Signal/Noise Analysis to Compare Tests for Measuring Visual Field Loss and Its Progression

Paul H. Artes and Balwantray C. Chauban

PURPOSE. To describe a methodology for establishing signal-to-noise ratios (SNRs) for different perimetric techniques, and to compare SNRs of frequency-doubling technology (FDT2) perimetry and standard automated perimetry (SAP).

METHODS. Fifteen patients with open-angle glaucoma (median MD, −2.6 dB, range +0.2 to −16.1 dB) were tested six times with FDT2 and SAP (SITA Standard program 24-2) within a 4-week period. Signals were estimated from the average superior-inferior difference between the mean deviation (MD) values in five mirror-pair sectors of the Glaucosa Hemifield Test, and noise from the dispersion of these differences over the six repeated tests. SNRs of FDT2 and SAP were compared by mixed-effects modeling.

RESULTS. There was was moderate correlation between the signals of FDT2 and SAP ($r^2 = 0.68, P < 0.001$), but no correlation of noise ($r^2 = 0.01, P = 0.16$). Although both signal as well as noise estimates were higher with FDT2 compared with SAP, 60% to 70% of sector pairs showed higher SNRs with FDT2. The SNRs of FDT2 were between 20% and 40% higher than those of SAP ($P = 0.01$). There were no meaningful differences between parametric and nonparametric estimates of signal, noise, or SNR.

CONCLUSION. The higher SNRs of FDT2 suggest that this technique is at least as efficient as SAP at detecting localized visual field losses. Signal/noise analyses may provide a useful approach for comparing visual field tests independent of their decibel scales and may provide an initial indication of sensitivity, or deviation from normal, between the superior and inferior areas that are asymmetries between the superior and inferior areas that are characteristic of glaucomatosus visual field loss.

Visual field loss and its progression are hallmarks in glaucoma, optic neuritis, and other diseases. However, clinically important “signals” in the visual field are often small compared with the variability between successive tests (“noise”). This applies to the detection of abnormalities with single examinations as well as to the measurement of change over time with serial examinations. Consequently, tests may have to be repeated several times before one can be certain that damage is either present or absent, and at least five tests are necessary to detect change with any confidence.2,5 With typical variability, at least 3 years of 6-monthly tests are needed to detect sight-threatening rates of visual field progression.8,9

Previous investigators have sought to reduce variability through optimized threshold strategies,10–14 closer control over human factors such as response bias, fatigue, and attention15–19; and most notably through new types of stimuli.20–24 Studies on retest variability, and on response variability estimated by psychometric functions, have shown that several of the newer techniques do not suffer from the large increase in variability in damaged fields seen with standard automated perimetry (SAP), but have nearly uniform variability across the dynamic range.25–29

However, it is challenging to compare visual field data between one technique and another. It is difficult, for example, to compare threshold estimates from SAP to those of motion perimetry, and although the technique’s scales can be made to appear similar by empiric correction factors, such adjustments are more likely to conceal the problem rather than to solve it. Second, when visual fields change over time, a 10-dB change with one technique may translate into a smaller or larger change with another technique, and this may vary with the degree of damage. Moreover, techniques differ in their measurement ranges; in a damaged area of the visual field one technique may still provide useful threshold estimates while another only measures absolute losses (0 dB) that are not informative for determining change.28 In combination, these issues limit the usefulness of current analyses for comparing different techniques.

Direct evidence that one technique performs better than another can only be obtained through comparative studies with substantial numbers of patients, but even these studies do not always give conclusive results. Because normal reference data for different techniques are often obtained from different samples of healthy controls, it can be difficult to compare probability maps from different techniques.30–31 Longitudinal studies are needed to compare effectiveness in measuring change over time, but such studies are costly and often last several years.32–34 They may also be difficult to interpret because no single technique provides an ideal reference standard.35 A method is therefore needed that can provide clues to the potential merit of a new technique within a relatively short time.

