Using data for continuous improvement in grading accuracy

Dr Wendy Gatling
Clinical Lead Dorset &
National Grading Resources Advisory Group

Improving Grading Quality in Diabetic Eye Screening

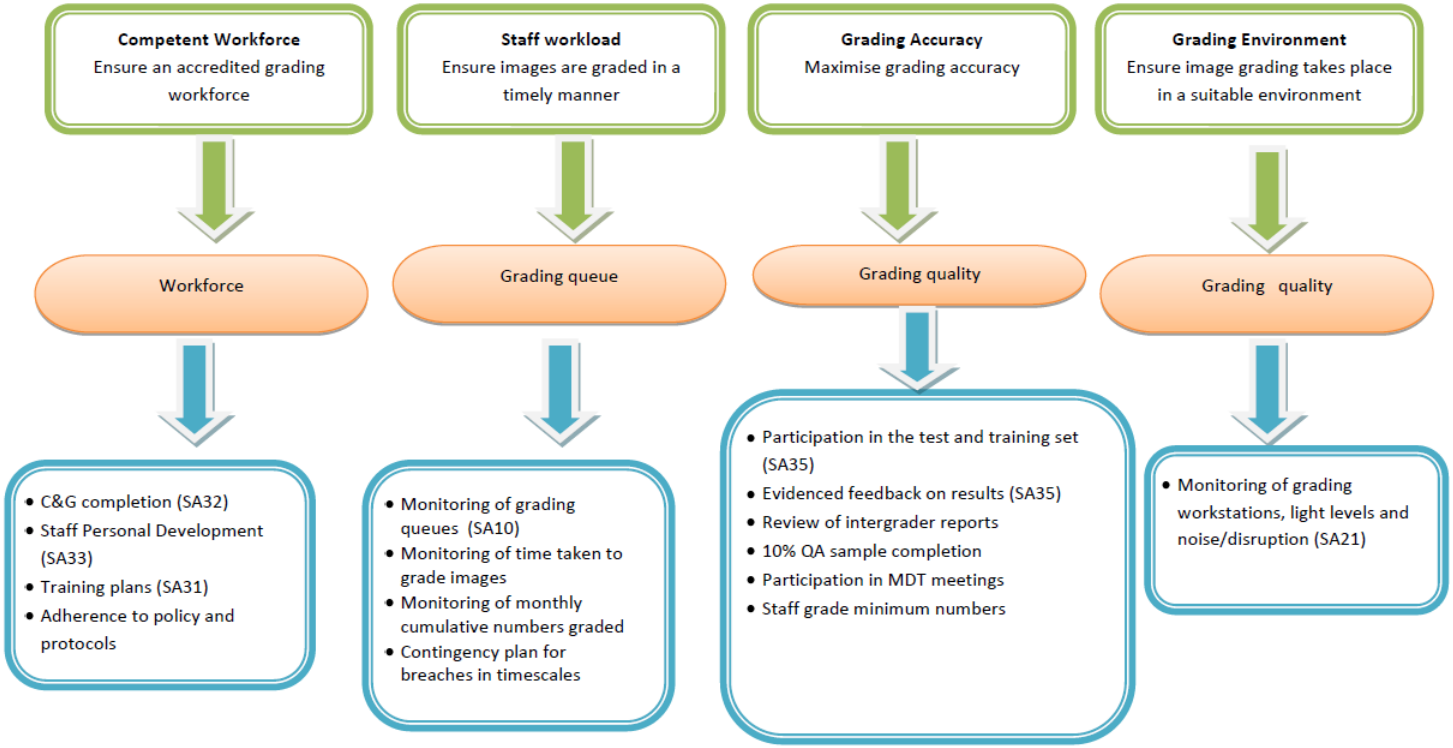
- Training & Accreditation
- Continuous Professional Development
- Internal Quality Assurance

What national guidance is available?

Process Three: Grading

Objective: Ensure a high quality service of reliable image grading results for patients

QA standards:
Objective 5: To ensure grading is accurate
Objective 14: To ensure that screening and grading of retinal images are provided by a trained and competent workforce
Objective 15: To ensure optimum workload for **all** graders in order to maintain expertise



http://diabeticeye.screening.nhs.uk/internal-qa

EQA visit document
Objective Measurement of Grading

- Regular MDTs with teaching
- Minimum no. images graded
- Time spent grading
- T&T results and feedback
- 1 in 10 R0M0 review
- Review & feedback of intergrader reports

- *UG rate*
- *Audit of ‘misses’, sight loss etc*
- *Reports to programme board*

Improve Grading Accuracy through *Feedback*

Individual graders

- Set standards
- Individual reports
- One to one review
- Grade or lesion training
- Buddy grader

