b, the Flicker method produced more specific results (negative designations that are truly negative) than slides. The slides method produced more sensitive results (positive designations that are truly positive) than Flicker, due in part to the low number of positive designations that were made. The low number of image pairs that were identified as showing glaucomatous progression (Table 3b) suggests that reviewers were less inclined to call the glaucomatous progression between a set of stereoscopic image pairs while using the slides method. For all 3 trainee observers, Flicker was more accurate for the true positive image pairs ($P=0.001$), less accurate for the true negative image pairs ($P=0.008$), and, regarding the full set of 100 unique image pairs, more accurate overall than stereoscopic slide review ($P=0.001$) (Table 4). The test-retest reproducibility of detecting optic disc deterioration from stereophotographs by the certified graders at the Optic Disc Reading Center (ODRC) in the Ocular Hypertension Treatment Study was reported to show a specificity of 98–100% and a sensitivity ranging from 64–81%.¹⁷ Our three trainee reviewers had a similar specificity of 95–100% and a sensitivity of 65–75% for the 40 duplicate slides they examined (Table 6).

With the demand for the clinician to see higher volumes of patients, the time taken to perform an accurate analysis is important. All three observers using Flicker were faster than using 35mm slides ($P=0.04$) (Table 2). Integration with the EMR will also be increasingly important over time. Scanning of digitized stereo images into the EMR and stereo viewing of these images usually requires specialized viewing devices and a specialized camera to obtain a simultaneous stereo image. The Flicker software has the advantage of using monoscopic images and does not require the use of stereo photography.

The current study only compared physicians at different stages of training. A future study comparing Flicker versus viewing of stereo-photos among glaucoma experts (sub-specialists with multiple years of experience using stereo images) to determine the accuracy of detecting glaucoma progression would be useful to fully explore the role of Flicker analysis in clinical practice.

Our study with ophthalmologists-in-training demonstrated that the use of MatchedFlicker® software is comparable or even better in accuracy than the traditional methods of stereo-image optic disc viewing and was faster to perform. By using a technique for evaluating optic discs that is easier, faster, and potentially more accurate than the current gold standard, and in particular does not require stereo imaging and is suitable for EMR use, it is felt that ophthalmology residents would be more likely to regularly track the optic disc structure of their glaucoma patients, as well as visual fields and RNFL OCT. It has the potential to assimilate more easily into a clinical practice setting with EMR, and in conjunction with other functional and structural tests would be a good compliment to the current methods of glaucoma progression detection.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Disclosure:

- A. This study was supported in part by an unrestricted grant from Research to Prevent Blindness (New York, NY), to the Department of Ophthalmology, University of Florida and NIH grant 1UL1TR000064 from the National Center for Advancing Translational Sciences (Bethesda, MD).
- B. No financial disclosures.
- C. There are no additional acknowledgments.

References

1. Quigley HA, Broman AT. The Number of People with Glaucoma Worldwide in 2010 and 2020. *Br J Ophthalmol*. 2006; 90(3):262–267. [PubMed: 16488940]
2. Mackenzie PJ, Mikelberg FS. Evaluating Optic Nerve Damage: Pearls and Pitfalls. *The Open Ophthalmol Journal*. 2009; 3:54–58.
3. Jamara RJ, Denial A, Valentini D, Thorn F. Clinical Quality Assessment Using Computer Monitor Photoimages of Optic Nerve Head Cupping. *Optom Vis Sci*. 2000; 77(8):433–436. [PubMed: 10966070]
4. Gordon MO, Beiser JA, Brandt JD, et al. The Ocular Hypertension Treatment Study: Baseline Factors that Predict the Onset of Primary Open-Angle Glaucoma. *Arch Ophthalmol*. 2002; 120(6): 714–720. [PubMed: 12049575]
5. Leske MC, Heijl A, Hyman L, Bengtsson. Early Manifest Glaucoma Trial: Design and Baseline Data. *Ophthalmol*. 1999; 106(11):2144–2153.
6. Musch DC, Lichter PR, Guire KE, Standardi CL. The Collaborative Initial Glaucoma Treatment Study: Study Design, Methods, and Baseline Characteristics of Enrolled Patients. *Ophthalmol*. 1999; 106(4):653–662.
7. Xu G, Weinreb RN, Leung CK. Optic Nerve Head Deformation in Glaucoma: The Temporal Relationship between Optic Nerve Head Surface Depression and Retinal Nerve Fiber Layer Thinning. *Ophthalmology*. 2014; 121(12):2362–2370. [PubMed: 25108319]