In this work, we propose a simple extension for analyses of retest data and demonstrate how perimetric techniques may be compared, even if they use different types of stimuli and do not share the same decibel scale. The underlying rationale of this analysis is to compare the ability of the techniques to measure systematic differences within a visual field—for example, the asymmetries between the superior and inferior areas that are characteristic of glaucomatosus visual field loss.36 Differences in sensitivity, or deviation from normal, between the superior and inferior mirror pairs of sectors within a visual field can be interpreted as a signal, and the variability of these differences from test to test can be interpreted as noise. By estimating the ratio between signal and noise (signal/noise ratio; SNR), the ability of the technique to identify localized visual field loss may be expressed independent of the decibel scale of the instrument, so that a paired comparison can be made between perimetric techniques.

From Ophthalmology and Visual Sciences, Dalhousie University, Halifax, Nova Scotia, Canada.

Presented at the 2006 meeting of the International Perimetric Society (IPS), Portland, Oregon.

Supported by an E. A. Baker Foundation Project Grant (PHA) and Canadian Institutes of Health Research Grant MOP-11357 (BCC).

Submitted for publication February 21, 2009; revised April 14, 2009; accepted August 6, 2009.

Disclosure: P. H. Artes, None; P. C. Chauhan, None

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be marked “advertisement” in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Corresponding author: Paul H. Artes, Ophthalmology and Visual Sciences, Dalhousie University, Room 2035, West Victoria, 1276 South Park Street, Halifax, Nova Scotia, B3H 2Y9, Canada; paul@dal.ca.

Copyright © Association for Research in Vision and Ophthalmology

4700
METHODS

Data

Fifteen patients with open-angle glaucoma (mean age, 66.3 years; range, 56.1–80.6) who had early to moderately advanced visual field loss with SAP (median mean deviation [MD], −2.6 dB, range +0.2 to −16.1) were recruited from the clinics of the Queen Elizabeth II (QEII) Health Sciences Centre (Halifax, Nova Scotia, Canada). Inclusion criteria were a clinical diagnosis of open-angle glaucoma, refractive error within 5 D equivalent sphere or 3 D astigmatism, visual acuity better than or equal to 6/12 (+0.3 logMAR), and prior experience with frequency doubling technology (FDT) perimetry (i.e., FDT1) and SAP. During a period of 4 weeks, one eye of each patient was tested six times with FDT2 (24-2 threshold test) and six times with SAP (SITA Standard; 24-2 test), in randomized order. The protocol was approved by the QEII Health Sciences Centre Research Ethics Committee; all participants gave written informed consent.

FIGURE 1. Stimulus locations of FDT2 and SAP. The five visual field sectors in the superior hemifield compared with their mirror images in the lower hemifield are outlined. Black square, white dot: location of the blind spot.

FIGURE 2. Example illustrating how signal and noise were derived. (a) In the upper and lower arcuate sectors, total deviations in each test were averaged to sector MDs. (b) From the six repeated tests, six sMDs were obtained in each of the sectors. (c) The distribution of differences between the upper and lower sectors was obtained by enumerating all 36 possible combinations of the six tests, and (d) estimates of signal (black arrowhead) and noise (curly bracket) were derived from the average and the dispersion of this distribution, respectively.

FIGURE 3. Relationship between sectoral MDs obtained with SAP and FDT2. Each data point represents the mean of six repeated tests with FDT2 and SAP. For sectors with SAP MDs better than −10 dB, RMA regression estimated a slope of 2.1 (95% CI, 1.9–2.3). The 1:1 line is shown for comparison (gray diagonal).
after full disclosure of the study protocol and risks, in accordance with the Declaration of Helsinki. Full details of this dataset are described in a previous paper on threshold and variability properties of the FDT2 perimeter (Humphrey-Matrix; Carl Zeiss Meditec, Inc., Dublin, CA).

