An Objective Evaluation of Gaze Tracking in Humphrey Perimetry and the Relation With the Reproducibility of Visual Fields: A Pilot Study in Glaucoma

Yukako Ishiyama, Hiroshi Murata, Chihiro Mayama, and Ryo Asaoka

Department of Ophthalmology, The University of Tokyo, Tokyo, Japan

Correspondence: Ryo Asaoka, Department of Ophthalmology, The University of Tokyo, Graduate School of Medicine, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8655 Japan; rasaoka-tky@umin.ac.jp.

Submitted: August 25, 2014
Accepted: October 22, 2014

Citation: Ishiyama Y, Murata H, Mayama C, Asaoka R. An objective evaluation of gaze tracking in Humphrey perimetry and the relation with the reproducibility of visual fields: a pilot study in glaucoma. *Invest Ophthalmol Vis Sci*. 2014;55:8149–8152. DOI:10.1167/iovs.14-15541

Purpose. To develop a novel method to evaluate gaze tracking (GT) results and to examine their relationship with test–retest reproducibility of visual field (VF) measurements.

Methods. Subjects comprised of 42 eyes of 42 glaucoma patients. Vision fixation during VF tests with the Humphrey Field Analyzer was evaluated using the gaze fixation line chart at the bottom of the VF printout. We defined some GT parameters as follows: average tracking failure frequency per stimulus (TFF), average frequency of eye movements between 1° and 2°, 3° and 5°, and more than 6°. Humphrey VFs (24-2 and 10-2 Swedish Interactive Threshold Algorithm [SITA] standard) were prospectively examined twice within a period of 3 months in 42 glaucoma patients. Mean absolute variability of total deviation (TD) values in the test–retest VFs was measured and its relationship to fixation losses (FLs), false positives (FPs), false negatives (FNs), mean deviation (MD), and pattern standard deviation (PSD) was investigated using the corrected Akaike Information Criterion (AICc) from linear modeling.

Results. The best model to predict test–retest variability in the 24-2 VF included PSD, TFF, and FNs as dependent variables, while the best model for the 10-2 VF included PSD and average frequency of eye movements between 3° and 5° ($P < 0.05$ for all coefficients).

Conclusions. Gaze tracking parameters are closely related to the reproducibility of VF results, and it would be beneficial to objectively use these parameters when estimating the reliability of VF tests.

Keywords: gaze tracking, glaucoma, visual field

The Humphrey Field Analyzer (HFA; Carl Zeiss Meditec, Dublin, CA, USA) is commonly used to monitor visual field (VF) damage in glaucoma patients around the world. The instrument includes several methods to estimate the reliability of VF tests. The frequency of false positive (FP) answers is estimated by maximum likelihood estimation based on the number of positive answers that occur during a "listening time," which starts shortly after the end of the response window and ends 180 ms after the onset of the next stimulus.[1] A false negative (FN) mainly occurs when a patient fails to respond to a much more intense stimulus than he/she had responded to previously, although FNs are also associated with VF deterioration.[2] A high rate of FP answers is thought to indicate "trigger-happy" patients and a high rate of FN responses is thought to represent inattention during an examination.[3–5] Fixation loss (FL) is recorded when a stimulus projected onto the area of the eye's blind spot is perceived and it indicates the test reliability and vision fixation. High rate of FPs can lead to a high FL rate. Fixation loss can also result from mislocalization of the blind spot.[6] Others have reported that fixational instability can be found even in well trained observers[7,8] and elevated FL can mask the presence of early scotoma.[7,9] While some past studies have reported on the usefulness of these indices,[10,11] more recent studies have pointed out their limitations.[2,12]

Gaze tracking (GT) is a method to monitor eye movements[13]; it measures the status of fixation during the VF test. Gaze tracking also records a count of the number of times an eye's position cannot be determined, usually because of ptosis obscuring the pupil, a presentation during a blink, or an irregular pupil that prevented accurate recording.[13] It has been reported that GT is useful for evaluating the quality of fixation, particularly when VF defects surround the blind spot.[14] However, its usage in clinical practice has been somewhat limited, since results are merely represented as a printed line

