Development and validation of a new photographic scale to grade lid wiper epitheliopathy

Chris W. Lievens a, *, Yvonne Norgett b, Peter M. Allen c, Marta Vianya-Estopa b

a The Eye Center, Southern College of Optometry Memphis, TN, USA
b Department of Vision & Hearing Sciences, Anglia Ruskin University, Cambridge, UK
c Vision and Eye Research Unit, Anglia Ruskin University, Cambridge, UK

ARTICLE INFO

Keywords:
Lid wiper epitheliopathy
Dry eye
Lissamine green
Grading scale
Soft contact lens

ABSTRACT

Purpose: Lid wiper epitheliopathy (LWE) is a clinical sign that has been associated with dry eye disease (DED) and contact lens discomfort (CLD). This study describes the development, validation and graders’ preference of a new photographic scale for LWE, the Photographic Lid Wiper Epitheliopathy (PLWE) scale.

Methods: The PLWE grading scale was developed using LWE images selected from 57 screened patients (≥18 years of age) with confirmed LWE in both eyes. To validate the PLWE scale, a set including 20 images showing varying degrees of LWE from none to severe was chosen. To assess grading validity and grading reliability, observers were asked to grade the selected images using the PLWE and another commonly used subjective LWE grading protocol (Korb) on two separate sessions.

Results: The mean grade (±SD) of all images was not statistically significant different between the PLWE scale (1.55 ± 0.44) and the alternative grading scale (Korb, 1.47 ± 0.54) (ANOVA F1, p > 0.05). The average difference from the mean of all graders was 0.03 ± 0.53 using the PLWE scale and 0.06 ± 0.57 when using the Korb protocol (ANOVA F1, p > 0.05). The Coefficient of Repeatability was 1.04 and 1.12 for the PLWE and Korb scales (p > 0.05). Ninety-five percent of the graders found PLWE easier to use than Korb and the same percentage would consider using the PLWE scale in clinical practice.

Conclusion: The format of the PLWE is similar to other anterior segment visual grading scales and this study revealed an ease of use preference for employing the PLWE by the graders. The presence of LWE has been associated with DED and CLD, and the addition of this new photographic scale could facilitate clinical judgement and record keeping of LWE in clinical practice.

1. Introduction

Anterior eye slit-lamp findings are commonly graded with the use of illustrative grading scales and should offer high discriminability, reliability and be simple to administer [1,2]. When grading scales for the eye first gained popularity in the 1990s, wordy descriptions were used [3]. Verbal descriptive scales can be overly subjective and arbitrary to determine selected levels of severity [4]. What one clinician finds to be unusual, another clinician may find to be typical, depending on the practice setting. Numeric scales are commonly used to indicate a stage of development, progression and regression as the clinician observes, evaluates, and then assigns a numeric grade. This grade then serves as a standard and baseline by which future assessments are then compared. Grades, like this, can be used to notate deviations from normality and as a factor contributing to treatment decisions [5]. Numeric scales can be created using standard reference photographs, artistic renderings or computer-generated images and can be combined with qualitative terminology for ease of use [6]. Visual scales may afford more consistency among clinicians and, when combined with a fine incremental scale, can be very sensitive to detecting clinical change [1,5]. Visual scales are now commonplace in the assessment of the anterior eye. Their use are considered to be a best practice to reduce inconsistencies between examiners and to encourage a more uniform grading approach [7]. Common examples, at present, are the Brien Holden Vision Institute scale and the Efron Grading Scale [8,9].

Wolffsohn et al. found that 84.5 % of respondents to a worldwide survey regarding the examination of the anterior eye (680 out of 809 practitioners) regularly used grading scales in practice [3]. Such...
illustrations can be derived from real ocular photos of patients’ eyes or can be artistic renderings [10–12]. These scales allow the clinician to directly classify and compare a condition based on the referenced levels of severity depicted by the visual images in the scales [4], improving standardization and reducing subjectivity [13]. Construction of photographic scales is reliant upon images being available in the needed breadth of severity that is reflective of the population. Naturally, magnification, lighting, and other photographic conditions need to be standardized [2]. Optimal scales should have adequate grading precision, reliability, and inter- & intra-observer agreement.