All programme

- Set standards
- Group report for bench marking
- Group MDT sessions
- Grade or lesion training

Regular review at local programme level

DDESP Annual Report data

	2013/14	2013/12	2012/11	2011/10
Any retinopathy	**30.3%**	**30.5%**	**30.1%**	**31.4%**
Referable Retinopathy	**5.1%**	**5.9%**	**5.1%**	**4.5%**
Ungradable	**1.7%**	**5.3%**	**5.7%**	**4.9%**

DDESP Annual Report data

Number of Grades Performed

Year	Primary	Secondary	%	Arbitration	%
2013/14	**31220**	**10462**	**33.5%**	**785**	**2.5%**
2012/13*	**21530**	**7871**	**36.5%**	**3979**	**18.4%**
2012/11	**27219**	**10228**	**37.5%**	**2230**	**8.1%**
2010/11	**26127**	**11223**	**42.9%**	**2543**	**9.7%**

* 9 months data

3-6 monthly Programme Performance Review

	Target or Standard	Dorset Jan – Jun 2014 RDS
Patients screened as Ungradable	2.5-6.3% <3.0%	1.9 %
1st, 2nd & Arbitration grading rates for all patients screened	1st -100% 2nd - 35-50% Arbitration - <10% with target <5%	RDS only 1st 100%, 2nd 36.7% Arbitration 11.3% RDS & DS Arbitration 10.4%

IQA: Software Grading Reports

- Grading numbers per grader
- Time spent grading
- Arbitration rate & agreement rate
- R0M0 1 in 10 agreement
- Grading accuracy & kappa score >0.8
- Ungradable rate

Standards for Accredited Graders in Dorset

Test & Training tests	Minimum: 10 sets per year & score >80% on 8 of 10
Grading numbers	> 500 per year for optometrists & >1000 for graders
Grading Accuracy on Intergrader agreement	GA ≥80% with kappa score ≥0.8
1 in 10 R0M0 QA	≥90%
Arbitration rate on primary grading	< 10% with final grade agreement >50%

Individual Grader Review

- Feedback to all
- No single result correctly identifies below standard grading
- Low score in one test / report – look for other evidence
- Be honest and open
- Give constructive feedback

- Provide anonymised data on all graders

Grading Accuracy:
inter grader agreement

Table b: Grader M

Total agreement:		526/562
Proportion agreement:		93.59%
Cohen's Kappa:		0.865
Confidence interval:		0.823 to 0.908

Grader Number 0

Final Grade	R0M0	R1M0	R3SM0	R1M1	R3SM1	R2M0	R2M1	R3AM0	R3AM1	U	Total
R0M0	72	9	0	1	0	0	0	0	0	11	93
R1M0	6	386	0	1	0	0	0	0	0	1	394
R3SM0	0	0	2	0	0	0	0	1	0	0	3
R1M1	0	2	0	28	0	0	0	0	1	0	31
R3SM1	0	0	1	0	4	0	0	0	0	0	5
R2M0	0	0	0	0	0	3	1	0	0	0	4
R2M1	0	0	0	1	0	0	7	0	0	0	8
R3AM0	0	0	0	0	0	0	0	0	0	0	0
R3AM1	0	0	0	0	0	0	0	0	1	0	1
U	0	0	0	0	0	0	0	0	0	23	23
Total	78	397	3	31	4	3	8	1	2	35	562

What does the kappa score mean?

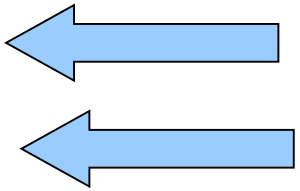
Grading Accuracy Data

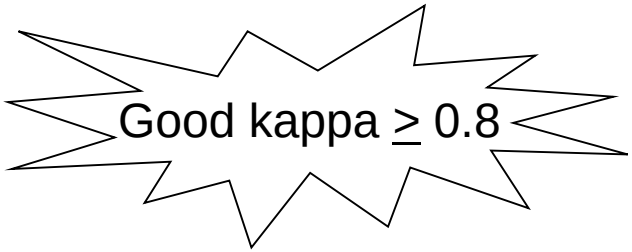
Kappa score

- Based on the proportion agreement %
- Takes into account the ‘case mix’
- Target ≥ 0.8

Table b: Grader M

Total agreement:		526/562
Proportion agreement:		93.59%
Cohen's Kappa:		0.865
Confidence interval:		0.823 to 0.908





