Optimising Subjective Anterior Eye Grading Precision

[1]Marta Vianya-Estopa PhD marta.vianya@anglia.ac.uk

[2]Manbir Nagra PhD manbir.nagra@port.ac.uk

[3]Arnold Cochrane BSc ajs.cochrane@ulster.ac.uk

[4]Neil Retallic BSc optomneil@outlook.com

[5]Dean Dunning MEd D.Dunning@bradfordcollege.ac.uk

[6]Louise Terry PhD TerryL1@cardiff.ac.uk

[7]Aoife Lloyd PhD aoife.lloydmckernan@dit.ie

[8]James S Wolffsohn PhD j.s.w.wolffsohn@aston.ac.uk

and members of the British & Irish University & College Contact Lens Educators (BUCCLE)

Affiliations

[1]Vision and Hearing Sciences, Anglia Ruskin University, Cambridge, UK

[2]Optometry, University of Portsmouth, Portsmouth, UK

[3]School of Biomedical Sciences, Ulster University, Coleraine, UK

[4]Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK

[5]Advanced Technology Centre, Bradford College, Bradford, UK

[6]School of Optometry and Vision Sciences, Cardiff University, Cardiff, UK

[7]School of Physics & Clinical & Optometric Sciences, Technological University Dublin, Ireland.

[8]Ophthalmic Research Group, Aston University, Birmingham, UK

Corresponding author: Prof James S Wolffsohn, Aston University, Aston Triangle, Birmingham, B4 7ET, UK j.s.w.wolffsohn@aston.ac.uk

Abstract

Purpose: To establish the optimum grading increment which ensured parity between practitioners while maximising clinical precision.

Methods: Second year optometry students (n=127, 19.5 ± 1.4 years, 55% female) and qualified eye care practitioners (n=61, 40.2 ±14.8 years, 52% female) had 30 seconds to grade each of bulbar, limbal and palpebral hyperaemia of the upper lid of 4 patients imaged live with a digital slit lamp under 16x magnification, diffuse illumination, with the image projected on a screen. The patients were presented in a randomised sequence 3 times in succession, during which the graders used the Efron printed grading scale once to 0.1 precision, once to 0.5 precision and once to the nearest integer grade in a randomised order. Graders were masked to their previous responses.

Results: For most grading conditions less than 20% of clinicians showed a ≤0.1 difference in grade from the mean. In contrast, more than 50% of the student graders and 40% of experienced graders showed a difference in grade from the mean within 0.5 for all conditions under measurement. Student precision in grading was better with both 0.1 and 0.5 grading precision than grading to the nearest unit, except for limbal hyperaemia where they performed more accurately with 0.5 unit precision grading. Limbal grading precision was not affected by grading step precision for experienced practitioners, but 0.1 and 0.5 grading precision were both better than 1.0 grading precision for bulbar hyperaemia and 0.1 grading precision was better than 0.5 grading precision and both were better than 1.0 grading precision for palpebral hyperaemia.

Conclusion: Although narrower intervals scales maximise the ability to detect smaller clinical changes, the grading increment should not exceed one standard deviation of the discrepancy between measurements. Therefore, 0.5 grading increments are recommended for subjective anterior eye physiology grading (limbal, bulbar and palpebral redness).

Keywords: grading; hyperaemia; student; eye-care practitioner; scale increments

Introduction

Since their initial introduction approximately thirty years ago, anterior eye grading scales have firmly established themselves as an essential part of the eye care practitioner's (ECPs) armamentarium. With usage reported at approximately 60-85% amongst ECPs [1,2] this seemingly low-tech approach has had a significant impact on clinical practice. Grading scales hold several advantages over the sole use of written descriptions and sketches that practitioners had previously relied upon. Grading scales are quantitative, simplify the monitoring and progression of pathological and physiological changes, are a universal familiar language so can be interpreted by different nationalities and across health care professionals, aid in medical legal cases, and ultimately facilitate patient management.

While grading scales are easy to use, widely available, and considered best practice [2], they are not without their limitations. Grading is subjective, associated with poor repeatability [3] and high variability amongst practitioners. Grading scales are not interchangeable and the scale range varies, thus grading scores will differ depending on scale used [4] with estimates reported to be higher for scales which have a shorter dynamic range. [5] Further, there are concerns about the grading reference images themselves. Wolffsohn [6] found grading scale images did not follow a linear increase in severity, but instead followed a quadratic pattern, such that precision is greater for lower severity reference images i.e. the increments between gradings are unequal. Digital versions of grading scales have been produced with morphing technology [7] used to generate reference images down to 0.1 scale grade increments, but any improvement in gradingvariability has not been published.