At present, the most commonly used grading scale for lid wiper epitheliopathy (LWE) does not include visual representations [14,15]. Instead, the grading protocol proposed by Korb [16,17] requires observers to visualize and mentally measure the curvilinear width and length of staining on the everted eye lid and then average these measurements together to derive a severity assessment. Yamamoto et al. used a modified Korb scale in conjunction with four representative images when grading the presence of LWE in the upper and lower eyelid margins [18]. Kunnen et al. compared the subjective assessment of LWE using Korb’s protocol with a semi-automated image analysis system and found that observers overestimated the height and underestimated the width of LWE staining [19]. To overcome the challenges associated with grading LWE staining, the present work aims to develop and validate a novel photographic LWE grading scale (referred as the PLWE) and assess graders’ preference between the Korb and the PLWE grading scales.

2. Material and methods

2.1. PLWE scale development

In the present study, the PLWE grading scale was developed using LWE images selected from a previous study that had screened 57 patients (≥18 years of age) for LWE using lissamine green (LG) dye [20]. Unbranded liquid LG pipetted dye was used to control for dye volume and to not distract by brand inconsistencies (Greenpark Compounding Pharmacy, Houston Texas, USA) [21]. A double instillation was used and photographs were taken no sooner than 1-minute post second instillation [20]. Images had been anonymized and permission had been granted for their use. In total, 855 images were reviewed and a semi-automated image system (ADCIS, Advanced Concepts in Imaging Software, Saint Contest, FR) was used make an initial, objective assessment of LWE area of staining [20,22,23]. The principal investigator verified every image to ensure that the computer analysis correctly identified LG staining and not LG pooling when measuring LWE area. Areas that were judged to be LG pooling were manually removed using the software tool (semi-objective methodology) [20]. Both eyes of the 57 patients were photographed at multiple time points after dye instillation to ultimately result in a large dataset of images. For ease of comparison the same eye images were used as potential images for the present study. The area of LWE staining was used in conjunction with the clinical expertise of a panel of optometrists that reviewed and selected the final images for the PLWE scale. The panel attempted to remove images of non-continuous LWE and emphasized images in which LG had as clear a staining in its intensity as possible. All images of the everted lid (resolution of 2000*1000 digitized on 8 bits, 12x magnification, Haag-Streit B900 LED Silit Lamp system and Canon EOS 60D digital camera) were captured in raw mode, and then converted into tiff-format images. The ADCIS software automatically detects LWE when using LG dye as previously described [20,22]. As LWE may have different presentations (continuous and non-continuous staining), the calculated area of lid wiper staining (mm²) used for analysis includes all stained regions as described by ADCIS on x-axis.

Staining patterns can differ in terms of intensity, area, shape and segmentation [25] and this was considered during the image selection process. Continuous patterns of LWE staining were preferred for the photographic scale as it was felt that progression would be more clearly ascertained. The final images were chosen such that the area of LWE staining had a linear relationship to the grade level as shown in Fig. 1. The final PLWE is included in Fig. 2.

2.2. Experimental protocol

Twenty graders of varying degrees of patient care experience were recruited and compensated for their time. Ten graders where final 4th year Optometry interns with <1-year clinical experience and the rest where qualified optometrists with at least 3 years of clinical experience. None of the graders reported to grade LWE in their clinical practice (Table 4). Written informed consent was obtained after explanation of the study and possible consequences of participation. Inclusion criteria for the graders included current engagement with eye care and documented near and intermediate corrected vision of 20/20 or better at the last eye examination within 12 months of the present study. None of the graders were excluded due to the near and/or intermediate vision requirements.

To validate the PLWE scale, a set of 20 images showing varying degrees of LWE from none to severe was chosen. To assess grading validity and grading reliability, observers were asked to grade the selected images, in randomized order, using the PLWE and Korb grading scales on two occasions 1-week apart. All graders attended a virtual meeting where the principal investigator (CL) explained the study processes. During this training session, they were asked to grade a demonstration example to establish familiarity with the study process before session 1. In addition, graders received hard copies of the grading scales. After study completion, all graders completed a short survey to assess their preference for the two grading scales used in the study.

Graders were asked to follow the steps described by Korb et al. [16,17] when grading LWE using Korb’s protocol (Table 1). Graders were instructed to visually estimate and record (grade 0–3) the horizontal length and width of staining (Steps 1 and 2 in Table 1) whilst disregarding the line of Marx. The final Korb score (Step 3 in Table 1, 0.5 steps) was automatically calculated for the graders.

When grading using the PLWE (Fig. 2), graders were instructed to judge the LWE grade in 0.5 grade increments through comparison with

![Fig. 1. Plot of area vs grade for the final images used in the photographic lid wiper epitheliopathy (PLWE) scale. Graph shows linear Pearson correlation (r = 0.1.0) of LWE severity with best fit line, y-axis is grade plotted against raw value of LWE area (mm²) as measured by ADCIS on x-axis.](image-url)
the PLWE scale. In both grading methods, graders were asked to exclude the line of Marx, mentally add separate areas of LWE if non-continuous staining was observed and were not permitted to return to previous slides in order to compare grades given with the alternate method. When using the PLWE scale, images judged to be more severe than grade 3 were recorded as 3.5.