8. Kass MA, Heuer DK, Higginbotham EJ, et al. The Ocular Hypertension Treatment Study: A Randomized Trial Determines that Topical Ocular Hypotensive Medication Delays or Prevents the Onset of Primary Open-Angle Glaucoma. *Arch Ophthalmol*. 2002; 120(6):701–713. [PubMed: 12049574]
9. Xin D, Greenstein VC, Ritch R, Liebmann JM, DeMoraes CG, Hood DC. A Comparison of Functional and Structural Measures for Identifying Progression of Glaucoma. *Invest Ophthalmol Vis Sci*. 2011; 52(1):519–526. [PubMed: 20847115]
10. 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(4):717–723. [PubMed: 20045571]
11. Breusegem C, Fieuws S, Stalmans I, Zeyen T. Agreement and Accuracy of Non-expert Ophthalmologists in Assessing Glaucomatous Changes in Serial Stereo Optic Disc Photographs. *Ophthalmology*. 2011; 118(4):742–746. [PubMed: 21055815]
12. American Academy of Ophthalmology Glaucoma Panel. Primary Open-Angle Glaucoma. San Francisco, CA: American Academy of Ophthalmology; 2010. Preferred Practice Pattern® Guidelines. Available at: www.aao.org/ppp. Accessed December 10, 2015
13. Hertzog LH, Albrecht KG, LaBree L, Lee PP. Glaucoma Care and Conformance with Preferred Practice Patterns. Examination of the Private, Community-Based Ophthalmologist. *Ophthalmology*. 1996; 103(7):1009–1013. [PubMed: 8684788]
14. Friedman DS, Nordstrom B, Mozaffari E, Quigley HA. Glaucoma Management Among Individuals Enrolled in a Single Comprehensive Insurance Plan. *Ophthalmology*. 2005; 112(9): 1500–1504.
15. Shaffer RN, Ridgway WL, Brown R, Kramer SG. The Use of Diagrams to Record Changes in Glaucomatous Disks. *Am J Ophthalmol*. 1975; 80(3Pt 1):460–464. [PubMed: 1163592]
16. Radcliffe NM, Smith SD, Syed ZA, et al. Retinal Blood Vessel Positional Shifts and Glaucoma Progression. *Ophthalmology*. 2014; 121(4):842–848. [PubMed: 24342023]
17. Parrish RK II, Schiffman JC, Anderson DR, et al. Test-retest Reproducibility of Optic Disk Deterioration Detected from Stereophotographs by Masked Graders. *Am J Ophthalmol*. 2005; 140(4):762–764. [PubMed: 16226544]

Figure 1.
Pentax Stereo Viewer

Figure 2.
The MatchedFlicker ® Software

Table 1
Comparison of individual Reviewer Percent Accuracy when comparing Glaucomatous Disc Change using Stereo-Disc viewing and the MatchedFlicker Software Program

A comparison of the accuracy of each individual reviewer to determine progression of the 100 slides (50 with progression and 50 without) while using either the stereoscopic slide-viewing technique or the Flicker program. The p-value was assessed with a 2-sided paired t-test comparing each reviewer's percent accuracy for each viewing method. A more precise analysis was also done comparing the individual slides, the method of which can be viewed in Appendix B.

Reviewer	Slides	Flicker	P-value
Reviewer 1	90.00%	80.71%	
Reviewer 2	84.29%	70.71%	
Reviewer 3	88.57%	76.43%	
Average	75.95%	87.62%	0.011

Table 2
Comparison of Individual Reviewer Evaluation Time per Assessment (Seconds) when comparing Glaucomatous Disc Change using Stereo-Disc viewing and the MatchedFlicker Software Program

A comparison of each individual reviewer's time, in seconds, spent determining progression on all 140 slides presented (50 with progression and 50 without, along with 20 repeated slides from each group) while using either the stereoscopic slide-viewing technique of the Flicker program. The difference, in seconds, between the time spent using each method is also presented. The p-value was assessed with a 2-sided paired t-test comparing the 3 individual reviewer times for each method.

Reviewer	Stereo-viewer	Flicker	Difference	P-value
Reviewer 1	33.60	20.58	13.02	
Reviewer 2	29.38	22.84	6.54	
Reviewer 3	39.15	31.29	7.85	
Average	34.1 (4.9) ^a	24.9 (5.6) ^a	9.2 (3.4) ^a	0.04

^a Parentheses () = Standard Deviation

Table 3a
Analysis of Sensitivity, Specificity, and Accuracy of Individual Reviewers using the MatchedFlicker Software Program to compare Glaucomatous Disc Change

An analysis of each reviewer's performance while using the Flicker program 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", and accuracy by the percentage of all designations that were correct out of the 100 slides.