**Analysis**

The thresholds of FDT2 and SAP were transformed to total deviations by using reference data from healthy volunteers. For each test, we then averaged the total deviations within the 10 visual field sectors of the Glaucoma Hemifield Test to obtain 10 sectoral (s)MDs (Fig. 1). Reduced major axis (RMA) regression was used to estimate the relationship between the sMDs of FDT2 and SAP. Unlike ordinary least-squares regression, which assumes that the x-values are from an independent variable measured without error, the RMA fit minimizes the residuals in both vertical and horizontal directions and is therefore a more appropriate method for establishing the slope of the relationship between FDT2 and SAP.

**Signal/Noise Analysis**

Because each patient had been examined six times, there were six sMDs, one for each sector, for both SAP and FDT2. For each patient, we calculated the superior–inferior difference in sMD between the mirror sectors in each test, and in all 30 combinations of the six tests, so that 36 differences were obtained for each patient, each sector, and each of the two techniques (SAP and FDT2). Estimates of signal and noise were then derived from the mean and the population standard deviation (SD) of the distribution of differences (n = 36), respectively (Fig. 2).

Because a single outlying data point can have an unduly large influence on both the mean and the SD, we also computed nonparametric estimates by replacing the mean with the median, and the SD

![Figure 4](https://example.com/figure4.png)

**Figure 4**: Example A: In this visual field with mild damage, a 2-dB asymmetry between the nasal superior and inferior sectors was measured with FDT2, but not with SAP. The dispersion of the differences (noise) was similar with SAP and FDT2 (0.9 and 0.7 dB, respectively), and FDT2 had a higher SNR (3.0 compared with 0.2 with SAP). Vertical gray bar (noise): ±2 SD to indicate whether the mean of the differences is significantly different from zero (dashed gray line).
with the median absolute deviation (MAD)\textsuperscript{39} of the differences (described in the Appendix, along with Bland-Altman comparisons of parametric and nonparametric estimates).

Signal, noise, and signal/noise ratios (SNRs) were then compared between SAP and FDT2, in each of the five pairs of sectors and each of the 15 patients. All analyses were performed in the freely available open-source environment R\textsuperscript{40,41}. The nlme library\textsuperscript{42} was used to estimate statistical significance and confidence intervals. Patients were treated as random factors, to adjust for the nonindependence of the five sector pairs within each patient.

**RESULTS**

The sMD in the 10 visual field sectors ranged from $-27.8$ to $+1.5$ dB (median, $-2.3$ dB) with SAP, and from $-26.2$ to $+3.3$ dB (median, $-2.4$ dB) with FDT2. The relationship between SAP and FDT accounted for $69\%$ of the variance in the data (Spearman rank correlation, $P < 0.001$, Fig. 3), and, for sectors with SAP sMD better than $-10$ dB, the relationship between SAP and FDT2 appeared linear (Tukey’s test for additivity,\textsuperscript{43} $P = 0.22$) with a slope of $2.1$ (RMA regression; $95\%$ confidence interval [CI], $1.9$–$2.3$). For sectors with more advanced damage, however, the relationship between the sMDs of both techniques became progressively weaker and deviated significantly from that observed in less damaged sectors ($P < 0.001$, Tukey’s test). Despite the averaging of six repeated tests with both techniques, the data exhibited a large degree of scatter (Fig. 3).

Signal and noise estimates were derived from the mean and the SD of the superior–inferior differences (Fig. 2), with both FDT2 and SAP. Three selected examples are shown in Figures 4, 5, and 6.

There was a moderately close relationship between the signals of both techniques (Fig. 7a, $r^2 = 0.52$, $P < 0.001$), but no such relationship between the noise estimates ($r^2 = 0.01$, $P = 0.16$; Fig. 7b).

**FIGURE 5.** Example B: In this extensively damaged visual field, both SAP and FDT2 revealed a small difference ($2.8$ and $4.2$ dB, respectively) between the superior and inferior paracentral sectors. With SAP, the large variability of both sectors (SD, $3.7$ dB) made it difficult to distinguish this signal (SNR, 0.8). With FDT2, the asymmetry was more clearly apparent (SNR, 3.1), chiefly because of lower noise (1.3 dB). Vertical gray bar (noise): as described in Figure 4.
Although both signal and noise appeared numerically larger with FDT2 compared with SAP (Table 1), a direct comparison between the estimates is difficult to interpret because they are expressed in instrument-specific decibel units. However, calculation of the ratio between signal and noise causes the instrument-specific units in numerator and denominator to cancel each other, such that the resultant SNR is independent of the decibel scale of the instrument.

