Validation of AOS software for anterior eye complications

Dr Byki Huntjens
City, University of London

21 December 2017 v1
25 May 2018 v2

This study was performed independently at City, University of London, by Dr Byki Huntjens Senior Lecturer in the Division of Optometry and Visual Sciences. It investigates the reliability (repeatability) of, and agreement (variability) between a series of grading scales, subjectively and objectively. Grading was completed by a student optometrist (MB) and an experienced optometrist (BH) to investigate the agreement between observers. The work was commissioned by AOS (Advanced ophthalmic Systems).

Grading systems included were Efron and CCLRU grading scales, both subjective, and objective AOS software.

Summary:

1. Excellent reliability of the objective and novel AOS grading system, which is significantly improved in comparison to the subjective grading systems Efron and CCLRU
2. Agreement between the AOS and Efron system was good, with an average difference between the gradings of 0.38 units and LoA of approximately ±1 unit.
3. Agreement between the AOS and CCLRU systems was reduced in comparison to Efron, with an average difference between the gradings of 1.25 units and LoA of approximately ±1.5 units.
4. Agreement between observers is significantly improved using the objective AOS software compared to subjective grading scales. We observed moderate agreement between a novel and an experienced grader when assessing for bulbar hyperaemia, while palpebral hyperaemia showed good agreement between the two types of graders.

In conclusion, the objective AOS system is more reliable than the subjective methods of grading; however, the three systems cannot be used interchangeably.

Introduction

A grading scale can be defined as 'A tool that enables quantification of the severity of a condition with reference to a set of standardised descriptions or illustrations' (Efron, 2012). The standardised descriptions can be written (Woods 1989), art illustrations (Efron, 2000 and 2012; Schnider, 1990), and photographic scales (McMonnies and Chapman-Davies, 1987; CCLRU, 1997; Anderson et al., 1996).

The two main grading scales used in an optometric clinical practice include the original CCLRU (rebranded IER and more recently published under the name of Brien Holden Vision Institute) and Efron scales. Both have advantages and disadvantages. The Efron grading scale consists of a series of artist illustrated depictions of 16 different conditions using a 0-4 scale, while the CCLRU consists of photographs of 6 conditions, two of which are presented in multiple manifestations. The latter have been criticised for the lack of consistency between images representing the same condition, due to the use of different eyes, illumination, and area under display. Although the Efron grading scale overcomes this due to artistic clarity between the grades, it is not representing a real-life situation.

Although the grading scales all vary in the descriptors adopted to denote the severity, all grading scales describe Grade 0 as 'absent', 'none' or 'normal'. Grade 1 represents 'slight', 'trace' or 'very slight'. Grade 2 is commonly known as 'mild' or 'slight'. And Grade 3 and 4 are described as 'moderate' and 'severe', respectively.

Despite the apparent consistency, there is large variation between the grading scales, and it has been recommended that these cannot be used interchangeably in a clinical setting without the use of a personalised correction factor (Efron et al., 2001). In addition, grading of clinical conditions should be completed to the nearest 0.1 grading scale unit (Bailey et al., 1991), which optimises grading sensitivity. However, this will be at the expense of observer concordance, i.e. agreement between repeated measurements by the same observer, and between measurements by different observers.

More recently, computerised image analysis techniques have been used for grading anterior eye characteristics. Wolffsohn (2001) reported that different studies have used a combination of thresholding (Fieguth et al., 2002; Owen et al., 1996; Papas 2000; Chen et al., 1987; Guillon and Shah 1996; Simpson et al., 1998), edge detection (Owen et al., 1996; Villumsen et al., 1991; Maldonado et al., 1997), smoothing (Fieguth et al., 2002; Owen et al., 1996; Villumsen et al., 1991; Willingham et al., 1995), colour extraction (Fieguth et al., 2002; Papas 2000; Simpson et al., 1998; Willingham et al., 1995), and morphometry and densiometry (Horack et al., 1996) to grade redness of the eye, clinically known as bulbar hyperaemia. One study suggesting that the number of vessels and the proportion of the image occupied by vessels are more important than relative colouration (Papas 2000) whereas another indicated both these factors were integral to grading (Fieguth et al., 2002). However, the correlation between the computer image analysis techniques used and clinician grading was not linear, and was more discrepant for higher grades of bulbar hyperaemia (Fieguth et al., 2002). Less research has been conducted on the objective grading of palpebral hyperaemia (redness under the eyelid) and corneal staining (damaged or displaced corneal cells, visible with the use of fluorescein sodium dye), although it has been noted that there are significant differences between observers in subjective grading of these features (Begley et al., 1996; Mackinven et al., 2001). It was concluded that the printed grading scales have a higher sensitivity for grading features of low severity. Grading features such as palpebral hyperaemia and corneal staining are complex and there is a compromise between the simplicity of a single scale and the ability to fully describe and monitor changes in the feature. Edge detection and colour extraction image analysis techniques were highly repeatable and offer the potential for more repeatable and sensitive grading than using printed subjective grading scales.

The AOS software was designed to objectively grade bulbar (eye) and palpebral (under the eyelid) redness based on the CCLRU grading scale. The software is also able to determine the number of corneal staining but this result does not represent a single corneal staining grading as per the printed grading scales.

The main aim of this study was to validate the objective grading software by AOS by comparing its results to two existing subjective grading scales Efron and CCLRU. We report grading reliability of all grading scales and agreement between the grading scales as well as different type of observers, and document expected performance of the AOS system in clinical practice.

Methods

A database of n=30 bulbar and n=26 palpebral conjunctival redness images were selected from a private database, the International Association of Contact Lens Educators slide collection, and the internet. The data set included a large variety of severities of both conditions, ranging from normal to severe. All images were labelled numerical, and displayed in full colour (24 bit) on a desktop computer with monitor resolution (1920 x 1080 pixels).

