Grading Bulbar Redness Using Cross-Calibrated Clinical Grading Scales

Marc M. Schulze,1,2 Natalie Hutchings,2 and Trefford L. Simpson2

PURPOSE. To determine the between-scale agreement of grading estimates obtained with cross-calibrated McMonnies/Chapman-Davies (MC-D), Institute for Eye Research (IER), Efron, and Validated Bulbar Redness (VBR) grading scales.

METHODS. Modified reference images of each grading scale were positioned on a desk according to their perceived redness (within a 0 to 100 range) as determined in a previous psychophysical scaling experiment. Ten observers were asked to represent perceived bulbar redness of 16 sample images by placing them, one at a time, relative to the reference images of each scale. Only 0 and 100 were marked on the scale, but not the numerical position of the reference images. Perceived redness was taken as the measured position of the placed image from 0 and was averaged across observers.

RESULTS. Overall, perceived redness depended on the sample image and the reference scale used (repeated measures ANOVA; P = 0.0008); six sample images had a perceived redness that was significantly different between at least two of the scales. Between-scale correlation coefficients of concordance ranged from 0.93 (IER vs. Efron) to 0.98 (VBR vs. Efron). Between-scale coefficients of repeatability ranged from five units (IER vs. VBR) to eight units (IER vs. Efron) of the 0 to 100 range.

CONCLUSIONS. The use of cross-calibrated reference grades for bulbar redness grading scales allows comparison of grading estimates obtained with different scales. Perceived redness is dependent on the dynamic range of the reference images of the scale, with redness estimates generally being found to be higher for scales with a shorter dynamic range. (Invest Ophthalmol Vis Sci. 2011;52:5812–5817 DOI:10.1167/iovs.10-7006)

In 1987, Charles McMonnies and Anthony Chapman-Davies introduced the first photographic bulbar redness grading scale in an attempt to improve the standardization of clinical procedures.1 Following this example, a number of scales have been developed since, including the Institute for Eye Research (IER; previously known as CCLRU) scale;2 the Efron scale that uses artist-rendered illustrations;3,4 or the validated bulbar redness (VBR) scale for which the reference levels had been validated objectively.5 Despite the advantages that illustrative grading scales possess over written observations or the use of purely descriptive scales,1,6–8 the still variable assessments frequently triggered criticism regarding their use. Typically, the variability of grading estimates was attributed to the subjectivity associated with the clinical application of grading scales, and was reported to occur likewise between observers or for the same observer over time.8–14

Leaving the observer-induced variability aside, grading estimates were also found to vary if different grading scales were used.12,15 Efron et al.12 reported that bulbar redness grades with the IER scale were on average 0.6 grading units higher than with the Efron scale, a finding that was later qualitatively demonstrated by Peterson and Wolffsohn.15 There are apparent differences between the four scales regarding the number of reference images, the scale levels and range, and the conjunctival region displayed. Objective techniques have been used to quantify these differences in the scale images for various physical redness characteristics,16–18 and confirmed the visual impression that the reference levels of these bulbar redness grading scales are not aligned (i.e., grade 1 in one scale does not necessarily display the same degree of redness in another scale19,20).

Because of these reasons it has been suggested that scales not be interchanged, and that grading estimates not be compared across scales.2–12,17 The ability to convert redness grades obtained with different grading scales would be particularly valuable for research settings, however, and it seems that if grading scales were better aligned, a comparison between grading estimates may be possible. In an attempt to achieve a better comparability of redness estimates, we have introduced a psychophysical scaling model21,22 that allowed the perceived redness of the McMonnies/Chapman-Davies (MC-D), IER, and Efron bulbar redness grading scales to be quantified for a 0 to 100 redness range relative to the reference images of the VBR scale.22 Similar to the studies using objective metrics,16–18 the perceived redness data also indicated a misalignment of the reference images between scales, and detected widely different dynamic ranges (i.e., the range that is covered by the reference images) between the scales. Table 1 shows the original grades for the MC-D, IER, Efron, and VBR scale and their associated calibrated grades for the 0 to 100 range as determined by psychophysical scaling.22

Reproducibility of measurements has been defined as the “closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement.”23 In the context of clinical grading, the use of different scales for the assessment of redness in the same eyes can be considered a change in the conditions of measurement. Based on this definition, the purpose of this study was to determine the between-scale agreement of the newly calibrated MC-D, IER, Efron, and VBR scales to estimate the reproducibility of grading estimates for a set of 16 sample images.

