



City Research Online

City, University of London Institutional Repository

Citation: Huntjens, B., Basi, M. and Nagra, M. (2019). Evaluating a new objective grading software for conjunctival hyperaemia. Contact Lens and Anterior Eye, doi: 10.1016/j.clae.2019.07.003

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/22543/>

Link to published version: <http://dx.doi.org/10.1016/j.clae.2019.07.003>

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

1 **Title: Evaluating a new objective grading software for conjunctival hyperaemia**

2
3 **Byki Huntjens^{a*}, Mandeep Basi^a, Manbir Nagra^b**

4 ^a Centre for Applied Vision Research, Division of Optometry and Visual Science,
5 City, University of London, United Kingdom

6 ^b School of Health Sciences and Social Work, Faculty of Science, University of
7 Portsmouth, United Kingdom

8
9
10 **Abstract**

11 *Background/ Aims:* Standardised numeric grading scales are used in ophthalmic
12 practice to improve consistency between clinicians in recording the severity of ocular
13 conditions and to facilitate the monitoring of such changes. We investigated the intra-
14 and inter-observer grading reliability and the agreement between subjective Cornea
15 and Contact Lens Research Unit (CCLRU) and Efron grading scales as well as a
16 new Advanced Ophthalmic Systems (AOS) software which uses an objective
17 approach to grading conjunctival hyperaemia.

18 *Methods:* One experienced observer graded n=30 bulbar and n=26 palpebral
19 conjunctival hyperaemia images to 0.1 increments. Masked grading of randomised
20 images was undertaken for all three methods, on two separate occasions. The
21 agreement within and between the grading methods was assessed between
22 sessions, and compared to the results of a novice observer.

23 *Results:* There were no statistically significant differences ($P > 0.05$) between test
24 and retest values. However, repeatability in the grading estimates of both bulbar and
25 palpebral conjunctival hyperaemia was improved using the AOS grading method
26 ($R^2=0.998$; Coefficient of Repeatability CoR 0.10–0.13), compared to Efron ($R^2 =$
27 0.926 ; CoR 0.62) and CCLRU ($R^2 = 0.885$ – 0.911 ; CoR 0.50–0.78). Intraclass
28 coefficient correlations (ICC) improved inter-observer agreement using objective (>
29 0.995) versus subjective methods (0.853 – 0.959).

30 *Conclusion:* These subjective and objective grading methods are not
31 interchangeable. Due to the excellent repeatability and improved agreement
32 between experienced and novice observers, the objective grading method provides a

33 more consistent approach when grading ocular abnormalities and may achieve
34 greater reliability in record keeping and clinical monitoring in the future.

35

36

37 Keywords: Objective grading, Subjective grading, Agreement, Bulbar, Palpebral,
38 Conjunctiva, Hyperaemia, Imaging

39

40

41

42 **Introduction**

43 A fundamental aspect of clinical practice is an eye care practitioner's (ECP's) ability to
44 record ocular conditions in an accurate and repeatable manner. Standardised numeric
45 grading scales are used by ECPs in an attempt to improve record keeping and have
46 been shown to make grading more consistent over time [1]. Grading provides
47 opportunities to assess deviations from normal or healthy appearances, to record
48 baseline measurements to which future observations can be compared, and facilitate
49 clinical decision making with respect to management and treatment options [2]. A
50 survey of Australian optometrists found grading scales were used extensively in
51 optometric practice and were considered standard contact lens practice [3]. Similarly,
52 a worldwide study involving primary and secondary ECPs found approximately 85%
53 of practitioners used grading scales [4]. Nevertheless, some ECPs prefer to rely upon
54 sketches, photographs, or descriptions instead of grading scales [3]. An extensive
55 review of grading scales was recently published by Begley *et al.* [5], highlighting the
56 lack of a universally accepted "gold-standard" grading scale for corneal and
57 conjunctival staining. Two of the most widely used grading scales are the Cornea and
58 Contact Lens Research Unit (CCLRU), more recently known as the Institute for Eye
59 Research or Brien Holden Vision Institute scale [6-7], and the Efron Grading Scales
60 for Contact Lens Complications [1,8]. Both the Efron and CCLRU grading scales are
61 inexpensive, portable, and available as hardcopies.

62

63 Grading reliability has been defined as the ability of the grader to give similar results
64 time after time [9]. It has been observed that grading estimate variability is due to the
65 subjectivity associated with grading scales and the variation that occurs between
66 different observers, as well as for the same observer on different occasions [10,11].