For valid comparison between the three grading scales, we investigated the following:

1. **Bulbar conjunctival hyperaemia**. This is called conjunctival redness in Efron grading scale; 5 images covering 0-4 grading from normal to severe (Figure 1). In the CCLRU grading scale this is called 'bulbar redness'; 4 images covering 1-4 grading from very slight to severe (Figure 2).
2. **Palpebral conjunctival hyperaemia**. This is not available in the Efron grading scale. In the CCLRU, we used the 'lid redness (area 2)'; 4 images covering 1-4 grading from very slight to severe (Figure 3). Area 2 represents the middle section under the lid, as shown in Figure 4.

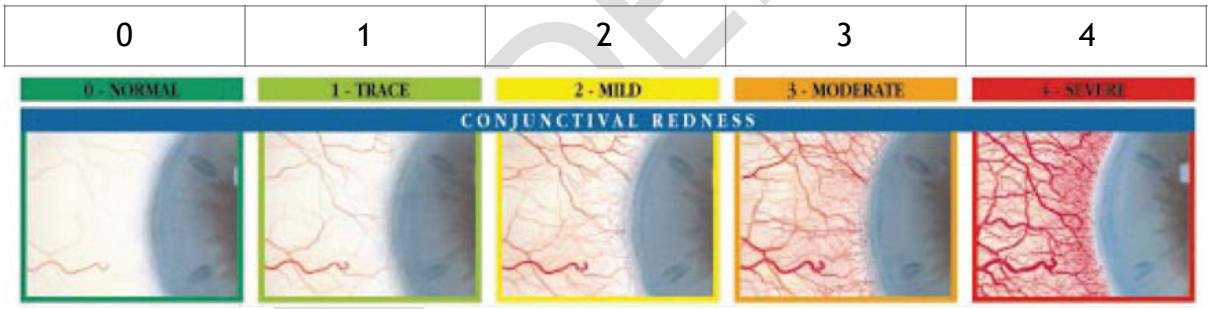

Figure 1. Efron grading scale for bulbar conjunctival hyperaemia (0-4 grade)

Figure 2. CCLRU grading scale for bulbar conjunctival hyperaemia (1-4 grade)

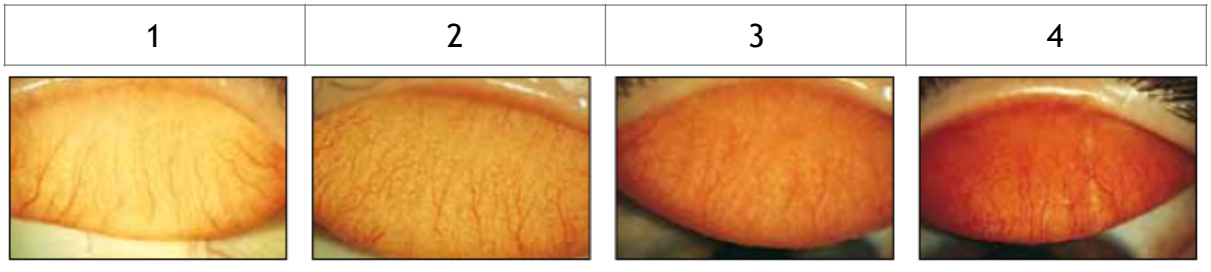

Figure 3. CCLRU grading scale for palpebral conjunctival hyperaemia (1-4 grade)

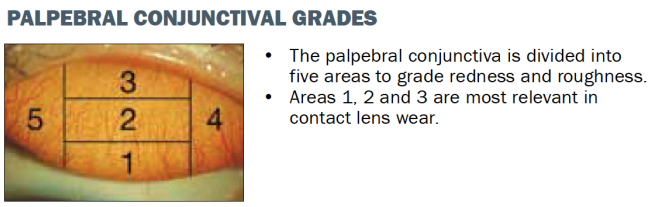

Figure 4. Overview of lid assessment zones when grading palpebral hyperaemia. Area 2 is used for grading.

Independently of each other, one student optometrist (MB) and one experienced clinical optometrist (BH) graded all bulbar hyperaemia images in a randomised order using the Efron grading scale (Figure 1). Randomisation was completed using an electronic software available online (https://www.random.org/integer-sets/). After a 1-hour break and masked to earlier results, all bulbar and palpebral hyperaemia images were randomised and graded using the CCLRU grading scale (Figure 2 and Figure 3). This was again repeated after one hour in a randomised order using the AOS grading software (Figure 5 and Figure 6). All steps as described above were repeated on a different day (visit 2). Each observer was therefore required to make a combined total of 336 grading estimates over both days: *bulbar hyperaemia* 30 images x 3 grading scales x 2 sessions plus *palpebral hyperaemia* 26 images x 3 grading scales x 2 sessions).