From the 1Centre for Contact Lens Research, School of Optometry, and the 2School of Optometry, University of Waterloo, Waterloo, Ontario, Canada.

Submitted for publication December 6, 2010; revised February 14 and March 7, 2011; accepted March 30, 2011.

Disclosure: M.M. Schulze, P; N. Hutchings, None; T.L. Simpson, P

Corresponding author: Marc M. Schulze, Centre for Contact Lens Research, School of Optometry, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1; m3schulz@sciborg.uwaterloo.ca.

Investigative Ophthalmology & Visual Science, July 2011, Vol. 52, No. 8
Copyright 2011 The Association for Research in Vision and Ophthalmology, Inc.
Materials and Methods

Sample Images

Sixteen sample images depicting bulbar redness of the temporal conjunctiva were selected from a database of photographs available at the Centre for Contact Lens Research (CCLR) that had been previously compiled during a clinical study from a group of experienced contact lens wearers. The previously collected redness data (subjective grading estimates and photometric chromaticity; CIE \( u' \)) were analyzed to select sample images for this study that represented a contact lens population that might typically be seen in clinical practice (Kolmogorov-Smirnov test; \( P > 0.20 \)). Regions of interest in a size of 250 \( \times \) 156 pixels were cropped out of the sample images, so that only conjunctival vascular detail, but no lids or lashes were visible.\(^{25}\) This corresponded to an area of approximately 1.1 \( \times \) 0.69 cm on the ocular surface. One eye showing an uncharacteristically high level of bulbar redness for this sample of contact lens wearers (due to a non-study related event) was included to evaluate how grading estimates with the newly calibrated scales were affected if higher degrees of redness were to be assessed. This eye, and the eye being perceived to be the least red were perceived to be the least red (left) and most red (right).

Subjective Redness Assessments

Redness was subjectively estimated on a table top within the same 1.5 m range that had been used for the psychophysical scaling\(^{21,22}\) and the subsequent calibration\(^{22}\) of the reference images of the MC-D, IER, Efron, and VBR scales. The start and end point of this 1.5 m range corresponded to the minimum and maximum redness level, and were labeled by 0 and 100, respectively. Within this range, the modified reference images of each scale (i.e., after removal of potentially confounding features such as lids or lashes)\(^{21}\) were presented so that their position matched their newly determined, calibrated reference grades (Table 1).\(^{22}\) The grading task took place under full room illumination (cool white fluorescent lighting; general color rendering index 84). Light meter (DVM 1300; Velleman, Gavere, Belgium) measures across the table surface ranged from 350 to 390 lux and were consistent for all grading assessments.

The same 10 participants that had also participated in the previous redness scaling experiments\(^{21,22}\) were asked to represent perceived bulbar redness of printed color copies of the sample images (5 \( \times \) 3 cm) by placing them relative to the unlabeled reference images of one of the four scales. No time constraints were imposed for completion of the task. The participants had no previous experience with grading of ocular conditions, other than than having completed the scaling of the reference images.\(^{21,22}\) After the placement of each image, its position was measured, translated into its corresponding redness grade between 0 and 100, and removed before the next sample image was presented for assessment. Each participant estimated the redness of the sample images four times, once per scale, with a break of at least two days between each grading session. The order of scale and sample image presentation was randomized for each participant. After completion of the experiment, the perceived redness of each sample image was averaged across observers to allow comparison between scales. The study followed the tenets of the Declaration of Helsinki and received ethics approval from the Office of Research Ethics at the University of Waterloo; informed consent was obtained from each participant before starting the study.

Data Analysis

Statistical analysis was performed using STATISTICA version 8 (StatSoft, Inc., Tulsa, OK); an alpha level of \( \leq 0.05 \) was considered statistically significant. Repeated measures (RM) ANOVA was used to determine whether the redness estimates for the sample images depended on the grading scale being used; the Bonferroni post hoc test was used to correct for multiple comparisons. The Pearson product-moment correlation coefficient (Pearson’s \( r \)) was used to evaluate the strength of linear association of the redness estimates between grading scales. Between-scales agreement of redness estimates was evaluated with the intraclass correlation coefficient (ICC),\(^{24-26}\) the correlation coefficient of concordance (CCC),\(^{27}\) the coefficient of repeatability (COR; 1.96 \( \times \) SD),\(^{6,9}\) and Bland-Altman’s limits of agreement (LOA; \( d \pm COR \)).\(^{28}\)

Results

Each participant required no more than five minutes for the 16 grading estimates with any scale, averaging to < 20 seconds for each sample image. Figure 2 shows the perceived redness for the sample images averaged across observers for each of the scales. Overall, there was a statistically significant interaction between scales and sample images (RM ANOVA; \( F_{(15, 405)} = 1.88; P < 0.001 \)).