67 To overcome the bias observed with subjective grading, objective grading techniques
68 e.g. Keratograph 5M (Oculus, Optikgerate, Germany) using digital software have been
69 developed to improve standardisation of grading [11-13]. Digital image analysis offers
70 a highly repeatable method of clinical monitoring and detection of changes in ocular
71 physiology over time, which often allow a continuous rather than discrete incremental
72 change in grading images. It has been reported that objective analysis can be 16
73 times more reliable than subjective analysis [11]. Given the likelihood of future
74 utilization of automated objective grading systems in clinical settings, validation of
75 such systems is desirable. One such novel automated objective grading software
76 (<https://aos-hub.com>) was designed by Advanced Ophthalmic Systems (AOS;
77 Weybridge, United Kingdom). The software can be used to assess a variety of anterior
78 and posterior ocular parameters including redness of the palpebral and bulbar
79 conjunctiva. Using Automated Intelligence to analyse the ocular surface in any digital
80 image, the software identifies all the vessels within the area selected (see Figure 1),
81 and an algorithm analyses environmental lighting of the conjunctiva while translating
82 the redness of the pixels into graded values. The system follows a grading scale format
83 resembling the Efron grading scale (grade 0 to 4) and the CCLRU grading scale (area
84 specific) in 0.1-unit increments. This study investigated by how much the digital AOS
85 method was likely to differ from the conventional subjective CCLRU and Efron grading
86 scales, whether the three scales could be used interchangeably, and whether
87 previously observed variability between experienced and novice observers could be
88 reduced, potentially improving clinical interpretation and management of the patient.

89

90 **Methods**

91 The study took place at the Division of Optometry and Visual Sciences, City, University
92 of London (United Kingdom) between December 2017 and March 2018. Ethical
93 approval for the study was obtained from the Optometry Proportionate Review
94 Committee. A series of anonymised images were taken from a private clinical
95 database, the International Association of Contact Lens Educators slide collection,
96 and from the internet. The images consisted of n=30 bulbar and n=26 palpebral
97 conjunctival hyperaemia of different eyes depicting various levels of redness
98 perceived ranging from none to severe. The raw images were numerically labelled and
99 displayed in full colour on a desktop computer with a monitor of resolution 1920 x 1080
100 pixels, while both subjective grading scales were used in printed version. The following

101 features were assessed for a valid comparison between the 3 grading methods:

102

- 103 1. **Bulbar conjunctival hyperaemia.** This is referred to as conjunctival redness in
104 Efron (Millennium Edition) grading scale and consists of five images depicting 0-4
105 grading ranging from normal to severe [1]. In the CCLRU grading scale, this is
106 known as ‘bulbar redness’ consisting of four images covering 1-4 grading, from
107 very slight to severe [6]. Bulbar redness was graded in the largest visible quadrant
108 (nasal, inferior, temporal or superior) depending on the subject’s position of gaze.
- 109 2. **Palpebral conjunctival hyperaemia.** Since grading of palpebral hyperaemia
110 cannot be differentiated from the grading of palpebral conjunctivitis on the Efron
111 grading scale, only the CCLRU scale was used. Using the CCLRU scale, ‘lid
112 redness” consists of 4 images covering 1-4 grading from very slight to severe. Lid
113 redness can be graded in 5 different areas of the palpebral conjunctiva: this study
114 graded area 2 representing the middle section under the eyelid [6].

115

116 Independently of one other, an experienced clinical optometrist (BH) and an optometry
117 student (MB) graded all bulbar and palpebral conjunctival hyperaemia images in a
118 randomised order on the same computer using the Efron grading scale (labelled as
119 *session 1*). To minimize a potential source of bias, randomisation was completed by
120 each observer using an electronic software available online
121 (<https://www.random.org/integer-sets/>), and graded to the nearest 0.1 [14]. Masked to
122 earlier results, all bulbar and palpebral hyperaemia images were randomised and
123 graded using the CCLRU grading scale on a separate day. The same method was
124 used for the AOS software whereby the area for grading was manually selected and
125 a grade between 0 and 4 was calculated by the software (Figure 1). All steps as
126 described above were then repeated approximately 1 week later (labelled *session 2*)
127 by both observers.

128

129 **Grading reliability**

130 Intra-observer variability is the ability of the grader to give similar results when the
131 process is repeated. For each grading scale, we calculated the numerical differences
132 between *session 1* and *session 2* grading estimates by the experienced optometrist
133 (BH). The standard deviation of this discrepancy distribution describes the grading
134 reliability.