Grader Number	0									
Final Grade	**R0M0**	**R1M0**	**R3SM0**	**R1M1**	**R3SM1**	**R2M0**	**R2M1**	**R3AM0**	**R3AM1**	**U**	**Total**
R0M0	72	9	0	1	0	0	0	0	0	11	93
R1M0	6	386	0	1	0	0	0	0	0	1	394
R3SM0	0	0	2	0	0	0	0	1	0	0	3
R1M1	0	2	0	28	0	0	0	0	1	0	31
R3SM1	0	0	1	0	4	0	0	0	0	0	5
R2M0	0	0	0	0	0	3	1	0	0	0	4
R2M1	0	0	0	1	0	0	7	0	0	0	8
R3AM0	0	0	0	0	0	0	0	0	0	0	0
R3AM1	0	0	0	0	0	0	0	0	1	0	1
U	0	0	0	0	0	0	0	0	0	23	23
Total	78	397	3	31	4	3	8	1	2	35	562

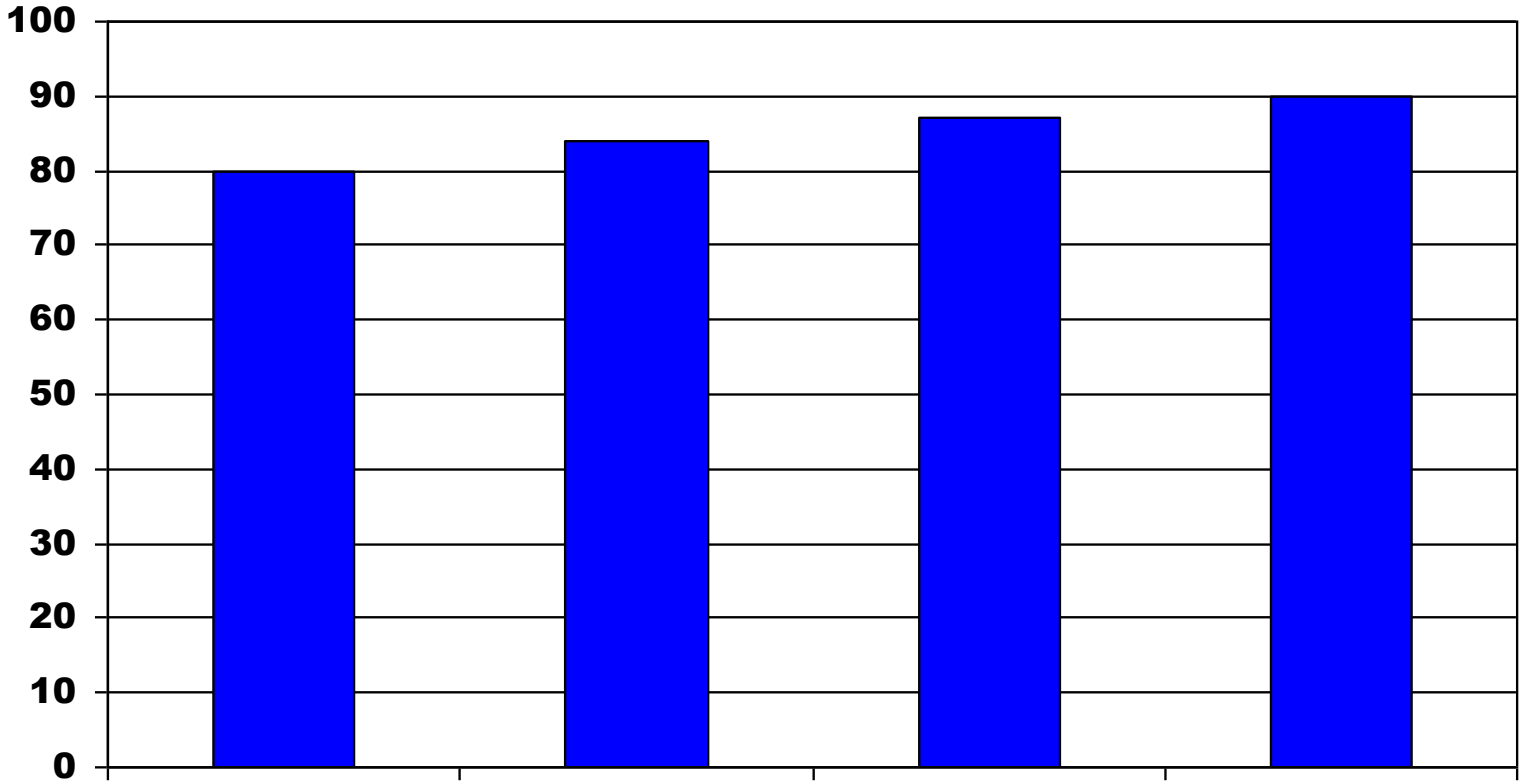
How do Test & Training Sets help assess grading accuracy?