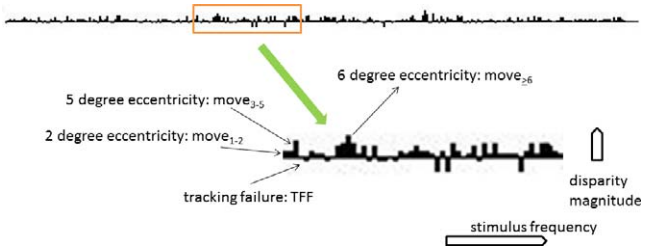

Figure. An example of a GT figure with the GT parameters. An *upward bar* in the chart indicates fixation disparity and the length of the bar represents the magnitude of disparity, from 1° to a maximum of 10°. A *short downward bar* represents tracking failure, while a *long downward bar* indicates eyelid closure. Gaze tracking parameters were calculated as follows: average TFF per stimulus, the average frequency of eye movement per stimulus between 1° and 2° (denoted $move_{1\text{-}2}$), 3° and 5° (denoted $move_{3\text{-}5}$), and more than 6° (denoted $move_{\geq 6}$).

Table 1. Parameters Used in Model Selection

Analyzed Parameters
VF parameters
MD, dB
PSD, dB
Traditional reliability parameters
FL, %
FP, %
FN, %
Age
Refractive error spherical equivalent, D
GT parameters
Average frequency of eye movement per stimulus between 1° and 2° ($move_{1\text{-}2}$)
Average frequency of eye movement per stimulus between 3° and 5° ($move_{3\text{-}5}$)
Average frequency of eye movement per stimulus of 6° or more degrees ($move_{\geq 6}$)
Average tracking failure frequency per stimulus (TFF)

diagram at the bottom of the VF printout; this forces clinicians to evaluate the results subjectively. In the current study, GT results were evaluated objectively and quantitatively to investigate their relationship with the test–retest reproducibility of VFs.

Methods

The study was approved by the research ethics committee of the Graduate School of Medicine and Faculty of Medicine at The University of Tokyo (Tokyo, Japan). Written consent was given by patients for their information to be stored in the hospital database and used for research. This study was performed according to the tenets of the Declaration of Helsinki.

Subjects

Forty-two eyes of 42 open-angle glaucoma patients (20 males and 22 females) were included in the study. All patients were prospectively recruited at the glaucoma clinic in The University of Tokyo Hospital. Only one eye in a patient was included in the current study and if both eyes satisfied the inclusion criteria, one eye was chosen at random. All the VFs were measured using the HFA (24-2 and 10-2 Swedish Interactive Threshold Algorithm [SITA] standard program). Each patient performed 24-2 and 10-2 VF tests twice within 3 months. One patient's 24-2 VF was excluded because GT could not be recorded, however 10-2 VF of this patient was included in the analysis.

All patients enrolled in the study fulfilled the following criteria: (1) glaucoma was the only disease causing VF damage, (2) patients were followed for at least 6 months at The University of Tokyo Hospital and have experienced at least two VF measurements prior to this study, and (3) all patients had glaucomatous VF defects in at least one eye defined as three or more contiguous total deviation points at P less than 0.05, or two or more contiguous points at P less than 0.01, or a 10 dB difference across the nasal horizontal midline at two or more adjacent points, or MD worse than −5 dB.[4] All of the visual acuities of the eyes examined were equal to or better than 6/6.

Gaze Tracking Measurements

The GT system monitors patients' gaze position at each stimulus presentation (Fig.).[15] An upward bar in the chart indicates fixation disparity and the length of the bar represents the magnitude of disparity, from 1° to a maximum of 10°. A short downward bar represents tracking failure, while a long downward bar indicates eyelid closure or tear film breakup due to dry eye during the test.