135

136 **Grading agreement**

137 Agreement between two methods of grading describes the extent to which both
138 methods give similar results. Due to differences in grading scale scoring, it was likely
139 that grading of the same image would produce different outcomes depending on the
140 scale used. To estimate agreement between the methods, we calculated the numeric
141 differences between two grading scales by an experienced optometrist (BH) measured
142 during *session 2*. Data obtained during *session 2* was selected for analysis as previous
143 reports have suggested clinical grading may improve towards the end of a study [15].
144 In addition, we investigated the agreement between the two observers in grades
145 obtained during *session 2* for all three grading methods.

146

147 **Statistical analysis**

148 All statistical analyses were performed using SPSS version 22.0 for Windows (SPSS
149 Inc., Chicago, USA). Values in the text and tables are presented as the mean grading
150 score \pm standard deviation (SD). Preliminary analyses ensured that there were no
151 violations of the assumptions of normality (Kolmogorov-Smirnov normality test;
152 $P > 0.05$). The Coefficient of Repeatability (CoR) was calculated as $1.96 * SD$ of the
153 difference between pairs of measurements [16]. Limits of agreement (LoA) were
154 calculated as the mean difference between two sets of data \pm CoR, indicating the
155 range in which 95% of the differences between measurements will lie [17]. We
156 determined the correlation between the various methods for grading bulbar and
157 palpebral hyperaemia using Pearson's Correlation Coefficient (r). A one-way repeated
158 measures ANOVA was used to assess differences between the three methods, while
159 a paired sample t-test was used to compare between sessions and observers.
160 Intraclass Correlation Coefficients (ICC) [18] and Concordance Correlation
161 Coefficients (CCC) [19] were calculated to express inter-observer and inter-method
162 agreements, respectively. Statistical significance was accepted at $P < 0.05$.

163

164 **Results**

165 Thirty images were graded for bulbar hyperaemia, and after deletion of 2 images due
166 to incomplete lid area 2 data, 24 images were graded for palpebral hyperaemia. All
167 images were only presented once for each grading scale.

168

169 *Intra-observer reliability*

170 The reliability data for all images per grading scale obtained by an experienced
 171 optometrist (BH) is shown in Table 1. The difference between *session 1* and *session*
 172 *2* was only statistically significant when grading bulbar hyperaemia using the CCLRU
 173 grading method ($t(29)=3.143$; $P=0.004$). Using Efron or AOS methods, grading was
 174 not statistically different between the two sessions for either type of hyperaemia
 175 ($P>0.05$). Reliability scores with the objective AOS system were lowest, indicating
 176 better reliability for bulbar as well as palpebral hyperaemia when compared to
 177 subjective grading (Table 1). Subjective grading of bulbar hyperaemia was less
 178 reliable than palpebral hyperaemia. Using the objective AOS grading system, there
 179 was little difference between the reliability of bulbar and palpebral hyperaemia.

180

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

182 *Data from experienced observer (BH).*

	Bulbar hyperaemia			Palpebral hyperaemia	
	Efron	CCLRU	AOS	CCLRU	AOS
Sample size	30	30	30	24	24
Mean \pm SD session 1	2.21 \pm 1.14	3.13 \pm 0.60	1.80 \pm 1.37	2.41 \pm 1.22	2.46 \pm 1.18
Mean \pm SD session 2	2.16 \pm 1.14	2.98 \pm 0.72	1.81 \pm 1.40	2.43 \pm 1.05	2.46 \pm 1.17
Mean difference	-0.05	-0.15	0.017	0.021	<0.001
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	P=0.423	P=0.004	P=0.169	P=0.800	P=1.000
R ² value	0.926	0.885	0.998	0.911	0.998

183

184 Bland-Altman plots (Figure 2 top) show the mean of the differences between two
 185 sessions for each of the grading scales and both areas of hyperaemia. The continuous
 186 line represents the mean of the differences, also known as the line of agreement,
 187 which represents the systematic difference or estimated bias between the two

188 methods. It is bound by two parallel dotted lines which represents the 95% LoA above
189 and below the line of agreement. A narrow LoA implies a better agreement between
190 the two sessions.

