

The evaluation of bulbar redness grading scales

by

Marc-Matthias Schulze

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Vision Science

Waterloo, Ontario, Canada, 2010

©Marc-Matthias Schulze 2010

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

ABSTRACT

The use of grading scales is common in clinical practice and research settings. A number of grading scales are available to the practitioner, however, despite their frequent use, they are only poorly understood and may be criticised for a number of things such as the variability of the assessments or the inequality of scale steps within or between scales.

Hence, the global aim of this thesis was to study the McMonnies/Chapman-Davies (MC-D), Institute for Eye Research (IER), Efron, and validated bulbar redness (VBR) grading scales in order to (1) get a better understanding and (2) attempt a cross-calibration of the scales. After verifying the accuracy and precision of the objective and subjective techniques to be used (chapter 3), a series of experiments was conducted.

The specific aims of this thesis were as follows:

- Chapter 4: To use physical attributes of redness to determine the accuracy of the four bulbar redness grading scales.
- Chapter 5: To use psychophysical scaling to estimate the perceived redness of the four bulbar redness grading scales.
- Chapter 6: To investigate the effect of using reference anchors when scaling the grading scale images, and to convert grades between scales.
- Chapter 7: To grade bulbar redness using cross-calibrated versions of the MC-D, IER, Efron, and VBR grading scales.

Methods:

- Chapter 4: Two image processing metrics, fractal dimension (D) and % pixel coverage (% PC), as well as photometric chromaticity (u') were selected as physical measures to describe and compare redness in the four bulbar redness grading scales. Pearson correlation coefficients were calculated between each set of image metrics and the reference image grades to determine the accuracy of the scales.
- Chapter 5: Ten naïve observers were asked to arrange printed copies of modified versions of the reference images (showing vascular detail only) across a distance of 1.5m for which only start and end point were indicated by 0 and 100, respectively (non-anchored scaling). After completion of scaling, the position of each image was hypothesised to reflect its perceived bulbar redness. The averaged perceived redness (across observers) for each image was used for comparison to the physical attributes of redness as determined in chapter 4.
- Chapter 6: The experimental setup from chapter 5 was modified by providing the reference images of the VBR scale as additional, unlabelled anchors for psychophysical scaling (anchored scaling). Averaged perceived redness from anchored scaling was compared to non-anchored scaling, and perceived redness from anchored scaling was used to cross-calibrate grades between scales.
- Chapter 7: The modified reference images of each grading scale were positioned within the 0 to 100 range according to their averaged perceived redness from anchored scaling, one scale at a time. The same 10 observers who had participated in the scaling experiments 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. Perceived redness was taken as the measured position of the placed image from 0 and was averaged across observers.

Results:

- Chapter 4: Correlations were high between reference image grades and all sets of objective metrics (all Pearson's r 's ≥ 0.88 , $p \leq 0.05$); each physical attribute pointed to a different scale as being most accurate. Independent of the physical attribute used, there were wide discrepancies between scale grades, with sometimes little overlap of equivalent levels when comparing the scales.
- Chapter 5: The perceived redness of the reference images within each scale was ordered as expected, but not all consecutive within-scale levels were rated as having different redness. Perceived redness of the reference images varied between scales, with different ranges of severity being covered by the images. The perceived redness was strongly associated with the physical attributes of the reference images.
- Chapter 6: There were differences in perceived redness range and when comparing reference levels between scales. Anchored scaling resulted in an apparent shift to lower perceived redness for all but one reference image compared to non-anchored scaling, with the rank order of the 20 images for both procedures remaining fairly constant (Spearman's $\rho = 0.99$).
- Chapter 7: Overall, perceived redness depended on the sample image and the reference scale used (RM ANOVA; $p = 0.0008$); 6 of the 16 images had a

perceived redness that was significantly different between at least two of the scales. Between-scale correlation coefficients of concordance (CCC) ranged from 0.93 (IER vs. Efron) to 0.98 (VBR vs. Efron). Between-scale coefficients of repeatability (COR) ranged from 5 units (IER vs. VBR) to 8 units (IER vs. Efron) for the 0 to 100 range.

Conclusions:

- Chapter 4: Despite the generally strong linear associations between the physical characteristics of reference images in each scale, the scales themselves are not inherently accurate and are too different to allow for cross-calibration based on physical redness attributes.
- Chapter 5: Subjective estimates of redness are based on a combination of chromaticity and vessel-based components. Psychophysical scaling of perceived redness lends itself to being used to cross calibrate the four clinical scales.
- Chapter 6: The re-scaling of the reference images with anchored scaling suggests that redness was assessed based on within-scale characteristics and not using absolute redness scores, a mechanism that may be referred to as clinical scale constancy. The perceived redness data allow practitioners to modify the grades of the scale they commonly use so that comparisons of grading estimates between calibrated scales may be made.
- Chapter 7: The use of the newly calibrated reference grades showed close agreement between grading estimates of all scales. The between-scale variability was similar to the variability typically observed when a single scale is repeatedly

used. Perceived redness appears to be dependent upon the dynamic range of the reference images of the scale.

In conclusion, this research showed that there are physical and perceptual differences between the reference images of all scales. A cross-calibration of the scales based on the perceived redness of the reference images provides practitioners with an opportunity to compare grades across scales, which is of particular value in research settings or if the same patient is seen by multiple practitioners who are familiar with using different scales.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Trefford Simpson, for his guidance and advice through all these years. Thank you for always encouraging me to also take a different perspective on my research.

I am sincerely grateful to Dr. Natalie Hutchings, for her guidance and support. Thank you for all the time you spent helping me understand the 'stats' and with editing my publications. I would like to thank Dr. Deborah Jones for getting me started on my research on bulbar redness grading scales and her continued support, and to Dr. Paul Stolee for introducing me to qualitative research on scales and Rasch Analysis. Thank you all for your help.

I am very grateful to Dr. Charles McMonnies, Dr. Nathan Efron, and the International Association of Contact Lens Educators (IACLE) for providing me with high resolution copies of the original scale reference images and for allowing me to use them in my research.

I would like to thank Dr. Des Fonn for allowing me to do my PhD research as a part of the Centre for Contact Lens Research. Thank you to everyone in the CCLR for making my stay so enjoyable.

I would also like to thank Kevin van Doorn, Raiju Babu, Dr Vivian Choh, and Dr. Jeff Hovis for their willingness to provide input on my research.

Thank you to all graduate officers and to Krista Parsons for your help with the administrative part of my studies. Thanks to Anne Weber for making TAing with

you so much fun and to all the participants of my experiments for your time and help. And to all fellow graduate students for making Waterloo a fun place to be at.

Finally, I would like to thank my parents, Ulli and Jo, for always believing in me and supporting me in my decisions.

DEDICATION

To Dörte.

I could not have done it without you!

Table of Contents

LIST OF FIGURES.....	XV
LIST OF TABLES	XVIII
1 LITERATURE REVIEW	1
1.1 METROLOGY	1
1.1.1 <i>History of Measurement</i>	1
1.1.1.1 <i>Theory of numbers</i>	1
1.1.1.2 <i>Early measurement systems</i>	2
1.1.1.3 <i>Helmholtz: Counting and measuring</i>	3
1.1.1.4 <i>Campbell: Measurements in physics</i>	4
1.1.1.5 <i>Stevens: Measurements in psychology</i>	5
1.1.1.6 <i>The representational theory of measurement</i>	7
1.1.1.7 <i>Soft metrology</i>	9
1.1.1.8 <i>Terminology</i>	11
1.2 CONTACT LENS COMPLICATIONS.....	13
1.2.1 <i>Conjunctiva</i>	13
1.2.1.1 <i>Anatomy</i>	13
1.2.1.2 <i>Conjunctival vasculature</i>	14
1.2.1.3 <i>Bulbar hyperaemia</i>	15
1.3 CLINICAL GRADING.....	16
1.3.1 <i>Types of Grading Scales</i>	16
1.3.1.1 <i>Descriptive grading scales</i>	17
1.3.1.2 <i>Illustrative grading scales</i>	19
1.3.1.3 <i>Computer-generated grading scales</i>	26
1.3.2 <i>Research on grading scales</i>	27
1.3.2.1 <i>Number of scale steps</i>	27
1.3.2.2 <i>Use of incremental scale steps</i>	28
1.3.2.3 <i>Criticism related to grading scales</i>	29
1.3.2.4 <i>Level of measurement</i>	30
1.4 OBJECTIVE TECHNIQUES TO ASSESS BULBAR REDNESS.....	32
1.4.1 <i>Colourimetry</i>	33
1.4.1.1 <i>Application in ocular research</i>	38

1.4.2	<i>Image processing</i>	39
1.4.2.1	<i>Application in ocular research</i>	42
1.4.3	<i>Fractal Analysis</i>	46
1.4.3.1	<i>What are fractals?</i>	47
1.4.3.2	<i>How are fractal dimensions calculated?</i>	50
1.4.3.3	<i>Application in ocular research</i>	54
2	RATIONALE	57
3	THE ACCURACY AND REPEATABILITY OF OBJECTIVE AND SUBJECTIVE TECHNIQUES TO ESTIMATE BULBAR REDNESS	61
3.1	INTRODUCTION.....	61
3.2	METHODS.....	66
3.2.1	<i>Image processing</i>	66
3.2.2	<i>Fractal analysis</i>	68
3.2.3	<i>Photometric chromaticity</i>	71
3.2.4	<i>Psychophysical scaling</i>	72
3.3	RESULTS.....	73
3.3.1	<i>Image processing</i>	73
3.3.2	<i>Fractal analysis</i>	74
3.3.3	<i>Photometric chromaticity</i>	76
3.3.4	<i>Psychophysical scaling</i>	78
3.4	DISCUSSION.....	83
3.4.1	<i>Image processing</i>	83
3.4.2	<i>Fractal Analysis</i>	84
3.4.3	<i>Photometric chromaticity</i>	85
3.4.4	<i>Psychophysical scaling</i>	88
4	THE USE OF FRACTAL ANALYSIS AND PHOTOMETRY TO ESTIMATE THE ACCURACY OF BULBAR REDNESS GRADING SCALES	90
4.1	OVERVIEW.....	91
4.2	INTRODUCTION.....	93
4.3	METHODS.....	95
4.3.1	<i>Grading scale images</i>	95
4.3.2	<i>Image processing and fractal analysis</i>	97
4.3.3	<i>Photometric measurements</i>	103

4.3.4	<i>Data analysis</i>	104
4.4	RESULTS	104
4.5	DISCUSSION	109
4.5.1	<i>Accuracy of the grading scales</i>	109
4.5.2	<i>Comparison and cross-calibration between the grading scales</i>	112
4.6	CONCLUSION	115
4.7	ACKNOWLEDGEMENTS	115
5	THE PERCEIVED BULBAR REDNESS OF CLINICAL GRADING SCALES	117
5.1	OVERVIEW	118
5.2	INTRODUCTION	119
5.3	METHODS	123
5.3.1	<i>Grading Scale Images</i>	123
5.3.2	<i>Participants</i>	124
5.3.3	<i>Psychophysical scaling method</i>	124
5.3.4	<i>Perceived vs. physical redness</i>	125
5.3.5	<i>Data analysis</i>	126
5.4	RESULTS	127
5.5	DISCUSSION	133
5.6	ACKNOWLEDGMENTS	141
6	THE CONVERSION OF BULBAR REDNESS GRADES USING PSYCHOPHYSICAL SCALING	142
6.1	OVERVIEW	143
6.2	INTRODUCTION	145
6.3	METHODS	148
6.3.1	<i>Psychophysical scaling</i>	148
6.3.2	<i>Conversion of grades between scales</i>	150
6.3.3	<i>Data Analysis</i>	151
6.4	RESULTS	151
6.5	DISCUSSION	158
6.5.1	<i>Non-anchored vs. anchored scaling</i>	160
6.5.2	<i>Conversion of grades between scales</i>	162
6.6	ACKNOWLEDGEMENTS	165

