Total Deviation Probability Plots for Stimulus Size V Perimetry

A Comparison With Size III Stimuli

Michael Wall, MD; Caridad F. Brito, PhD; Kimberly R. Woodward, MS; Carrie K. Doyle, MS; Randy H. Kardon, MD, PhD; Chris A. Johnson, PhD

Objective: To compare empirical probability plots in patients with glaucoma for size V and III perimetry testing.

Methods: We computed empirical probability plot percentile limits after testing 60 age-matched controls tested with both size III (Swedish interactive thresholding algorithm) and size V (full threshold) perimetry twice. Probability plots of 120 patients with glaucoma tested in the same way were computed. We compared the number of abnormal test locations in the 2 stimulus sizes; we then compared these results with those from size III StatPac software (Zeiss Humphrey Systems, Dublin, California) using 2-way repeated-measures analysis of variance.

Results: We found a similar number of abnormal test locations (P ≤ .05) for the size III and size V testing conditions identified by the probability plots (no significant difference); there were significantly fewer abnormal locations using StatPac (size III) than from our size III database. When results were stratified by mean deviation, the mild visual loss group again did not show any significant differences between sizes III and V.

Conclusions: Size V full-threshold testing gives a similar number of abnormal test locations in patients with glaucoma compared with the size III Swedish interactive thresholding algorithm standard test. Size V testing, with its greater dynamic range and lower variability, may be a viable alternative to size III testing in patients with glaucoma.

Arch Ophthalmol. 2008;126(4):473-479

Automated static perimetry, since its introduction in 1979, has used the 0.43°, size III, stimulus. This was chosen as a compromise between the effects of blur and reduced dynamic range associated with the small (0.11°), size I, stimulus and the apparent loss of resolution of detecting small scotomata with the larger (1.72°), size V, stimulus. Since that time, it has been shown that detection of defects from glaucoma and other optic neuropathies can be done at least as well with stimuli larger than 0.43°. For example, the 10°×10° stimulus of frequency-doubling technology perimetry is as sensitive as size III stimuli in detecting glaucoma; similar findings have been reported for nonglaucomatous optic neuropathies.

The size III stimulus used in initial studies in healthy controls showed low variability. However, it was later reported that with this stimulus size, variability was high in areas of visual field damage. Heijl and coworkers investigated this retest variability in 51 eyes of 51 individuals with glaucoma experienced with perimetry representing all stages of damage. The patients were tested 4 times in a 4-week period using size III full-threshold testing with the Humphrey field analyzer (Zeiss Humphrey Systems, Dublin, California). Test locations with 8 to 18 dB of loss initially had a 95% prediction interval that nearly covered the full measurement range of the instrument (0-40 dB). An important finding of Heijl and coworkers, which has been observed by others, is that pointwise intertest variability increases dramatically with decreasing sensitivity of the test location. A major ramification of this finding is that areas with the most visual loss have the highest variability. Therefore, the most clinically important regions with highest variability are precisely those in which determination of change is most difficult. Of the various strategies available to lower variability, use of larger stimuli has shown promise.

In a study comparing variability of size III and V, we used frequency of seeing curves in patients with glaucoma that were generated by using a custom test program to evaluate these patients at 2-dB intervals across a 2- to 3-log unit range. The

Author Affiliations: Departments of Neurology (Dr Wall) and Ophthalmology (Drs Wall, Kardon, and Johnson, and Mss Woodward and Doyle), University of Iowa, College of Medicine, Veterans Administration Hospital, Iowa City; Department of Psychology, Eastern Illinois University, Charleston (Dr Brito); and Discoveries in Sight, Devers Eye Institute, Portland, Oregon (Dr Johnson).
patients were tested with size III and size V stimuli in areas of normal sensitivity and areas of 10 to 20 dB loss. The same test locations were used in the same sitting for both sizes. As shown by a steepening of the slope of the frequency of seeing curves, variability substantially decreased when the size V stimulus was used. These findings of lower variability have led some to use size V stimuli in standard automated perimetry to observe patients with moderate to severe visual loss. However, studies by Zulauf and Caprioli, Wilensky et al, and Gramer et al suggest a loss of resolution using size V compared with size III stimuli.