191

192 *Between-method agreement*

193 Agreement between the three grading scales by an experienced optometrist (BH)
194 measured during *session 2* is presented in Table 2 and Figure 2 (middle). A one-way
195 repeated measures ANOVA was conducted to compares scores between the three
196 methods for bulbar hyperaemia. There was a statistically significant difference
197 between the three methods ($F(2,28)=40.34$, $P<0.0005$, multivariate eta squared =
198 0.74), whereby post hoc analysis revealed that the mean (\pm SD) grades using the AOS
199 method (1.81 ± 1.39) were significantly lower than the Efron (2.19 ± 1.13 ; $P=0.01$) and
200 CCLRU scale (3.06 ± 0.65 ; $P<0.0005$). In addition, the results from the Efron grading
201 scale were significantly lower than those from the CCLRU ($P<0.0005$). All showed a
202 large effect size (partially eta squared in Table 2). A paired sample t-test was
203 conducted to evaluate the agreement between CCLRU and AOS grading methods for
204 palpebral hyperaemia, which was not statistically significant different ($t(23)=-0.355$,
205 $P=0.726$).

206

207 **Table 2. Grading agreement data between methods.** The average grade between
 208 two sessions was used to calculate the differences between the methods.

	Bulbar hyperaemia			Palpebral hyperaemia
	Efron (method 1) vs CCLRU (method 2)	Efron (method 1) vs AOS (method 2)	CCLRU (method 1) vs AOS (method 2)	CCLRU (method 1) vs AOS (method 2)
Sample size	30	30	30	24
Mean \pm SD method 1	2.16 \pm 1.14	2.16 \pm 1.14	2.98 \pm 0.72	2.42 \pm 1.12
Mean \pm SD method 2	2.98 \pm 0.72	1.81 \pm 1.40	1.81 \pm 1.40	2.46 \pm 1.17
Mean difference	0.82	-0.35	-1.25	0.04
95% LoA	1.90 to -0.26	0.86 to -1.56	0.56 to -2.90	1.11 to -1.03
CCC	0.603	0.850	0.436	0.899
Confidence Intervals CCC	0.444 to 0.725	0.730 to 0.919	0.273 to 0.575	0.787 to 0.954
T-test	P<0.0005	P=0.004	P<0.0005	P=0.726
Effect size (partial eta squared)	0.73 (large effect)	0.26 (large effect)	0.67 (large effect)	0.005 (small effect)
R ² value	0.856	0.810	0.614	0.788

209
 210
 211 Mean grades for bulbar hyperaemia using the CCLRU scale produced a grade 1.17
 212 units higher than the objective AOS system. Bland-Altman plots (Figure 2 middle)
 213 showed that the two subjective grading scales differed on average by approximately
 214 1 grade (0.82 units) which may be due to the variation in their presentation of the eye,
 215 as well as the small shift in range between the scales (the CCLRU scale offers 4
 216 images while Efron presents a 5-point scale). Increased grading units were noted
 217 using CCLRU compared to the Efron gradings scale, which was more apparent in
 218 images showing less severe bulbar hyperaemia. As a result, a slanted difference
 219 versus mean plot was observed, whereby the agreement between the two methods
 220 improved for images of increasing severity. Similarly, Figure 2 (middle) shows that
 221 agreement between the subjective Efron grading method agrees and AOS method
 222 improved with increasing condition severity. Mean bulbar hyperaemia grading using

223 the Efron grading scale produced a grade 0.35 units higher than the AOS system. The
224 agreement between the CCLRU and AOS also improved for images of increasing
225 severity. For palpebral hyperaemia, mean difference between the CCLRU and AOS
226 methods was found to be close to zero, indicating that a subjective grade using the
227 CCLRU is on average increased by 0.04 in comparison to the objective AOS software
228 over the whole range of severities. Overall, we observed 95-100% of the variability
229 observed for bulbar and palpebral hyperaemia were within a total of 2 grading units.

230

231 *Inter-observer agreement*

232 Table 3 and Figure 2 (bottom) show data for inter-observer agreement. The difference
233 between the two observers was statistically significant when grading bulbar and
234 palpebral hyperaemia using the Efron and CCLRU grading systems, whereby the
235 experienced optometrist graded higher than the student optometrist ($P < 0.05$). Using
236 the AOS grading method, there was no statistical difference between the experienced
237 and the novice observer; although the experienced observer did record slightly higher
238 grades for both palpebral and bulbar hyperaemia (0.017 and 0.05 units, respectively).
239 Subjective and objective grading of bulbar hyperaemia was more variable between
240 observers than palpebral hyperaemia, although 92-97% of the variability observed
241 were within maximum one grading unit. The reliability and agreement using the AOS
242 method was much improved for bulbar as well as palpebral hyperaemia when
243 compared to the subjective methods of grading.