7	GRADING BULBAR REDNESS USING CROSS-CALIBRATED CLINICAL GRADING SCALES	166
7.1	OVERVIEW	166
7.2	INTRODUCTION	168
7.3	METHODS.....	171
7.3.1	<i>Sample images.....</i>	<i>171</i>
7.3.2	<i>Subjective redness assessments.....</i>	<i>172</i>
7.3.3	<i>Objective redness measurements.....</i>	<i>173</i>
7.3.3.1	<i>Image processing and fractal analysis.....</i>	<i>174</i>
7.3.3.2	<i>Photometric chromaticity measurements.....</i>	<i>175</i>
7.3.4	<i>Data Analysis.....</i>	<i>176</i>
7.4	RESULTS	177
7.5	DISCUSSION	184
7.5.1	<i>Agreement between scales.....</i>	<i>184</i>
7.5.2	<i>Physical redness attributes vs. subjective grading estimates.....</i>	<i>188</i>
7.6	CONCLUSION.....	189
8	CONCLUSIONS AND FUTURE WORK.....	191
8.1	FUTURE WORK.....	198
	REFERENCES	201
	APPENDIX.....	230

List of Figures

Figure 1-1:	The McMonnies/Chapman-Davies (MC-D) scale. ⁵²	19
Figure 1-2:	The Institute for Eye Research (IER) scale for bulbar redness. ^{76,77}	21
Figure 1-3:	The Efron scale for conjunctival hyperaemia. ⁵⁶	22
Figure 1-4:	The Validated Bulbar Redness (VBR 5) scale. ²⁵	25
Figure 1-5:	The VBR 10 scale. ²⁵	25
Figure 1-6:	CIE Luminosity functions.	34
Figure 1-7:	The 2° 1931 CIE Standard Observer.	35
Figure 1-8:	The CIE 1931 x,y chromaticity diagram.	36
Figure 1-9:	The CIE 1976 u',v' chromaticity diagram.	37
Figure 1-10:	The Cantor Dust.	47
Figure 1-11:	The Koch Snowflake.	49
Figure 1-12:	The coastline of Britain.	50
Figure 1-13:	The Sierpinski Carpet.	52
Figure 1-14:	The box-counting method.	53
Figure 1-15:	Log-log plot of count and scale (ϵ).	54
Figure 3-1:	The target analogy for accuracy and precision.	63
Figure 3-2:	Generated line art images with known pixel count to evaluate the accuracy of ImageJ.	67
Figure 3-3:	The Sierpinski Triangle.	70
Figure 3-4:	View through the eyepiece of the photometer.	72
Figure 3-5:	Result of image subtraction.	74
Figure 3-6:	Limits of agreement for photometric chromaticity, u'	77

Figure 3-7:	Averaged perceived redness plots for test vs. retest.	80
Figure 3-8:	Limits of agreement for psychophysical scaling.	81
Figure 3-9:	Variability between observers relative to averaged perceived redness.	82
Figure 4-1:	Simulated fractal dimensions (D) representing different degrees of vascular branching on the conjunctiva.	95
Figure 4-2:	The bulbar redness grading scales analyzed. ³⁻⁶	96
Figure 4-3:	Image pre-processing steps.	99
Figure 4-4:	Standardized photometric setup.	103
Figure 4-5:	Resulting images from fractal analysis.	105
Figure 4-6:	Scale grades vs. physical metrics.	108
Figure 5-1:	The McMonnies/Chapman-Davies (MC-D), Institute for Eye Research (IER), Efron, and Validated Bulbar Redness (VBR) grading scales.	121
Figure 5-2:	Averaged perceptual scores for session 1 vs. session 2	128
Figure 5-3:	The perceptual scores for the references images.	129
Figure 5-4:	The effect of type of color information.	130
Figure 6-1:	The modified reference images of the MC-D, IER, Efron, and VBR scales.	147
Figure 6-2:	Experimental setup for anchored redness scaling.	149
Figure 6-3:	Shift to lower perceived redness with anchored (filled squares) compared to non-anchored scaling (open circles).	152
Figure 6-4:	Non-anchored vs. anchored redness scaling (solid fit line) for the MC-D, IER, and Efron scales.	155
Figure 6-5:	Perceptually scaled reference images within the 0 to 100 range for the anchored scaling experiment.	156
Figure 7-1:	The sample images perceived to be the least (left) and most red (right).	172
Figure 7-2:	Image pre-processing steps for the sample images.	175
Figure 7-3:	Setup for photometric measurement of the sample images.	176

Figure 7-4:	Grading estimates compared between scales.....	177
Figure 7-5:	Images with significantly different grading estimates between scales.	178
Figure 7-6:	Between-scales concordance of grading estimates.....	180
Figure 7-7:	Between-scales limits of agreement (LOA).....	181

List of Tables

Table 1-1:	Scales of measurement.	6
Table 1-2:	Slit Lamp Findings Classification Scale, scale for injection. ⁷²	17
Table 1-3:	Slit lamp classification system after Mandell, scale for injection. ⁵⁵	18
Table 3-1	Accuracy of fractal analysis.	75
Table 3-2:	Repeatability coefficients for photometric chromaticity, u'	76
Table 3-3:	Repeatability coefficients for psychophysical scaling.	78
Table 4-1:	Image size and resolution.	97
Table 4-2:	Signal-to-noise ratio.	98
Table 4-3:	Pearson correlation coefficients between scale grades and their associated physical attributes.....	106
Table 5-1:	COR, CCC, and Pearson's r for the three image sets.	127
Table 5-2:	Pearson correlation matrix between perceived and physical redness; all $p < 0.001$	131
Table 5-3:	Partial correlation coefficients.	132
Table 5-4:	Multiple regression analysis.	133
Table 6-1:	The perceived redness of each reference image for non-anchored (top) or anchored (bottom) psychophysical scaling.....	153
Table 6-2:	Conversion table between scales.	157
Table 7-1:	Calibrated scale grades from psychophysical scaling experiment. ²²	169
Table 7-2:	ICC, CCC, COR, and mean of the differences for each pair of scales.....	179
Table 7-3:	Pearson correlation matrix.	182
Table 7-4:	Partial correlation coefficients.	183
Table 7-5:	Stepwise multiple regression analysis.....	184

*“When you can measure what you are speaking about, and express it in numbers,
you know something about it; but when you cannot measure it,
when you cannot express it in numbers,
your knowledge is of a meagre and unsatisfactory kind.”*

Lord Kelvin

Lecture to the Institution of Civil Engineers, May 3, 1883¹

1 Literature Review

1.1 Metrology

The 'Bureau International des Poids et Mesures' (BIPM), or 'International Bureau of Weights and Measures', has defined metrology as "*the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology*".² The term metrology stems from the ancient Greek 'metron' (measure) and 'logos' (word), and was derived from 'metrologia', the theory of ratios.³ The following section will review the history of measurement theory and science, and address the question of measurability, of what can or cannot be measured.

1.1.1 History of Measurement

1.1.1.1 Theory of numbers

As the concept of measurement requires an understanding of the different types of numbers, this section will start with a brief explanation of the number types. *Natural* numbers such as 1, 2, 3 etc. are the most basic form of numbers, and are the numbers that children first use when they learn how to count.^{4,5} If we include zero, we have the *whole* numbers, and with the negative numbers we reach the next level, the *integers*. It is important to note that each new type of numbers always contains the previous level within it, which means that integers include the natural numbers, zero, and the negatives of the natural numbers. The *rational* numbers represent the next level of the number types, and use fractions to combine an integer numerator and a non-zero integer denominator. Rational numbers can also be expressed in decimal notation with infinite repetitions, such

as $\frac{5}{3} = 1.66666667$. Rational numbers are the common output for any type of measurement.⁵ *Irrational* numbers are numbers that cannot be expressed as a fraction of two integers, and may be written in decimal notation, their digits however are non-repeating. Examples for irrational numbers are π (i.e. 3.1412592654.....) or the square root of natural numbers such as $\sqrt{3}$ (i.e. 1.732050808.....). All the number types described above make up the *real* number system^{4,5}, which can be extended to the *complex* numbers by the addition of an imaginary part to the real numbers. A more detailed description would go beyond the scope of this thesis, however, it shall be mentioned that complex numbers are required for the creation of certain fractals such as the Mandelbrot set.⁶

1.1.1.2 Early measurement systems

The roots of the history of measurement go back as far as to the bartering of goods in prehistoric time, when the value of goods was measured against each other. During the Urban Revolution in Egypt and Mesopotamia, this system was further modified by using tokens of different values to pay for certain goods.⁷ The ancient Greeks investigated measurement with a more philosophical approach, by trying to establish a relationship between numbers and the real world. Pythagoras for example tried to establish arithmetics as the fundamental study in physics.⁸ The term arithmetics is derived from Greek 'arithmos' (number) and 'arithmein' (to count), and is a first indication of how measurement is involved with numbers and counting, a concept later more explicitly elaborated by Helmholtz.⁹ The concept of measurement and counting was also used by the ancient Arabs, who defined the qirat, the seed of a coral tree, as a unit of weight. Using a balance scale, the

number of qirats required to balance the scale was counted and thus represented a measure of weight.^{7,10}

Another early measurement system was based on human morphology, and used human dimensions for measuring distances.^{2,7} Some of these units are still being used today, such as the inch (end of the thumbnail to the first knuckle), the foot (heel to toe), or the yard (nose to the end of the middle finger on a laterally stretched arm).⁷ Despite using human morphology as basis for this measurement system, its measurement units were not fixed, and differed between geographical regions, between occupations, or simply because the human dimensions differ between individuals.² The first unified system of measurement was the decimalized metric system, which was introduced in 1795 and has since become the commonly accepted measurement system in most parts of the world.² Since 1960, the International System of Units (SI), which is based on the metric system, is the recommended practical system of units of measurement², and is accepted as official system of measurement for all nations except for Myanmar, Liberia, and the United States.¹¹

1.1.1.3 Helmholtz: Counting and measuring

In 1887, Helmholtz discussed the foundations of measurement in his book ‘Zählen und Messen, erkenntnis-theoretisch betrachtet’, in which he established an analogy between counting and measuring.^{8,9,12} His approach to measurement was based on the “*fact that we express as quantities, through concrete numbers, situations of real objects*”.^{9,12} In his understanding, measurements were concerned with the determination of quantities that could be regarded as the sum of the

associated parts. Based on his analogy, he derived measurability conditions which related the outcomes of measurements to each other. In other words, the same conditions that are the basis for counting, order and addition, are required to allow for measurement.¹² With his understanding of measurement, von Helmholtz set the stage for what is known today as the representational theory of measurement (see below).⁸

1.1.1.4 Campbell: Measurements in physics

During the 20th century, the question of measurability, of what can or cannot be measured, was the focus of discussion between physicists and psychologists. Campbell discussed measurability in his book 'Physics - The Elements', and extended Helmholtz' theory further.^{8,12} Campbell agreed that the measurability of a property required two conditions to be satisfied. First, an empirical order relation had to be established, which was considered the basic requirement for measurement. Second, the measures had to satisfy the logical properties of addition.^{8,9,12} This means that measures could either be obtained through direct measurements, for example by comparison of objects to a measurement scale (*fundamental* quantities, e.g. the length of objects), or by measurement of other quantities so that the measures of interest could be derived (*derived* quantities, e.g. density as a function of mass and volume).¹³

Campbell was part of the 'Committee of the British Association for the Advancement of Science' that discussed the possibility of quantitative estimates of sensations as determined with psychophysical measurements. The Committee, which consisted of members of the physical and psychological sciences, disagreed

on the measurability of psychological sensations, and in the end failed to recognize psychophysical measurements as being valid.^{8,12} In the final report of the Committee, Campbell defined measurement *“in the broadest sense, as the assignment of numerals to objects or events according to rules”*.¹⁴ According to the physicists’ view in the Committee, these rules were not conformed to by psychophysical measurements because those would not allow for addition operations, which were considered to represent a fundamental requirement for measurement in general.^{8,12}

1.1.1.5 Stevens: Measurements in psychology

At about the same time, Stanley Smith Stevens, a psychologist, addressed the issue of measurability by suggesting that a more general theory of measurement was needed.^{12,14} By doing so he tried to circumvent the measurability condition as proposed by the physicists half of the Committee^{8,12} that empirical addition operations were required, and based his approach on the use of the equality of ratios as fundamental requirement for measurement.¹² Therefore he suggested to classify scales of measurement in terms of the possible transformations that leave the scales invariant, and proposed four levels of measurement: nominal, ordinal, interval, and ratio (Table 1-1).¹²⁻¹⁴

Table 1-1: Scales of measurement.