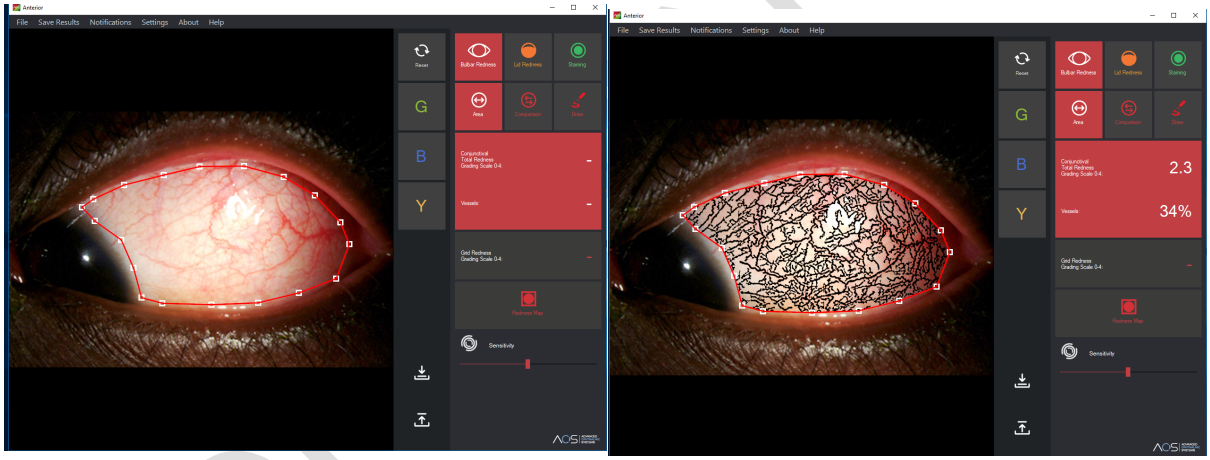

Figure 5. Selection of the area of interest using the AOS software for grading bulbar hyperaemia (left image). Bulbar conjunctival hyperaemia grade is displayed on right hand side (2.3 units; right image).

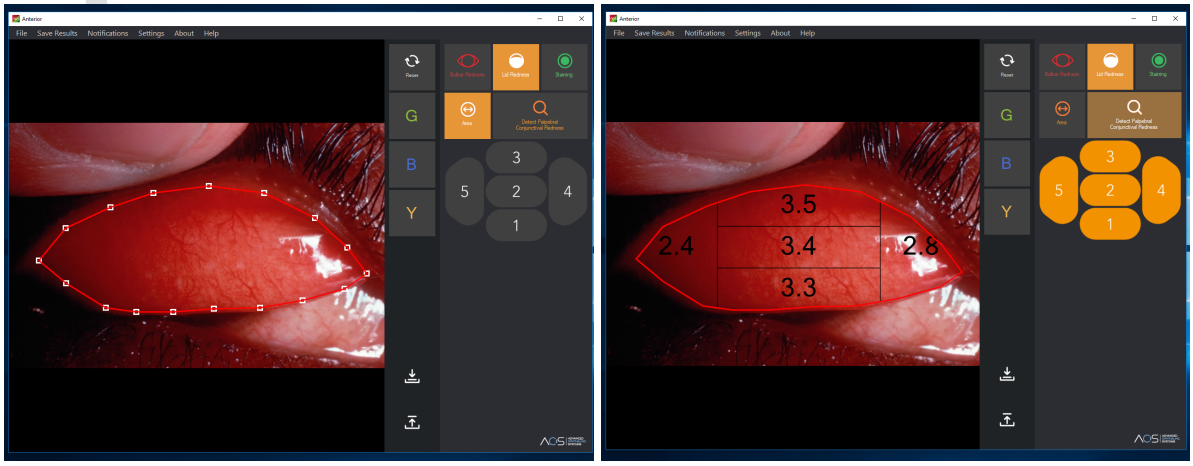

Figure 6. Selection of the area of interest using the AOS software for grading palpebral hyperaemia (left image). Palpebral conjunctival hyperaemia grades over 5 areas are displayed directly on the image (area 2: 3.4 units; right image)

Grading reliability
Grading reliability is the ability of the grader to give similar results when the process is repeated. This represents the intra-observer variability in grading. We calculated the numeric differences between test and retest grading estimates, between repeat measurements using the same grading scale. The standard deviation of this discrepancy distribution describes grading reliability. The 95% confidence intervals within which grading estimate cannot be considered to differ is taken as (1.96 x reliability). This is known as the coefficient of repeatability (Bland and Altman, 1986).

Grading agreement
It is most unlikely that different methods will agree exactly by giving the identical result for all individuals. We are interested to know by how much the new method is likely to differ from the old: if this is not enough to cause problems in clinical interpretation we can replace the old method (subjective grading) by the new (objective grading) or use the two interchangeably.

Agreement between two methods of grading is when both methods give similar results. This represents the between-method variability. To estimate agreement between the methods, we calculated the numeric differences between the three grading scales (Efron versus AOS; CCLRU versus AOS; and CCLRU versus Efron) when measured during the second session. The 95% confidence intervals within which grading estimate cannot be considered to differ is taken as (1.96 x variability; Bland and Altman, 1986).

For all three grading methods, we also investigated the between-observer variability in grades obtained during visit 2 between one novel (student MB) and one experienced (optometrist BH) observers.

None of the data sets were found to be statistically significant from a normal distribution, as checked with the Kolmogorov-Smirnov test ($p>0.05$).

Results

Thirty images were graded for bulbar hyperaemia, and after deletion of 2 images due to incomplete lid area 2, 24 images were graded for palpebral hyperaemia.

Grading reliability

Intra-observer reliability

The reliability data for all images per grading scale is shown below. The difference between the two sessions was only statistically significant when grading bulbar hyperaemia using the CCLRU grading system ($t(29)=3.143$; $p = 0.004$). All other results were not statistically different between session 1 and 2 for bulbar or palpebral hyperaemia ($p>0.05$). Reliability scores with the AOS system were lowest, indicating better reliability for bulbar as well as palpebral hyperaemia when compared to those graded subjectively (see Table 1). Subjective grading of bulbar hyperaemia was less reliable than palpebral hyperaemia. Using the objective AOS grading system, there was little difference between the reliability of bulbar and palpebral hyperaemia.

Table 1. Grading reliability data per grading method (between sessions).