244

245 **Table 3. Grading reliability data per grading method (between observers).** Data
 246 collected during session 2 by the experienced optometrist (BH) were compared to
 247 those collected independently by the optometry student (MB). ICC = Intraclass
 248 Correlation Coefficient

	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
ICC	0.959	0.853	0.995	0.944	0.999
95% Confidence Intervals ICC	0.798 to 0.986	0.293 to 0.950	0.989 to 0.997	0.850 to 0.977	0.999 to 1.000
T-test	P<0.0005	P<0.0005	P=0.177	P=0.023	P=0.162
R ² value	0.904	0.802	0.982	0.829	0.998

249

250

251 Discussion

252 This study investigated the reliability and agreement between a novel objective,
 253 automated ocular grading software and two 'gold-standard' subjective grading
 254 methods commonly used by ECPs, to determine if objective image analysis of bulbar
 255 and palpebral hyperaemia was more reliable than subjective grading.

256

257 *Intra-observer reliability*

258 Objective grading of bulbar as well as palpebral hyperaemia showed substantially less
 259 variation between sessions as indicated by its narrow LoA (Table 1). We did note

260 statistically significant differences in grading bulbar hyperaemia between two different
261 sessions using the CCLRU grading scale ($P=0.004$), although the mean difference of
262 0.15 units suggests that this was not considered clinically significant [20]. It is possible
263 that the intra-observer variability for CCLRU especially in the higher severities is
264 caused by the lack of reference images for the more severe degrees of redness [15].
265 Schulze *et al.* found that the CCLRU reference images were perceived to cover only
266 the lower half of the total range of bulbar hyperaemia available [21]. Furthermore,
267 similar to Wolffsohn [12], our data showed that severity did not support linear grading;
268 particularly in the low range of hyperaemia (<2.5 units) sensitivity between the
269 sessions increased and a difference >1.0 units was observed. For bulbar hyperaemia,
270 there were two occasions (out of 30) whereby these lower range grading scores were
271 *reduced* by approximately 1 grading unit during the second session, while for the lower
272 severities of palpebral hyperaemia three (out of 24) grading scores *increased*
273 approximately 1 unit during the second session (Table 1). The underestimation of
274 palpebral hyperaemia during the first session (or overestimation during the second
275 session) may be explained by the learning effect or grading confidence of selecting
276 area 2. The AOS grading software only expressed a mean difference of 0.017 units
277 between visits with narrow LoA (0.14 to -0.11), whereas Efron varied on average 0.05
278 units and wide LoA (0.57 to -0.66). The ranges imply that 95% of the differences
279 between measurements varied >1 grade for bulbar hyperaemia using the Efron or
280 CCLRU scales and about 1.5 grades for palpebral hyperaemia using CCLRU, while
281 this was only 0.25 grade using the AOS method. Using the objective AOS software,
282 any variability observed between sessions was attributed to the manual area selection
283 for image analysis by the software. In addition, the correlation coefficient identified an
284 improved repeatability of the AOS grading system compared to Efron and CCLRU,
285 with a R^2 value close to 1, showing that for nearly every ocular image the grading
286 estimate was the same on visit 1 and on visit 2. CCLRU showed the lowest
287 repeatability between visit 1 and visit 2, with an R^2 value of 0.72. Poor repeatability of
288 the subjective gradings may be attributable to inconsistencies in image resolutions.
289 The images were obtained from a variety of databases, and viewed under the same
290 conditions including image size which may have decreased visible resolution. This has
291 shown to be a particular advantage of the objective grading method, which seems to
292 overcome this limitation unless the resolution of the image falls below 150 by 150
293 pixels.

294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313

Between methods agreement

Our data showed a lack of agreement between subjective and objective grading systems for bulbar, but not palpebral, hyperaemia. This may be attributable to reflectivity of different ocular surfaces, contrast levels i.e. red on red vs red on white grading, or differences in surface area sampled. In addition, the two subjective grading scales differ on average by approximately 1 grade (0.82 units) mainly due to the disagreement in presentation (drawing versus photographs). Additionally, the absence of a zero scale in CCLRU means that this method presents 4 images for the whole range of severities while the Efron grading scale uses 5 images. This may have caused a small shift in range of scales particularly in the lower severities of hyperaemia. In line with previous studies [11,21], we did indeed observe differences between grading systems to be non-linear whereby the agreement between the two subjective scales seems to improve for images of increasing severity. This reduces the possibility of applying a simple correction factor to interchangeably use different grading systems. However, it has been shown that cross-calibrated scales (after applying a correction factor) can lead to repeatable results between different scales [10]. On the other hand, for palpebral hyperaemia, the agreement between CCLRU and the objective AOS grading methods was excellent with a linear mean difference of 0.04 unit.

