EXTENDED REPORT

Performance of the 24-2-5 frequency doubling technology screening test: a prospective case study

P G D Spry, H M Hussin, J M Sparrow

Br J Ophthalmol 2007;**91**:1345–1349. doi: 10.1136/bjo.2007.115436

See end of article for authors' affiliations

Correspondence to:
Dr Paul G D Spry, Bristol Eye Hospital, Lower Maudlin Street, Bristol, BS1 2LX; paul.spry@ubht.nhs.ukAccepted 16 March 2007
Published Online First
27 March 2007**Aim:** To evaluate the performance of the frequency doubling technology (FDT) 24-2-5 screening test by comparison with the established N-30-5 FDT screening test for detection of glaucoma.**Method:** A prospective random sample of individuals referred for possible glaucoma were tested with FDT screening tests 24-2-5 and N-30-5 using the Humphrey Matrix perimeter in addition to standard clinical examination relevant to glaucoma detection. Discriminatory power, reliability and test time of these tests were assessed and compared. The case definition for glaucoma was made by patient according to the established clinical diagnosis.**Results:** Of 63 referred eligible individuals, 53 (84%) were recruited. Sensitivity and specificity for the N-30-5 screening test was 78 and 85% respectively, compared with 83% and 75% for the 24-2-5 with areas under a receiver operator characteristic curve being 0.87 and 0.92. Differences between these indices were not statistically significant. For a specificity of 95%, sensitivity values were 76% and 56% for the 24-2-5 and N-30-5 respectively. Mean (standard deviation) test duration for the FDT 24-2-5 and N-30-5 screening tests were 111 (13) and 39 (10) seconds respectively ($p < 0.001$). A total of 19 subjects (36%) produced unreliable test results in one or both eyes when tested with the 24-2-5 screening test compared with 5 subjects (9%) with the N-30-5 ($p < 0.0005$).**Conclusion:** Minimal discriminatory power differences existed between the two screening tests evaluated, with both screening tests exhibiting high discriminatory power for detection of individuals with glaucoma. More individuals produced unreliable results on the 24-2-5 screening, which also took longer to perform.

Suprathreshold visual field tests can be considered appropriate for use in glaucoma screening programmes, population studies and case detection strategies because they are quick, simple to administer, relatively easy to undertake and generally acceptable to those being tested. Suprathreshold test strategies have been popular for opportunistic case finding of patients with glaucoma in the United Kingdom for a number of years.¹

At present, a number of perimetric instruments are available that offer suprathreshold strategies, including instruments that use frequency doubling technology (Welch-Allyn, Skaneateles, New York, USA). The first such instrument available, the Frequency Doubling Technology (FDT) Perimeter (Zeiss Humphrey Systems, Dublin, California, USA and Welch-Allyn, Skaneateles, New York, USA) uses suprathreshold strategies ('screening' tests), with either 17 or 19 stimulus locations within the central 20 or 30 degrees of the visual field respectively. These tests have been evaluated by many investigators in a variety of scenarios.^{2–9} Although the consensus is that screening test performance is good, it is possible that the large 10 degree target size could limit the performance of the suprathreshold strategy because small scotomata that can occur early in disease can be 'averaged out' by surrounding areas of normal sensitivity. For example, threshold FDT performed on a group of high-risk glaucoma suspects and early glaucoma patients identified a greater proportion of abnormal test locations with 54 small 4 degree test targets in a 24-2 stimulus test pattern than when the same subjects were tested with 17 larger 10 degree stimuli of the FDT perimeter.⁴

In the second generation FDT instrument, the Humphrey Matrix, a suprathreshold program is available that employs smaller test targets in the well known 24-2 test pattern. However, the impact of testing a larger number of test locations

using a smaller stimulus size in a suprathreshold screening strategy remains unknown.

The objective of this study was to prospectively evaluate the performance of the Humphrey Matrix 24-2-5 FDT (suprathreshold) screening test by comparison with the established N-30-5 FDT screening test in a hospital eye service clinic into which glaucoma suspects are referred for initial assessment.