There was a moderately close association between the SNRs of SAP and FDT2 ($r^2 = 0.68, P < 0.001$, Fig. 8). Of the 75 sector pairs, 46 (61%) had a larger SNR with FDT2. On average, the SNRs of FDT2 were 19% larger (95% CI, −6% to 52%; $P = 0.14$) than the corresponding SNRs of SAP.

Sector pairs within which there were no differences in damage contributed little information, but they may have diluted genuine differences between the techniques if included in an overall comparison. To reduce this dilution while avoiding selection bias, we compared the SNRs of SAP and FDT2 for all those pairs in which either SAP or FDT2 had an SNR greater than 0.5, 1.0, or 2.0 (Table 2).

Of the sector pairs with SNRs >0.5 with either FDT2 or SAP, between 68% and 93% had larger SNRs with FDT2 compared to SAP (Table 2, column 3). However, these findings were statistically significant ($P < 0.05$) only for the subset of SNRs which were greater than 1.0. In these sector pairs, the SNRs of FDT2 were approximately 40% larger than the corresponding SNRs of SAP (Table 2, column 2).

**DISCUSSION**

The purpose of this study was to develop an approach for comparing perimetric techniques independent of their measurement scales and to apply this methodology to visual field data from the FDT2 perimeter.

In a previous study, we demonstrated that the variability characteristics of both techniques were qualitatively different. In this visual field, both SAP and FDT indicated substantial damage. With SAP, an asymmetry of 6 dB was clearly apparent in the nasal sectors (SNR > 2), despite large variability in the more damaged inferior sector. Despite lower variability (1.3 dB), there was no detectable signal with FDT2 (SNR, 0.3). Vertical gray bar (noise): as described in Figure 4.
ent—threshold estimates from FDT2 had nearly uniform variability across the measurement range of the instrument, whereas those of SAP showed an exponential increase in variability with decreasing sensitivity. Similar results were obtained in other studies, with both FDT1 and FDT2. We were then, however, unable to make a quantitative comparison of the variability, because both instruments use different types of stimuli and different definitions of the decibel scales. The lack of a general method of comparing threshold data from different perimetric tests motivated the signal/noise analyses performed in this study. By relating systematic differences within a visual field (signal) to the precision with which such differences can be measured (noise), we can compare techniques independent of their underlying measurement units.

Our data showed a substantial correlation between the signals of the two techniques ($r^2 = 0.52$), but no such correlation for the noise. The lack of a relationship between the noise estimates of FDT2 and SAP clarifies why, in some patients, either technique may have true advantages for measuring losses that are less detectable with the other. Although our dataset of 15 patients is too small for a meaningful subgroup analysis, we suggest that the signal/noise methodology proposed in this article provides a useful framework for studying systematic differences between different perimetric techniques. It is particularly important to establish factors that contribute to the large scatter apparent in Figure 5. Because six tests had been averaged for each data point, it is unlikely that this scatter can be explained solely by measurement variability of FDT2 and SAP.

For sectors with SAP MDs better than $-10\, \text{dB}$, the slope of the relationship between the sectoral MDs of FDT2 and SAP was $2.1$ (Fig. 5). This closely mirrors the findings reported in our earlier paper in which we compared threshold estimates from individual test locations, and it is in agreement with the slope of $2.0$ expected from the different definitions of the decibel scale of both techniques. With FDT2, a decibel is defined as $-20\, \log_{10}$ of Michelson contrast such that a change of $20\, \text{dB}$ corresponds to a change of $2\, \log$ units. Of importance, the empirically determined slope $-2$ suggests that the magnitude (in decibel) of early and moderate visual field losses, and changes over time, could be up to twice as large with FDT2 than with SAP. In our data, the signals (superior–inferior differences between the mirror pairs of visual field sectors), were, on average, only $40\%$ higher with FDT2 than with SAP, but this average would have been reduced by sector pairs within which there were no meaningful differences in damage and therefore no measurable signals.