Some of the shortcomings may be attributable to the process of grading itself; typically, anterior eye grading involves the application of a discrete scale (a limited fixed number of grades) to a continuous variable (the severity of a particular ocular condition). [8] Several sources [2,8] have advocated the reduction of grading scale increment size to increase clinical precision i.e. grading to the nearest integer should produce poorer clinical precision than grading to the nearest 0.5 or 0.1. Nevertheless, achieving adequate clinical precision may not necessitate use of the smallest grading increment possible. Peterson and Wolffsohn [3] showed a mean difference of approximately 0.70-1.03 bulbar redness (Efron) image grades was needed for it to be discernible by subjective grading. Given the widespread use of grading scales, and their vulnerability to subjective bias, it is of clinical interest to establish an evidence base for a best practice approach to grading. The aim of this study was to establish the optimum grading increment which ensured parity between practitioners while maximising clinical precision. Based on previously published data, it is hypothesised that whole integer grading will be less accurate (a larger absolute deviation from the mean practitioner grade) than grading to the nearest 0.5 or 0.1 unit.

Method

The study was granted a favourable ethical opinion by Ulster University (practitioner study) and Aston University (student study) ethics committees and followed the tenets of the Declaration of Helsinki. Participants gave written informed consent to take part after an explanation of the study.

The graders were 2nd year undergraduate optometry students enrolled at Aston University (n=127, 19.5 ± 1.4 years, 55% female) and qualified eye care practitioners (at least 2 years) attending the BCLA UK conference in June 2018 (n=61, 40.2 ±14.8 years, 52% female) all familiar with using grading scales with the Efron grading scale. Data collection for the two cohorts occurred on separate occasions.

The ocular surface of 4 patients with no ocular pathology were observed live under 16x magnification, diffuse illumination, with a digital slit-lamp (Keeler, Windsor, UK) and the image projected on a screen. The patients were presented in a randomised sequence 3 times in succession during which the graders used the Efron printed grading scale once to 0.1 increments, once to 0.5 increments and once to the nearest integer grade in randomised order. They had 30 seconds to grade each of bulbar, limbal and upper lid palpebral hyperaemia each time, and were masked to their previous grades.

Statistical Analysis

The absolute average difference from the mean of all graders, for each grader with each increment level was calculated for each of the 4 patients. As the data was not normally distributed, non-parametric statistics were applied (Friedman test repeated measure analysis of variance with Wilcoxon signed-rank test post-hoc pairwise comparison where significance was identified). In addition the discrepancies between pairs of observers were assessed for each of the 4 patients and the standard deviation calculated.

Based on a standard deviation of 0.4 [9] for subjective grading, a clinically significant difference ($p<0.05$) of 0.2 units between groups could be detected with 80% power with a sample size of 61 participants in each group and 0.15 units with 127 participants in each group. https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html

Results

Across the 4 patients examined, the average bulbar grade ranged from 0.8-1.5, average limbal grade ranged from 0.4 to 1.2 and the average palpebral grade ranged from 0.4 to 1.6 and was similar between patients used for the student grading and practitioner grading sessions. The distribution of the difference from the mean is shown in Figure 1 for student graders and Figure 2 for qualified eye care practitioners. The mean of these differences for each feature is shown in Table 1, along with statistical significance. There was a significant difference (p<0.001) across all grade increment comparisons except practitioner graded limbal hyperaemia (p=0.478). The percentage of clinicians increased for all conditions within a greater difference in grade from the mean. For most conditions less than 20% of clinicians showed a ≤0.1 difference in grade from the mean. In contrast, more than 50% of the student graders and 40% of experienced graders showed a difference in grade from the mean within 0.5 for all conditions under measurement.

Grading Increment		0.1	⇔	0.5	⇔	1.0	⇔ 0.1
		mean	p	mean	p	Mean	P
Student n=127	Bulbar	0.40±0.30	0.342	0.42±0.31	**<0.001**	0.51±0.29	**<0.001**
	Limbal	0.44±0.31	0.156	0.42±0.31	**0.001**	0.47±0.36	0.259
	Palpebral	0.35±0.26	0.645	0.34±0.32	**<0.001**	0.49±0.27	**<0.001**
Practitioner n=61	Bulbar	0.58±0.50	0.633	0.58±0.53	**0.004**	0.64±0.45	**0.001**
	Limbal	0.54±0.46	0.790	0.54±0.49	0.940	0.53±0.52	0.874
	Palpebral	0.71±0.64	**0.026**	0.75±0.67	**<0.001**	0.82±0.64	**<0.001**

Table 1: Mean grade difference (± S.D.) from mean and significance between grading increments. The arrows above the significance (p) values point to the two increments being compared.