Scale	Conditions for scale transformation	Mathematical group structure	Permissible statistics (invariance)	Examples
Nominal	One-to-One e.g. A, B, C → i, ii, iii	Permutation	Number of cases Mode	Numbering of players Labels
Ordinal	Preserve Order e.g. 1,2,3,4 → 1,5,7,10	Isotonic	Median Percentiles	Hardness of minerals Beaufort scale Intelligence
Interval	Preserve equal intervals e.g. 1,2,3,4 → 2,4,6,8	Linear	Mean Standard deviation Pearson's r	Celsius temp. scale Fahrenheit temp. scale Position
Ratio	Preserve equal ratios e.g. 10:5 → 2:1	Similarity	Coefficient of variation	Length Mass Density Kelvin temp. scale

Note: modified after Stevens¹⁴, with examples taken from multiple sources.¹²⁻¹⁷

The empirical operations that are involved with the process of measurement are reflected in the four scale types that Stevens proposed. Their classification is therefore meant to be cumulative, in the sense that e.g. the permissible statistics for interval scales also allow the use of the statistics applicable for nominal and ordinal scales.¹⁴ The first empirical operation in measurement is the identification and classification of objects that have a certain property of interest in common, with the numerals being assigned for better discrimination of these objects only. The numbering of players on a sports team is the prime example, as the numbers on the players backs are no indication of ability (as in higher number, better player), but represent a means to identify individual players on the team.

The next higher level of measurement is the ordinal scale, which introduces the empirical operation of rank-ordering. After having classified the objects based on a particular property they possess (nominal scale), they are now ranked to

represent their relative order¹³ with respect to this property. The ranks are no indication of differences or magnitudes, however.¹²

The equality of intervals or differences is an attribute of interval scales.¹²⁻¹⁴ Any linear transformation of interval scales by multiplication by and addition of a constant will leave these scales invariant^{12,14,15}, such as in the conversion of temperature from the Fahrenheit (F) to the Celsius (C) scale. Therefore the zero point is arbitrary for interval scales, as 0 °C corresponds to 32 °F, despite both scales measuring temperature. Since there is no absolute zero for interval scales, it is not possible to form ratios between two measures on the scale. This means that the difference between 10 °C and 20 °C is equivalent to a difference between 20 °C and 30 °C, however, 20 °C can only be considered to be 10 °C warmer than *but not* twice as warm as 10 °C.

The most powerful measurement scale is the ratio scale, which is the most restricted when it comes to possible transformations as it requires an absolute zero point.¹⁴⁻¹⁶ Ratio scales can be transformed by multiplying by a constant only, as in the transformation from inches to centimetres. The most prominent ratio scale is the number scale itself that we use for everyday counting, with other ratio scales being used to measure length or weight (fundamental ratio scales), or for density or force (derived ratio scales).

1.1.1.6 The representational theory of measurement

A measurement approach that embraces and tries to combine the viewpoints of both physical and psychological sciences is the representational theory of measurement¹² as defined by Suppes and Zinnes¹⁸, Pfanzagl¹⁷, and later by

Finkelstein.^{8,19,20} In the representational theory of measurement, Finkelstein* has defined measurement in the wide sense as “*the assignment of numbers to properties of objects or events in the real world by means of an objective empirical operation, in such a way as to describe [or to represent] them*”.^{8,19,20} In other words, numbers are mapped to manifestations of objects, so that the relation between the manifestations is *represented* by the numbers.

Finkelstein further discriminates between strongly and weakly defined measurement.⁸ Strongly defined measurement conforms to the paradigm of the physical sciences and provides measures that are based on empirical observations. It is supported by a theory that explains the relations that are expressed in the numerical mapping. Measures that are based on this theory are thus related to other measures for which that same theory holds, for example the measures of force relate to measures of mass and acceleration. Weakly defined measurements are also based on an empirical process, but do not conform to the paradigm of the physical sciences and are not supported by a strong theory. Weakly defined measurements are generally found in the social and psychological sciences; according to this definition, psychophysical measurements are weakly defined measurements as well.⁸

Finkelstein’s definition of measurement implies that any measurement, either in the physical (strongly defined) or in the psychological sciences (weakly

* Note: Throughout his publications, Finkelstein used slightly modified versions of his definition of measurement based on the representational approach. In general, all definitions had the same meaning, but certain terminology was used interchangeably (e.g. attribute and property). The definition as used here represents a combination of his definitions that best fits into the context and terminology of this thesis.

defined), is possible as long as certain measurability conditions are satisfied.^{8,12} These conditions include that the measurements are based on an empirical process which is the result of an observation, not of a thought process only, and that the measurement is independent of the observer, i.e. it is objective.^{8,12} This view is extended by Rossi, who suggests that a measurement scale may also be constructed by assigning numbers to manifestations of a property based on functional relations only (i.e. no empirical observations for the construction of the scale are required). In Rossi's view, the more important part of measurement is to test the new scale, independent whether it is developed from empirical observations or from functional relations, with objects that exhibit the same property of interest but are not included in the newly developed scale in order to evaluate if the scale measures what it intends to measure. Therefore, "*a characteristic may be considered measurable if, after trying to measure it, we succeed*".¹²

1.1.1.7 Soft metrology

The irreconcilable opinions between physical and social scientists in the 1940s and 1950s (see 1.1.1.4) with regard to the measurability of psychological sensations led to a strict separation of measurements in physics and psychology.¹² A recent call for papers, entitled 'Measuring the Impossible'²¹, is an example that shows the emerging interest in the measurability of sensory events. The rationale for this call is based on the (new) understanding that scientific research is interested in a number of phenomena that could not be assigned to a single field only, but that were rather seen as being "*multidimensional and multi-disciplinary, with strong cross-over between physical, biological and social sciences*".²²

The term soft metrology refers to the branch of metrology that is concerned with the measurement of appropriate physical parameters that correlate with perceptual quantities.²² To be specific, soft metrology has been defined as *“measurement techniques and models which enable the objective quantification of properties which are determined by human perception”*, where *“the human response may be in any of the five senses: sight, smell, sound, taste and touch”*.^{12,22}

The concept of the measurement of appearance is closely related to the science of psychophysics.²³ An example of how perceptual quantities for the measurement of visual appearance can be related to physical parameters is the measurement of body height. If a group of children is positioned for a photograph so that the smallest child stands in the front, and the tallest in the back, we order them based on how we perceive their body height to differ.²² Perceptually, the differing heights of the children could be estimated by a psychophysical technique known as magnitude estimation, which requires that a number is assigned to each child that relates to the perception of its height.²⁴ If the estimated heights correlate with physical measurements of the children’s height, for example by means of a measuring tape, a measurement scale (i.e. the measuring tape) is found that allows physical measurement of the perceptual sensation.²²

Another, more complex aspect of the measurement of visual appearance relates to what we perceive as red when we assess an eye, and which physical characteristic(s) might best be attributed to this perceptual response. The clinical grading of bulbar redness can be considered another type of magnitude estimation²⁴, where eyes are compared to a standard, in this case a reference scale. But what do individuals look for when they assess the redness of the conjunctiva?

Do they look for changes in colour and luminance? Or are assessments based on spatial criteria rather, such as how many vessels can be seen, how large the vessels are, or how much area is covered by the vessels? A number of studies have reported on the measurement of such physical characteristics of redness that may be related to subjective estimates, such as the use of a photometer to measure chromaticity or image processing techniques that determine the area covered by vessels (see section 1.4 for details). The Validated Bulbar Redness (VBR) scale²⁵ is an example for how a measurement scale was established that was based on quantifying the perception of redness (using psychophysical scaling) and correlating it with a physical characteristic of redness, chromaticity.

1.1.1.8 Terminology

To conclude this section on measurement theory, the terminology that will be used in this thesis will be briefly summarized. In general, the terminology is following closely the definitions of the International Vocabulary of Metrology (VIM) which was developed by eight measurement and standardization organizations (e.g. BIPM or ISO). It is meant to be a common reference for scientists and researchers in all fields of study by trying to harmonize the language used in the field of metrology.²⁶

First it is important to clarify that measurements always refer to a specific *property*, or to a number of properties of an object or event, but not to the object or event itself.^{8,12,17,18,20,27} The terms *property*, *characteristic*, *attribute*, and *metric* will be used interchangeably in this thesis when measurements are discussed. In many cases we want to measure the *quantity* of a certain characteristic⁹, which will

then allow us to compare between certain *manifestations* (with different quantities) of this object property. According to VIM²⁶, a quantity is defined as “*property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference*”. The actual numerical outcome of any measurement that is described in this thesis will be referred to as *measure*, *score*, or, as this thesis focuses on grading scales, as *grade*. This assignment of numbers to the various manifestations of any characteristic of an object then allows discrimination between the measures, and to establish empirical relations between them. This terminology is in agreement with the current understanding of measurement theory, and has been described in more detail elsewhere.^{8,12,17,20,27}

1.2 Contact Lens Complications

Contact lens wear is often associated with the formation of clinical changes in ocular structures that may require clinical management or even medical treatment. These changes or possible complications may affect the cornea, the limbus, the conjunctiva, the sclera and the eyelids, and may be due to multiple causes.²⁸⁻³⁰ Complications may affect different tissue layers, and show a diversity of clinical signs, such as conjunctival redness or corneal staining.³¹ To ensure an exact diagnosis and corresponding management, practitioners commonly refer to aids such as slit lamp biomicroscopes and grading scales to facilitate clinical decision making. Bulbar hyperaemia, a vasodilation of the blood vessels in the bulbar conjunctiva, is one of the possible complications associated with contact lens wear, and will be the focus of this section.

1.2.1 Conjunctiva

1.2.1.1 Anatomy

The conjunctiva is a transparent, vascular mucous membrane that covers the sclera and the inner surface of the eyelids.³¹⁻³⁸ Anatomically, the conjunctiva can be divided into three parts according to their location: palpebral, forniceal, and bulbar.^{31-34,36,37} The palpebral conjunctiva is firmly attached to the tarsal plate of the eyelids. The forniceal conjunctiva connects the palpebral conjunctiva as a ring-like pouch of loose tissue with the bulbar conjunctiva.^{32,34,37} The bulbar conjunctiva is loosely attached to the anterior layer of the sclera, the episclera, and ends at the limbal conjunctiva.^{31-34,37,38} Histologically, the conjunctiva consists of two layers, the epithelium and the underlying stroma.^{33,37,38}

1.2.1.2 Conjunctival vasculature

The ophthalmic artery supplies all arteries of the eye.^{32,34,39} The peripheral palpebral artery and the anterior ciliary arteries are responsible for the blood supply of the conjunctiva.⁴⁰ Branches of the peripheral palpebral artery form the posterior conjunctival arteries that supply the peripheral conjunctiva.⁴⁰ Two sets of arterial branches continue on from the anterior ciliary arteries, a deeper one penetrating the sclera, and more superficial episcleral arteries that continue on to form the episcleral circle. Two sets of vessels branch from the episcleral circle: recurrent conjunctival vessels that supply the superficial bulbar conjunctiva, and small arterioles that proceed towards the limbus to supply the limbal or corneal arcades.^{34,39-41} The blood is drained from the conjunctiva and the corneal or limbal arcades through a system of episcleral veins that return the blood towards the rectus muscles.^{34,39,41}

The bulbar conjunctival vessels represent the most superficial layer of vasculature, with the deep episcleral vascular plexus lying posterior to the conjunctival vessels.^{31,34} Accordingly, ocular redness can be present in both the conjunctival and the episcleral vasculature, and its discrimination is required to allow for appropriate clinical decision making.³⁸ The bulbar conjunctival vessels consist of tortuous arteries and straight veins³³, and of capillaries and post-capillary venules.³⁹ They are only found in the stromal layer of the conjunctiva, while the conjunctival epithelium does not contain any vessels.^{33,38} Vasodilation of the conjunctival blood vessels is referred to as bulbar hyperaemia.^{28,38,42} Because of their superficial location, conjunctival vessels are brighter red⁴⁰ and show concurrent movement with the conjunctiva, for example when pushed with a cotton tip.^{31,33,40}