Approximately $70\%$ of sector pairs with an SNR $>1.0$ with either SAP or FDT2 had higher SNRs with FDT2 (Table 2). This result is evidence of an overall gain and confirms that the benefits of the higher FDT2 signals are not offset by the larger variability of this technique (Table 1). For pairs with SNRs $>1.0$, the SNRs of FDT2 were approximately $40\%$ larger than those of SAP. This difference is similar in magnitude to the improvement in SNR that would be expected from repeating a test, since performing the same test twice can reduce the variability, in theory, by a factor of $\sqrt{2}$ ($1.41$). A difference of this magnitude would mean a substantial net improvement in the detectability of early changes, for cross-sectional detection of visual field loss as well as for longitudinal measurement of visual field progression. For the former, empiric investigations on total and pattern deviation probability maps with FDT2 and SAP, as well as global visual field indices such as pattern standard deviation, are in agreement with our results.

How may SNRs, calculated from retest data, help to estimate performance in measuring progression? The rationale for using gradients in space as a surrogate for changes over time is illustrated in Figure 9. Progression of visual field loss is a change in sensitivity over time, and the usefulness of a test for following patients over time depends on how well its data reflect these changes. In contrast, SNRs estimated from tests performed within a short period of time express the detectability of differences within a visual field at that particular time. SNRs express how reliably a technique reflects gradients of damage within a visual field, and this is a function of the depth

### Table 1. Distribution of Signal and Noise Estimates from FDT2 and SAP

<table>
<thead>
<tr>
<th></th>
<th>FDT2 (dB)</th>
<th>SAP (dB)</th>
<th>FDT2&gt;SAP</th>
<th>FDT2/SAP (CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal: mean, median (range)</td>
<td>3.5, 1.9 (0.1–15.4)</td>
<td>2.8, 1.2 (0.0–16.6)</td>
<td>51/75 (68%)</td>
<td>1.40 (1.06–1.85)</td>
<td>0.02</td>
</tr>
<tr>
<td>Noise: mean, median (range)</td>
<td>1.7, 1.6 (0.6–3.0)</td>
<td>1.5, 1.3 (0.5–5.2)</td>
<td>48/75 (64%)</td>
<td>1.17 (0.97–1.42)</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Probabilities and confidence intervals were established by mixed-effects modeling, because each patient contributed five estimates.
of loss, the variability of the measurements, and the dynamic range of the technique. A larger SNR therefore does not necessarily mean that one technique is more sensitive than another, nor does a more sensitive technique necessarily provide a larger SNR (Fig. 10).

The signal/noise methodology proposed in this article has several limitations. To obtain robust estimates of signal and noise, multiple tests have to be performed. Nevertheless, a rigorous protocol with at least five examinations per eye has advantages also for the derivation of test–retest intervals, because a large number of combinations of test–retest examinations can be analyzed. SNRs depend on the sample of patients and therefore cannot be compared across different studies. They also depend on the somewhat arbitrary choice of where in the visual field the signal and noise distributions are derived from. In this study, we used the superior–inferior sectors of the Glaucoma Hemifield Test, and therefore our finding of larger SNRs with FDT2 may strictly apply only to those analyses that make use of a similar clustering. In principle, however, other pairs of test locations or pairs of clusters could be chosen. Finally, SNRs can be estimated only if focal losses are present in the visual field; diffuse reductions in sensitivity do not contribute a signal. As a consequence, the method is unsuitable for evaluating techniques that predominantly uncover diffuse loss.

The assumption that gradients in space can be used as a first approximation for change over time appears reasonable but is, as yet, untested. Signal/noise estimates from test–retest studies will therefore not replace longitudinal studies for investigating new visual field tests’ ability to monitor patients with glaucoma, but they may provide early insight into properties that cannot be gained solely from analyses of test–retest variability. They may help in hypothesis-building and in planning effective longitudinal studies of new visual field tests.