Reviewer	Sensitivity (True Positive) ^a	Specificity (True Negative) ^b	Accuracy of all Diagnoses
Reviewer 1	42/47 (89%)	45/53 (85%)	87/100 (87%)
Reviewer 2	40/46 (87%)	44/54 (81%)	84/100 (84%)
Reviewer 3	44/51 (86%)	43/49 (88%)	87/100 (87%)

^aSensitivity = the percentage of positive diagnoses that were correct, or "true positive"

^bSpecificity = the percentage of negative diagnoses that were correct, or "true negative"

Table 3b
Analysis of Sensitivity, Specificity, and Accuracy of Individual Reviewers using Stereo-Disc viewing to compare Glaucomatous Disc Change

An analysis of each reviewer's performance while using the stereoscopic slide viewer in the three categories of sensitivity, specificity, and accuracy. Sensitivity was determined by the percentage of positive diagnoses that were correct, or "true positive", specificity by the percentage of negative diagnoses that were correct, or "true negative", and accuracy by the percentage of all designations that were correct out of the 100 possible slides.

Reviewer	Sensitivity (True Positive) ^a	Specificity (True Negative) ^b	Accuracy of all Diagnoses
Reviewer 1	32/32 (100%)	50/68 (74%)	82/100 (82%)
Reviewer 2	21/21 (100%)	50/79 (63%)	71/100 (71%)
Reviewer 3	29/34 (85%)	45/66 (68%)	74/100 (74%)

^aSensitivity = the percentage of positive diagnoses that were correct, or "true positive"

^bSpecificity = the percentage of negative diagnoses that were correct, or "true negative"

Table 4
Comparison of Accuracy of all Reviewers using Stereo-Disc viewing and the MatchedFlicker Software Program to compare Glaucomatous Disc Change in the True Positive, True Negative, and Full Image Sets, using a Logistic Regression Model

A comparison of the likelihood that a slide from a certain group will be designated correctly while using either the Flicker program or the stereoscopic slide-viewing method. Three groups of slides were analyzed: all 100 unique slides, the 50 True Positive slides, and the 50 True negative slides.

Subject Set	Odds Ratio ^a	95% Confidence for OR ^b	P-value
(a) All	1.983	(1.308, 3.038)	0.001
(b) True Positive ^b	4.482	(2.621, 7.886)	< 0.001
(c) True Negative ^c	0.248	(0.080, 0.646)	0.008

^a An adjusted odds ratio above 1.0 favors Flicker for accuracy, while a ratio below 1.0 favors the slides

^b OR: Odds Ratio

^c True Positive = the slides or image pairs whose correct diagnosis is positive

^d True Negative = the slides or image pairs whose correct diagnosis is negative

Table 5
Comparison of Inter-Reviewer Variability by analyzing the Discordant Pairs between Reviewers while using Stereo-Disc viewing and the MatchedFlicker Software Program to compare Glaucomatous Disc Change

A comparison of each reviewer's performance against that of his peers for both the stereoscopic slide-viewing technique and the Flicker program, by determining who had the correct designation when two reviewers had disagreeing designations for a slide. P-values were assessed by McNemar's chi-squared 2-sided test.

Reviewers Compared	Ratio ^a (Flicker)	P-Value (Flicker)	Ratio ^a (Slide)	P-Value (Slide)
1 : 2	14:11	0.55	15:4	0.012
1 : 3	6:6	1.00	16:8	0.10
2 : 3	8:11	0.49	10:13	0.53

^aRatio = Discordant pairs correct with first reviewer : Discordant pairs correct with second reviewer

Table 6
Comparison of Intra-Reviewer Variability for each Individual Reviewer while using Stereo-Disc viewing and the MatchedFlicker Software Program to compare Glaucomatous Disc Change

An analysis of each individual reviewer's designation variability for the 40 slides that were repeated randomly throughout the trials, for the Flicker program and for the stereoscopic slide-viewing technique.

Method	Reviewer	Repeated Negative controls (20)	Repeated Positive Controls (20)	Total
Flicker	Reviewer 1	17/20	15/20	32/40
	Reviewer 2	15/20	12/20	27/40
	Reviewer 3	15/20	16/20 (1) ^a	31/40 (1) ^a
Slide	Reviewer 1	19/20	15/20 (5) ^a	34/40 (5) ^a
	Reviewer 2	20/20	13/20 (8) ^a	33/40 (8) ^a
	Reviewer 3	20/20	14/20 (5) ^a	34/40 (5) ^a

^a Parentheses = the number of times there was agreement on the incorrect diagnosis. It occurred 18 times for the slides and only once for Flicker.