Gaze tracking data were exported as JPEG images from the Beeline (Tokyo, Japan) data filing system. Then summary GT parameters were calculated as follows by simply calculating the frequency of the upward and downward bars with each length in the GT records: average tracking failure frequency per stimulus (TFF), the average frequency of eye movement per stimulus between 1° and 2° (denoted $move_{1\text{-}2}$), 3° and 5° (denoted $move_{3\text{-}5}$), and equal to or more than 6° (denoted $move_{\geq 6}$). The three levels of $move_{1\text{-}2}$, $move_{3\text{-}5}$, and $move_{\geq 6}$ were chosen following a previous paper.[14]

Statistical Analysis

The mean absolute variability of the 52 or 68 total deviation values (24-2 and 10-2 VFs, respectively) in the two test–retest VFs were calculated (i.e., the mean absolute variability was calculated as average test–retest variance of total deviation values between the two tests). The relationship between variability and the GT summary parameters, FP, FN, and FL reliability indices as well as MD, PSD, refractive error, and age were analyzed using linear models (Table 1). The best linear model was then selected among all possible combinations of predictors based on the second order bias corrected Akaike Information Criterion (AICc) index. The AIC is a well-known statistical measure used in model selection, and the AICc is a corrected version of the AIC, which provides an accurate estimation even when the sample size is small.[16] All predictors included in the model were calculated based on their mean values in the two VF tests.

Table 2. Patient Demographics

	24-2, $N = 41$	10-2, $N = 42$
Age, mean ± SD (range)	62.5 ± 11.6 (31–80)	61.6 ± 11.6 (31–80)
Sex (male:female)	22:19	22:20
First VF		
MD, dB, mean ± SD (range)	−10.7 ± 7.5 (−27.9–2.0)	−12.2 ± 7.6 (−27.0–1.7)
PSD, dB, mean ± SD (range)	11.2 ± 4.1 (1.9–16.8)	10.8 ± 4.5 (1.2–15.8)
Test duration, s, mean ± SD (range)	481 ± 111 (265–765)	470 ± 104 (298–712)
Second VF		
MD, dB, mean ± SD (range)	−10.9 ± 7.7 (−28.6–2.2)	−11.8 ± 7.9 (−27.5–1.1)
PSD, dB, mean ± SD (range)	11.2 ± 4.0 (2.1–16.7)	10.6 ± 4.6 (1.32–15.97)
Test duration, s mean ± SD (range)	420 ± 63.5 (306–557)	425 ± 74.2 (296–582)

Table 3. Frequency of Eye Movement Between 1° and 2°, 3° and 5°, and More Than 6°

	24-2	10-2
$Move_{1-2}$, mean± SD (range, per stimulus)	0.62 ± 0.18 (0.07–0.93)	0.64 ± 0.15 (0.21–0.92)
$Move_{3-5}$, mean± SD (range, per stimulus)	0.12 ± 0.08 (0.004–0.41)	0.11 ± 0.08 (0.004–0.33)
$Move_{\geq 6}$, mean± SD (range, per stimulus)	0.10 ± 0.19 (0–0.75)	0.079 ± 0.12 (0–0.58)
TFF, mean± SD (range, per stimulus)	0.07 ± 0.13 (0.001–0.61)	0.10 ± 0.15 (0.0003–0.73)

All analyses were performed using the statistical programming language 'R' (R version 2.15.1; The Foundation for Statistical Computing, Vienna, Austria).

Results

Characteristics of the study subjects are summarized in Table 2. The mean (±SD) age of the patients was 62.5 ± 11.6 years for 24-2 VF testing and 61.6 ± 11.6 years for 10-2 VF testing, ranging from 31 to 80 years in both groups (one patient's 24-2 VF was excluded because GT was accidentally not recorded). The MD values of the initial VFs were −10.8 ± 7.5 (mean ± SD, [range, −28.2–2.1]) dB in the 24-2 VFs and −12.0 ± 7.7 (27.2–1.4) dB in the 10-2 VFs. Pattern SD values were initially 11.2 ± 4.1 (1.9–16.8) dB in the 24-2 VFs and 10.7 ± 4.6 (1.3–15.9) dB in the 10-2 VFs.