Vessels in the deeper episcleral vascular plexus have a duller red colour⁴⁰ and do not move with the conjunctiva.^{31,33}

1.2.1.3 Bulbar hyperaemia

One of the most prominent clinical signs of ocular irritation is bulbar hyperaemia, a vasodilation of the conjunctival blood vessels.^{28,42} In its normal state, the conjunctiva is a transparent tissue with subtle vessels in front of the white sclera.⁴⁰ Upon ocular irritation, the circumference of the vessel enlarges, resulting in increased blood flow that gives the eye a red appearance. Because of this red appearance, bulbar hyperaemia is commonly referred to as bulbar redness.^{28,43-45}

Ocular discomfort is commonly associated with some level of bulbar redness.²⁸ Bulbar redness may be associated with various factors, including exposure to environmental stimuli such as allergens or air pollutants⁴⁵, foreign bodies^{31,45}, dry eye⁴⁶, hypoxia^{30,47-50}, diurnal variations⁵¹, and contact lens wear.^{28,30,42,49,52} Bulbar redness may be associated with virtually every adverse response to contact lens wear.²⁸ The causes of contact lens induced bulbar redness include metabolic influences such as hypoxia, mechanical irritation (e.g. a poorly fitting or a damaged lens), or toxic reactions to contact lens solutions (e.g. hydrogen peroxide), among others.^{28,30,42} Because of the multiplicity of possible causes, it is quite likely that nearly all contact lens wearers will exhibit a certain level of hyperaemia at some point.²⁸ Therefore, the implementation and recording of baseline measurements and the subsequent monitoring of changes is crucial for the successful management of ocular redness.²⁸⁻³⁰

1.3 Clinical Grading

The clinical care for patients and contact lens wearers is a crucial part of the daily routine for eye care professionals in clinical practice and research settings. The assessment and recording of the current ocular status represents a legal requirement in the ophthalmic field.^{53,54} To allow the detection of possibly clinically significant changes, for example due to contact lens wear, reliable record keeping is imperative.^{53,55} Before the introduction of grading scales, the assessment and recording of ocular signs was often based on the use of descriptive (qualitative) terms such as ‘absent’, ‘normal’, ‘slight’, ‘mild’, ‘moderate’ or ‘severe’. While the use of descriptive terms allows flexibility for the practitioners in the assessment of clinical presentations, they are non-systematic and are dependent on individual (practitioner-dependent) interpretations of their meaning.^{53,56}

1.3.1 Types of Grading Scales

To overcome possible inappropriate clinical decision making due to qualitative terms only⁵³, numeric grading systems have been recommended for better standardization of patient records and to reduce the subjectivity inherent in clinical assessments.^{52,55,57} The use of numeric grades serves as a standard by which previous and current ocular states can be compared, facilitates the detection of possible change, and provides a basis for statistical analysis of the changes expressed by the numeric grades.^{53,58,59} Based on these recommendations a number of clinical grading scales have been introduced in the ophthalmic field. Grading scales employ grades that are systematically assigned to terms or illustrations in order to *“enable the quantification of the severity of a condition with reference to a*

set of standardized descriptions or illustrations".⁵⁷ Particularly in the field of contact lens research and practice, for which the detection of small changes is required so that possible treatment may be initiated, grading scales are commonly used to aid with these assessments.^{25,43,44,60-71}

1.3.1.1 Descriptive grading scales

One of the first grading scales for the assessment of the anterior segment of the eye was a slit lamp classification scale developed by the 'Food and Drug Administration' (FDA) of the United States of America.⁷² The scale employed grades from 0 to 4 to classify different levels of severity that were each linked to a single term followed by a more specific description of the associated ocular state. Scales for five ocular conditions (edema, corneal neovascularization, corneal staining, injection and tarsal abnormalities) were provided, with the scale for injection being shown Table 1-2.

Table 1-2: Slit Lamp Findings Classification Scale, scale for injection.⁷²

Grade	Descriptive term	More specified description of the ocular state
0	NONE	No injection present
1	TRACE	Slight limbal (mild segmented), bulbar (mild regional), and/or palpebral injection
2	MILD	Mild limbal (mild circumcorneal), bulbar (mild diffuse), and/or palpebral injection
3	MODERATE	Significant limbal (marked segmented), bulbar (marked regional or diffuse), or palpebral injection
4	SEVERE	Severe limbal (marked circumcorneal), bulbar (diffuse episcleral or scleral), or palpebral injection

Robert Mandell presented a similar descriptive slit lamp classification system to be used for the assessment of contact lens complications which expanded the number of categories compared to the FDA system.⁵⁵ Decimalized increments were used as references to describe the type of problem or its respective location, and were not meant to represent incremental steps corresponding to an increase in severity. Recommendations regarding possibly required interventions (e.g. temporary cessation of contact lens wear) were included to allow for appropriate patient management if contact lens induced complications occurred. The classification system for injection is shown in Table 1-3.

Table 1-3: Slit lamp classification system after Mandell, scale for injection.⁵⁵

	Classification	Grade
A	None	0
B	Mild conjunctival hyperaemia which is likely due to excess lacrimation and/or adaptation	
	a) palpebral	1.1
	b) palpebral and/or bulbar	1.2
C	Mild circumcorneal injection	1.3
D	Moderate conjunctival hyperaemia	
	a) palpebral	2.1
	b) palpebral and/or bulbar	2.2
E	Moderate circumcorneal injection	2.3
F	Severe conjunctival hyperaemia	
	a) palpebral	3.1
	b) palpebral and/or bulbar	3.2
G	Severe circumcorneal injection	3.3
H	Other. (Grade by severity as either 1.9, 2.9, 3.9, 4.9.)	

1.3.1.2 Illustrative grading scales

The use of photographs for the assessment of ocular conditions has been suggested by numerous authors to further standardize clinical procedures.^{58,73-75} Following recommendations by Kahn et al.⁷³ that the use of photographs would help to further reduce the impact of subjectivity in clinical observations, McMonnies and Chapman-Davies introduced the first photographic grading scale for the assessment of bulbar redness (MC-D scale).⁵² To develop the scale, 20% hypertonic saline solution was instilled into the lower conjunctival sac of one eye to artificially induce hyperaemia, and sequential photographs of the inferior conjunctiva were taken during the recovery of the eye. After recovery to baseline, a vasoconstricting agent was used to reduce visible hyperaemia to its minimum, and further photographs were taken. The selection of the photographs to be used as scale reference images was subjective. The images corresponding to the minimum (grade 0) and maximum (grade 5) level of redness were selected first, and four additional images were subsequently selected to represent equally distributed intermediate steps. A modified version of the MC-D scale is shown in Figure 1-1, with the original scale being shown in Appendix A.



Figure 1-1: The McMonnies/Chapman-Davies (MC-D) scale.⁵²

The original format⁵² (Appendix A) of the MC-D scale was modified for the purpose of this thesis with regard to the arrangement of the images and by adding the respective reference grades to the associated images.

To evaluate its clinical performance, the new photographic scale was used by McMonnies and Chapman-Davies in a number of studies for which the impact of contact lens wear on the development of bulbar hyperaemia was investigated.^{49,52} They demonstrated that the new scale was capable of detecting statistically significant changes in bulbar hyperaemia between hard, soft, and non-contact lens wearers, and found high agreement between different observers using the scale (inter-observer) and for the same observer at different time points (intra-observer). Based on these findings they concluded that the inclusion of photographs appeared to reduce the subjectivity of redness assessments.⁵²

Following recommendations by Terry et al.⁷⁵ regarding the advantages of using a photographic reference system in contact lens research, the Cornea and Contact Lens Research Unit (CCLRU) developed a photographic set of grading scales for ten common contact lens complications⁷⁶, which is currently available as '*Institute for Eye Research*' (IER) grading scales⁷⁷ (Appendix A). Each of the ten scales displays four levels of severity ranging from 'very slight' (Grade 1) to 'severe' (Grade 4). According to the instructions for the application of the scales, a change of more than one grade is clinically significant, with grades greater than 2 to be considered outside normal limits.⁷⁷ Murphy et al.⁴³ and Pult et al.⁴⁴ however independently determined that, when using the IER scale, a bulbar redness grade greater than 2.6 should be considered abnormal, and suggested that this might be due to the unusually white appearance for the reference image corresponding to grade 1 that might have caused the dynamic range of the scale to shift.⁴³



Figure 1-2: The Institute for Eye Research (IER) scale for bulbar redness.^{76,77}

It is important to note that the bulbar redness scale does not display the same eye at different stages of severity (Figure 1-2), but is composed of images depicting three different eyes, with only the reference images corresponding to grades 1 and 2 showing the same eye. A further noteworthy feature is that the set of scales, despite only depicting four levels of severity from grade 1 to 4, is intended to represent 5-step scales including an ‘imaginary’ grade 0 (corresponding to absent) to be considered when assessing ocular conditions.^{67,71}

Suggesting that the use of grading scales would not only be an asset for research but also for individual practitioners, Nathan Efron introduced a set of grading scales that used artist-rendered illustrations instead of photographs for the eight most commonly seen contact lens complications.⁵⁶ The set of illustrations was designed so that the four key tissue types affected by contact lens wear were covered, and included epithelial staining and microcysts, stromal edema and neovascularization, endothelial polymegathism and blebs, conjunctival hyperaemia, and papillary conjunctivitis. For each set of illustrations, traffic-light colouring was used to emphasize increases in severity over a range of five stages from normal (Grade 0, green framing) to severe (Grade 4, red framing). The scale for conjunctival hyperaemia is shown in Figure 1-3.

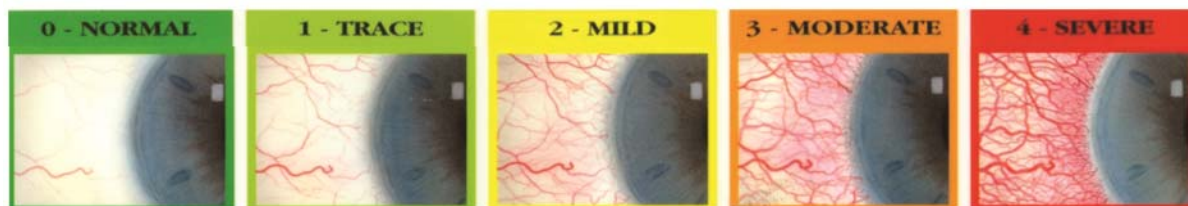


Figure 1-3: The Efron scale for conjunctival hyperaemia.⁵⁶

Efron preferred artist-rendered illustrations since those, despite not depicting ‘real’ conditions, allowed for better standardization of the images (with respect to eye, illumination, magnification, or angle of view) and for a systematic advance of severity from each stage to the next. In addition, Efron suggested that painting of the severity levels allowed highlighting specific features associated with each condition, while other, confounding artefacts, could be avoided in painted scales.^{56,60,78} After reviewing feedback from practitioners who had been using the scales in clinical practice, Efron introduced the Millennium edition of grading scales that included eight further complications of contact lens wear that had been missing in the first set of scales.⁷⁸ Efron established criteria for the use of his scales, and pointed out that changes in the ocular state of a patient of more than 0.7, or grades greater than 2 were to be considered abnormal and required clinical action.⁷⁸ The complete Millennium edition including all 16 contact lens complications can be found in Appendix A.