A problem with these investigations has been the lack of a standard with which to compare size III and V testing. The gray-scale printout used is methodologically identical for both tests, yet sensitivities for size V (with a 31.5-apostilb background) are about 4 dB higher in healthy controls. Test locations are identified on the size V Humphrey field analyzer printout if they are more than 5 dB from the expected normal value and no statistical package is available for size V. Lastly, retest variability is lower with size V stimuli; thus, 5 dB is probably too conservative of a cutoff. Therefore, use of the currently available Humphrey field analyzer display of the results makes comparisons of stimulus sizes problematic.

Because larger perimetric stimuli can have greater dynamic range and lower variability, it is important to know their sensitivity to detect defects. To test the hypothesis that defect detection is no different for size III and size V in patients with glaucoma, we developed databases from the same set of healthy controls and generated empirical probability plots for both size III and V using the same methodology. Our aim was to compare our patients with glaucoma with the database from the common set of controls and with the StatPac (Zeiss Humphrey Systems, Dublin, California) size III (Swedish interactive thresholding algorithm [SITA]) database of the Humphrey field analyzer.

METHODS

PARTICIPANTS

The visual testing protocol was approved by the University of Iowa’s institutional review board. The tenets of the Declaration of Helsinki were followed. Sixty healthy controls and 120 patients with glaucoma were tested at baseline and again at a separate appointment within 1 to 8 weeks. They all gave informed consent to participate in the study. The controls were volunteers, paid in accordance with the institutional review board, who answered advertisements inviting them to participate in research. The patients with glaucoma were invited from the glaucoma clinic at the University of Iowa Department of Ophthalmology and Visual Sciences. The mean (SD) age of controls was 57.24 (7.85) years, with a range of 41 to 78 years. Thirty-eight of the volunteers were women and 22 were men. The mean (SD) age of the patients with glaucoma was 64.95 (9.48) years, with a range of 38 to 81 years; their mean deviation from normal (SD) was −6.67 (4.4) dB loss.

Controls were included if they had (1) no history of eye disease except refractive error (no more optical correction than 5 diopters of sphere or 5 diopters of cylinder), (2) no history of diabetes mellitus or systemic arterial hypertension, and (3) normal ophthalmologic examination results, including 20/25 or better corrected Snellen visual acuity. The participants either had undergone a complete eye examination within 12 months before this study or were examined by an ophthalmologist on the day of testing.

Patients from the University of Iowa Hospitals and Clinics’ Glaucoma Clinic were offered admission if they met study entry criteria. They were enrolled if they had glaucomatous optic disc changes with abnormal standard automated perimetry results (glaucomatous visual field defects, ie, ≥ 3 adjacent abnormal test locations in a clinically suspicious area at the $P < .05$ level or 2 adjacent abnormal locations with $\geq 1$ at the $P < .01$ level; and a mean deviation in the range of 0 to −19.94 dB on standard automated perimetry). Because the visual field defects are similar for all types of glaucoma and the rates of progression are similar, we included patients with primary, secondary, or normal-tension glaucoma. The patients did not have another disease affecting vision and were capable of reliably performing standard automated perimetry and returning for follow-up visits. Patients were excluded if they had cataracts causing visual acuity worse than 20/30, their pupils were smaller than 2.5 mm, they were younger than 19 years, or they were pregnant. The first 120 consecutive patients who agreed to enter the study constituted the glaucoma cohort.

VISUAL TESTING

All participants first underwent testing with size III stimuli using the standard 24-2 SITA, followed by testing with size V stimuli. The size III stimulus has an area of 4 mm$^2$ or is 0.43° in diameter and the size V has an area of 16 mm$^2$ or is 1.72° in diameter. Since there is no SITA strategy available for size V stimuli, the participants were tested with the Humphrey 24-2 full-threshold algorithm. We followed the manufacturer’s recommendations and used a corrective lens when necessary. Care was taken to prevent lens rim artifact. The healthy controls had testing performed in 1 eye chosen at random, which was used for all tests. In the participants with glaucoma, if both of their eyes qualified for the study, 1 eye was chosen at random. All visual field examinations met the following reliability criteria: fixation losses less than 20% or normal gaze tracking, a false-positive rate of less than 10%, and a false-negative rate of less than 33%. We define normal gaze tracking as the presence of only an occasional upward deflection representing an eye movement.