### Table 2. Comparison between SNRs of FDT2 and SAP

<table>
<thead>
<tr>
<th>SNR</th>
<th>Ratio of SNR (FDT2/SAP CI)</th>
<th>SNR FDT2 &gt; SAP/total</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;0.5</td>
<td>1.20 (0.93–1.56)</td>
<td>66/71 (93%)</td>
<td>0.15</td>
</tr>
<tr>
<td>&gt;1.0</td>
<td>1.43 (1.05–1.94)</td>
<td>34/50 (68%)</td>
<td>0.05</td>
</tr>
<tr>
<td>&gt;2.0</td>
<td>1.39 (1.04–10.6)</td>
<td>22/30 (73%)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

A ratio > 1 in column 2 indicates that FDT2 provided a larger SNR than SAP. Column 3 gives the number of pairs in which the SNR of FDT2 was greater than that of SAP. P values and confidence intervals were established by mixed-effects modeling, because each patient contributed five estimates.

### Figure 8.
The relationship between SNRs with SAP and FDT2. The points labeled A, B, C correspond to the examples in Figures 4, 5, and 6, respectively. Axes are drawn on a square-root scale to emphasize mid-range values.

### Figure 9.
(a) The SNR measured between locations A1 and B1 measures how reliably the technique represents the gradient of damage in space (vertical arrow). (b) If B deteriorates over time, such that its deviation at B2 becomes equal to that of A1, the gradient in time between B1 and B2 is equal to that between A1 and B1. The ability to detect the change from B1 to B2 should be similar to that measured by the SNR between A1 and B1.

### Figure 10.
(a) The relationship between signals of two techniques, A and B, when A is more sensitive than B. (a) Technique A reveals superior loss (signal at nasal step, vertical arrow). There is no signal with B. (b) Technique A reveals more extensive loss than does technique B but has reached the limit of its dynamic range; its signal is small compared to B. (c) Owing to its larger dynamic range, B continues to provide signal even though both superior and inferior sectors are damaged.
Mean and SD are parametric estimates of central tendency and dispersion and can be highly affected by outliers, particularly in small samples. Compared with mean and SD, the nonparametric median and MAD are more robust to outliers, but they are also less efficient (more variable) when there are no outliers.

To investigate whether there were meaningful differences between parametric and nonparametric estimates of signal, noise, and SNRs, we computed nonparametric estimates of signal and noise from the median and MAD. The MAD was scaled by a factor of 1.483 to make it similar to the SD in a normal distribution.49

It should be noted that the enumeration of all possible 36 differences between the superior and inferior sMDs as explained in Figure 2 of the Methods section needs to be performed only to compute the nonparametric estimates. For the parametric estimates, the mean of the 36 differences is identical with the difference between the means of the six sMDs in the superior and inferior sectors (equation 1), and the SD of the differences is identical with that obtained from pooling the SDs of the superior and inferior sectors (equation 2):

\[ \bar{x}_{\text{diff}} = \bar{x}_{\text{sup}} - \bar{x}_{\text{inf}} \]  
\[ \sigma_{\text{diff}} = \sqrt{(\sigma_{\text{sup}})^2 + (\sigma_{\text{inf}})^2} \]

The Bland-Altman plots in Figure A1 show good overall agreement between the parametric and nonparametric estimates of signal and noise. With both FDT2 and SAP, noise estimates >2 dB were systematically smaller with the nonparametric method (Figs. A1c, A1d; \( P < 0.10 \)). The discrepancies observed in the SNR ratios of FDT2 increased with the magnitude of the SNR (Fig. A1f). However, none of our findings changed substantially when the analyses were performed with the nonparametric SNRs instead of the parametric alternatives.

References

45. Sun HAO, Dul MW, Swanson WH. Linearity can account for the similarity among conventional, frequency-doubling, and gabor-based perimetric tests in the glaucomatous macula. *Optom Vis Sci*. 2006;83:E455.