Test & Training Sets

- All test cases have 3 out of 3 Grading College agreement (2012)
- Clear cut cases
- Demonstrate grader can grade accurately in a test

NB New software for better images soon

Dorset Diabetic Eye Screening Programme
Average Test and Training System Score since 2010



Number of graders	42	34	35	35

Test & Training Sets *but*

- Provides only correct score
- R1/R0 error same score as R3/R0
- No set standards
- 80% (8 of 40 wrong) but miss R3 cases
- Is it training or testing?
- Is it robust to detect sub-standard grading?
- It’s not real life grading

What do I want from external QA test?

- Robust, fair test for graders
- Participation is straight forward
- Provide good feedback
- Reassurance of good grading accuracy
- Detect sub – standard grading
- Pin point grading problems
- Bench mark grading accuracy across programme – local & national

Ask the statistician!

- How many cases need to be graded to detect sub-standard grading?
- Use data from 7 test sets from 2012/13
- Since 1.4.12, all TAT cases graded by grading college

Sensitivity

Sensitivity – ability of the test to **correctly** identify a positive case

$$\frac{\text{Positive cases correctly detected}}{\text{Positive cases detected + Positive cases missed}} \times 100$$

Low sensitivity: cases missed

Specificity

Specificity – ability of the test to **correctly** identify people who do not have the condition

Negative cases correctly graded

All negative cases (including false positive cases) x 100

Low specificity means high false positives

Review of T&T results 2012/13:

7 complete test sets

Graders assessed on ability to detect correct outcome not actual DR grade ie Any DR= all grades except R0M0; Referable retinopathy= R2, R3 or M1; and fast track=R3