	Bulbar hyperaemia			**Palpebral hyperaemia**	
	Efron	**CCLRU**	**AOS**	**CCLRU**	**AOS**
Sample size	30	30	30	24	24
Mean ± SD session 1	2.21 ± 1.14	3.13 ± 0.60	1.80 ± 1.37	2.41 ± 1.22	2.46 ± 1.18
Mean ± SD session 2	2.16 ± 1.14	2.98 ± 0.72	1.81 ± 1.40	2.43 ± 1.05	2.46 ± 1.17
Mean difference	-0.05	-0.15	0.017	0.021	<0.0005
Reliability	0.31	0.26	0.06	0.40	0.05
Coefficient of Repeatability	0.62	0.50	0.13	0.78	0.10
95% LoA	0.57 to -0.66	0.35 to -0.65	0.14 to -0.11	0.80 to -0.76	0.10 to -0.10
T-test (between sessions)	P=0.42	**P=0.004***	P=0.17	P=0.80	P=1.00
R^2 value of regression equation (between sessions)	0.926	0.885	0.998	0.9101	0.998

Grading agreement

Inter-method agreement

Agreement between the three grading scales is shown below. Paired sample t-tests were conducted to evaluate the agreement between two different grading systems (see Table 3). A one-way repeated measures ANOVA was conducted to compares scores between the three methods for bulbar hyperaemia. There was a statistically significant difference between the three methods ($F(2,28)=40.34$, $p<0.0005$, multivariate eta squared = 0.74). Post hoc analysis revealed that the mean (± SD) grades using the AOS grading scale (1.81 ± 1.39) was significantly lower than the Efron (2.19 ± 1.13; $p=0.01$) and CCLRU (3.06 ± 0.65; $p<0.0005$). The results from the Efron grading scale were significantly lower than those from the CCLRU ($p<0.0005$). All showed a large effect size (partially eta squared in Table 3),

A paired sample t-test was conducted to evaluate the agreement between two different grading methods for palpebral hyperaemia. There was no statistically significant difference between the two methods (t(23)=-0.355, p=0.73; Table 3).

Table 3. Grading agreement data per grading method (between systems). The average grade between two sessions was used to calculate the differences between the systems.

	Bulbar hyperaemia			**Palpebral hyperaemia**
	Efron (method 1) vs CCLRU (method 2)	**CCLRU (method 1) vs AOS (method 2)**	**Efron (method 1) vs AOS (method 2)**	**CCLRU (method 1) vs AOS (method 2)**
Sample size	30	30	30	24
Mean ± SD method 1	2.16 ± 1.14	2.98 ± 0.72	2.16 ± 1.14	2.42 ± 1.12
Mean ± SD method 2	2.98 ± 0.72	1.81 ± 1.40	1.81 ± 1.40	2.46 ± 1.17
Mean difference	0.82	-1.25	-0.38	0.040
95% LoA	1.90 to -0.26	0.56 to -2.90	0.86 to -1.56	1.11 to -1.03
T-test (between 2 methods)	**P<0.0005***	**P<0.0005***	**P=0.004***	P=0.73
Effect size (partially eta squared)	0.73 (large effect)	0.67 (large effect)	0.26 (large effect)	0.005 (small effect)
R^2 value	0.856	0.614	0.810	0.788

Bland and Altman plots are shown below, to visualise the mean of the differences between two grading scales including the 95% limits of agreement (LoA). The continuous red line represents the mean of the differences, also known as the line of agreement; it is the systematic difference or *estimated bias* between the two methods. It is bound by two parallel dotted lines which represents the 95% LoA above and below the line of agreement. The LoA shows how far apart measurements by the two methods are likely to be for most individuals. Narrow LoA imply a better agreement between the two methods.

Bulbar hyperaemia - CCLRU versus AOS

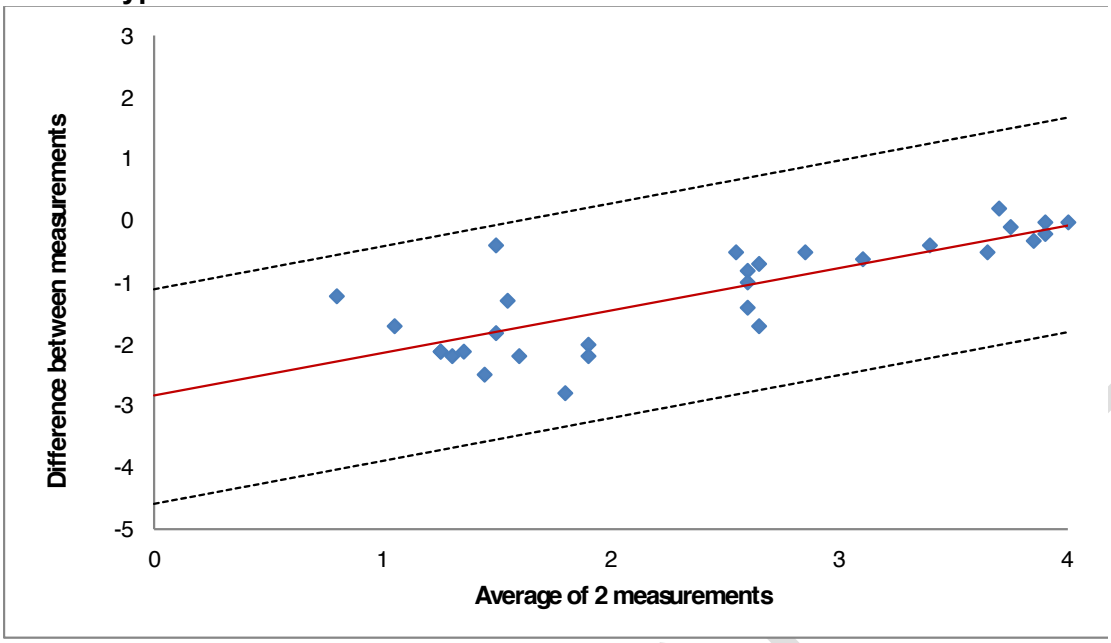

Figure 7. Difference versus means plot between CCLRU and AOS grading results for bulbar hyperaemia.