The ‘Validated Bulbar Redness’ scale (VBR) is a 100-point photographic scale that was developed at the School of Optometry in Waterloo, Canada, as part of my undergraduate research project.^{25,79} To acquire the images for the scale, standardized settings regarding illumination, magnification, and gaze were

established. Photographs of the right nasal conjunctiva of 15 participants were taken with a digital camera attached to a zoom photo slit lamp that was interfaced to a personal computer. To achieve a wider range of redness, bulbar hyperaemia was induced by instillation of 5% hypertonic saline into the eye of one participant, and photographs of the recovering eye were taken. Twenty-one images that were estimated to cover a bulbar redness range of about 0 to 70 were selected from the set of all images, and four additional images were modified using Adobe PhotoShop 5.0 to further extend the redness range.²⁵

Psychophysical scaling was used to select the reference images for this scale. The 25 images were randomly presented on a tabletop, and nine observers (separated into two groups consisting of four optometrists and five optometry students) were asked to position the images within a designated 1.5m range so that the separation reflected the observer's perception of redness. Only the start and end points of this range were labelled with 0 and 100 to represent minimum and maximum redness, respectively. Images with a low amount of redness were to be placed closer to 0 with increased redness being identified by positioning closer to the 1.5 m endpoint. After completion of the task, the position of each image represented its perceived redness with respect to the other 24 images within the 1.5m range. There were high linear associations between the arrangements of all observers, independent of their level of experience in the assessment of redness.²⁵

The spectrophotometer SpectraScan PR650 (Photo Research Inc., Chatsworth, CA, USA) was used to determine various photometric quantities for the photographed bulbar conjunctivae. The photometric quantities were then compared to the associated perceived redness as determined by psychophysical scaling. There

was a very strong linear association between averaged perceived redness and photometric chromaticity.²⁵ This objective validation sets the VBR scale apart from the other three illustrative scales (MC-D, IER, and Efron), for which the reference image selection had been based on clinical experience and subjective judgments only.^{52,56,60}

Based on the perceived redness data of the 25 images, two 100-point grading scales were developed, using five (Figure 1-4) and ten reference images (Figure 1-5), respectively. There was no image representing grade zero for both versions of the scales, as none of the 25 images was perceived to be absolutely free from redness. Since the bulbar conjunctiva usually exhibits some level of redness, it was decided that the exclusion of an image representing grade zero represented only a minor limitation to the scales. Approximately equal perceptual and physical scale steps for both scales were established by selecting the images closest to each 10-point step. The linear association between these interval scale steps and photometric chromaticity was found to be $r \geq 0.993$ (Pearson's product moment correlation coefficient) for both the 5-picture and the 10-picture scale.²⁵ However, it needs to be mentioned that despite both scales being mainly composed of images depicting the same eye at increasing levels of severity, the images corresponding to grades VBR 10 and VBR 20 were taken from different eyes.

Validated Bulbar Redness (VBR 5) Scale

© Schulze M, Jones D, Simpson T

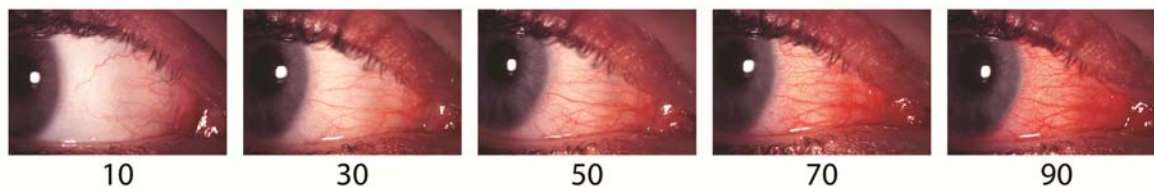


Figure 1-4: The Validated Bulbar Redness (VBR 5) scale.²⁵

Validated Bulbar Redness (VBR 10) Scale

© Schulze M, Jones D, Simpson T

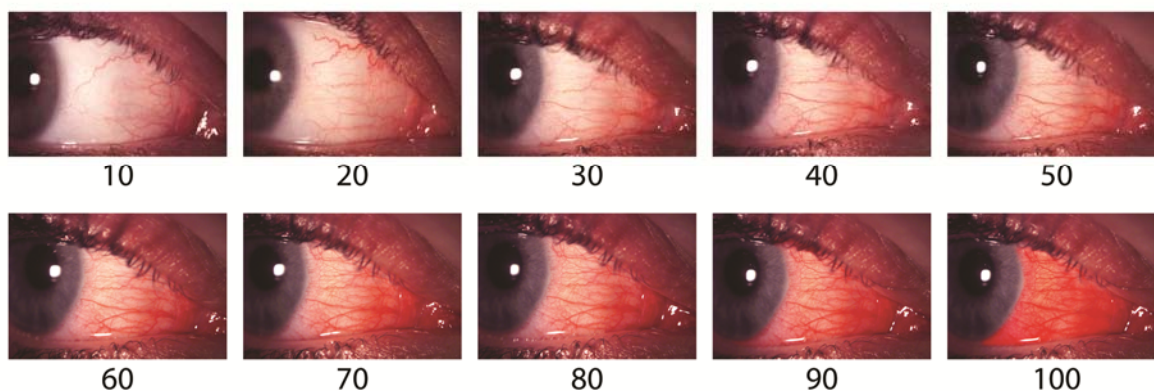


Figure 1-5: The VBR 10 scale.²⁵

The performance of the newly developed grading scales was evaluated by asking 19 observers with three different levels of experience (none, basic, and high) to use the scales for the assessment of 30 photographic slides of bulbar redness. Each participant used each of the scales twice, with the 30 sample images being presented in randomized order for each of the four sessions. Test/retest repeatability of the assessments was very high for both of the scales, independent of the level of experience.²⁵

1.3.1.3 Computer-generated grading scales

A different type of illustrative grading scales that has been introduced in recent years are computer-generated continuous grading scales that take advantage of so-called morphing software.^{25,61,80,81} Morphing software allows the continuous transformation of images displaying different degrees of severity into each other, so that a series of intermediate scale steps can be generated.^{25,61,81} In use, continuous grading scales allow the depiction of virtually every possible intermediate step by adjustment of a slider in order to find the best match to the current ocular state a patient presents with. There is disagreement regarding the repeatability that can be obtained by means of continuous scales, however, with one study favouring continuous scales⁶¹, one study finding superior repeatability for discrete scales²⁵, and another study finding no significant difference between the discrete and continuous scales.⁸⁰

Because of the possibility to depict any intermediate scale step, continuous grading scales represent highly sensitive tools for the detection of clinical change^{25,61,81}, and would therefore represent a valuable tool particularly in research settings. For individual practitioners, however, their application is limited, as a photo slit lamp system is required so that the eye of the patient can first be captured and then simultaneously displayed on the computer screen to be analyzed with the continuous scale. This more involved setup is also perhaps the reason for the limited use of continuous scales in clinical research settings, with only very few studies reporting their use for the assessment of changes over time.^{80,82-84} Since discrete illustrative scales on the other hand may be easily used during the

assessment at the slit lamp biomicroscope, this might according to Efron⁸⁰ explain the more frequent use of discrete scales for such studies.

1.3.2 Research on grading scales

Research on grading scales in the past has focused on evaluating their application in clinical practice or research settings, and has addressed a number of factors that may affect the scales' performance and repeatability.^{49,52,53,58,61,63-65,67,79,85-88} The repeatability of assessments between observers (inter-observer) or for the same observer at different time points (intra-observer) has received the most attention in this context, and is affected by factors such as the number of scale steps^{25,52,59,89} or the fineness or coarseness of the scales.^{25,58,59}

1.3.2.1 Number of scale steps

Although a number of clinical decisions are of binary nature only, for example regarding the presence or absence of a pathological finding, these dichotomous decisions carry limited information.⁸⁹ Since the assessment of clinical conditions and the management of patients is in part concerned with the detection of change⁷⁷, binary decisions may not be sufficient for some clinical purposes. Therefore, five to seven reference steps have been recommended as optimum number for clinical grading scales.^{25,53,60,89} Most of the grading scales for bulbar redness are based on this design, with five reference images for the Efron⁵⁶ and VBR²⁵ scale and six for the MC-D⁵² scale. The IER⁷⁷ scale, although only showing four reference images, is also based on a 5-step design, as the 'absent' condition (representing grade 0) is not shown but was recommended to be used for assessments as well.⁷¹ Grading scales with a limited number of reference steps do

also seem to better suit the users of the scales, as implied by the large majority of users preferring the 5-picture over the 10-picture version of the VBR scale due to its smaller size, even though the repeatability of both scales was almost identical.²⁵

1.3.2.2 Use of incremental scale steps

The performance and repeatability of grading scales do not only depend on the number of reference steps provided but also if, and how many, incremental steps are used. Interpolation of reference steps has been achieved by decimalization of 0 to 4 scales, e.g. by using 0.5 or 0.1 incremental steps, or by using integers for 100-point scales such as the VBR scale. It has been shown that the repeatability and concordance of assessments are closely connected to the number of incremental steps used, with a fine line between how coarse a scale should be to provide acceptable sensitivity to detect change and the concordance between repeated assessments. Coarse scales, for example scales that employ five reference steps but do not allow the use of intermittent steps, are likely to produce highly concordant results but the high levels of concordance are at the expense of a reduced ability of the scale users to detect change.^{58,59} It has therefore been recommended that coarse scales may be more appropriately used in studies where concordance between observers, for example in multi-centre studies, is of interest.⁴² Scales with very fine incremental steps, on the other hand, were suggested for studies for which all assessments are made by a single observer only, and where the detection of very small changes, for example for the assessment of tissue reactions to different contact lens materials, is required.^{42,61} To estimate the sensitivity of scales, Bailey et al.⁵⁸ proposed criteria that related the size of the scale increments to the standard deviation of the discrepancy (s_d) between repeated

assessments. Based on these criteria, a scale provides fine sensitivity to detect change if the size of the scale increment did not exceed $\frac{1}{3} s_d$, while moderate sensitivity was defined to not exceed $1 s_d$.

The interpolation of scale steps has been adopted in multiple clinical research settings, with different scale increments selected depending on the purpose of the respective study.^{43,44,60,61,63-68,86,87,90-93} Nevertheless, it appears that practitioners are somewhat reluctant to take advantage of all possible steps within the scale range and resort to primarily using 'round' numbers (e.g. multiples of 5 for 100-point scales)^{25,87,94,95}, so that interpolation of scale steps, but with a limited number of increments only, has been suggested to achieve a compromise between concordance and sensitivity of assessments.^{25,94}

1.3.2.3 Criticism related to grading scales

The development of quantitative grading scales and the addition of illustrations to further improve the scales have perhaps contributed to a better standardization of clinical assessments. Despite these putative advances, grading scales are – and will be – used for the subjective assessment of clinical conditions, and the resulting variability of assessments between observers or for the same observer at different time points represents a major criticism to their use.^{61,67,68,75,87,94}

The variability between assessments can be quite extensive, as Fieguth and Simpson have reported that the variability of subjective grading estimates for thirty sample images was found to be at least 25% and on average even 55% of the whole scale range. To overcome the subjectivity inherent in grading, automated grading systems have been recommended that estimate redness based on physical

characteristics. The objective techniques to estimate redness are summarized in Chapter 1.4.

Aside from subjectivity being a factor for the variability of the assessments, the grading scales themselves have been focus of criticism, which included unequal distribution of scale steps^{43,58} or differences in the scale range covered.^{43,70,87} Inspection of Figures 1-3 to 1-6 shows that there are also differences in the way bulbar redness is displayed in the MC-D, IER, Efron, and VBR scales, for example regarding the number of scale steps, the scale range or the conjunctival region selected to display the different stages of bulbar redness. A number of grading scales have also been analyzed objectively using digital image processing.^{88,96,97} The physical metrics that were used to describe the change in redness across the scale range were found to be different between scales as well (also see section 1.4).^{88,96} Efron reported that grading estimates obtained for the same eye varied depending on the scale being used, and that the repeatability of these assessments was affected as well.⁸⁷ Because of these differences between the scales, it has been recommended that scales not be interchanged, or grading estimates of different scales not be compared.^{61,87,88,96} However, it would be beneficial for clinicians to be able to compare grading estimates obtained with different scales.