As shown in Table 3, the average eccentricity of eye movements throughout the VF measurement was 2.3 ± 1.8° (0.92–9.6; mean ± SD [range]) for 24-2 VFs and 2.1 ± 1.4° (0.88–7.4) for 10-2 VFs. In the 24-2 VF tests, $move_{1-2}$, $move_{3-5}$, $move_{\geq 6}$, and TFF results were 0.62 ± 0.18 (0.07–0.93) per stimulus, 0.12 ± 0.08 (0.004–0.41) per stimulus, 0.10 ± 0.19 (0–0.75) per stimulus, and 0.07 ± 0.13 (0.001–0.61) per stimulus, respectively, while those of the 10-2 VFs were 0.64 ± 0.15 (0.21–0.92) per stimulus, 0.11 ± 0.08 (0.004–0.33) per stimulus, 0.079 ± 0.12 (0–0.58) and 0.10 ± 0.15 (0.003–0.73) per stimulus, respectively.

The mean frequencies of FL, FP, and FN are shown in Table 4. Both in the 24-2 and 10-2 VFs, a significant relationship was not observed between the GT parameters and FL; the correlation coefficients between $move_{1-2}$ and FL was 0.14 ($P = 0.38$), $move_{3-5}$ and FL was 0.0014 ($P = 0.99$), $move_{\geq 6}$ and FL was −0.12 ($P = 0.46$) and TFF and FL was −0.19 ($P = 0.25$) in the 24-2 VF. Also, the correlation coefficients between $move_{1-2}$ and FL was 0.0055 ($P = 0.97$), $move_{3-5}$ and FL was 0.020 ($P = 0.90$), $move_{\geq 6}$ and FL was 0.019 ($P = 0.90$) and TFF and FL was −0.068 ($P = 0.67$) in the 10-2 VF.

Pattern SD, TFF, and FN were selected as significant predictors of variability in the best model for 24-2 VFs, while PSD and $move_{3-5}$ were selected in the best model for 10-2 VFs (Table 5). The coefficients of the best-selected models were as follows: mean absolute variability = 0.11 × PSD + 3.6 × TFT + 12.6 × FN (24-2 VF model) and mean absolute variability = 0.087 × PSD + 3.5 × $move_{3-5}$ (10-2 VF model); P less than 0.05 for all coefficients.

Discussion

In the current study, a novel method was developed to evaluate GT results quantitatively and objectively in VF tests. The relationship between GT results and test–retest reproducibility of 24-2 and 10-2 VFs was then investigated. The TFF GT index appeared to be particularly important for VF reproducibility in 24-2 and 10-2 VFs. In contrast, FNs were a significant predictor of VF variability in 24-2 VFs, but not in 10-2 VFs. Age was not related to VF reproducibility, in agreement with a previous report.[17]

In the current study, PSD, which measures the amount of unevenness of the VF, was selected as an important predictor of VF variability in both 24-2 and 10-2 VF models, while MD was not selected. This is probably because the majority of patients included in the study were in a mild to moderate stage of the disease. Early glaucomatous damage is often reflected more sensitively using the PSD index rather than the MD statistic; this is because early focal VF change can be masked by the averaging carried out in the calculation of MD.[18] Previous studies have reported that VF reproducibility is relatively good in early glaucoma and worsens with the progression of the disease, then becomes good again when glaucoma reaches an advanced stage because of the 'floor effect' of VF sensitivity.[19] On the other hand, PSD is high in moderate glaucoma and lower in early and advanced stages, which may explain why PSD was selected in the best models, instead of MD.

In the best model for variability in 24-2 VFs, only TFF was selected among all GT parameters; this may be because, in 24-2 VFs, test points are located in 6° intervals, and hence the influence of eye movements within 6° may have a limited effect. Indeed, $move_{3-5}$ was selected as a significant predictor in the best model for 10-2 VFs while $move_{1-2}$ was not selected, which is noteworthy when we consider that test points are located in 2° intervals in the 10-2 VF test pattern. Nonetheless, $move_{\geq 6}$ was not selected in the best model for either 24-2 or 10-2 VFs. This may be because the eye tracking system is unable to track eye movements at this resolution and so TFF is a more useful predictor. Furthermore, as shown in Table 3, patients' eye movements may not often exceed 6°, in which case the influence of this predictor in the model will be diminished.