The mean difference between the two methods was found to be -1.17 units of grading, indicating that the average subjective grade using CCLRU is approximately 1 higher in comparison to the objective AOS software. As shown in Figure 7 by the slope of the red line, the methods do not equally agree through the whole range of bulbar hyperaemic severities: that there is a tendency for the mean difference to improve with increased grades. This could also be explained by the fact that the AOS grading method varies between 0 and 4 units, while CCLRU is based on grades 1 to 4. It is unclear how much each of these factors contribute to this finding.

Bulbar hyperaemia - Efron versus AOS

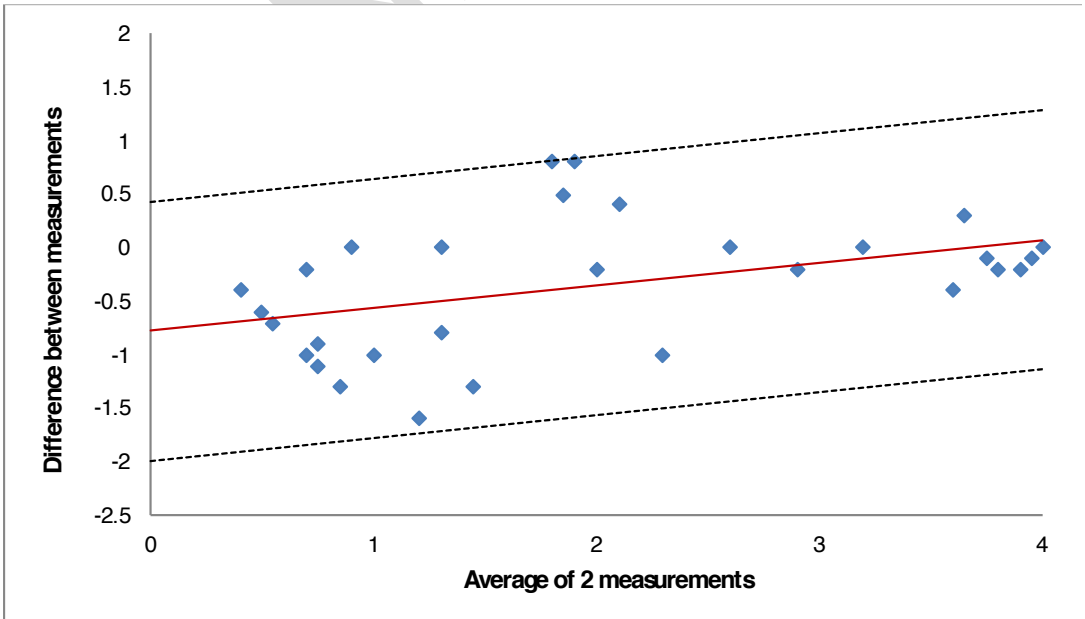

Figure 8. Difference versus means plot between Efron and AOS grading results for bulbar hyperaemia.

The mean difference between the two methods was found to be 0.35 units of grading, indicating that the average grade is approximately 0.4 grade higher using subjective Efron grading scale in comparison to the objective AOS software.

The data implies that the agreement between the AOS software and the Efron grading scale is closer than compared to CCLRU. Because both grading systems use a scale from 0 to 4 units, it is not surprising to observe improved agreement between the two methods. However, as a result it can be said with more certainty that these two methods do not equally agree through the whole range of gradings, as the mean difference improves with increased severity of bulbar hyperaemia. As can be seen from the slope of the red line, this increase is significantly less steep compared to the red line in Figure 7.