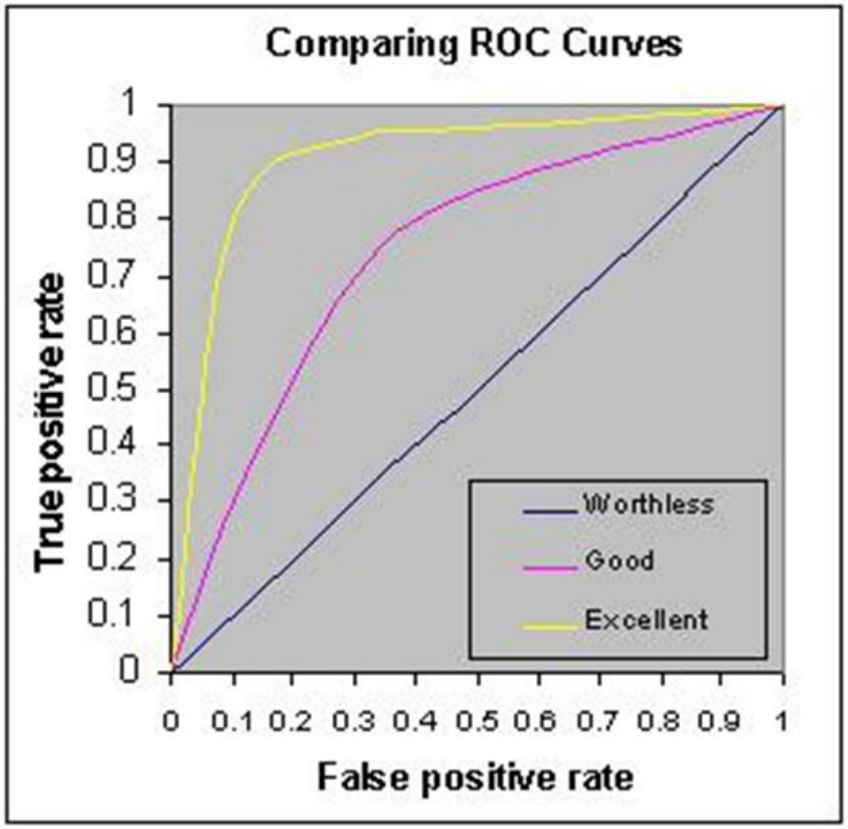
Trainee graders excluded

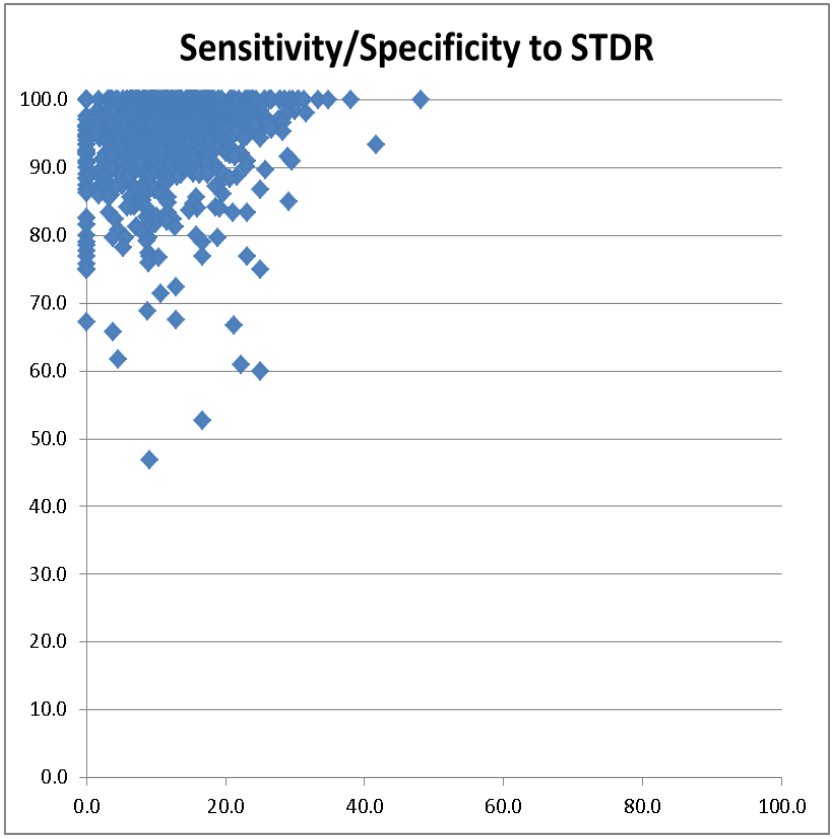
Level of DR	Image-sets completed	Image-sets with correct level of DR	Overall Sensitivity	Overall Specificity
Any DR	**104,575**	**102,181**	**97.7%**	**95%**
Referable DR	**60,386**	**60,155***	**95.1%**	**88%**
Fast track DR	**16,870**	**12,208****	**72.4%**	**72%**

*99.6% image-sets had at least any DR identified so would go to second disease grading in a live programme setting

**99.9% had at least any DR and 93.5% had at least referable DR identified

Test characteristics: STDR





Plot of test results for 1215 Graders

Statistician asks us!

- What is the target sensitivity for good grading accuracy?
- What is the threshold sensitivity for sub-standard grading?
- What proportion of poor graders do you want to detect?
- What will you tolerate in categorising sub-standard grading when it is just chance?

1	2	3	4	5	6	7
P value	**Power**	**Cut off sensitivity level in T&T**	**Target Sensitivity**	**Per cent of cases with Referable DR**	**Required size of test set**	**Comments**
1%	**80%**	**80%**	**95%**	**30%**	**218**	**Starting point**
1%	**70%**	**80%**	**95%**	**40%**	**146**	
1%	**80%**	**80%**	**95%**	**40%**	**164**	
1%	***90%***	***80%***	***95%***	***40%***	***191***	***Best option***
1%	**90%**	**80%**	**95%**	**30%**	**254**	
1%	**90%**	**85%**	**95%**	**30%**	**479**	

Test and Training Scheme

Participation

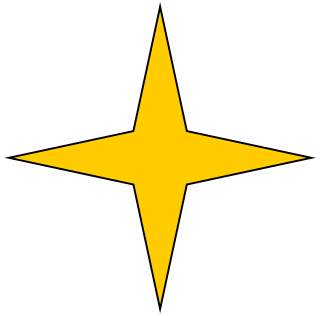
- 6 tests per year

- From 1.4.14, 10 tests per rolling 12 months

TAT: Participation 2013/14 & 2014

Year 2013 / 2014 - 1336 full disease graders registered

- 1164 (87%) participating in 2013/14
- 75% had taken $\geq$ 6 sets
- 49% $\geq$ 10 sets
- 172 (13%) no sets

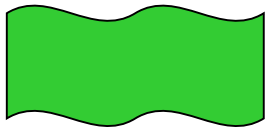
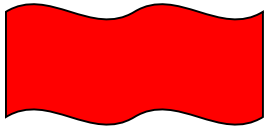


By August 2014

- 1132 (85%) participating
- 208 (16%) no sets

Test & Training Scheme: Participation

- All graders including ROG
- ROG is basis of internal QA assessments
- Minimum 10 sets i.e. 200 image sets per year

- Flag system for participation
  - Green $\geq$ 10 sets in last 12 months 
  - Red < 10 sets in last 12 months 