The specific aims were: (1) to quantify the discriminatory power of the 24-2-5 FDT screening test and compare with the existing FDT N-30-5 screening test in a clinical environment where patients considered to exhibit possible glaucoma are initially examined; (2) to measure the time taken to perform the 24-2-5 test and compare this with the existing FDT screening test, the N-30-5 and; (3) to quantify the proportion of reliable test results achieved by study subjects with the 24-2-5 strategy and compare this with the proportion achieving reliable results with the established FDT screening test, the N-30-5.

MATERIALS AND METHODS

Study design and sample selection

A prospective case series study design was used. The reference population was individuals referred to the Hospital Eye Service (HES) from any source because of suspected glaucoma. The sampling frame was chosen to be representative of this reference population and therefore comprised the list of all individuals referred for further examination because of initial findings suggestive of glaucoma. At the Bristol Eye Hospital, it is usual practice for such individuals to be examined in dedicated 'new glaucoma patient' clinics. Eight consecutive new patient clinics were sampled between March 2004 and

Abbreviations: FDT, frequency doubling technology; SAP, standard automated perimetry

June 2004. Of all individuals receiving appointments, 25% were prospectively selected using a simple random sampling strategy (random number tables), this proportion being dictated by physical capacity of the investigators to test participants within the normal clinic duration. Aside from referral to the clinic, no further eligibility criteria were imposed.

The study was approved by the local ethics committee and adhered to the tenets of the declaration of Helsinki. Every effort has been made to report this study to defined standards.¹⁰

Clinical examination

Standard examination comprised assessment of corrected Snellen visual acuity and standard automated perimetry (SAP) using the Humphrey Field Analyzer (HFA) Program 24-2 SITA-Fast (Carl Zeiss Meditec, Dublin, California, USA) followed by ocular examination (anterior segment examination, gonioscopy, Goldmann applanation tonometry, dilated posterior segment examination including optic nerve head examination with Volk binocular indirect ophthalmoscopy.) The SITA-Fast thresholding strategy was employed in this study as it is the default test used in the clinic.

SAP tests were performed by a pool of clinic staff trained in visual field testing. Subjects wore near spectacle correction if appropriate. For patients habitually wearing bifocal, varifocal or tinted spectacles, the near correction was entered as full aperture trial lenses.

FDT suprathreshold screening tests

The Humphrey Matrix (Welch Allyn, Skaneateles Falls, New York, USA and Carl Zeiss Meditec, Dublin, California, USA) suprathreshold ('screening') tests N-30-5 and the 24-2-5 FDT were performed in addition to standard ophthalmic examination within the routine outpatient clinic environment. FDT tests were performed by a single investigator (HMH). Individuals wore their distance refractive correction for the test where appropriate. For patients habitually wearing bifocal, varifocal or tinted spectacles, distance correction was supplied using wide aperture correcting lenses designed for perimetry.¹¹

The FDT Matrix suprathreshold tests and SAP were carried out in no particular order. A rest interval of at least 10 min was provided between SAP and FDT Matrix tests. All visual field examinations tested the right eye first.

The results of the first type of visual field tests (FDT or SAP) were not available to the individual operating the second test type. The two FDT screening tests were performed in a randomised order such that 50% of subjects performed each screening test first. A rest interval was given between these two screening tests.

The stimulus locations for the FDT screening tests are shown in fig 1. Both N-30-5 and 24-2-5 tests employ square stimuli, with diameters of 5° and 10° respectively. Stimuli of 10° has a spatial frequency of 0.5cycles/° and a temporal frequency of 18Hz, whilst these properties are 0.25 cycles/° and 25Hz for 5° stimuli.

Both FDT screening tests were run at a suprathreshold increment designed to optimise sensitivity by selection of the 'dash five' (eg, N-30-5) screening option such that stimuli were presented at contrast levels equivalent to the 95% contrast sensitivity level (5% probability level) of an age and location matched normal contrast sensitivity distribution, rather than the available alternative of 99% employed by the 'dash one' test (eg, N-30-1). Although the selected probability level was the same, the manner in which unseen stimuli were handled differed between the tests. For both tests, all unseen test presentations are repeated. If missed on a second presentation, the 24-2-5 test recorded this location as unseen, unlike the N-30-5 test that proceeds to present a higher contrast (third)

stimulus at the 2% probability level. If still unseen, a final (fourth) presentation is made at contrast equivalent to the 1% probability level.