Palpebral hyperaemia - CCLRU versus AOS

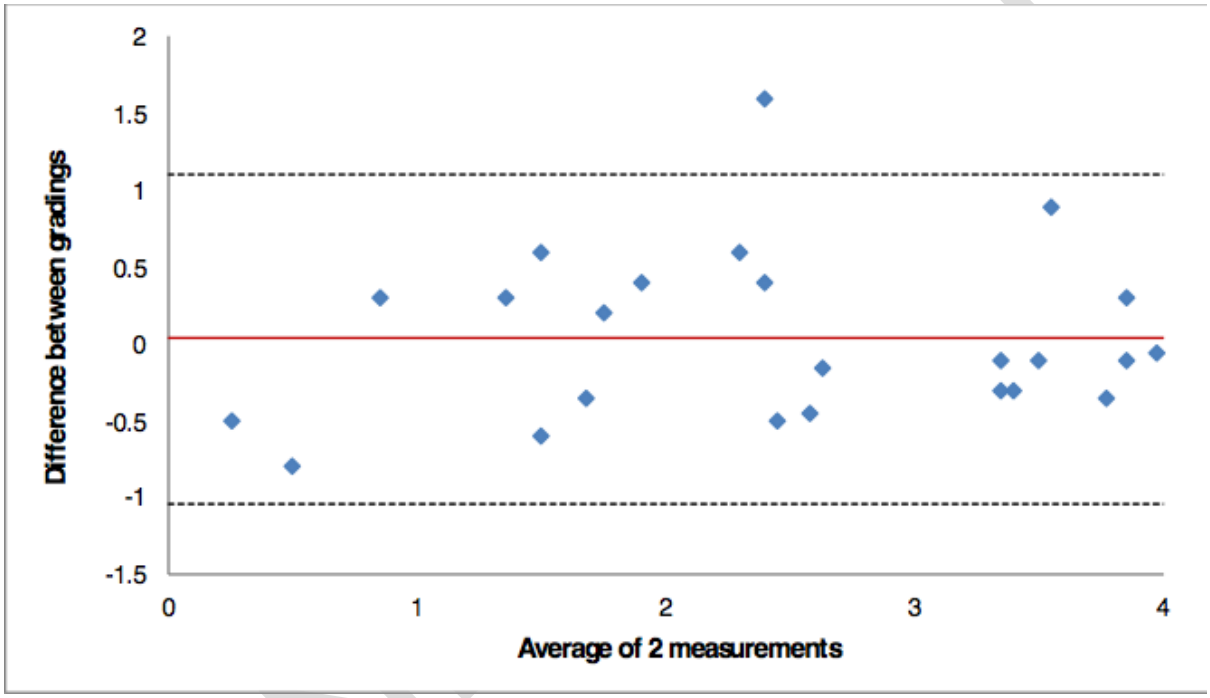

Figure 9. Difference versus means plot between CCLRU and AOS grading results for palpebral hyperaemia.

The mean difference between the two methods was found to be close to zero, indicating that a subjective grade using the CCLRU is systematically higher by 0.04 in comparison to the objective AOS software. This represents excellent agreement between the two methods, and similar variability to bulbar hyperaemia results: 95% of the results were spread over a total of 2 grading units (Figure 9).

Inter-observer agreement

The difference between the two observers was statistically significant when grading bulbar and palpebral hyperaemia using the Efron and the CCLRU grading systems (Table 2), whereby the experienced observer graded higher than the novel (student) observer. There was no significant difference between the two observers when using the AOS grading method for either form of hyperaemia (palpebral and hyperaemia ($P>0.05$), although the experienced observer did record slightly higher grades for both palpebral and bulbar hyperaemia.

Table 2. Grading reliability data per grading method (between observers).

	Bulbar hyperaemia			Palpebral hyperaemia	
	Efron	**CCLRU**	**AOS**	**CCLRU**	**AOS**
Sample size	30	30	30	24	24
Mean ± SD experienced	2.16 ± 1.14	2.98 ± 0.72	1.81 ± 1.40	2.43 ± 1.05	2.46 ± 1.17
Mean ± SD student	1.86 ± 1.2	2.52 ± 1.00	1.76 ± 1.32	2.21 ± 1.08	2.45 ± 1.15
Mean difference	0.30	0.47	0.05	0.08	0.017
Reliability	0.37	0.48	0.20	0.78	0.06
Coefficient of Repeatability	0.73	0.95	0.39	1.54	0.11
95% LoA	1.03 to -0.42	1.41 to -0.48	0.44 to -0.34	1.61 to -1.46	0.13 to -0.09
Agreement κ	0.037 Poor	0.109 Poor	0.401 Moderate **P<0.0005***	0.067 Poor	0.646 Good **P<0.0005***
T-test (between observers)	**P<0.0005***	**P<0.0005***	P=0.18	**P=0.023***	P=0.162
R^2 value	0.904	0.802	0.982	0.829	0.998

Subjective and objective grading of bulbar hyperaemia was more variable between observers than palpebral hyperaemia. However, the reliability of the AOS method was improved for bulbar as well as palpebral hyperaemia when compared to the subjective methods of grading. In addition, the agreement between the two observers improved significantly when using an automated method of grading, resulting in moderate agreement for bulbar hyperaemia (κ=0.401; P<0.0005) and good agreement for palpebral hyperaemia (κ=0.401; P<0.0005).

Agreement in gradings between the two observers are shown in the figures below:

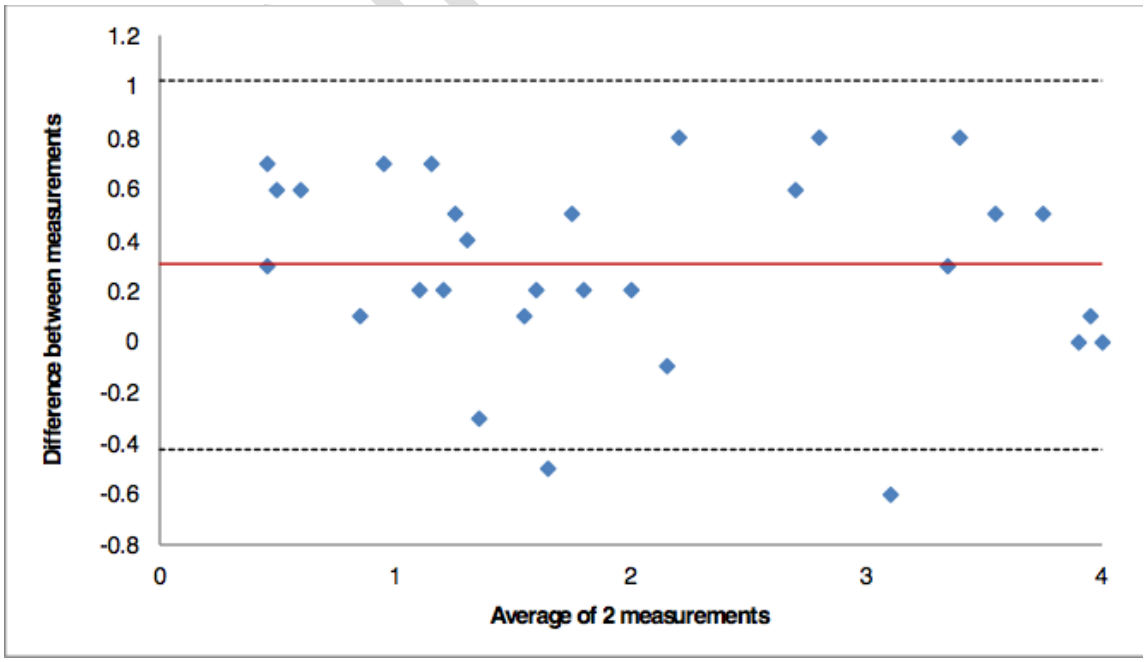

Figure 10. Inter-observer agreement of bulbar hyperaemia using Efron grading scale (data from visit 2).