Setting a standard

So Rates *could be*

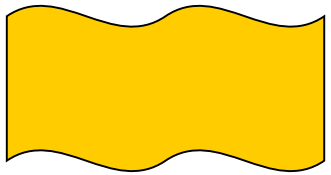
<u>STDR sensitivity</u>

95% Achievable

90% Expected

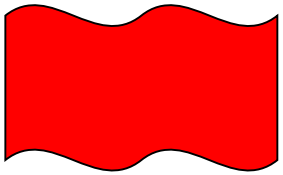
< 85% Yellow Flag Warning

[40 (3.3%) of 1215]



<80% Red Flag remedial threshold

[20 (1.6%) of 1215]



Setting a standard

So Rates *could be*

Specificity

90% Achievable

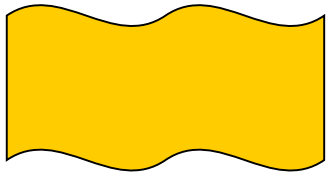
[539 (44%) of 1215]

85% Expected

[903 (74%) of 1215]

<80% Yellow Flag

[126 of 1215]



Test and Training Scheme: the future

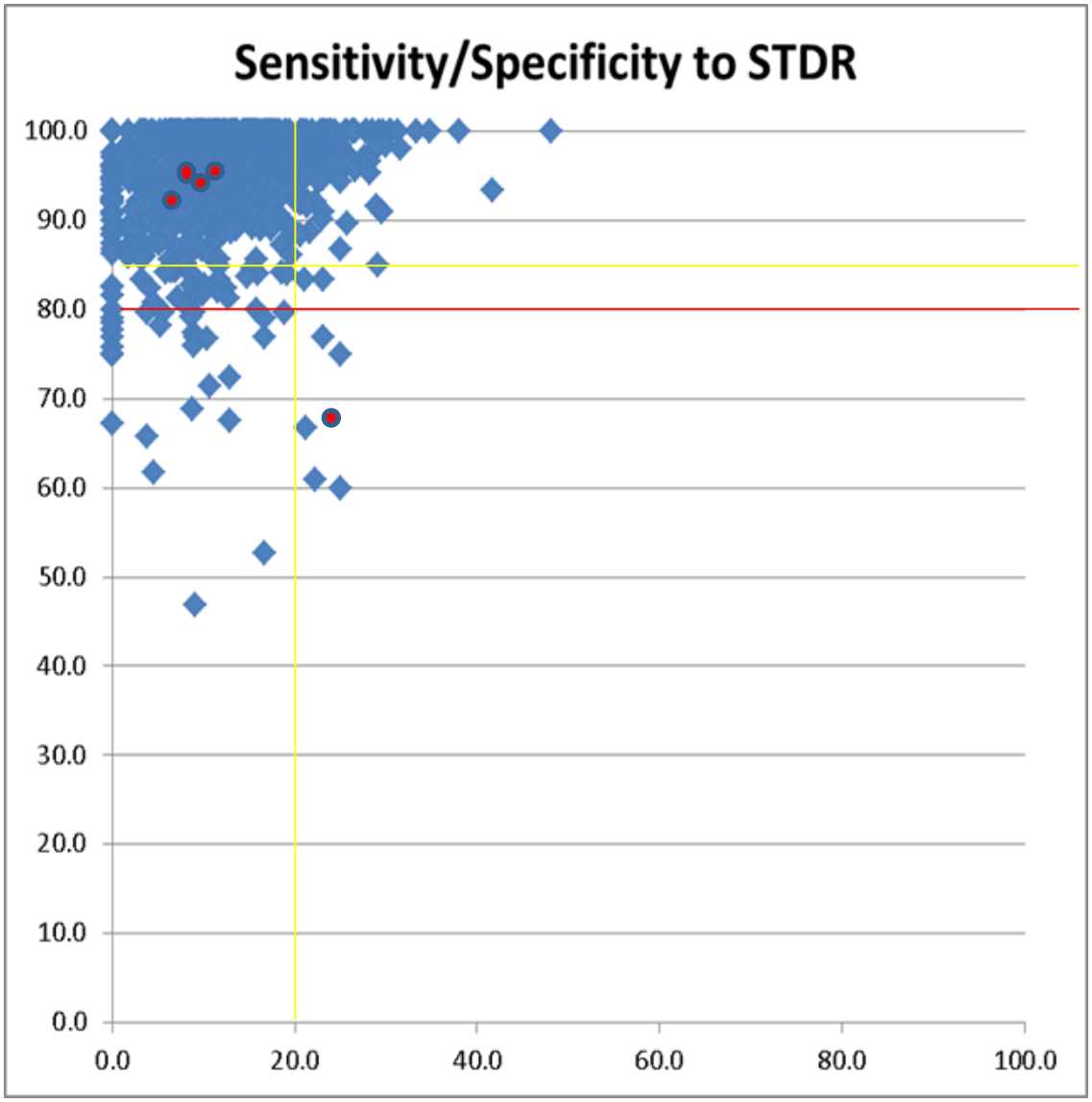
- Flag system for participation: red & green
- Standards for sensitivity & specificity for STDR and flags if below standards
- Reports to clinical lead
- Improved participant feedback report
- Anonymised reports to Programme Board & Regional QA team