1.3.2.4 Level of measurement

A further aspect that is of importance in the field of grading scales is the level of measurement¹⁴ that can be achieved with a particular scale. For clinical practice, the use of interval scales was recommended because of the inherent equality of intervals or differences between scale steps.^{62,67} There are different

options to evaluate if scales that are being used in the field of health sciences can be considered to be interval scales. In psychology and behavioural sciences, where rating scales or questionnaires are frequently used to assess character traits or abilities of patients, Rasch Analysis is considered the method of choice to determine if a scale or questionnaire allows measurement at the interval level.^{7,98} Rasch Analysis has also found its application in eye care, for example for the measurement of vision disability⁷ and the development of a new anxiety scale aimed at patients in the optometric practice.⁹⁹ Rasch Analysis is a mathematical model which applies statistical tests (fit statistics) to determine the appropriate items to be used on a questionnaire (e.g. which tasks a cataract patient is capable of doing), so that redundant or confounding items can be identified and removed. Rasch Analysis utilises logit values (i.e. logarithmic odds $[\ln\left(\frac{p}{1-p}\right)]$) of a patient affirming that statements about tasks in a questionnaire apply to himself/herself¹⁰⁰ that allow the ordering of items according to their difficulty and patients according to their ability on a single interval scale.^{7,99,100}

In contact lens research, the assessment of patients is most commonly done using illustrative scales.^{43,44,47,49,52,67,91,93,101-103} As these illustrative scales have a physical basis, the linearity of their steps may be evaluated by comparison to physical metrics that measure the same property.^{88,96} The reference images of the VBR scale for example were selected using psychophysical scaling, and its scale grades were subsequently validated by demonstrating their strong linear correlation to photometric chromaticity.²⁵ The evaluation of the accuracy of the available bulbar redness scales was the focus of one part of the research conducted for this thesis.

1.4 Objective Techniques to Assess Bulbar Redness

There have been a number of attempts to automate the clinical grading of redness, particularly for research settings when the detection of small changes is of interest.^{42,62,67-70,88,94,97,104-109} In general, this automation involved the establishment of physical metrics to describe bulbar redness that were based on possible subjective strategies perhaps applied during the clinical assessment of redness.^{67,107,110} The assessment of redness is likely based on at least two general strategies which may also be used conjointly, the first being estimating chromaticity/luminance and the second one based on describing the spatial structure or pattern of the visible conjunctival vessels.^{23,94} Accordingly, the objective quantification of redness was attempted by photometric techniques or image processing that produced variables based on colour (e.g. photometric chromaticity^{70,107} or relative redness^{67,97}) or on spatial structures (e.g. area of vessel coverage^{42,67,88,97}, vessel calibre, or the number of vessels⁶⁷). In the experiments that were conducted for this PhD, photometry and image processing techniques were used to objectively quantify redness of the reference images of the MC-D, IER, Efron, and VBR grading scales. A technique that has been applied to objectively quantify the retinal vasculature is fractal analysis, which, among other things, estimates fractal dimension, a measure of the complexity of structures. Fractal analysis was used during this research as a new method to quantify bulbar redness, and will thus be reviewed in the last part (1.4.3) of this section.

1.4.1 Colourimetry

Radiometry is the science of measuring the electromagnetic radiation for the frequency range between 3×10^{11} and 3×10^{16} Hz, which corresponds to a range of wavelengths between 0.01 and 1000 μm that include the ultraviolet, the visible, and the infrared part of the electromagnetic spectrum.^{23,111} Photometry is concerned with the measurement of the visible part of the spectrum only (380 nm to 770 nm).²³ The spectral response of the eye varies depending on the wavelength of the light.^{23,111,112} The 'Commission Internationale d'Eclairage' (CIE) developed luminosity functions for standard observers that describe the spectral sensitivity of the eye depending on wavelength. The cones, the photoreceptors of the eye that are responsible for high luminance and colour vision, have the highest sensitivity at 555 nm, while the rods, the photoreceptors for vision at low luminance, have their highest sensitivity at 505 nm.^{23,111,113} The photopic luminosity function $V(\lambda)$ for photopic vision and the scotopic luminosity function $V'(\lambda)$ for scotopic vision are shown in Figure 1-6.

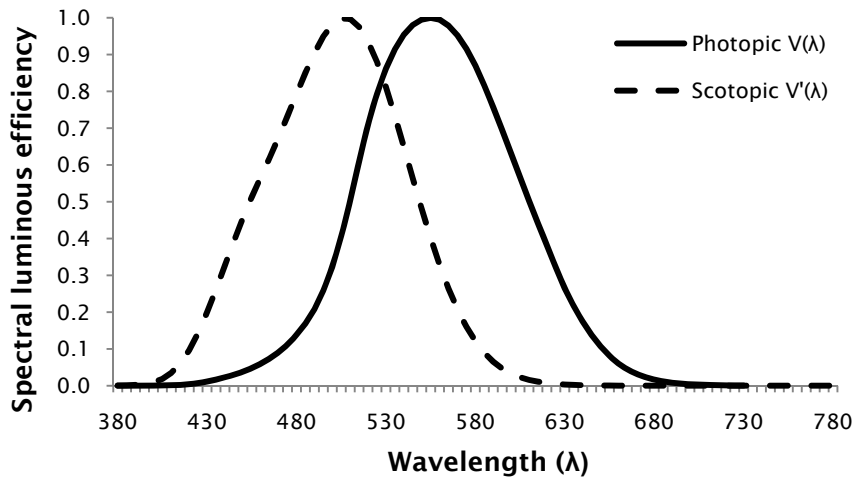


Figure 1-6: CIE Luminosity functions.

The CIE 1924 photopic ($V(\lambda)$; solid line) and CIE 1951 scotopic luminosity functions ($V'(\lambda)$; dashed line). The functions were created based on data provided at <http://www.cvrl.org/>.

Colourimetry is the science of the measurement of colour.^{23,114} There are a number of colour measurement systems that attempt to describe colour numerically. Most of these measurement systems consider colour as being composed of three attributes, one relating to luminance, and the other two to saturation and hue that correspond to the chromaticity of the colour.^{23,114} In independent attempts to quantify colour, Guild (in 1928) and Wright (in 1931) performed experiments in which observers were asked to visually match each wavelength of light in the visible spectrum by additive mixture of three primary lights.^{23,113} Based on these experimentally obtained data, and after the application of linear transformation operations to achieve a better representation of colour, the CIE derived the 2° 1931 CIE Standard Observer.^{23,113,115} The 2° 1931 CIE Standard Observer employs a set of colour matching functions (\bar{x} , \bar{y} , and \bar{z}) that can be thought of as spectral weighting functions that allow the modeling of any colour

(Figure 1-7).^{23,113,115} It is worth noting that they only represent linear transformations of the experimentally obtained cone sensitivities of an average observer, with the \bar{y} colour matching function being transformed in a way so that it matched the 1924 CIE luminosity function for photopic vision (Figure 1-6).¹¹⁵

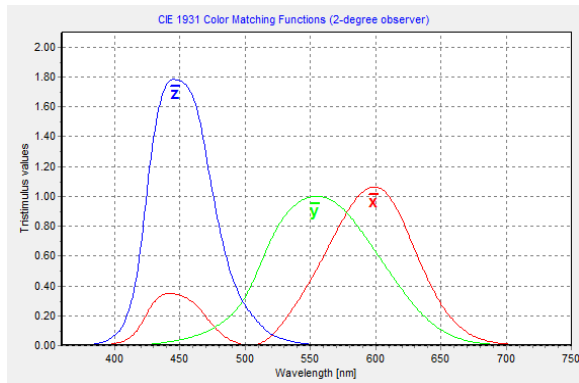


Figure 1-7: The 2° 1931 CIE Standard Observer.

Image created with software available at <http://www.efg2.com/Lab/Graphics/Colors/Chromaticity.htm>.

The colour matching functions in Figure 1-7 show the amount that each of three imaginary primary lights is required to contribute so that the colour of one unit of radiant power of the respective wavelength can be matched.^{23,113,115} Based on these colour matching functions it was possible to derive X , Y , Z tristimulus values (i.e. integrals of \bar{x} , \bar{y} , and \bar{z} each multiplied by the spectral distribution of the colour stimulus)²³ that were used to develop the X , Y , Z colour space. Despite being composed of three components, only Y has a perceptual correlate in lightness (i.e. the perception of luminance), while X and Z do not (directly) perceptually correspond to hue and saturation.^{22,23} Therefore, the CIE recommended the use of the chromaticity coordinates x , y , and z (obtained by transformation from the tristimulus values) to describe colour chromatically in the CIE 1931 x,y chromaticity

diagram (Figure 1-8).²³ The horseshoe-shaped curve in the CIE 1931 x,y chromaticity diagram represents the spectrum locus that contains all monochromatic lights with wavelengths between 380 nm and 770 nm and is enclosed by the line of purples that includes the saturated non-spectral purple colours. The saturation of the colours in the diagram depends on their location, with colours lying on the spectrum locus being most saturated, and colours lying closer to the white point being less saturated. Within these boundaries, each colour can be described by the combination of its x,y chromaticity coordinates.^{23,113-115} The xyY colour space is based on the x,y chromaticity diagram, and allows to describe colour not only chromatically but also with respect to luminance.^{23,113-115}

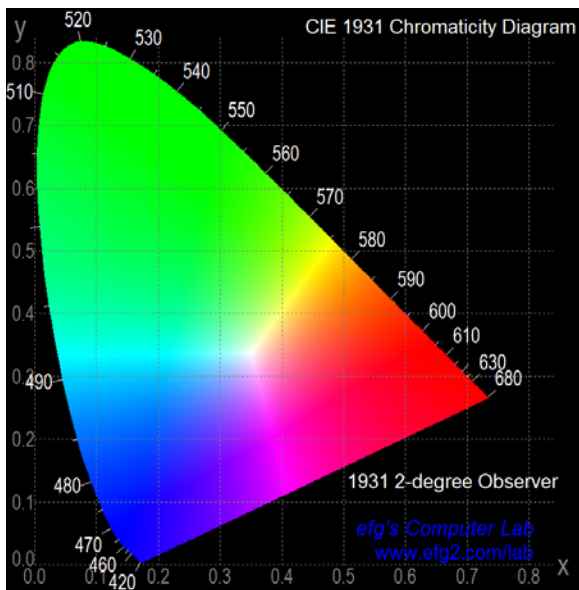


Figure 1-8: The CIE 1931 x,y chromaticity diagram.

Image created with software from <http://www.efg2.com/Lab/Graphics/Colors/Chromaticity.htm>.

A downside to the CIE 1931 x,y chromaticity diagram is that it is not perceptually uniform across the whole diagram.^{22,23,113-115} This means that the sensitivity of the eye to perceived differences in colour varies depending on the location of the colours on the chromaticity diagram. While the green region in the CIE x,y chromaticity diagram is fairly large compared to the red and blue regions, the perceptual differences between colours in the green region are much smaller than for red or blue.²² To overcome this non-uniformity, the CIE introduced the CIE 1976 u',v' chromaticity diagram, in which the physical distance between two colours represents approximately the same colour differences, independently of where two colours in the diagram are compared (Figure 1-9).^{22,23,25} A 3-dimensional expression of colour is the CIE 1976 $L^*u^*v^*$ space which includes lightness (L^*), the perception of luminance.

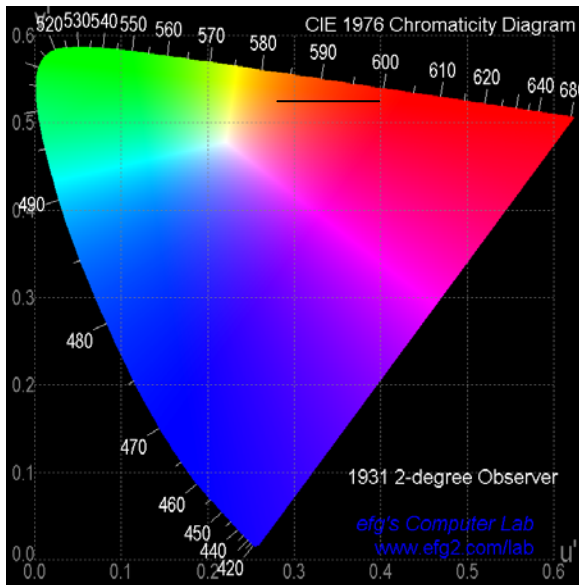


Figure 1-9: The CIE 1976 u',v' chromaticity diagram.