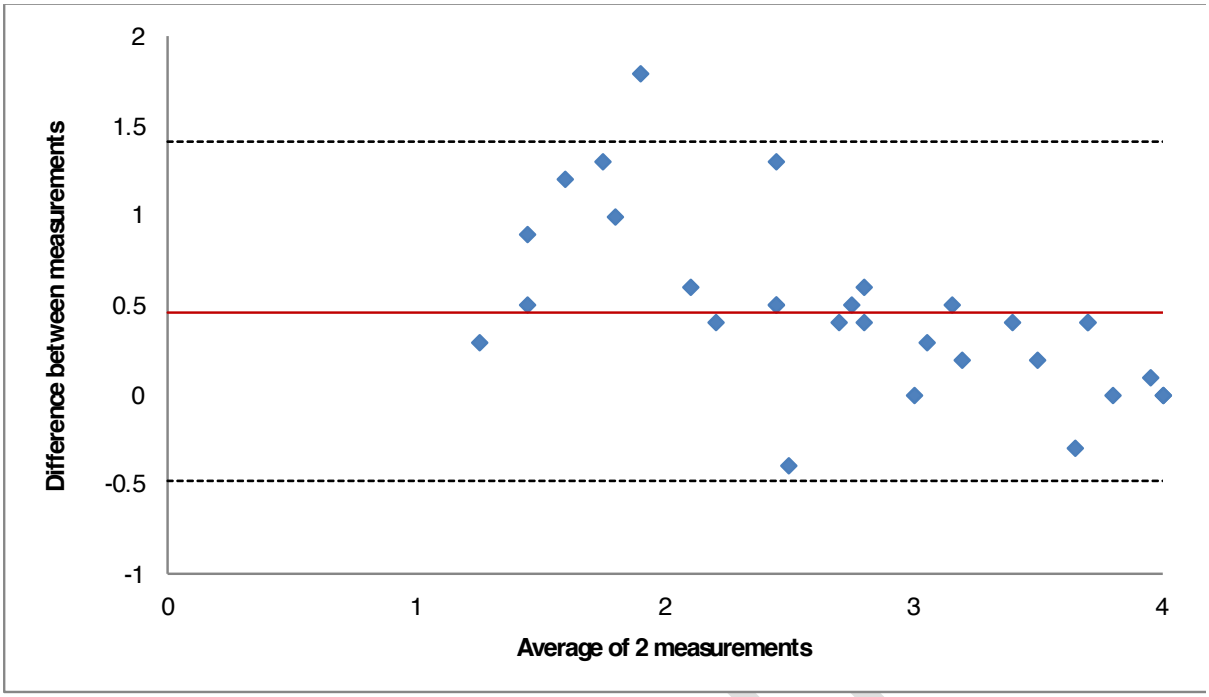

Figure 11. Inter-observer agreement of bulbar hyperaemia using CCLRU grading scale (data from visit 2).

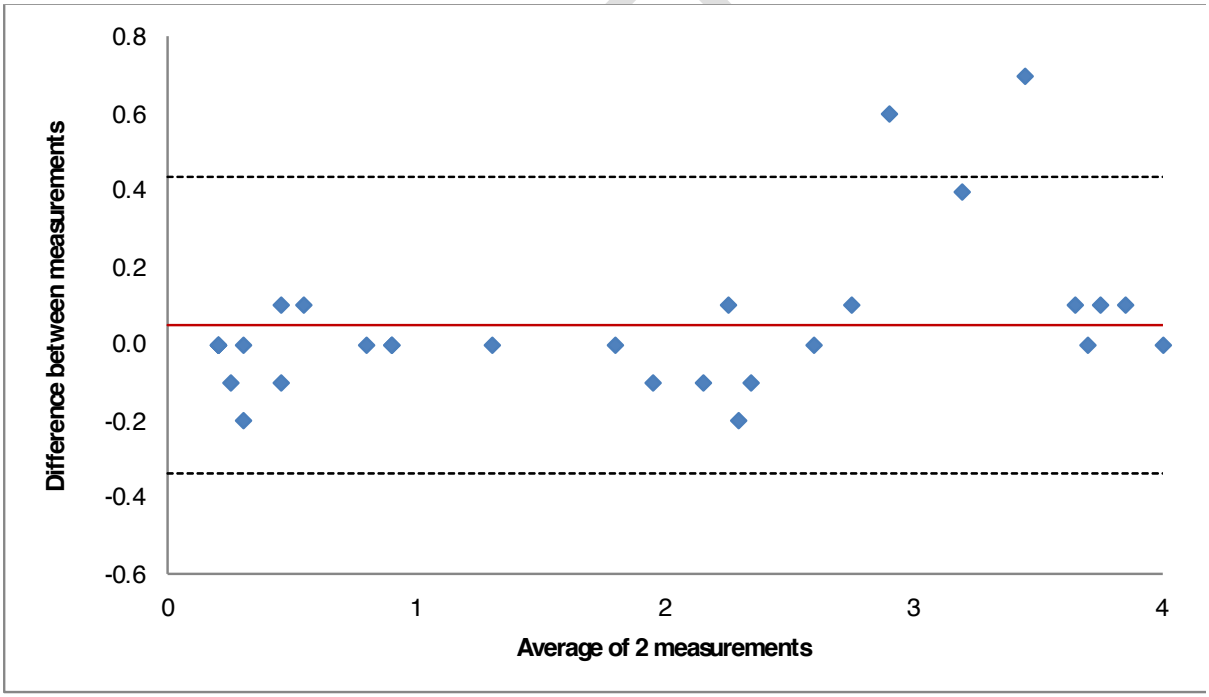

Figure 12. Inter-observer agreement of bulbar hyperaemia using AOS grading method (data from visit 2).