314
315

Inter-observer agreement

The onset of conjunctival hyperaemia can indicate a range of ocular conditions varying from dry eye to scleritis. Therefore, it is important that ECPs are able to evaluate any subtle variations in the anterior eye with confidence [11]. Our findings show that intra-observer repeatability is generally (clinically) acceptable for both the subjective and objective methods of anterior eye grading (bulbar and palpebral hyperaemia), although the objective method produced significantly less disparity between observers with different levels of experience. This was apparent from the statistically significant differences in grading both palpebral as well as bulbar hyperaemia between observers (Table 3). Several reports have shown that experience improves an observer's ability to grade [11,22]. In accordance with such reports, we found significant differences between the experienced and novice observers for the subjective grading methods, and that the novice clinician used a wider range of the subjective scales. High

328 agreement between subjective and objective methods have been reported previously
329 [14,23-24] particularly with higher number ($n>5$) of graders [25-26]. Critically, over the
330 full range of severities, the objective method of grading (AOS) did express excellent
331 reliability without significant disparities between our two observers, demonstrating its
332 potential as a tool for inexperienced practitioners and/or teaching purposes. Using this
333 objective grading system, experienced ECPs can rely with confidence on the grading
334 recorded by a novice.

335

336 Intra-observer reliability and inter-observer agreement were most favourable using the
337 objective AOS system, suggesting that objective methods of grading may establish
338 themselves as the new gold-standard in ocular grading. The software allows for instant
339 analysis of any digital image using a desktop or mobile phone application, providing
340 an opportunity for consistent and extensive (5 separate areas resembling CCLRU plus
341 a combination of vascular presentations including hue, visibility, width of vessels etc)
342 grading with minimal effort.

343

344 One limitation of our study was that images were sourced from a variety of databases
345 and so aspects such as magnification and image quality were not standardised.
346 Furthermore, larger-scale studies are required to understand the potential benefits and
347 shortcomings of such objective systems. In particular, ocular characteristics such as
348 disease specific hyperaemia (e.g. allergic or bacterial conjunctivitis, infectious
349 keratitis, or dry eye) and/or corneal staining and lid roughness should be included in
350 future studies. Consideration must be given to whether practice investment in
351 objective grading systems will bring about a significant improvement to clinical
352 diagnosis, monitoring, and quality of patient care.

353

354 **Conclusion**

355 Although all three methods showed acceptable repeatability, the novel automated
356 AOS system used for objective grading of bulbar and palpebral hyperaemia was
357 substantially more reliable than the subjective methods of grading using Efron and
358 CCLRU grading scales. Practitioners ought to be dissuaded from attempting to use
359 multiple systems interchangeably to prevent large variability in clinical interpretation
360 and management of the patient over time.

361

362

363 **Acknowledgements**

364 A portion of this work was conducted as part of an unrestricted grant from Advanced
365 Ophthalmic Systems (AOS), who provided a copy of the software. This work was
366 completed independently by MB as part of her undergraduate studies research
367 project.

368

369

370

371 **Legends Figures**

372 Figure 1. Objective grading method using the AOS software. Manual selection of the
373 area of interest using the AOS software for grading bulbar hyperaemia (A). Bulbar
374 conjunctival hyperaemia grade is displayed as 2.3 units (B). Image C shows manual
375 selection of the area of interest while grading palpebral hyperaemia. Palpebral
376 conjunctival hyperaemia gradings over 5 areas are displayed directly on the image (D)
377 Area 2 is shown as 3.4 units of palpebral hyperaemia.

378

379 Figure 2. Bland and Altman plots comparing sessions, methods, and observers for
380 bulbar (left) and palpebral (right) conjunctival hyperaemia.

381

382 **References**

383

384 [1] Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens
385 complications. *Ophthalmic Physiol Opt* 2001;21(1):17-29.

386 <https://doi.org/10.1046/j.1475-1313.1999.00420.x-i1>

387 [2] Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects
388 of scaling. *Invest Ophthalmol Vis Sci* 1991;32(2):422-32.

389 [3] Efron N, Pritchard N, Brandon K, Copeland J, Godfrey R, Hamlyn B, Vrbancic V.

390 A survey of the use of grading scales for contact lens complications in optometric
391 practice. *Clin Exp Optom* 2011;94(2):193-9. <https://doi.org/10.1111/j.1444->

392 [0938.2010.00549.x](https://doi.org/10.1111/j.1444-0938.2010.00549.x)

393 [4] Wolffsohn JS, Naroo SA, Christie C, Morris J, Conway R, Maldonado-Codina C.