Case definition

Because the primary objective of this investigation was to compare the ability of the two FDT Matrix screening tests to identify glaucoma cases, a pragmatic clinical case definition for glaucoma was employed that is routinely used for determination of patients requiring immediate glaucoma treatment. This definition comprised the results of examinations routinely performed in the new patient glaucoma clinic, primarily based on a combination of stereoscopic optic nerve head evaluation and standard (white-on-white) automated perimetry. Individuals were therefore classified as either having glaucoma, being glaucoma suspects or no glaucoma present. Diagnosis was by patient, rather than by eye. Optic nerve head evaluation was performed by a single experienced consultant ophthalmologist (JMS) with a specialist interest in glaucoma. This examining ophthalmologist was masked to FDT Matrix test results.

Reliability

Reliability was assessed for the N-30-5 and 24-2-5 screening tests using the Matrix's proprietary catch trials for false positives and fixation losses. The number of catch trials for each of these indices is fixed for each test type. For the N-30-5 three catch trials are used for each index, with 10 being used in the 24-2-5. In this study, test reliability was defined by patient rather than by eye using the rationale that reliability (1) is a patient dependent, rather than eye dependent property; and (2) catch trials constitute a very small sample of total stimuli and have a relatively low likelihood of detecting unreliable observers. For both test types, any subject with a proportion of fixation loss or false positive catch trials exceeding 33% of presentations in either eye were therefore categorised as unreliable.

Discriminatory power

Two approaches were adopted to quantify the ability of each test type to correctly identify subjects meeting the glaucoma case definition.

The first approach was criterion based. The results of each test type were categorised into a binary outcome variable for each patient that comprised either passing (a 'normal' result) or failing the test (an 'abnormal' result). The criteria for this approach were chosen to represent those typically applied in clinical situations as likely to represent abnormal test results. For the N30-5 test type the criterion selected consisted of two or more unseen test stimuli at any test location in either eye. For the 24-2-5 test type the criterion used was presence of a cluster of 3 or more locations in either hemifield of either eye where stimuli were unseen, of which at least one location was a non-edge point.

The second approach was criterion-free and treated data as continuous. With this approach, levels of sensitivity and specificity were calculated for each increment on the scale of total number of points missed (that is, from zero to all points missed). The relationship between sensitivity and specificity were then used to construct receiver-operator characteristic (ROC) curves. The area under the ROC curve and its associated 95% confidence interval gives an estimate of the discriminatory power for populations to which this dataset is generalisable. This second approach was adopted; (1) to facilitate direct comparison of discriminatory power between test types; (2) to directly compare test types without the influence of any specific arbitrary cut-off criteria such as those selected in the first

Figure 1 Diagram showing FDT screening test stimulus locations for FDT suprathreshold 'screening' tests. Large squares show locations of 10° test targets of the N-30-5 stimulus pattern, with smaller squares being the 5° targets of the 24-2-5. Circles represent 10° eccentricity.

analyses approach; and (3) to enable comparison of sensitivity at a defined level of specificity (95%).

Statistical analysis

Where required, analyses were performed using Intercooled Stata version 7 (Stata Corporation, College Station, Texas, USA). Graphs were drawn in Sigmaplot 2001 for Windows version 7.0 (Jandel Corporation, San Rafael, California, USA).

RESULTS

General

Of 63 subjects randomly selected for inclusion, 53 individuals (84%) were recruited into the study with complete ascertainment of study data. Of the 10 individuals that were not recruited, reasons for non-participation comprised non-attendance at clinic ($n = 7$), inability to perform any type of visual field test due to poor cognition ($n = 1$) and investigator failure to identify and test patient ($n = 2$) during the course of clinic attendance.

The mean participant age (standard deviation (SD)) was 60 (13.7) years and 24 subjects were male (45%). For those individuals selected but not recruited, mean (SD) age was 60 (10.9) yrs and 40% were male.