Student precision in grading was better with both 0.1 and 0.5 grading increments than grading to the nearest unit, except for limbal hyperaemia where it was only better with 0.5 unit increment grading (there was no significant difference between the 0.1 and 1.0 increments for this feature). Limbal grading precision was not affected by grading step increment for experienced practitioners, but 0.1 and 0.5 grading increments were both better than the 1.0 grading increment for bulbar hyperaemia. For palpebral hyperaemia, the 0.1 grading increment was more accurate than the 0.5 grading increment and both were better than 1.0 grading increment (Table 1). The standard deviation of discrepancies between observers was 0.65-0.87 across the students and was 0.72 to 0.84 across experienced practitioners.

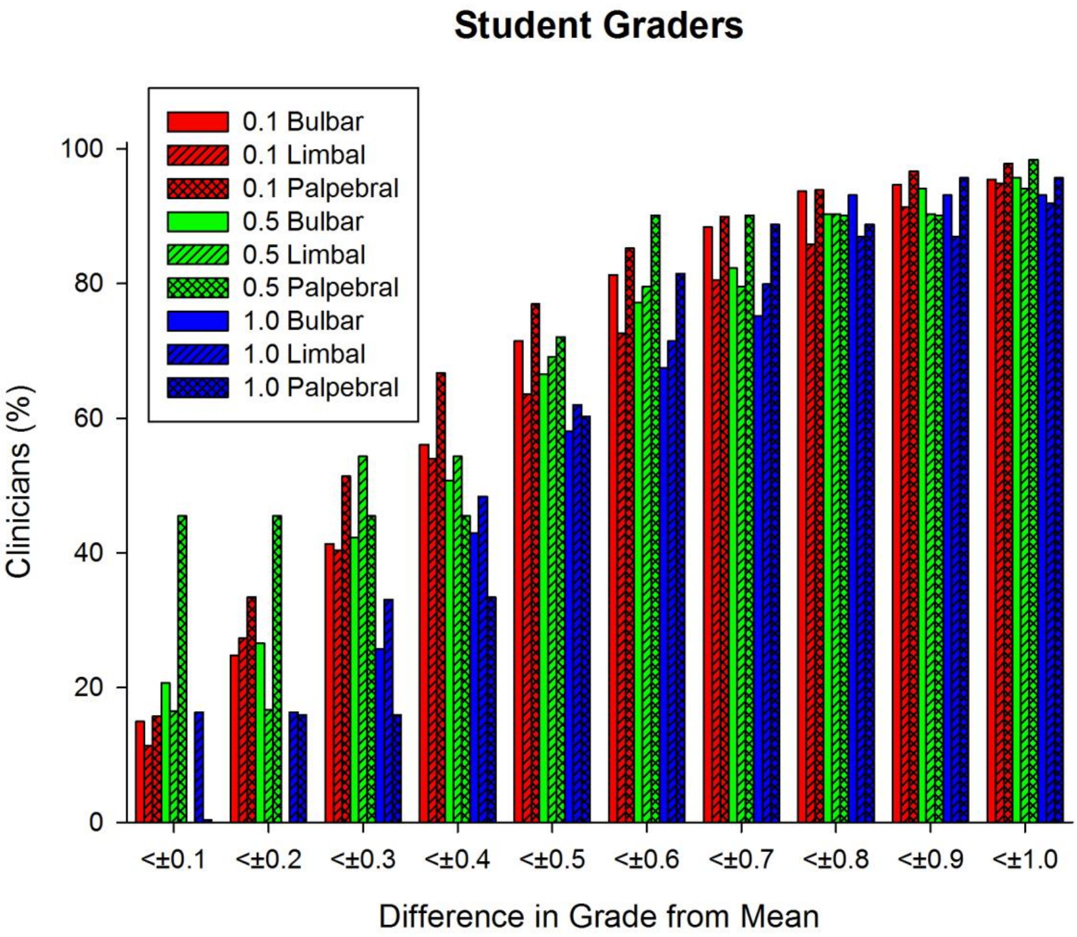

Figure 1: What proportion of student clinicians were within 0.1 to 1.0 grades different from the mean of all clinicians for bulbar, limbal, and palpebral hyperaemia with each of the grading increments. N=127.