Treating images as main effect, RM ANOVA showed statistically significant differences for 6 of the 16 sample images between at least two of the scales (Fig. 3; significant differences after Bonferroni correction for multiple comparisons are indicated by the respective scale names). For completeness, the \( P \) values for the remaining images were \( P = 0.06 \) (image 4) and \( P \leq 0.17 \) for all other images.

Table 2 shows the between-scale ICCs, CCCs, and the CORs for each pair of scales. The last column shows the mean of the differences (d) between each pair of scales; the scale that produced higher grades is indicated next to the associated mean difference. There was very strong linear agreement between grading estimates for each pair of scales (all Pearson’s \( r = 0.98 \) except IER vs. Efron \( r = 0.96 \)).

Table 1. Original and Calibrated Scale Grades from Psychophysical Scaling Experiment\(^{22}\)

<table>
<thead>
<tr>
<th>Original</th>
<th>Calibrated</th>
<th>Original</th>
<th>Calibrated</th>
<th>Original</th>
<th>Calibrated</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC-D</td>
<td>IER</td>
<td>Efron</td>
<td>VBR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>13</td>
<td>0</td>
<td>7</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>3</td>
<td>36</td>
<td>3</td>
<td>71</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>4</td>
<td>43</td>
<td>4</td>
<td>88</td>
</tr>
<tr>
<td>5</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4 shows the concordance between grading estimates for each pair of scales. The solid line represents the best linear fit between grading estimates, and the dashed line corresponds to the 45° line indicating perfect concordance.

The between-scales limits of agreement ($d_{\text{COR}}$) are shown in Figure 5 for each combination of scales. The dashed line near zero represents the mean of the differences between each combination of two scales; the solid lines show the limits of agreement ($d \pm \text{COR}$).

**DISCUSSION**

The purpose of this study was to determine the between-scale agreement of grading estimates obtained with the cross-calibrated MC-D, IER, Efron, and VBR bulbar redness grading scales.

Overall, the perceived redness depended on the sample image and the reference scale that was used (Fig. 2; RM ANOVA; $P < 0.001$). The perceived redness of 6 of the 16

| Table 2. ICC, CCC, COR, and Mean of the Differences for Each Pair of Scales |
|-------------------|-------------------|-------------------|-------------------|
| IER vs. MC-D      | 0.99              | 0.97              | 6.4 (IER)         |
| IER vs. Efron     | 0.96              | 0.93              | 7.8 (IER)         |
| IER vs. VBR       | 0.97              | 0.94              | 5.0 (IER)         |
| MC-D vs. Efron    | 0.97              | 0.94              | 7.1 (MC-D)        |
| MC-D vs. VBR      | 0.97              | 0.94              | 5.6 (MC-D)        |
| Efron vs. VBR     | 0.99              | 0.98              | 5.9 (IER)         |

**FIGURE 2.** Grading estimates compared between scales. *Denotes significant differences between scales.

**FIGURE 3.** Images with significantly different grading estimates between scales; Bonferroni corrected significant differences between scales are indicated by the respective scale names.
images was significantly different between at least two of the four grading scales (Fig. 3). In general, sample images were perceived to be different between the IER or MC-D scale versus the Efron or VBR scale; only image 16, representing the eye that was perceived to be the most red of all images, deviated from this trend. This finding suggests that redness estimates are dependent on the dynamic range of the scale being used (Table 1), as the scales having a shorter dynamic range (IER and MC-D) generally resulted in higher redness estimates than the scales with a wider dynamic range (Efron and VBR).

Despite these differences between single images, there was close agreement for the grading estimates between all scales (Table 2; Fig. 4 and Fig. 5). There were very high levels of linear association for each combination of scales (all Pearson’s \( r > 0.96 \)). The ICC represents a measure of the variability of scores between test and retest sessions to the overall variability. In this particular case, the ICC was used to quantify the reproducibility of grading estimates obtained with different scales. ICC (2,k) was selected because it estimates the agreement between assessments for a random sample of raters that can be generalized to other raters within some population, and represents an indicator of the interchangeability of the grading scales. Averaged across observers, between-scale ICCs were found to be at least 0.96, indicating very low variability between grading estimates with different scales.