The black line indicates the range of u' values (0.276-0.397) for the images of the VBR 10 scale²⁵; $v'=0.528$ (sd=0.001). Diagram created with software from <http://www.efg2.com/Lab/Graphics/Colors/Chromaticity.htm>.

1.4.1.1 Application in ocular research

A number of studies have quantified bulbar redness by means of colourimetric parameters.^{25,51,70,107,116,117} Simpson et al.¹⁰⁷ used the Minolta 100CS spectrophotometer to investigate if bulbar redness can better be described by first order (chromaticity and luminance) or second order (spatial structure) measurements that were obtained by image processing. Colourimetric (luminance and CIE x , y , z chromaticity) and vessel-based metrics (fractal dimension) were obtained for a set of 32 sample photographs of bulbar redness and subsequently compared to subjective grading estimates for these images. Out of all objective metrics, chromaticity (CIE x) was found to be best correlated to subjective estimates of redness, while the vessel-based characteristics contributed relatively little information.

In his diploma thesis, Schaefer¹¹⁶ compared clinical grading to colourimetric measures obtained with a different spectrophotometer, the SpectraScan PR650, and found a high linear association between subjective grading and CIE chromaticity (CIE u^*). Using the same instrument, Situ et al.¹¹⁷ found good repeatability for measurements of photometric chromaticity (CIE u^*) in patients of a clinical study and found good agreement with clinical grading estimates, concluding that the use of a spectrophotometer in clinical settings was recommendable. In a similar study, Sorbara et al.⁷⁰ compared subjective grading estimates using a modified version of the IER scale to photometric measurements with the SpectraScan PR650 for a group of 24 silicone hydrogel wearers who wore their lenses overnight for a period of 6 months. They were able to demonstrate high levels of repeatability when measuring CIE u' for the reference images of the IER scale, and found moderate agreement to

subjective grading estimates. The moderate level of agreement was ascribed to a fairly low range of bulbar redness in the patients and to somewhat variable subjective estimates. Interestingly, photometric chromaticity (CIE u^*) did not change over the 6 months wearing period, while subjective estimates were significantly higher after the same period compared to baseline.

Duench et al.⁵¹ used the SpectraScan PR650 to demonstrate significant diurnal changes for bulbar redness (as expressed by CIE u') in a group of non-contact lens wearers for multiple time points during a span of 24 hours. Similar diurnal changes were detected for two other objective measures, conjunctival surface temperature and blood flow.

The SpectraScan PR650 was also used for the development of the VBR grading scales.²⁵ After the perceived redness of 25 photographs of various degrees of bulbar redness had been determined by psychophysical scaling, the validity of these subjective redness estimates could be demonstrated by their high linear association to photometric chromaticity (CIE u^*).

1.4.2 Image processing

Digital image processing and image analysis have been used to derive physical attributes that describe bulbar redness chromatically and spatially. The goal of image processing is to transform or enhance an image.¹¹⁸ To allow digital processing, images need to be transformed into a digital form by a process called digitization.¹¹⁹ The digitization of images is based on two separate but conjoint processes, sampling and quantization. By sampling, the original scene is sampled into a rectangular or square array of picture elements (pixels). Each pixel has a

square shape with equal width and height dimensions, and represents a specific location in the sampled image. The spatial resolution of the sampled image depends on the number of pixels and its size. The higher the number of pixels in the image of a certain size (e.g. 8"x10"), the higher its spatial resolution, and the more spatial image detail can be reproduced.¹¹⁹

Quantization is concerned with assigning a specific brightness value to each sampled pixel in the image, so that information about the image content can be obtained.¹¹⁹ An image that only consists of black and white pixels is referred to as 1-bit or binary image, where a grey level of 0 corresponds to a black, and a grey level of 1 corresponds to a white pixel. The number of grey levels contained in an image depends on the number of bits used, and can be calculated by 2^b , where b represents the number of bits.¹¹⁹⁻¹²¹ Thus, a 3-bit image for example consists of eight grey levels (2^3), where black corresponds to 0 and white to 7, while 256 grey levels represent an 8-bit image. A brightness histogram is commonly used for graphical representation of the number of pixels at each grey level, and provides information about the contrast of the image.¹¹⁹ An 8-bit image with a lower dynamic range has grey level values that are clustered in a certain region of the histogram, resulting in low contrast, while an 8-bit image that has grey level intensities that are distributed over the full range of the histogram (high dynamic range) corresponds to an image with well-balanced contrast.¹¹⁹

For some image processing procedures, however, the use of binarized images may be required, for example to derive physical attributes of redness such as the area covered by vessels^{67,97,105} or the number of vessels.⁶⁷ Binarized images can be achieved by thresholding procedures, for which a grey level cut-off point is

either manually or automatically selected. This means that all grey level values below the cut-off point will be assigned to 0 (i.e. black), and all grey levels above the cut-off point to 1 (i.e. white).

A 24-bit colour image is typically based on three 8-bit colour components, red, green, and blue, each of which contains intensities from 0 to 255. If a colour image is split into its three colour components, the information that is contained in each of these channels depends on the colour of the original scene displayed in the image. The red channel of an image displaying conjunctival vasculature for example will contain almost no information since both background and vessels will have similar luminance resulting in very low contrast. The green channel on the other hand provides improved visibility and detail of the blood vessels, since little green light is reflected from the haemoglobin compared to the sclera (background) and this therefore gives the vessels a darker appearance in front of the background.^{97,106} Therefore, the appropriate selection of the channel that contains the highest contrast is required before vascular-based attributes of redness can be derived.^{106,122} The intensities of the three colour components have been used in a variety of attempts to objectively quantify redness, for example to derive physical attributes of redness such as the relative redness of the image.^{67,105,106}

Image size and storage space are important factors to consider in medical digital imaging. Since storage space is usually limited, various studies have tried to determine the smallest spatial resolution of the image and the best image compression level to allow for minimal image size in conjunction with perceptually lossless image compression (i.e. the person looking at the image cannot distinguish between the original and the compressed image).^{123,124} With respect to the number

of bits for medical imaging, the use of 8-bit greyscale and 24-bit colour images was recommended as these have been suggested to allow for smooth transition between grey levels and may provide an accurate representation of brightness differences.^{119,120} The question about whether images can be saved in greyscale rather than colour is important within this context as well, as the required storage space for such images differs significantly. A 24-bit (3 bytes) RGB image with three colour channels (red, green, blue) requires three times as much storage space (e.g. 1280x1024x3 bytes = 3.84 megabytes) as an 8-bit (1 byte) greyscale image with the same spatial resolution (1.28 megabytes). If assessments of ocular conditions are independent of the type of colour or greyscale information, as was shown by Papas⁶⁷ for grading estimates of bulbar redness, storage of these images in greyscale format will significantly reduce the required storage space.

1.4.2.1 Application in ocular research

Image processing and analysis have been used for the objective quantification of redness. In general, these studies analyzed sample images using image processing techniques, and compared the physical attributes to subjective estimates using grading.

The image processing operations that were commonly applied to enhance the photographs of bulbar redness aimed at the removal of noise from the images and at separating the vessels from the background so that spatial attributes of the vessels could be quantified. In general, noise reduction was achieved by masking (filtering) operations that are typically applied locally (i.e. for kernels of pixels in a size of 3x3 or 5x5).^{67,97,104,105,108} The advantage of using localized filtering operations is that small capillaries, which may have very similar grey level intensities as the

scleral background, are not removed from the image.¹⁰⁴ Examples of filtering operations are mean or median filtering, for which the center pixel of a particular kernel of pixels gets assigned to either the mean grey level intensity of all neighbouring pixels, or to the median grey level intensity. Therefore, these filtering operations allow for a smoother transition of grey levels in the background of the images.^{104,105,108}

Following noise reduction, the separation of the conjunctival vasculature was achieved by a number of procedures including contrast enhancement between vessels and background^{104,105} and edge detection algorithms.^{94,108,109} Edge detection algorithms are based on the analysis of the grey level intensities in an image, typically for kernels in a size of 3x3 or 5x5 pixels. By means of edge detection algorithms it is possible to highlight and enhance sharp changes of grey levels that indicate boundaries of objects, for example between the conjunctival vasculature and the scleral background. Edge detection operations that were previously used for this purpose were Canny^{94,109} edge detection or Sobel¹⁰⁹ edge detection.

Chen et al.¹⁰⁴ used a combination of localized filtering and contrast enhancement operations to objectively quantify morphometric variables of the conjunctival vasculature (vessel length and diameter, number of vessel segments, intervascular spacing) for a group of 25 study participants. They were able to demonstrate that their automated image processing technique was able to quantify morphometric variables in a repeatable manner that was superior to manual techniques. Villumsen et al.¹⁰⁸ used an automated image processing technique (smoothing and edge detection) to quantify the number of pixels representing vessels in a set of 12 images showing mild to moderate degrees of bulbar

hyperaemia. Similar to Chen¹⁰⁴, they could demonstrate that their automated results were in good agreement with a manual point counting technique but required significantly less time. Willingham et al.⁹⁷ used the reference images of the MC-D scale to develop an automated technique to measure relative redness and % vessel coverage. Despite demonstrating high levels of linear agreement between scale grades and objective redness characteristics, their approach was criticised because only a single eye was analyzed.⁶⁷ Guillon and Shah¹²⁵ used a line sampling technique to objectively measure characteristics of the conjunctival vasculature (number of vessels, average vessel width, and % vessel coverage). They concluded that their automated technique provided a precise measure of conjunctival redness that was superior to subjective grading estimates obtained with a 0 to 4 scale with 0.5 increments. Based on this finding they recommended adjusting the number of incremental steps for grading scales according to their intended use. Owen et al.¹⁰⁵ used a number of different thresholding procedures to determine the best suitable cut-off point for separation of conjunctival vessels from the background so that the area covered by vessels could best be quantified, and concluded that their new technique was sufficiently sensitive to detect differing degrees of redness in contact lens wearers. Papas⁶⁷ used image processing techniques to derive three morphometric and seven colour-based characteristics of redness in order to identify the parameters that were most closely associated to subjective judgments of bulbar redness. The best correlations were found between subjective estimates (obtained from seven optometrists experienced in the use of the IER scale) and two morphometric parameters, the number of vessels and percent area covered by vessels, while all colour-based parameters showed only poor to moderate linear

associations to redness grading. Since subjective grading estimates in a supplementary experiment were found to be very similar independent of whether colour or greyscale images were assessed, he concluded that redness estimates appeared to be essentially based on vascular information. Fieguth and Simpson⁹⁴ used an internet survey to obtain redness estimates for thirty sample images in order to develop an algorithm that was capable of predicting bulbar redness objectively. In contrast to Papas, they found that both morphometric and colour information played a role in the subjective assessment of redness, and suggested that grading estimates of lower redness levels were best described by vessel parameters, while higher degrees of redness were highly associated to a colourimetric component. Wolffsohn and Purslow¹⁰⁹ used different image processing procedures to quantify redness and concluded that a 3x3 edge detection (Sobel) algorithm was most sensitive to detect changes and most robust to differences in image luminance compared to other image processing procedures evaluated. Recently, Peterson and Wolffsohn⁶⁹ investigated whether subjective grading estimates obtained using the Efron and IER grading scales could be predicted by image analysis techniques (edge detection and relative colour extraction). In agreement with Fieguth and Simpson⁹⁴, they found that a combination of morphometric and colour-based information was best suitable to predict subjective redness estimates.

Two studies have analyzed the IER and Efron grading scales by means of image processing.^{88,96} Perez-Cabre et al.⁹⁶ developed an objective, fully automated technique to derive morphometric characteristics of redness (percent area covered by vessels, number of vessel intersections, and vessel segment length). They

claimed good agreement to the scale grades but they did not elaborate on the detected differences for the physical measures between the scales. Wolffsohn⁸⁸ used previously validated 3x3 edge detection and colour extraction techniques¹⁰⁹ to determine the incremental nature of four bulbar redness grading scales (Annunziato, Vistakon, Efron, and IER). He found that the scale grades were better described by quadratic rather than linear relationships to the objective characteristics, and recommended that the scales were more sensitive at the low end. However, inspection of his graphs suggests that 80% of the ROI analyzed for Efron grade 4 was covered by vessel edges, which indicates that his edge detection technique might overestimate the actual degree of redness.