DESP Test System Report - All grammes - Rolling 12 months

Report Date Aug 2012 - Aug 2

	Participation			Previous tests STDR Sensitivity				Cumulative sensitivity for the past 12 tests completed	
Grader ID	Programme	Status	Completed last 12 months	T-10 to -12	T-7 to -9	T-4 to -6	Latest 3	For STDR	
								sensitivity %	specificity %
1210	Somewhere	Grader	10	91%	92%	90%	90%	90%	95%
1211	Somewhere	Grader	12	95%	92%	93%	93%	93%	90%
1212	Somewhere	Grader	10	90%	87%	80%	80%	**79%**	82%
1213	Somewhere	Grader	10	90%	92%	95%	95%	95%	95%
1214	Somewhere	Grader	12	94%	95%	97%	97%	95%	95%
1215	Somewhere	Grader	**9**	null	87%	91%	91%	93%	90%
1216	Somewhere	Grader	12	86%	78%	88%	88%	**84%**	95%
1217	Somewhere	Grader	10	100%	100%	98%	98%	95%	95%
1218	Somewhere	Grader	11	98%	97%	95%	95%	98%	100%
1219	Somewhere	Grader	**5**	null	null	89%	89%	92%	90%
1220	Somewhere	Grader	11	95%	97%	94%	94%	92%	95%
1221	Somewhere	Grader	10	100%	98%	98%	98%	100%	**65%**
1222	Somewhere	Trainee	6	null	null	86%	86%	82%	80%
1223	Somewhere	Trainee	3	null	null	null	null	57%	90%

Table must be interpreted according to guidance documentation


Sensitivity/Specificity to STDR
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0.0
20.0
40.0
60.0
80.0
100.0

Screenshot

DESP Test Grade Agreement Report

Programme Name	**Somewhere**	**Start Date**	**1st Mar 2014**
Grader ID	**1124**		
Number of tests	**60**	**End date**	**30th June 2014**

Level	Times Agreed	Times Disagreed	Overgraded		Undergraded	
R0	10	3	as R1	2		
	77%	23%	as R2	0		
			as R3A	1		
R1	22	4	as R2	1	as R0	2
	85%	15%	as R3A	1		
R2	5	8	as R3A	1	as R0	0
	39%	61%			**as R1**	**7**
R3A	5	3			**as R0**	**0**
	62%	33%			**as R1**	**1**
					as R2	2
All R Levels	**42**	**18**	**6**		**12**	
	70%	**30%**	**10%**		**20%**	
M0	34	2	as M1	2		
	94%	6%				
M1	18	6			**as M0**	**6**
	75%	25%				
All M levels	**52**	**8**	**2**		**6**	
	87%	**13%**	**3%**		**10%**	

Summary for Period

Sensitivity to DR	**96**	**Specificity to STDR**	**77**
Sensitivity to STDR	**91**	**Specificity to DR**	**75**
Sensitivity Fastrack	**63**	**Specificity Fastrack**	**94**
Red Box R	**8**	**Red Box M**	**6**

Recommended Pathway for Training Support for Graders

Additional Training Material on
T&T website

TAT & External QA

- Remember TAT only 1 part of the system
- Seek other evidence
- Review internal GA reports

Improving Grading Accuracy: summary

- Set Programme standards for graders
- Set national standards for T&T sets
- Provide feedback
- Provide supportive training
- Ensure continuous CPD

Useful Information Sources

1. Internal Quality Assurance Guidance & Best Practice Toolkit http://diabeticeye.screening.nhs.uk/internal-qa

What are confidence intervals?