The CCC is a specific type of ICC that describes the departure from concordance of repeated measurements, with a CCC of 1 representing identical scores. There was high concordance between grading estimates for each combination of scales, with levels of CCC of at least 0.93. Figure 4 provides a qualitative representation of this relationship, and shows that there were only slight deviations from perfect concordance (dashed 45° line) for each pair of scales (solid fit line). Closer inspection shows that the higher redness for the MC-D and IER scale compared with the Efron and VBR scale appears to subside with increasing redness, as indicated by the converging solid fit line toward the 45° line of equality. Overall, the highest levels of between-scale ICC and CCC were found for the MC-D and IER scale and for the Efron and VBR scale, while combinations of scales with different dynamic ranges (e.g., Efron with MC-D) resulted in weaker correlations.

In terms of grading units, the variability of grading estimates between any pair of scales was very low, as indicated by the between-scale CORs (Table 2) and LOAs (Fig. 5). The between-scale LOAs show the range of grading estimates that can be expected 95% of the time when two different scales are used. It is quantified by the COR as a measure of the variability of the grades relative to the mean of the differences which indicates if there is systematic bias in the grading estimates between scales. There was a small but systematic bias toward higher grades for scales with shorter dynamic range (MC-D and IER), while scales with similar dynamic range showed no such trend (Table 2 and Fig. 5, dashed horizontal line). Overall, the between-scale CORs were small (indicating low variability and
good repeatability) and ranged from five (IER vs. VBR) to eight grading units (IER versus Efron) for the 0 to 100 bulbar redness range. In terms of grading units, the variability of assessments did not seem to be dependent on the dynamic range of the scales; it appeared, however, that CORs were slightly higher when grading estimates with the pictorial Efron scale were compared with the photographic scales. Overall, these findings suggest that there is close agreement between the grading estimates with the newly calibrated scales. In particular, it appears that grading scales with similar dynamic range provide closer agreement of grading estimates.

There is only one study that quantitatively compared subjective grades between bulbar redness scales. Efron et al. reported that the mean bulbar redness grades (across all observers) were approximately 0.6 grading units higher (for a 0 to 4 range) with the IER scale compared with the Efron scale for the same set of sample images. Proportionally, this means that grades were approximately 15% higher on average when the IER scale was used, whereas mean redness grades were only different up to 4% between any pair of the newly calibrated scales (Table 2). In general, CORs are typically used to quantify the variability of grades for test/retest settings with a single scale, while for this study CORs were calculated to estimate the differences of grading estimates between scales. This complicates a direct comparison with other studies, however, it allows to estimate how the variability between scales compares with the test/retest variability that is typically present with subjective grading. In this study, the between-scale CORs (Table 2) were found to be similar or even smaller than within-scale test/retest CORs that were previously reported. Therefore, the calibration of the grading scales appears to provide closer agreement between grading estimates than previously reported when different scales were used, which implies that the newly calibrated grading scales may be used interchangeably. This finding may provide great potential for application in research settings. In general, if comparisons of grading estimates are required, the use of the same scale is encouraged for every clinician involved; however, this may not always be possible. Wiegleb and Sickenberger have reported that different scales are popular in different parts of the world. Therefore it would be of particular benefit to researchers of geographically disparate research centers (e.g., involved in a multi-center study), who could continue using the reference images to which they are accustomed while assigning cross-calibrated reference grades which provide better agreement between scales than the original scale steps.

In conclusion, the newly calibrated grading scales were capable of producing highly reproducible redness estimates across scales. There were differences for the redness estimates...
between scales for some of the sample images only, and if images were found to be different, these differences appeared to be dependent on the dynamic range provided by the reference images of the respective grading scale. Redness estimates tended to be higher for scales with a comparatively short dynamic range (MC-D and IER) than found for the scales with wider dynamic ranges (Efron and VBR); scales with similar dynamic ranges showed closer agreement between grading estimates than scales with different dynamic ranges. Overall, there was very high agreement between the grading estimates of all scales, and it appears that using the newly calibrated grading scales might reduce the between-scale variability when subjectively estimating redness. The use of the newly calibrated scales in a more typical grading setting and with more experienced observers seems to be the logical next step to further evaluate this hypothesis.

References