The Figures 10-12 visualise the narrow LoA observed when using the automated AOS software compared to larger LoA when using the subjective grading scales, indicating much improved agreement between observers when grading images independently.

Note

Although corneal staining was not included in this study, it is expected that the ability of the software to count corneal staining would be hugely beneficial to clinicians, with respect to follow up of ocular conditions and evaluation of its management plan.

Conclusion

This study investigated the reliability and agreement between multiple subjective and one novel objective grading system. Images of palpebral (eye lid) and bulbar (eye) hyperaemia (redness) were graded by two independent observers (novel MB and experienced BH) during two different sessions using all three grading systems (Efron, CCLRU, AOS). Images were randomised, and breaks were introduced to reduce bias.

The study found that there is:

1. Excellent reliability of the AOS grading system, which is significantly improved in comparison to the subjective grading systems Efron and CCLRU
2. Agreement between the AOS and Efron system was good, with an average estimated bias between the gradings of 0.38 units and LoA of approximately ±1 unit.
3. Agreement between the AOS and CCLRU systems was reduced in comparison to Efron, with an average estimated bias between the gradings of 1.25 units and LoA of approximately ±1.5 units.
4. Agreement between observers is significantly improved using the objective AOS software compared to subjective grading scales. We observed moderate agreement between a novel and an experienced grader when assessing for bulbar hyperaemia, while palpebral hyperaemia showed good agreement between the two types of graders.

Although according to the manufacturer, the AOS algorithms are based on both Efron and CCLRU grading scale. However, the CCLRU grading scale starts at grade 1 while the AOS software produces gradings on a scale from 0 to 4 for both bulbar and palpebral hyperaemia. This explains why the agreement between Efron and AOS is improved compared to AOS and CCLRU.

In conclusion, the objective AOS system is more reliable than the subjective methods of grading; however, the three systems cannot be used interchangeably.

References

- Andersen JS, Davies IP, Kruse A, Lofstrom T, and Ringmann LA. 1996. *A handbook of contact lens management*. Vistakon, Jacksonville, FL, USA.
- Begley CG, Barr JT, Edrington TB, et al. 1996. Characteristics of corneal staining in hydrogel contact lens wearers. Optom Vis Sci, 73:193–200.
- Bland JM, Altman D. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476):307-10.
- CCLRU. 1997. Centre for Contact Lens Research Unit grading scales (Appendix D). In: Contact Lenses (eds AJ Phillips ans L Speedwell), Butterworth-Heinemann, Oxford, UK.
- Chen PCY, Kovalcheck SW, Zweifach BW. Analysis of microvascular network in bulbar conjunctiva by image processing. Int J Microcirculation Clin Exp 1987;6:245–55.
- Efron N. 2000. *Efron grading scales for contact lens complications.* The Millenium Edition, Hydron, Farnborough, UK.
- Efron N. 2012. *Contact Lens Complications E-Book*. Elsevier Health Sciences.
- Efron N, Morgan PB, and Katsara SS. 2001. *Validation of grading scales for contact lens complications.* Ophthalmic and Physiological Optics, *21*(1), pp.17-29.
- Fieguth P, Simpson TL. 2002. Automated measurement of bulbar redness. Invest Ophthalmol Vis Sci, 43:340–7.
- Guillon M , Shah D. Objective measurement of contact-lens induced conjunctival redness. Optom Vis Sci 1996;73:596–605.
- Horak F , Berger U, Menapace R, et al. Quantification of conjunctival vascular reaction by digital imaging. J Allergy Clin Immunology 1996;98:495–500.
- Mackinven J , McGuinness CL, Pascal E, et al. 2001. Clinical grading of the upper palpebral conjunctiva of non-contact lens wearers. Optom Vis Sci, 78:13–18.
- Maldonado M , Arnau V, Martinez-Costa R, et al. Reproducibility of digital image analysis for measuring corneal haze after myopic photorefractive keratectomy. Am J Ophthalmol 1997;123:31–41.
- Mcmonnies CW, Chapman-Davies A. 1987. *Assessment of conjunctival hyperemia in contact lens wearers. Part I.* Optometry and Vision Science, 64(4):246-50.
- Owen CG, Fitzke FW, Woodward EG. A new computer assisted objective method for quantifying vascular changes of the bulbar conjunctivae. Ophthal Physiol Opt 1996;16:430–7.
- Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. Invest Ophthalmol Vis Sci 2000;41:687–91.
- Schnider CM. 1990. Rigid gas permeable extended wear. Contact Lens Spectrum, 5(9):101-6.
- Simpson TL, Chan A, Fonn D. Measuring ocular redness: first order (luminance and chromaticity) measurements provide more information than second order (spatial structure) measurements. Optom Vis Sci 1998;75:279.
- Villumsen J , Ringquist J, Alm A. Image analysis of conjunctival hyperaemia: a personal computer based system. Acta Ophthalmol 1991;69:536–9.
- Willingham FF, Cohen KL, Coggins JM, et al. Automatic quantitative measurement of ocular hyperaemia. Curr Eye Res 1995;14:1101–8.
- Wolffsohn JS. Incremental nature of anterior eye grading scales determined by objective image analysis. British journal of ophthalmology. 2004 Nov 1;88(11):1434-8.
- Woods R. 1989. Quantitative slit lamp observations in contact lens practice. Journal of the British Contact Lens Association, 12:42-5.