PROBABILITY MAPS

We computed empirical probability plots from data collected from testing 60 controls (aged, 40-78 years) on size III and size V testing. Since the data set had participants both experienced with and naive to perimetry, we used the results of both the first and second visual field examinations of the controls (120 tests) and converted the data to a spreadsheet format using PeriData (Huerth, Germany). There was a modest learning effect from test 1 to test 2. With size III, mean (SD) sensitivity improved from 29.98 (0.97) dB to 30.02 (1.21) dB. With size V, sensitivity improved from 33.85 (0.94) dB to 34.12 (1.12) dB.

We then found the effect of age by regressing mean score on age. We found a loss of 0.051 dB per year for size III and 0.057 dB per year for size V. We then adjusted all threshold values to age 45 years. The 120 values were then ranked from highest to lowest in a spreadsheet for each test location. The 95th, 98th, and 99th percentiles were then empirically determined. We realize that by combining the results of the first and second tests, there are correlations within participants. However, because our population was 60 people, we used the 2 examination results to allow noninterpolated calculation of the percentile values.
To evaluate our method, we calculated the number of normal and abnormal test locations in 22 controls at their first and second visits who were not in our original data set. We compared the number of abnormal test locations, counting the number of test locations with loss at less than 1%, less than 2%, and less than 5% levels, with the number of test locations that we expected to be abnormal. For example, at a $P = .05$ level, we expected 1 in 20 test locations to be abnormal.

We then computed total deviation probability plots for each of the 120 participants with glaucoma representing all stages of glaucoma from $-0.12$ to $-19.94$ dB of mean deviation. We analyzed the probability plots of the 120 patients with glaucoma by counting the number of test locations with loss at less than 1%, less than 2%, and less than 5% levels and compared the number of normal and abnormal test locations using size III vs size V and size III results with StatPac software results. Since failure to detect a difference could be caused by patients with moderate or severe visual field loss having many damaged test locations, we stratified the patients into 3 equal groups by mean deviation. Group 1 had a mean deviation of $-3.9$ to $-0.12$; group 2 had a mean deviation of $-7.25$ to $-3.91$; and group 3 had a mean deviation of $-19.94$ to $-7.26$ dB. We then recounted the abnormal test locations by mean deviation group.

### Table 1. Perimetry Results of 22 Healthy Controls Tested Twice

<table>
<thead>
<tr>
<th>Measure</th>
<th>Size III Stimulus</th>
<th>Size V Stimulus</th>
<th>Expected Test Locations</th>
<th>Difference of Size III $a$</th>
<th>Difference of Size V $a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of normal test locations</td>
<td>1008</td>
<td>1059</td>
<td>1029.6</td>
<td>-21.6</td>
<td>29.4</td>
</tr>
<tr>
<td>No. of abnormal test locations</td>
<td>85</td>
<td>55</td>
<td>34.32</td>
<td>50.68</td>
<td>20.68</td>
</tr>
<tr>
<td>95th percentile</td>
<td>22</td>
<td>14</td>
<td>11.44</td>
<td>10.56</td>
<td>2.56</td>
</tr>
<tr>
<td>99th percentile</td>
<td>29</td>
<td>16</td>
<td>11.44</td>
<td>17.56</td>
<td>4.56</td>
</tr>
<tr>
<td>No. of normal test locations</td>
<td>1070</td>
<td>1069</td>
<td>1029.6</td>
<td>40.4</td>
<td>39.4</td>
</tr>
<tr>
<td>No. of abnormal test locations</td>
<td>43</td>
<td>51</td>
<td>34.32</td>
<td>8.68</td>
<td>16.68</td>
</tr>
<tr>
<td>95th percentile</td>
<td>12</td>
<td>12</td>
<td>11.44</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>99th percentile</td>
<td>19</td>
<td>12</td>
<td>11.44</td>
<td>7.56</td>
<td>0.56</td>
</tr>
</tbody>
</table>

$^a$ Difference in number of locations identified as being abnormal at each probability level compared with that expected based on calculating the random probability of the events occurring.

$^b$ There are fewer abnormal test locations on visit 2, likely because of perimetric learning effect. The results show the numbers of abnormal test locations are close to expected (we expect 1 in 20 test locations to be abnormal by chance alone [$P = .05$]).