394 Anterior eye health recording. *Cont Lens Anterior Eye* 2015;38(4):266-71.

395 <https://doi.org/10.1016/j.clae.2015.03.001>

396 [5] Begley C, Caffery B, Chalmers R, Situ P, Simpson T, Nelson JD. Review and
397 analysis of grading scales for ocular surface staining. *Ocular Surface* 2019.

398 <https://doi.org/10.1016/j.jtos.2019.01.004>

399 [6] Terry RL, Schnider CM, Holden BA, Cornish R, Grant TI, Sweeney D, La DH,
400 Back A. CCLRU standards for success of daily and extended wear contact lenses.

401 *Optom Vis Sci* 1993;70(3):234-43. <https://doi.org/10.1097/00006324-199303000->

402 [00011](https://doi.org/10.1097/00006324-199303000-00011)

403 [7] Dundas M, Walker A, Woods RL. Clinical grading of corneal staining of non-
404 contact lens wearers. *Ophthalmic Physiol Opt* 2001;21(1):30-5.

405 [https://doi.org/10.1016/S0275-5408\(00\)00011-9](https://doi.org/10.1016/S0275-5408(00)00011-9)

406 [8] Efron N. Clinical application of grading scales for contact lens complications.

407 *Optician* 1997;213:26-34.

408 [9] Bullimore MA and Bailey IL. Considerations in the subjective assessment of

409 cataract. *Optom Vis Sci* 1993;70(11):880-5. <https://doi.org/10.1097/00006324->

410 [199311000-00004](https://doi.org/10.1097/00006324-199311000-00004)

411 [10] Schulze MM, Hutchings N, Simpson TL. Grading bulbar redness using cross-

412 calibrated clinical grading scales. *Invest Ophthalmol Vis Sci* 2011;52(8):5812-7.

413 <https://doi.org/10.1167/iovs.10-7006>

414 [11] Peterson RC and Wolffsohn JS. Sensitivity and reliability of objective image
415 analysis compared to subjective grading of bulbar hyperaemia. *Br J Ophthalmol*
416 2007;91(11):1464-6. <https://doi.org/10.1136/bjo.2006.112680>

417 [12] Yoneda T, Sumi T, Takahashi A, Hoshikawa Y, Kobayashi M, Fukushima A.
418 Automated hyperemia analysis software: reliability and reproducibility in healthy
419 subjects. *Jpn J Ophthalmol* 2012 56(1):1-7. [https://doi.org/10.1007/s10384-011-](https://doi.org/10.1007/s10384-011-0107-2)
420 [0107-2](https://doi.org/10.1007/s10384-011-0107-2)

421 [13] Wu S, Hong J, Tian L, Cui X, Sun X, Xu J. Assessment of bulbar redness with a
422 newly developed keratograph. *Optom Vis Sci* 2015;92(8):892-9.
423 <https://doi.org/10.1097/OPX.0000000000000643>

424 [14] Wolffsohn JS. Incremental nature of anterior eye grading scales determined by
425 objective image analysis. *Br J Ophthalmol* 2004;88(11):1434-8.
426 <https://doi.org/10.1136/bjo.2004.045534>

427 [15] Murphy PJ, Lau JS, Sim MM, Woods RL. How red is a white eye? Clinical
428 grading of normal conjunctival hyperaemia. *Eye* 2007;21(5):633.
429 <https://doi.org/10.1038/sj.eye.6702295>

430 [16] Bland JM and Altman D. Statistical methods for assessing agreement between
431 two methods of clinical measurement. *Lancet* 1986;327(8476):307-10.
432 [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8)

433 [17] Bland JM and Altman DG. Measuring agreement in method comparison studies.
434 *Stat Methods Med Res* 1999;8(2):135-60.
435 <https://doi.org/10.1177/096228029900800204>

436 [18] Bartko JJ. Measures of agreement: a single procedure. *Stat Med* 1994;13(5-
437 7):737-45. <https://doi.org/10.1002/sim.4780130534>