At session 1, graders were presented with the set of LWE images and completed the grading using the scales in randomized order. At session 2 (1 week later ± 1 day), the same graders graded the images in reverse method order (image order was randomized at both visits). To assist with environmental conditions, all graders used the same computer monitors under the same room lighting. Monitors were set to maximum display brightness (Dell Optiplex 3011 AIO with resolution 1600 × 900 pixels) to view the images.

2.3. Grading precision

Grading precision represents the inter-observer differences in absolute value from the mean of all graders. In other words, we are interested in comparing how close each individual grading compares to the mean of all the graders for each image using the PLWE and the Korb grading scales. The mean grading score ± standard deviation (SD) is reported.

2.4. Grading reliability

Grading reliability refers to the ability of graders to give similar values when the process is repeated. In other words, grading reliability evaluated intra-observer variability between session 1 and session 2 grading estimates for each scale. Mean values between sessions are reported and the standard deviation of this discrepancy describes the grading reliability.

2.5. Statistical analysis

All statistical analyses were performed using Analyse-it for Microsoft Excel 5.68 build 7620.32918 (Microsoft Corporation, Microsoft Excel, 2018. Available from: https://office.microsoft.com/excel). Kolmogorov-Smirnov tests were used to assess normality of the PLWE and Korb data (all p = 0.10), and parametric analysis were employed as the data were normally distributed. Values in the text and tables are presented as the mean grading score ± standard deviation (SD). To assess intra-observer differences between session 1 and 2, we evaluated the absolute mean difference, reliability and Coefficient of Repeatability. A one-way repeated measures ANOVA was used to assess differences between the grading methods. Agreement between sessions and grading scales was also assessed using Bland and Altman plots. For all analyses, the PLWE grades that were > 3.0 were considered in a bin with grade 3 in order to allow comparison with the Korb scale which had a max grade of 3.0.

3. Results

3.1. Grading precision

The average of all graders for each of the 20 images and grading scale is shown in Fig. 5, indicating the spread in the severity of LWE. The mean (±SD) of all images was 1.55 ± 0.44 for the PLWE scale and 1.47 ± 0.54 for the Korb scale. One-way repeated measures ANOVA did not show a statistically significant difference between these methods (F1, 19 = 2.37, p = 0.14).

The absolute average difference from the mean of all graders indicate how close clinicians were to the mean values given by other graders. The absolute average difference from the mean of all graders was 0.32 ±
0.38 using the PLWE scale and 0.40 ± 0.37 when using the Korb scale. One-way repeated measures ANOVA showed a statistically significant difference between these methods ($F_{1, 39} = 6.97, p = 0.01$).

### 3.2. Grading reliability

The mean difference between sessions and the grading reliability (SD of the difference between measurements) for both PLWE and Korb scales are presented in Table 2. Reliability scores were slightly higher for the PLWE scale compared to Korb. The Coefficient of Repeatability (COR) was calculated as 1.96 * SD of the difference between measurements, indicating the value where the difference between repeated measurements will lie in 95% of the cases. No statistically significant differences between mean differences between sessions, grading reliability and COR scores were found for the PLWE and Korb scales ($F_{1, 39} = 0.07, p = 0.79$).

Fig. 4 shows Bland-Altman plots of the difference between the sessions against their mean for the PLWE and Korb scales. In the PLWE plot, the bias is shown with a yellow line (mean difference between the sessions = 0.03) and the continuous blue lines represent the 95% limits of agreement (-1.01 to +1.07). Similarly, Fig. 4b, shows a mean difference between sessions of 0.06 for the Korb scale and -1.06 to +1.18 limits of agreement. No bias was found with increasing severity in either scale.

### 3.3. Between methods agreement

Analysis of the frequency distribution of all grading estimates shows the highest agreement when no LWE is present, followed by Grade 1 – slight LWE changes (Table 3). Graders scored lower with Korb grading than PLWE in the more severe cases (Grade 3, Table 3). Thus, grades are not interchangeable between PLWE and Korb scales.

Fig. 5 shows Bland-Altman plots comparing the two grading methods. The regression line shown as a continuous red line indicates a very slight bias towards higher grading with the PWLE scale than the Korb scale with increasing level of severity.