None of the traditional reliability indices were selected as significant predictors of variability in the best models for either 24-2 or 10-2 VFs, except for FNs in the 24-2 VF model. It has been reported that FNs increase with the progression of glaucoma, which itself is associated with lower reproducibility.[2] On the other hand, Bengtsson and colleagues[20] investigated the relationship between reproducibility and FLs, FPs and FNs, and found that only FNs were significantly associated with reproducibility. The results in the current study suggest that FNs are related to the reproducibility of VF sensitivity in addition to the disease status, as represented by PSD. It is worth noting that FPs are calculated differently in the SITA algorithm than they are in the Full-Threshold test in which classic catch trials are employed. In the SITA algorithm any response prior to the minimum response time (~180 ms), adjusted according to the patient's individual mean response time, is considered a FP error.[1] This may suggest that all actual FP responses after the minimum response time are ignored in the FP calculation. On the contrary, GT parameters reflect the status of eye position directly during the actual threshold measurements. In addition, there is a previous report which

Table 4. Rate of FL, FP, and FN

	24-2	10-2
FL, %, mean ± SD (range)	6.4 ± 5.9 (0–26.3)	4.2 ± 5.6 (0–23.3)
FP, %, mean ± SD (range)	3.8 ± 3.9 (0–21.5)	1.2 ± 1.3 (0–6)
FN, %, mean ± SD (range)	3.4 ± 2.9 (0–11.5)	3.2 ± 4.4 (0–19.5)

Table 5. Selected Parameters for the Best Model in 24-2 and 10-2 VF Tests

24-2			10-2		
Selected Parameters	**Coefficient**	***P* Value**	**Selected Parameters**	**Coefficient**	***P* Value**
PSD	0.11	0.007	PSD	0.087	0.005
TFF	3.6	0.005	$move_{3\text{-}5}$	3.5	0.034
FN	12.6	0.023			

suggested the FPs with the SITA algorithm are underestimated compared with those in the Full-Threshold test, which uses the classic catch trials.[21] Also, as shown in Table 4, the mean rates of FL, FP, FN were low compared with GT indices. This may have contributed to the small effect of traditional indices and the selection of GT indices in the best models.

Our results do not deny the perception that traditional indices are important factors when investigating the reliability of VF measurements. A possible interpretation is that FN and FP are good indices of accurate VF measurements through the prediction of over- or underestimation of VFs but not so much through the prediction of test–retest reproducibility. Gaze tracking parameters could have been more useful for the prediction of reproducibility.

One of the possible caveats of the current investigation is the limited range of glaucomatous disease observed in the study. Most patients were in an early to moderate stage of glaucoma, and so an assessment of the usefulness of the GT parameters should also be carried out in patients with advanced disease in a future study. One of the difficulties in performing this analysis is that a degradation of VFs is often accompanied by a deterioration of visual acuity, which can cause poor VF reproducibility, in addition to eye movements during the VF measurement. Nonetheless, reproducibility of VFs is equally important in this population, and hence further investigation is required. Furthermore, GT results should be investigated in a larger population, including healthy controls and in patients with other ocular disorders; thus, this research should be considered a pilot study.

In the current study, GT data were exported as JPEG images from the Beeline data filing system and various GT parameters were simply calculated by reading the JPEG image. Thus, GT parameters could be obtained on a personal computer; clinicians would then be able to estimate the reliability of patient's VF at a clinical setting.

In conclusion, we have developed a method to quantitatively investigate the GT record on HFA VF tests. Moreover, the GT parameters derived in this study are significant predictors of reproducibility in both 24-2 and 10-2 VF tests.

Acknowledgments

Supported in part by Japan Science and Technology Agency (JST) CREST (RA and HM) Grants, 25861618 (HM), 50570701 (CM), and 26462679 (RA) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (Tokyo, Japan).