### Table 2. Number of Test Locations of Patients With Glaucoma

<table>
<thead>
<tr>
<th>Measure</th>
<th>Size III Stimulus</th>
<th>Size V Stimulus</th>
<th>StatPac Size III Stimulus $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of normal test locations</td>
<td>2417</td>
<td>2397</td>
<td>3021</td>
</tr>
<tr>
<td>No. of abnormal test locations</td>
<td>640</td>
<td>689</td>
<td>647</td>
</tr>
<tr>
<td>95th percentile</td>
<td>334</td>
<td>344</td>
<td>427</td>
</tr>
<tr>
<td>99th percentile</td>
<td>2849</td>
<td>2810</td>
<td>2145</td>
</tr>
<tr>
<td>No. of normal test locations</td>
<td>2468</td>
<td>2437</td>
<td>3008</td>
</tr>
<tr>
<td>No. of abnormal test locations</td>
<td>618</td>
<td>663</td>
<td>608</td>
</tr>
<tr>
<td>95th percentile</td>
<td>286</td>
<td>342</td>
<td>456</td>
</tr>
<tr>
<td>99th percentile</td>
<td>2868</td>
<td>2798</td>
<td>2168</td>
</tr>
</tbody>
</table>

$^a$ Zeiss Humphrey Systems, Dublin, California.

$^b$ Note how the number of normal (and hence total number of abnormal) test locations are nearly the same for sizes III and V when the Iowa databases are used to generate the total deviation probability plots.

#### STATISTICAL ANALYSIS

Three-way split-plot analysis of variance was used to compare the number of normal and abnormal test locations among the 3 probability levels, groups, and visits. Wilcoxon signed rank tests were performed on the medians. Differences between groups of all test results were interpreted as significant if the probability of their occurrence was $P < .05$.

#### RESULTS

We tested 22 healthy controls twice to check and validate our Iowa probability plot software. We found close to the expected number of abnormal test locations (Table 1) on both visit 1 and visit 2.

In the patients with glaucoma, we found, using the Iowa database of size III and V controls, that there were no significant differences in number of abnormal test locations between these 2 stimulus sizes ($P < .01$) (Table 2 and Figure 1). In addition, the median number of normal test locations (for group 1 with mild
loss at visit 1) was 30 for size III and 27 for size V ($P = .11$); the medians were almost the same on the second visit: $-30.5$ and $30.0$ for size III and V, respectively ($P = .43$).

The StatPac results from the Humphrey field analyzer database showed about 6 fewer abnormalities per size III test, confirming a difference in our databases (median number of normal test locations was 34.5 on visit 1 and 37.5 on visit 2). The 3-way mixed analysis of variance evaluated the effects of mean deviation group (3 groups), visit (2 visits), and test (III, V, and StatPac) for normal and 5%, 2%, and 1% levels of loss separately. When we compared the results from the Iowa databases with those from StatPac software, the model indicated no significant differences in the number of test locations counted owing to the visits at any of the levels of loss; nor did visit interact with any of the other factors in the model to affect the number of test locations counted. Overall counts were significantly different among the 3 mean deviation patient groups at the normal, 5%, and 1% levels of loss ($P < .001$); however, the 3 groups did not differ significantly at the 2% level of loss ($P = .08$) (Figure 1). For the normal ($P = .01$) and 1% loss ($P = .03$) test locations, there was a significant interaction between the test and group factors in the model. Follow-up simple effects tests and Bonferroni pairwise comparisons showed that for the normal loss group, StatPac evaluation for size III had significantly fewer abnormal test locations than III and V from the Iowa databases ($P < .001$ for both); however, the results from the Iowa databases of size III and V did not differ significantly from each other.

Figure 1. Number of normal (normal probability) and abnormal (5%, 2%, and 1% probabilities) test locations per visual field of 52 non-blind spot locations examined in 120 patients with glaucoma. Note the similarity of size III and size V results when the same healthy eyes from the Iowa database are used for the probability plots. Results from StatPac (Zeiss Humphrey Systems, Dublin, California) for size III show fewer abnormalities owing to broader confidence limits.

Figure 2. Example of typical results from a patient with glaucoma using the Swedish interactive thresholding algorithm (SITA) size III test and StatPac (Zeiss Humphrey Systems, Dublin, California) analysis showing manufacturer’s output. FL indicates fixation losses; FN, false-negative catch trials; GHT, Glaucoma Hemifield Test; MD, mean deviation; PSD, pattern standard deviation.
Figures 2, 3, and 4 show typical examples of the standard Humphrey field analyzer printout and our custom printout that uses the percentile values from the common set of Iowa controls. Both the number of abnormalities and the shape of the visual field defects are similar. This similarity of defect shape by probability plots was nearly always present when comparing size III and V.