### 3.4. Graders subjective preference between grading methods

The subjective preference was captured using a short survey at the end of the session 2 (Table 4). Graders reported evaluating lid eversion in 73.3% of their contact lens patients but only 10% routinely examined lid wiper and 20% routinely used grading scales to record anterior eye findings. 95% of the graders found PLWE easier to use than Korb and the same percentage would consider using the PLWE in clinical practice.

### 4. Discussion

This study aimed to develop and validate a new grading scale to support recording of LWE staining in clinical practice. The present study used digital image analysis in the selection of images and proposed a
scale with a linear progression of LWE. All of the images included in this study were sourced from the same bank of images, taken by the same camera/instrumentation to prevent issues with magnification, lighting and image quality. In selecting images for the PLWE scale, the goal was to identify eyes in which the severity of LWE differed while the other image characteristics remained as consistent as possible. The agreement of the PLWE was compared against the commonly used Korb scale for the subjective evaluation of LWE [15]. Grading precision and reliability indicated that the novel PLWE scale was comparable to the Korb scale. The reliability (i.e. standard deviation of the discrepancy) was found to be >0.1 for both scales (see Table 2). Based on this, the authors suggest that clinicians should use a 0.5 grading step to monitor change when using the PLWE. Clinicians should also be instructed to extrapolate their grading estimates beyond the illustrated limit of the PLWE scale if LWE appears to be greater than grade 3.0. For example, if a patient presents with a continuous band of LWE encompassing the whole length and width of the lid wiper area the grade should be recorded as 3.5. Similarly, Vianya-Estopa et al. [26] recently suggested that a 0.5 grading step is adequately precise in evaluating hyperemia in the anterior eye using a visual grading scale, and previous work has also graded LWE using 0.5 steps [21]. It is noteworthy that even when previous studies have encouraged graders to use 0.1 steps, there was a noticeable aggregation of scoring around whole-integer or half-integer steps [2]. Kunnen et al. noticed objective analysis was more accurate than human observers when grading LWE, yet, objective analysis is more costly and clinicians still rely on grading scales in contact lens practice [27].

Previous studies have not reported intra-observer or inter-observer LWE reliability data. The grading reliability reported in this study (Table 2) is similar to that reported by Efron et al. [2] for grading conjunctival redness using the Brien Holden Vision Institute scale. Moreover, the COR values reported in this study also appear to be in line with the 95 % confidence limits reported by Efron et al. for corneal staining, conjunctival redness and papillary conjunctivitis (±1.2 grading scale units when using 0.1 increments) in a group of inexperienced graders [2]. In contrast, Huntjens et al. reported a slightly lower COR of 0.78 for palpebral hyperemia for an experienced grader [8]. COR data indicates the size of the change or difference in severity that can be considered statistically significant and is dependent on clinician experience and/or training with the grading system. Typically, when using anterior eye grading scales a change of 1.0 grading scale represents a clinically significant changes [11]. In this case, when using either of the LWE scales a change of 1.0 should be taken to be both clinically and statistically significant. Additionally, it is worthwhile to note that the Korb scales uses 1-point increments and PLWE allows for 0.5-point increment. As such, it is expected to find a slightly larger COR with Korb as compared to the PLWE.

A primary goal of the present study was to determine any difference between the Korb and the novel PLWE scale. Unsurprisingly, the greatest agreement between the scales was shown on the extremes (around scores of 0.0 and 3.0 as shown in the Bland and Altman plots). This suggests that graders accurately assess a lack of LWE (with only the line of Marx picking up lissamine green staining) and a large extent of LWE. Some disparity of grading scores was noted mostly near the middle of both scales where graders underestimate the amount of LWE when using the Korb scale as compared to the PLWE scale. In fact, clinicians should be aware that grades are not interchangeable between PLWE and Korb scales. In support of these findings, Kunnen et al. noted that observers tended to overestimate the height and underestimate the width of LWE staining. Because the lid wiper region is not well defined, it is a difficult process for human observers to judge the stained region as a proportion of the lid wiper total region [19]. In particular, the Korb grading protocol requires the grader to mentally measure the length and width of the curvilinear ocular anatomy (whilst ignoring the line of Marx) which can pose a challenge and can explain differences when comparing the two methods. In contrast, the PLWE scale shows real ocular images, in which the line of Marx is present in all severity grades, including grade 0. As such, the line of Marx does not have to be mentally subtracted and the images can be assessed in their natural state. The use of the Korb

Table 4

Descriptions and opinions of those graders enrolled in the present study. Lid wiper epitheliopathy scale (LWE), Photographic lid wiper epitheliopathy scale (PLWE).