Disclosure: **Y. Ishiyama**, None; **H. Murata**, None; **C. Mayama**, None; **R. Asaoka**, None

References

1. Newkirk MR, Gardiner SK, Demirel S, et al. Assessment of false positives with the Humphrey Field Analyzer II perimeter with the SITA algorithm. *Invest Ophthalmol Vis Sci*. 2006;47:4632–4637.
2. Bengtsson B, Heijl A. False-negative responses in glaucoma perimetry: indicators of patient performance or test reliability? *Invest Ophthalmol Vis Sci*. 2000;41:2201–2204.
3. Fankhauser F, Spahr J, Bebie H. Some aspects of the automation of perimetry. *Surv Ophthalmol*. 1977;22:131–141.
4. Anderson DR, Patella VM. *Automated Static Perimetry.* 2nd ed. St. Louis: Mosby; 1999.
5. Johnson CA, Sherman K, Doyle C, et al. A comparison of false-negative responses for full threshold and SITA standard perimetry in glaucoma patients and normal observers. *J Glaucoma*. 2014;23:288–292.
6. Sanabria O, Feuer WJ, Anderson DR. Pseudo-loss of fixation in automated perimetry. *Ophthalmology*. 1991;98:76–78.
7. Demirel S, Vingrys AJ. Eye movements during perimetry and the effect that fixational instability has on perimetric outcomes. *J Glaucoma*. 1994;3:28–35.
8. Demirel S, Vingrys AJ. Fixational instability during perimetry and the blindspot monitor. In: Mills RP, ed. *Perimetry Update, 1992/1993*. Amsterdam: Kugler Publications; 1992:515–520.
9. Demirel S, Vingrys AJ. The effect of fixational loss on perimetric thresholds and reliability. In: Mills RP, ed. *Perimetry Update, 1992/1993*. Amsterdam: Kugler Publications; 1992: 521–526.
10. McMillan TA, Stewart WC, Hunt HH. Association of reliability with reproducibility of the glaucomatous visual field. *Acta Ophthalmol (Copenh)*. 1992;70:665–670.
11. Katz J, Sommer A. Screening for glaucomatous visual field loss. The effect of patient reliability. *Ophthalmology*. 1990;97: 1032–1037.
12. Henson DB, Evans J, Chauhan BC, et al. Influence of fixation accuracy on threshold variability in patients with open angle glaucoma. *Invest Ophthalmol Vis Sci*. 1996;37:444–450.
13. *Humphrey Field Analyzer Series 700 Service Guide*. Dublin, CA: Carl Zeiss Meditec; 1994.
14. Kunimatsu S, Suzuki Y, Shirato S, et al. Usefulness of gaze tracking during perimetry in glaucomatous eyes [in Chinese]. *Nihon Ganka Gakkai Zasshi*. 1999;103:748–753.
15. *Humphrey Field Analyzer II-i Series, User Manual*. Dublin, CA: Carl Zeiss Meditec; 2010.
16. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*. 2004;33:261–304.
17. Blumenthal EZ, Sample PA, Berry CC, et al. Evaluating several sources of variability for standard and SWAP visual fields in glaucoma patients, suspects, and normals. *Ophthalmology*. 2003;110:1895–1902.
18. Reddy GR. *A Visual Field Evaluation With Automated Devices*. New Delhi: Jaypee Brothers; 2006.
19. Artes PH, Iwase A, Ohno Y, et al. Properties of perimetric threshold estimates from Full Threshold, SITA standard, and SITA fast strategies. *Invest Ophthalmol Vis Sci*. 2002;43:2654–2659.
20. Bengtsson B. Reliability of computerized perimetric threshold tests as assessed by reliability indices and threshold reproducibility in patients with suspect and manifest glaucoma. *Acta Ophthalmol Scand*. 2000;78:519–522.
21. Wall M, Doyle CK, Brito CF, et al. A comparison of catch trial methods used in standard automated perimetry in glaucoma patients. *J Glaucoma*. 2008;17:626–630.