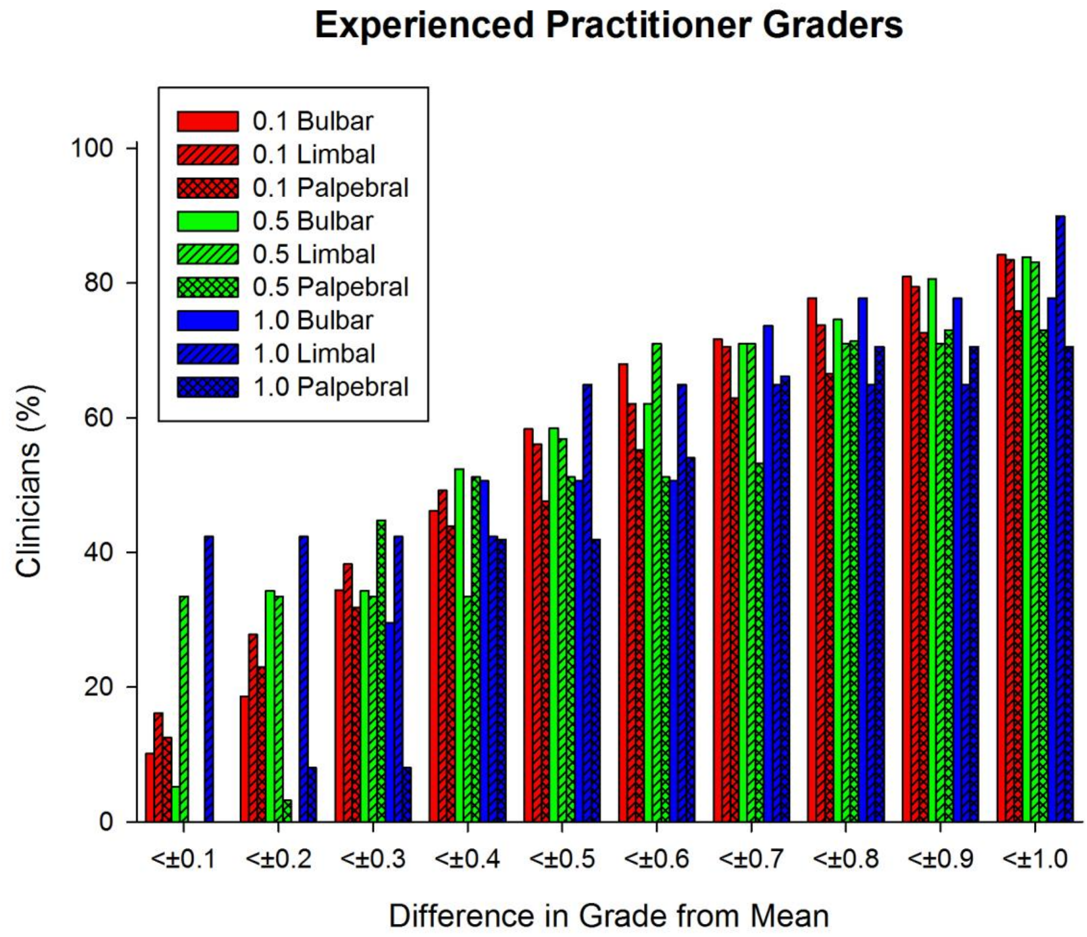

Figure 2: What proportion of experienced practitioners within 0.1 to 1.0 grades different from the mean of all clinicians for bulbar, limbal and palpebral hyperaemia with each of the grading increments. N=61.

Discussion

This study set out to show that smaller grading increment steps would lead to more accurate grading compared to the mean. In practical terms, the grades recorded by a practitioner should be as close as possible to the mean of other practitioners (average difference) rather than the discrepancy analysis (the difference between 2 practitioners) as modelled by Bailey et al. [8]. However, while this was the case for 0.5 grading units compared to whole integer grading, this was generally not the case for 0.1 grading units compared to 0.5. As shown in Table 1, the average difference from the mean was around 0.30 for student graders and 0.55 for experienced graders. The standard deviation between random pairs of observers was higher, as expected, being 0.72 for student graders and 0.78 for experienced graders. Bailey et al. [8] suggested that if the scale increment exceeds the standard deviation of the discrepancy this will result in a sharp broadening of the confidence limits. Thus, these findings suggest that a 0.5 grading step might be as precise as is possible to get when evaluating hyperaemia in the anterior eye using the Efron printed grading scale.