<table>
<thead>
<tr>
<th>Questions</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>During a typical contact lens appointment do you use visual grading scales to record anterior eye findings?</td>
<td>Yes: 4 No: 16</td>
</tr>
<tr>
<td>On what percentage of contact lens wearing patients would you perform upper eye lid eversion?</td>
<td>73.3 % average response (range 0 % to 100 %)</td>
</tr>
<tr>
<td>Do you typically examine the lid wiper?</td>
<td>Yes: 2 No: 18</td>
</tr>
<tr>
<td>Do you typically grade LWE?</td>
<td>Yes: 0 No: 20</td>
</tr>
<tr>
<td>Which grading protocol did you find easier to use when grading LWE?</td>
<td>Korb: 1 PLWE: 19</td>
</tr>
<tr>
<td>Would you use the new PLWE scale if available in practice?</td>
<td>Yes: 19 No: 1</td>
</tr>
</tbody>
</table>

Fig. 5. Bland-Altman analysis of the difference between severity estimates using the Korb scale (Korb) versus the proposed PLWE scale. Continuous yellow line represents the mean difference (PLWE scores 0.08 in excess of Korb). The continuous red line is the linear regression fitted to the data (y = 0.10 + 0.12x, r = 0.79). The upper and lower limits of agreement are shown by the continuous blue lines (-1.24 to +1.40). Photographic lid wiper epitheliopathy scale (PLWE). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
The protocol showed that clinicians underestimated LWE compared to PLWE particularly for greater levels of severity. This might be partly explained because graders were encouraged to use a 0.5 step when using the PLWE method which included extrapolation to 3.5.

The format of the PLWE is similar to other anterior segment visual grading scales and this study showed a clear preference for employing the PLWE as compared to Korb by the graders. The presence of LWE has been associated with dry eye symptomatology [28], and the addition of this new visual scale might facilitate the examination of LWE in clinical practice. The linear design of the PLWE affords an appropriate level of discrimination for the clinician to aptly determine LWE severity. Visual scales, like this one, are simple to administer and require little to no instruction. Since the average time to report the entirety of anterior eye health has been reported as 6.8 ± 5.7 min and only a few seconds is reported all that is required for precise grading of complications, efficiency and ease of use is key [3,29]. Grading scales remain to be the most widely used tool in clinical practice to gauge findings and change over time [4]. Delaveris et al. proposed an alternative to Korb with a simplified visual grading for LWE but their work lacked a full validation analysis [21]. In line with the present work, Delaveris et al. also suggested that a simplified scoring process using four images might be sufficient to assess LWE when using lissamine green vital dye [21].

Limitations of the present study include the use of images rather than viewing an eye on the slit lamp and potential grader fatigue. However, these should not compromise the validity of the findings as similar limitations have also been highlighted by Efron et al. in validation of grading scales for contact lens complications [2]. As mentioned previously, differences in clinicians’ experience and/or familiarity with clinical grading systems might affect grading reliability [2,26]. To overcome this limitation, the present study included a balanced number of graders with different levels of clinical experience. Despite this attempt, all of them reported to not routinely assess the lid wiper and for this reason they might be considered ‘inexperienced’ in LWE assessment.

Future work should find ways to support novel and experienced clinicians establish their grading reliability as this is invaluable in clinical decision making and sharing clinical data with colleagues. Inexperienced graders should familiarize themselves with the line of Marx and LWE staining patterns as much as possible prior to implementing the PLWE to provide the greatest level of consistency. LWE has also been reported to have varied clinical presentations [30]. The PLWE images chosen had continuous patterns. Non-continuous patterns could be more difficult to judge as they would have to be mentally added prior to matching to the visual scale. Neither contact lens wearers nor dry eye patients were specifically enrolled, nor excluded, in the photographic process of image collection. A separate investigation using the PLWE in these patient bases would be of interest to determine the scale’s utility. Lower-LWE was not assessed in the present study and the PLWE is primarily intended to be used to assess the severity of upper-LWE.

In conclusion, this study proposed and validated a novel PLWE. The scale was found to be repeatable and a reliable method to assess the severity of LWE when compared to the existing Korb protocol. In addition, graders with a range of clinical experience showed a strong preference for the use of this photographic scale when evaluating LWE staining.

Commercial relationship disclosure

Financial research support from Alcon (CL), AbbVie-Allergan (CL) and Transitions (CL) in the past three years.

Funding disclosure

No specific grant from funding agencies in the public, commercial, or not-for-profit sectors was provided for this study.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Nancy Briggs, PhD: consultation in project design, Mark Wainwright Analytical Centre, University New South Wales, Sydney, Australia.

References