Of the sample of 53 participants, 9 (17%) met the case definition for glaucoma, 27 (51%) were considered to be glaucoma suspects that required ongoing observation due to risk of glaucoma development and 17 (32%) were considered normal. Further diagnostic information is provided in fig 2. A frequency distribution of HFA 24-2 mean deviation values among the study participants is provided in fig 3, demonstrating a skew towards normal test results or early deficits.

Time

The distribution of test times is given in fig 4. The mean (standard deviation) test duration for the FDT 24-2-5 and N-30-5 screening tests were 111 (13) and 39 (10) seconds respectively. This difference was statistically significant ($p < 0.001$, Z-test). Although not the subject of comparison in this study, for reference purposes the mean (SD) duration of the SAP Humphrey Field Analyzer 24-2 SITA-Fast test was 215

(54) seconds, significantly longer than both FDT screening test types ($p < 0.001$, Z-test).

Reliability

A total of 19 subjects (36%) produced unreliable test results in one or both eyes when tested with the 24-2-5 screening test compared with five subjects (9%) with the N-30-5. This difference in reliability was statistically significant (Z test for difference in proportions, $p < 0.0005$). Four patients were found to be unreliable with both screening test types and there was a borderline significant association between obtaining unreliable test results with each screening test type (Fishers Exact test, $p = 0.050$).

Of the 19 subjects that produced unreliable 24-2-5 test results, 4 (21%) were unreliable in both eyes and 15 (79%) in only one eye. All 5 subjects producing unreliable N-30-5 test results were unreliable in one eye only.

Although not the subject of comparison in this study, for reference purposes 10 (18%) patients produced unreliable HFA results when the same criteria of greater than 33% fixation losses or false positives were applied.

Discriminatory power

The ability of both screening tests to discriminate reliable subjects with a clinical diagnosis of glaucoma from those without a glaucoma diagnosis (ie, glaucoma suspects and no glaucoma present) is provided in table 1. For analyses where screening tests results were categorised into a binary outcome, as either a passed or failed screening test, sensitivity and NPV estimates were higher for the 24-2-5 than the N30-5 whilst specificity and PPV estimates were better (ie, higher) for the N-30-5 than the 24-2-5. In spite of these trends, none of these differences were of statistical significance.

Criterion free (ROC) analysis where each screening test result was analysed as a continuous variable is also summarised in table 1. This analysis revealed that better differentiation of individuals with glaucoma from non-glaucomatous subjects was achieved with the 24-2-5 screening test than the N-30-5 in this sample, with a higher area under the ROC curve (AUROC) for this test. However, in a similar manner to the binary analyses, this difference in estimate was not statistically significant.

Using the ROC curve to determine the sensitivity-specificity relationship, it was found that when specificity was fixed at 95%, sensitivity (95% CI) of 24-2-5 and N-30-5 tests were 67% (61 to 73%) and 56% (49 to 63%) respectively.

DISCUSSION

In this study, the performance of 24-2-5 suprathreshold 'screening' FDT has been evaluated in a hospital setting using a random sample obtained from a sequential case series of patients presenting with possible glaucoma. This sample represents a population with an enriched glaucoma prevalence and as such is generalisable to groups of patients identified as having features suggestive of glaucoma on primary eye examination, such as referral refinement programs or Hospital Eye Service (HES) new glaucoma patient clinics in the United Kingdom, or referrals from primary care ophthalmologists or optometrists to glaucoma specialist clinicians in the United States.

The performance of the suprathreshold tests was assessed on the basis of three major parameters relevant to use of visual field tests in clinical environments; (1) the discriminatory power, or ability of the test to successfully differentiate those who required treatment for glaucoma from those who did not; (2) the time required to complete the test; and (3) the proportion of patients that produced reliable test results. The main findings were that when compared with the more established N-30-5, the newer 24-2-5 test exhibited non-significant

Figure 2 Frequency distribution of diagnoses occurring within study sample ($n=53$).

differences in discriminatory power, took about 1 min longer per eye, and produced a higher proportion of unreliable results. It is important to interpret these comparisons in the context that both suprathreshold test types exhibited high levels of discriminatory power ($\text{AUROC} \geq 0.87$) and test times within a range suitable for routine use at an average of less than 2 min/eye.