It is worth noting that limbal hyperaemia grading was more variable in grade than bulbar and palpebral redness. This finding is not surprising as the exact extent of the limbal region is not clearly defined clinically and graders might have been influenced by nearby conjunctival redness. Yet, observers need to ensure enough attention is given to this structure given the response between limbal hyperaemia and contact lens wear. For instance, several studies have shown that hydrogel lens wear results in significantly greater levels of limbal hyperaemia compared to silicone hydrogel lens wear for both daily and extended wear modalities, whereas bulbar redness is not significantly affected. [10-13]

Efron et al [4] suggested that grading of contact lens complications would be expected to improve with experience. His group also found grading variability improves statistically (but not clinically significant) with some experience, but no added benefit could be derived from supplemental training [14]. However, this study found experienced practitioners were less accurate than second year undergraduate optometry students. Similar findings between students and experienced practitioners were also noted by Wolffsohn et al [4]. Although a priori one might expect experienced practitioners to show greater precision than students, this might no longer be the case as the importance of grading in the assessment of anterior eye is currently emphasised to undergraduate optometry students. Similarly, Cardona and Serés [15] noted that contact lens knowledge improved grading precision in optometry students. The students taking part in this study had received a 1 hour seminar on the principals behind grading and had used the Efron grading scales in 5 weekly 2 hour clinics. Differences in the data projection of the images, such as screen resolution and ambient brightness could have made a difference between cohorts, but the student graders used the same conditions as half of the experienced graders and the difference between them was still evident. Future work should further explore the relationship that knowledge, training and experience might have on the uses of grading scales in anterior eye and contact lens assessment. A survey of UK practitioners in 2015 [2] indicated that 91.6% of respondents used grading scales for bulbar conjunctival hyperaemia and 77.8% and 63.4% for limbal and palpebral hyperaemia respectively. It could be hypothesised that less familiarity of usage might lead to more variability with grading and this seems to be the case with practitioners.

Recently, alternative methods to subjective assessment of bulbar and limbal hyperaemia have been proposed using software such as Keratograph 5M (Oculus) that objectively detects hyperaemia. [16] Artificial intelligence learning algorithms have been applied to retinal images, demonstrating their ability not just to quantify disease changes, but also to identify other features that might differentiate disease and its progression such as tortuosity, pallor and blood flow, not traditionally utilised by clinicians. [17] However, technological advances are not yet readily available by most clinicians. In addition, the results of this new technology might not be interchangeable with results obtained using subjective grading scales. [18-19] Thus, it is important to continue to support clinicians using grading scales optimally, although, digital photography can allow direct comparison at subsequent visits and is preferable to grading.

It is important to note that this study was conducted using projected slit lamp videos of eyes without pathology. The patients examined were different between the students and experienced practitioners, but the average grades were similar for each of the ocular anterior eye features examined and the comparison was the individual's difference from the mean, so the actual mean should not have a significant effect on the results. The mean grade of each feature was ≤2 for each participant; the entire range of the grading scale used was not included in the study. Therefore the conclusions cannot be extended to grading precision for more severe hyperaemic cases.

In conclusion, this study showed that 0.5 grading increments should be recommended when assessing anterior eye grading (limbal, bulbar and palpebral hyperaemia). This contradicts previous recommendation by Efron et al [4] and Wolffsohn et al. [2] of recording clinical signs using 0.1 increments between grades. Although narrower intervals scales maximise the ability to detect smaller clinical changes, Bailey et al [8] also indicated that for moderate precision the grading increment should not exceed one standard deviation of the discrepancy between measurements. Although narrower increments have been recommended in clinical practice, Efron et al [4] and Wolffsohn et al [2] found graders tended to grade using whole and half-digits indicating a reluctance to use finer increments. Thus, this research provides the evidence for clinicians to adopt 0.5 increments in their clinical grading alongside previous research highlighting the importance of recording the scale used and having the scale present when grading. [2,6]

References

[1] Efron N, Pritchard N, Brandon K, Copeland J, Godfrey R, Hamlyn B, Vrbancic V. A survey of the use of grading scales for contact lens complications in optometric practice. Clin Exp Optom 2011;94:193-9.