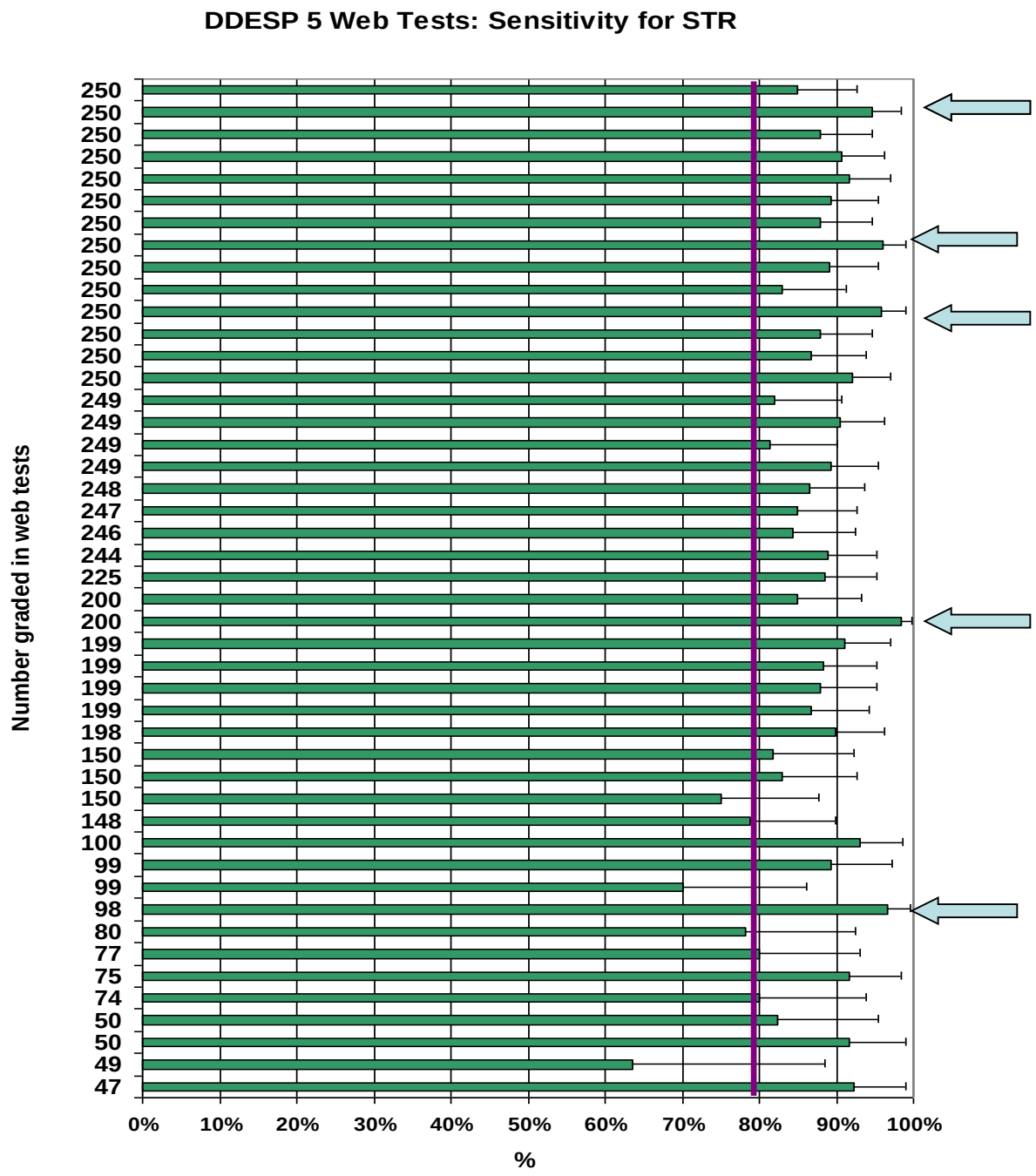

DDESP 5 Web Tests: Sensitivity for STR
Number graded in web tests
250
250
250
250
250
250
250
250
250
250
250
250
250
250
249
249
249
249
248
247
246
244
225
200
200
199
199
199
199
198
150
150
150
148
100
99
99
98
80
77
75
74
50
50
49
47
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
%

Table b: Grader M

Total agreement:		526/562
Proportion agreement:		93.59%
Cohen's Kappa:		0.865
Confidence interval:		0.823 to 0.908

Grader Number	0									
Final Grade	**R0M0**	**R1M0**	**R3SM0**	**R1M1**	**R3SM1**	**R2M0**	**R2M1**	**R3AM0**	**R3AM1**	**U**	**Total**
R0M0	72	9	0	1	0	0	0	0	0	11	93
R1M0	6	386	0	1	0	0	0	0	0	1	394
R3SM0	0	0	2	0	0	0	0	1	0	0	3
R1M1	0	2	0	28	0	0	0	0	1	0	31
R3SM1	0	0	1	0	4	0	0	0	0	0	5
R2M0	0	0	0	0	0	3	1	0	0	0	4
R2M1	0	0	0	1	0	0	7	0	0	0	8
R3AM0	0	0	0	0	0	0	0	0	0	0	0
R3AM1	0	0	0	0	0	0	0	0	1	0	1
U	0	0	0	0	0	0	0	0	0	23	23
Total	78	397	3	31	4	3	8	1	2	35	562