To date the authors are not aware of any reports on the performance of the FDT 24-2-5 screening test in the literature. For the well established N-30-5 test, there are also no directly comparable published studies reporting sensitivity and specificity estimations. The values for these parameters obtained in this study are broadly similar to those reported by other investigators for the C20-5 test,^{3-5,9} which is similar to the N-30-5 with the difference being that the latter test also examines two additional locations extending between 20 and 30 degrees nasally, one each side of the horizontal midline.

The logic underlying the use of a larger number of smaller test targets is that subtle deficits are more likely to be identified using smaller stimuli and that defects may be better characterised with such stimuli.^{4,12} This study provides some support for both these hypotheses. In terms of identification of earlier (subtle) defects, a trend was found towards higher sensitivity with the 24-2-5 test on both analysis approaches: using a pragmatic clinical SAP definition of visual field defect, sensitivity estimates were 83% and 78% for the 24-2-5 and N-30-5 respectively and from criterion free analysis estimates of 67% and 56% respectively were achieved when specificity was fixed at 95%. This finding therefore provides evidence that, within this dataset, some small glaucomatous defects were not identified using 10 degree suprathreshold stimuli. However, because the sensitivity difference did not reach statistical significance, the generalisability of this finding is uncertain. Whilst the lack of statistical significance could have been due to inadequate sample power, it is also possible that the difference in sensitivity estimates is truly not significant in the reference population. Whilst at face value it might appear surprising that only a 5% improvement in sensitivity was obtained when increasing the number of stimuli from 19 to 54. However, similar experiences of declining sensitivity improvement per additional stimulus have previously been reported for suprathreshold tests using a method whereby the number of stimuli of the same size can be freely varied.¹³ The non-significant

Figure 3 Frequency distribution of HFA 24-2 SITA-Fast mean deviation values for both eyes of all study participants ($n=106$ eyes from 53 individuals, right and left eye data pooled.)

difference between these two FDT test types also serves to confirm the high performance of suprathreshold testing using 10° sized FDT stimuli and demonstrates that performance cannot be dramatically improved by switching to a 24-2-5 test.

There are some examples within our dataset to anecdotally support improved defect characterisation when the higher resolution grid afforded by the 24-2 test is used for suprathreshold examination (fig 5). In the context of suprathreshold testing, this attribute might be qualitatively useful to determine defect shape and as such might be informative for differential diagnosis or detection of future defect progression.

Because this study was not performed in a true screening environment, such as a programme actively seeking disease in individuals presumed to be healthy,¹⁴ the estimates of discriminatory power might differ from those that would be achieved in such environments. Nonetheless, these data do provide an indication to inform debate about use of FDT tests in screening programmes. Published criteria on tests suitable for use in screening programmes¹⁵ require information including test validity and acceptability. Validity of both tests have been discussed above in terms of discriminatory power. Proxy measures contributing to acceptability include both test timing and reliability. This study has demonstrated that the 24-2-5 test took on average less than 2 min/eye to complete, which is only

Figure 4 Frequency distribution of test time per eye for both FDT screening tests ($n=106$ eyes). Data for HFA 24-2 SITA-Fast tests on the same patient group is included for context.

Table 1 Discriminatory power estimates of both FDT screening test types for differentiation between study participants with a clinical diagnosis of glaucoma in either eye from those without glaucoma (glaucoma suspects; no glaucoma detected)

	N-30-5 (n = 48)	24-2-5 (n = 34)
Sensitivity, % (95% CI)	78 (72 to 84)	83 (77 to 89)
Specificity, % (95% CI)	85 (79 to 91)	75 (68 to 82)
Positive predictive value, % (95% CI)	54 (47 to 61)	42 (34 to 50)
Negative predictive value, % (95% CI)	94 (91 to 97)	95 (91 to 99)
Area under ROC curve (95% CI)	0.87 (0.82 to 0.92)	0.92 (0.87 to 0.97)

Only individuals where reliable test results were obtained from both eyes of each test are included.