[2] Wolffsohn JS, Naroo SA, Christie C, Morris J, Conway R, Maldonado-Codina C. Anterior eye health recording. Contact Lens Ant Eye 2015;38:266-71.

[3] Peterson RC, Wolffsohn JS. Sensitivity and reliability of objective image analysis compared to subjective grading of bulbar hyperaemia. Br J Ophthalmol 2007;91:1464-6.

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

[5] Schulze MM, Hutchings N, Simpson TL. Grading bulbar redness using cross-calibrated clinical grading scales. Invest Ophthalmol Vis Sci 2011;52:5812-7.

[6] Wolffsohn JS. Incremental nature of anterior eye grading scales determined by objective image analysis. Br J Ophthalmol 2004; 88:1434-8.

[7] Efron N, Morgan PB, Jagpal R. Validation of computer morphs for grading contact lens complications. Ophthalmic Physiol Opt 2002;22:341-49.

[8] Bailey IL, Bullimore MA, Raasch TW and Taylor HR. Clinical Grading and the Effects of Scaling. Invest Ophthalmol Vis Sci 1991;32:422–32.

[9] Efron N. Grading scales for contact lens complications. Ophthalmic Physiol Opt 1998;18:182-6.

[10] Brennan NA, Coles ML, Connor HR, McIlroy RG. A 12-month prospective clinical trial of comfilcon A silicone-hydrogel contact lenses worn on a 30-day continuous wear basis. Cont Lens Anterior Eye 2007;30:108-18.

[11] Dumbleton K, Keir N, Moezzi A, Feng Y, Jones L, Fonn D. Objective and subjective responses in patients refitted to daily-wear silicone hydrogel contact lenses. Optom Vis Sci 2006; 83:758-68.

[12] Maldonado-Codina C, Morgan PB, Schnider CM, Efron N. Short-term physiologic response in neophyte subjects fitted with hydrogel and silicone hydrogel contact lenses. Optom Vis Sci 2004; 81:911-21.

[13] Dillehay SM, Miller MB. Performance of Lotrafilcon B silicone hydrogel contact lenses in experienced low-Dk/t daily lens wearers. Eye Contact Lens 2007; 33:272-77.

[14] Efron N, Morgan PB, Farmer C, Furuborg J, Struk R, Carney LG. Experience and training as determinants of grading reliability when assessing the severity of contact lens complications. Ophthalmic Physiol Opt 2003;23:119-24.

[15] Cardona G & Serés C. Grading Contact Lens Complications: The Effect of Knowledge on Grading Accuracy. Curr Eye Res 2009; 34: 1074–81.

[16] Wu S, Hong J, Tian L, Cui X, Sun X and Xu J. Assessment of Bulbar Redness with a Newly Developed Keratograph. Optom Vis Sci 2015; 92: 892-899.

[17] Varadarajan AV, Poplin R, Blumer K, Angermueller CA, Ledsam J, Chopra R, Keane PA, Corrado GS, Peng L and Webster DR. Deep learning for predicting refractive error from retinal fundus images. Invest Ophthalmol Vis Sci 2018;59: 2861-8.

[18] Perez-Bartolome F and Garcia-Feijoo J. Assessment of ocular redness measurements obtained with keratography 5M and correlation with subjective grading scales. Journal Français d'Ophthalmologie 2018; 41 (9): 836-46.

[19] Huntjens B, Basi M, Nagra M. Evaluating a new objective grading software for conjunctival hyperaemia. 2019. Contact Lens Ant Eye (in press)

Acknowledgements

BUCCLE's mission is to enhance optometry education in the UK and in this pursuit is funded by industry including Alcon, Bausch and Lomb, Coopervision, No7 Contact lenses and David Thomas/Menicon. This work was also supported by Cheryl Donnelly and the BCLA. Other BUCCLE members include Alison Alderson and Graham Mouat (University of Bradford), Claire McDonnell and Orla Murphy (TU Dublin), Byki Huntjens (City, University of London), Mark Mayhem (City, University of London), Eilidh Martin (Glasgow Caledonian University), Laura Sweeney (Glasgow Caledonian University), Katherine Evans (Cardiff University), Shehzad Naroo (Aston University), Robert Conway (Anglia Ruskin University), Luisa Simo (Plymouth University), Carole Maldonado-Codina and Claire Mallon (The University of Manchester), Jo Underwood (Association of British Dispensing Opticians), Kishan Trivedy (University of Portsmouth) and Mahesh Joshi (University of Plymouth).