Our results show that there are no significant differences in the total number of abnormal test locations when comparing SITA size III and full-threshold size V stimuli. Since various reports suggest slightly more abnormal test locations identified with SITA compared with full-threshold methods, it is unlikely that the thresholding method is the reason that the results of these 2 sizes are so similar.

The lack of any difference between size III and V results could be because of the large number of patients with moderate to severe visual loss. We therefore split the sample into thirds by mean deviation. When we compared the sizes in the group with a mean deviation of −0.12 to −3.90 dB, we found, on average, 1.33 more abnormal test locations per visual field examination (average of visits 1 and 2) with size V (Figure 5). We believe this slightly better performance of size V in identifying more abnormal locations is not clinically significant.

Bengtsson and Heijl compared short-wavelength automated perimetry, SITA short-wavelength automated perimetry, and SITA Fast results in 101 patients with glaucoma. They also defined reference limits for abnormality using a common set of 53 healthy participants and derived normal limits for total- and pattern-deviation probability maps. They compared the median number of significantly depressed test locations per eye and test program, analyzed paired comparisons of the number of significantly depressed test locations, and compared the number of eyes with clusters of significantly depressed points. Although, based on previous reports, one would expect more abnormal test locations in short-wavelength automated perimetry testing, Bengtsson and Heijl concluded that no significant difference in number of abnormalities could be detected among the 3 testing methods. Their study was the first to compare short-wavelength automated perimetry and standard automated perimetry with a common set of healthy participants for the purpose of comparing results of pointwise abnormalities.

While there are many studies comparing perimetry types, investigators typically use the manufacturer's software output rather than a common set of healthy participants to define the reference standards of normality. Our study and
the one by Bengtsson and Heijl\(^1\) suggest that any study claiming higher sensitivity of defect detection with a specific perimetry type should be viewed with caution unless a common database of controls is used. Moreover, it raises the question of whether all threshold perimetry types, whether they be differential light-sensitivity perimetry, motion perimetry, frequency-doubling perimetry, short-wavelength automated perimetry, flicker perimetry, color perimetry, or other types, are similar in their rates of defect detection. It will take studies of each perimetry type with a common comparison type, such as standard automated perimetry using a common set of healthy controls, to answer this question. We found on average 5 more abnormal test locations per study using our databases compared with the StatPac evaluation. For the Iowa normative database, we accepted all visual field examinations from participants who were free of eye disease except for refractive error.

**Figure 4.** Example of typical results from a patient with glaucoma using the Swedish interactive thresholding algorithm (SITA) size III and full-threshold size V tests using Iowa database analysis. Note how similar the probability maps appear for size III and V. FL indicates fixation losses; FN, false-negative catch trials; FP, false-positive catch trials; MD, mean deviation.
sults from participants with excessive fixation losses were included if gaze tracking was normal. We required less than a 10% false-positive rate. Otherwise, our entry criteria were similar to those for inclusion into StatPac. One difference is that we included the first and second study from each healthy control for our analysis and we analyzed both the first and second visual field. However, this was done for both size III and size V, so it would not affect this comparison of stimulus sizes, only the comparison with the StatPac results. However, use of this second test and that all perimetry was done at 1 center by 2 highly trained perimetrists may explain why our database may have less interparticipant variability and identified more abnormal test locations than the commercially available program.

Size V full-threshold testing has some advantages over size III SITA testing. The dynamic range is greater and the variability is lower. Our results show that there is little difference in detection of defects with these 2 methods. However, a longitudinal study would be required to know how size V full-threshold testing compares with size III SITA for investigation of visual field progression.

When a common set of healthy participants is used to develop a normative database, there is little difference in the number of probability plot abnormalities identified between size III and size V stimuli in patients with glaucoma. Because size V has a larger dynamic range and lower variability, it may be a preferable stimulus for standard automated perimetry.

Submitted for Publication: July 25, 2007; final revision received October 9, 2007; accepted October 19, 2007.

Correspondence: Michael Wall, MD, Department of Neurology, University of Iowa, College of Medicine, 200 Hawkins Dr, Ste 207, Roy Carver Pavilion, Iowa City, IA 52242-1053 (michael-wall@uiowa.edu).

Financial Disclosure: None reported.

Funding/Support: This study was supported by a VA Merit Review Grant and an unrestricted grant to the Department of Ophthalmology at the University of Iowa from Research to Prevent Blindness, New York, New York.

REFERENCES