Figure 5 Example of screening test results for the right eye of one patient within the study sample. It can be seen that in this case, greater spatial resolution of the 24-2-5 test findings (A) has better characterised the result of the early inferior arcuate defect than the N-30-5 (B).

approximately 50% of the time requirement of the HFA 24-2 SITA-Fast threshold test used for diagnostic purposes. In spite of this brevity, on average the 24-2-5 test took over 1 min longer per eye than the N-30-5. With regard to reliability, 24-2-5 yielded the highest proportion of unreliable test results with four times as many patients being unreliable compared with the N-30-5. Although it should be considered that this statistically significant and large difference could be due to longer test time, this is not supported by the fact that more individuals produce reliable HFA tests according to identical reliability criteria. Because HFA test order was controlled in the study design, it is unlikely that fatigue or learning effect might have introduced bias. The cause of this difference is therefore unclear. Sampling error cannot be completely excluded.

Authors' affiliations

P G D Spry, H M Hussin, J M Sparrow, Bristol Eye Hospital, Bristol, UK
P G D Spry, H M Hussin, University of Bristol, Division of Ophthalmology, Bristol, UK

Competing interests: PGDS has received research funds from Welch-Allyn, Skaneateles, New York, USA. JMS and HMH have no competing interests.

REFERENCES

- 1 Tuck MW, Crick RP. Use of visual field tests in glaucoma detection by optometrists in England and Wales. *Ophthalmic Physiol Opt* 1994;14:227-31.
- 2 Quigley HA. Identification of glaucoma-related visual field abnormality with the screening protocol of frequency doubling technology. *Am J Ophthalmol* 1998;125:819-29.
- 3 Yamada N, Chen PP, Mills RP, et al. Screening for glaucoma with frequency-doubling technology and Damato campimetry. *Arch Ophthalmol* 1999;117:1479-84.
- 4 Johnson CA, Cioffi GA, Van Buskirk EM. Frequency doubling technology perimetry using a 24-2 stimulus presentation pattern. *Optom Vis Sci* 1999;76:571-81.
- 5 Cioffi GA, Mansberger S, Spry PGD, et al. Frequency doubling perimetry and the detection of eye disease in the community. *Trans Am Ophthalmol Soc* 2000;98:195-9.
- 6 Burnstein Y, Ellish NJ, Magbalon M, et al. Comparison of frequency doubling perimetry with Humphrey visual field analysis in a glaucoma practice. *Am J Ophthalmol* 2000;129:328-33.
- 7 Cello KE, Nelson-Quigg JM, Johnson CA. Frequency doubling technology perimetry for detection of glaucomatous visual field loss. *Am J Ophthalmol* 2000;129:314-22.
- 8 Casson R, James B, Rubinstein A, et al. Clinical comparison of frequency doubling technology perimetry and Humphrey perimetry. *Br J Ophthalmol* 2001;85:360-2.
- 9 Johnson CA, Cioffi GA, Van Buskirk EM. Evaluation of two screening tests for frequency doubling technology perimetry. *Perimetry Update* 1998/1999. The Hague, The Netherlands: Kugler Publications, 1999:103-109.
- 10 Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic. *Accuracy Clin Chem*, 2003;49, 1-6.
- 11 Henson DB, Earlam RA. Correcting lens system for perimetry. *Ophthalmic Physiol Opt* 1995;15:59-62.
- 12 Johnson CA, Gunvant P, Fingeret M, et al. A primer for frequency doubling technology (FDT) perimetry using the Humphrey matrix. Welch-Allyn/Zeiss, 2004.
- 13 Henson DB, Chauhan BC, Hopley A. Screening for glaucomatous visual field defects: the relationship between sensitivity, specificity and the number of test locations. *Ophthalmic Physiol Opt* 1988;8:123-127.
- 14 Holland WW, Stewart S, Masseria C. Policy Brief. *Screening in Europe*. Geneva: World Health Organization, 2006.
- 15 UK National Screening Committee. Criteria for the appraising the viability, effectiveness and appropriateness of a screening programme. <http://www.nsc.nhs.uk> (accessed 12 December 2006).