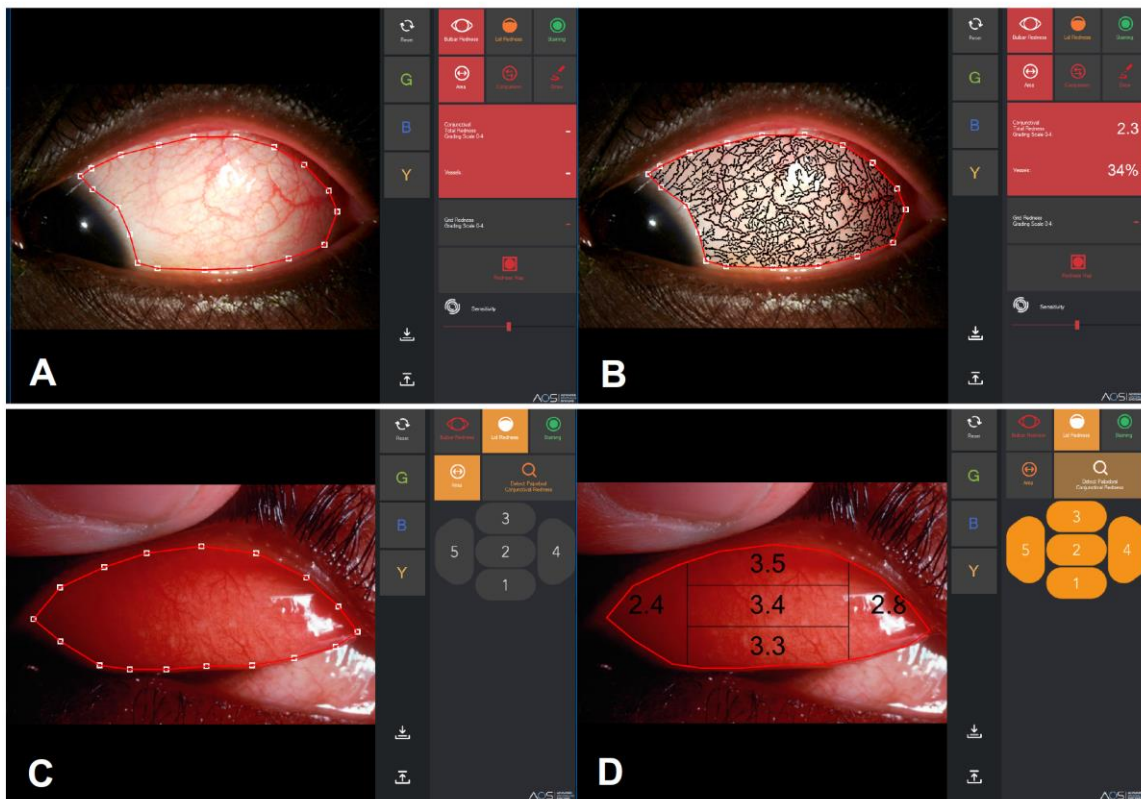
438 [19] Lawrence I and Lin K. A concordance correlation coefficient to evaluate
439 reproducibility. *Biometrics* 1989:255-68.

440 [20] Chong T, Simpson T, Fonn D. The repeatability of discrete and continuous
441 anterior segment grading scales. *Optom Vis Sci* 2000;77(5):244-51.
442 <https://doi.org/10.1097/00006324-200005000-00011>

443 [21] Schulze MM, Hutchings N, Simpson TL. The perceived bulbar redness of clinical
444 grading scales. *Optom Vis Sci* 2009;86(11):E1250-8.
445 <https://doi.org/10.1097/OPX.0b013e3181bb4225>

446 [22] MacKinven J, McGuinness CL, Pascal E, Woods RL. Clinical grading of the
 447 upper palpebral conjunctiva of non-contact lens wearers. *Optom Vis Sci*
 448 2001;78(1):13-8. <https://doi.org/10.1097/00006324-200101010-00009>
 449 [23] Papas EB. Key factors in the subjective and objective assessment of
 450 conjunctival erythema. *Invest Ophthalmol Vis Sci* 2000;41(3):687-91.
 451 [24] Schulze MM, Hutchings N, Simpson TL. The use of fractal analysis and
 452 photometry to estimate the accuracy of bulbar redness grading scales. *Invest*
 453 *Ophthalmol Vis Sci* 2008;49(4):1398-406. <https://doi.org/10.1167/iovs.07-1306>
 454 [25] Fieguth P and Simpson T. Automated measurement of bulbar redness. *Invest*
 455 *Ophthalmol Vis Sci* 2002;43(2):340-7.
 456 [26] Peterson RC and Wolffsohn JS. Objective grading of the anterior eye. *Optom*
 457 *Vis Sci* 2009;86(3):273-8. <https://doi.org/10.1097/OPX.0b013e3181981976>

458
 459
 460
 Figure 1



461
 462

Figure 2