1.4.3 Fractal Analysis

Fractal analysis is the analysis of shapes and objects which are detailed at all scales.^{6,126-128} To understand the concept of fractal analysis, it is important to first discriminate between Euclidean and non-Euclidean geometry. Euclidean objects are based on the geometric conventions introduced by the Greek mathematician Euclid of Alexandria at about 300_{BC}. Euclid established the basic principles of what is today known as classical geometry, and which is concerned with ideal shapes such as points, lines, circles, squares, cubes, etc. Euclidean objects are defined by having an integer-based dimension, d , such as one-, two-, or three-dimensional.

However, more and more shapes and structures were discovered that could not be explained by Euclidean geometry. To describe these shapes, mathematicians introduced new types of geometry that have been combined to the overall body of non-Euclidean geometry. One of those new types of geometry, fractal geometry, is

concerned with objects that are embedded in shapes of Euclidean dimension, but that cannot be described by an integer dimension. In 1918, Felix Hausdorff derived a novel measure to describe more complicated shapes, the ‘Hausdorff dimension’, which gave these shapes a fractional dimension. It was only in 1975 that the term that is currently used for these objects, fractal, was introduced by Benoit Mandelbrot.¹²⁷ It is based on the Latin word fractus, which can be translated as broken or fragmented.^{6,126} Thus, fractal geometry is concerned with objects or shapes that lie in-between Euclidean dimensions. The complexity of the object of interest, the way the object fills up the space, is quantized by a fractional or fractal dimension which is smaller than the Euclidean dimension it is embedded in.^{6,126}

1.4.3.1 What are fractals?

Fractals are objects that display self-similarity independent of magnification or scale.^{6,126-128} An example of self-similarity is the Cantor Dust, one of the first fractal objects ever described (in 1883), although not termed fractal at the time (Figure 1-10). The Cantor Dust consists of a set of lines which contain two one-third sized copies of themselves for each level of scale except at the highest level.^{6,126} Independent of the magnification used to look at the Cantor Dust (or any fractal object), it will look the same – in other words, it is self-similar.^{6,127-129}

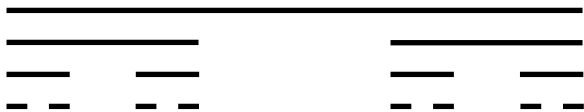


Figure 1-10: The Cantor Dust.

When it comes to the term self-similarity, it is important to discriminate between regular fractals such as the Cantor Dust, and random fractals, the fractals that occur in nature.^{6,127} There are multiple examples of random fractals around us, including trees, fern, leaves, river systems, mountains, clouds, or lightning bolts. However, as opposed to regular or computer generated fractals which are exactly self-similar¹³⁰, they have more irregular shapes that may include rough edges, and only look the same over a finite range of scale. Therefore, fractals that can be found in nature are only statistically self-similar, and are perhaps better described by the term scale-invariant.^{126-128,131}

As opposed to the fractals that occur in nature, regular fractals as the Cantor Dust may be generated on a computer by the process of iteration. The process of iteration can be defined as a feedback process that repeats an n number of times, and which always uses the result of a mathematical operation as the new starting point for the infinite repetition of the same operation.⁶ The process of iteration can be understood when looking at the Koch Snowflake (also referred to as Koch Curve; Figure 1-11). Similar to the Cantor Dust, the single line is scaled down by a factor of three, but instead of removing the middle piece an additional piece is added, making four new equally-sized scaled down copies of the original line. Each of the scaled down pieces is then used as starting point for the next iteration, which are again scaled down by a factor of three to produce four new equally-sized scaled down copies.

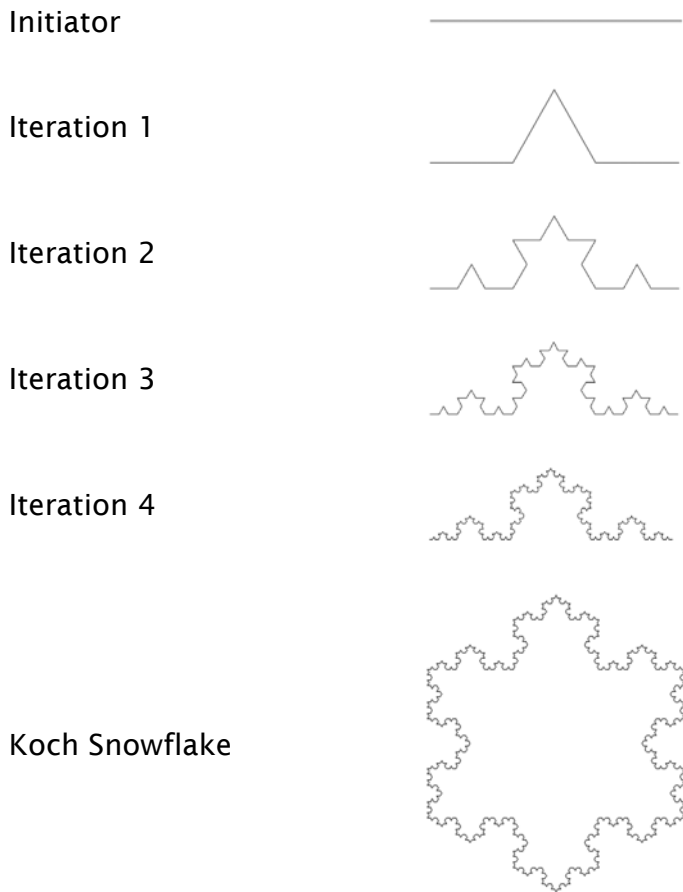


Figure 1-11: The Koch Snowflake.

The single line (initiator) is broken down into four copies of itself each one-third in size of the original line (iteration 1). For each iteration, this process is repeated for each of the four single lines, so that the number of lines increases by a factor of four, while their size goes down by a factor of three. *The Fractals shown were generated with software available at <http://www.efg2.com/Lab>.*

1.4.3.2 How are fractal dimensions calculated?

But how can the dimension of fractals like the Koch Snowflake be described? It is not 1-dimensional as a straight line, but also not 2-dimensional as a square, it is somewhere in-between. In 1961, Lewis Fry Richardson described a scaling relation for the length of geographical borders and noticed that the length of these borders depended on the scale of the measurement.^{128,131} Instead of representing straight lines, some of these borders are so rugged that they rather represent a set of geographical curves.¹³¹ One of the borders he investigated, the coastline of Britain, shows this dependence quite clearly (Figure 1-12). It becomes quite obvious simply by inspection that the overall length of the coastline of Britain increases with decreasing scale of measurement (i.e. measurement stick), and even approaches infinity the closer the scale of measurement comes to zero.¹²⁶



Figure 1-12: The coastline of Britain.

The length of the coastline of Britain increases significantly with decreasing length of the measuring stick. The grey bars correspond to measurement units of 200km, 100km, and 50km, resulting in an overall coastal length of 2400km, 2800km, and 3400km, respectively (left to right). *Figure reproduced from <http://commons.wikimedia.org/wiki/File:Britain-fractal-coastline-combined.jpg>.*

By plotting the logarithm of the measured lengths for the coastlines and borders vs. the logarithm of the scale of measurement, Richardson suggested an entirely empirical equation¹³⁷ which related the measured length $L(G)$ to the scale of measurement G :

$$L(G) = MG^{1-D}$$

where M and D are constants that were not further defined, however.¹³¹

A further discussion of Richardson's work by Benoit Mandelbrot provided this explanation. Mandelbrot related Richardson's graphs of the different individual coastline lengths to the length metric.^{126,131} He showed that D was the slope of the log/log relationship in Richardson's graph, and that it actually was the Hausdorff dimension, previously believed to be a purely technical contrivance and not a concrete notion.¹²⁶ Based on this finding, Mandelbrot suggested that a better way to describe geographical curves would be by means of this 'fractional' dimension which should be regarded as an indicator of the 'wrinkliness' of the respective structure.¹³¹ As a side note, the fractal dimension of the coastline of Britain corresponds to $D=1.25$, about the same as the Koch Snowflake (Figure 1-11).

The calculation of the fractal dimension of shapes and objects is performed by taking advantage of the knowledge of this scaling relationship. If the (embedding) dimension of an object is known, exponents can be used to calculate the number of new pieces we get when we reduce the size of the object by a certain factor (m). For an n -dimensional object, the number of new pieces corresponds to m^n $1/m$ -sized copies of itself. The Sierpinski Carpet will be used to further describe this relationship (Figure 1-13).



Figure 1-13: The Sierpinski Carpet.

The Sierpinski Carpet is a filled square that is scaled down by a factor of three, so that $3^2 \cdot \frac{1}{3}$ -sized copies of itself are created of which the one at the centre is removed. The same scaling procedure is infinitely repeated for each of the remaining eight copies, so that the carpet gets more detailed with every repetition. The fractal dimension of the Sierpinski Carpet can then be determined by using a re-arranged form of Richardson's equation based on the log-log relation of count (8) and scale (3).¹²⁶

$$D = \frac{\log(\# \text{ of new pieces})}{\log(\text{scaling factor})} = \frac{\log 8}{\log 3} = 1.893$$

Using this general rule, the fractal dimension of any fractal object for which the embedding dimension and the scaling factor are known can be calculated. Thus, the fractal dimensions of the Cantor Dust (Figure 1-10) and the Koch Snowflake (Figure 1-11) correspond to $D=0.631$ ($\log 2/\log 3$) and $D=1.262$ ($\log 4/\log 3$), respectively.

The determination of fractal dimensions becomes more involved when information about the fractal to be analyzed is limited. There are a number of techniques available today which facilitate this task, such as the box-counting, the

mass-radius relation, and the pair correlation function method.¹²⁸ The public domain Java image-processing program ImageJ 1.38x¹³² was used in combination with the ImageJ plug-in FraCLac (ver. 2.5 Rel.1b5i)¹³³ to calculate fractal dimensions for this work. Since FraCLac uses box-counting algorithms to derive fractal dimensions, this technique will be described in more detail in the following section.

Box-counting algorithms are closely related to Richardson's procedure of counting the number of new pieces and relating it to the change of scale. The box-counting method is based on placing a series of grids of boxes of decreasing size over the structure of interest, for which the number of boxes containing detail is counted. Figure 1-14 shows examples for different box sizes within a single scan of the same image. It is obvious that as the boxes become smaller, the number of boxes containing detail increases.

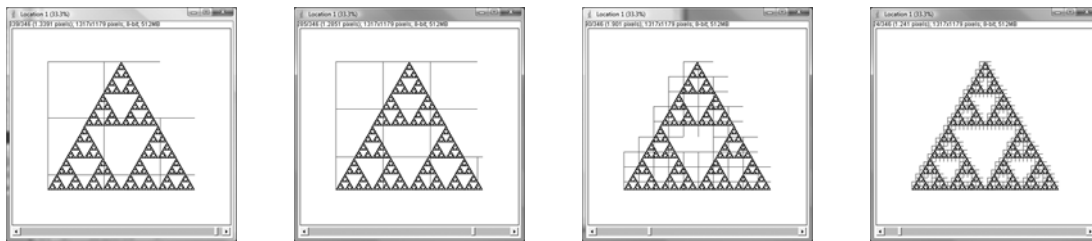


Figure 1-14: The box-counting method.

By changing the size of the boxes, the scaling relation between box count and scale can be approximated.¹³³ To derive the fractal dimension of the fractal, FraCLac calculates the slope of the regression line for the log-log plot of count and scale (ϵ), which is equivalent to the fractal dimension, D , of the fractal. Figure 1-15 shows this relationship for the regression line of a regular fractal, the Sierpinski

Triangle, which contains three $1/2$ -sized copies of itself ($\log 3 / \log 2$; $D=1.585$). The natural logarithm of scale (ϵ ; abscissa) is plotted vs. the natural logarithm of the count of boxes (ordinate) that contained detail. It can be seen that with almost every change of box size there was also a change in count. The slope of the corresponding regression line is given at the top of the plot, and was determined to be $D=1.5877$, which is very close to the expected fractal dimension for the Sierpinski triangle.

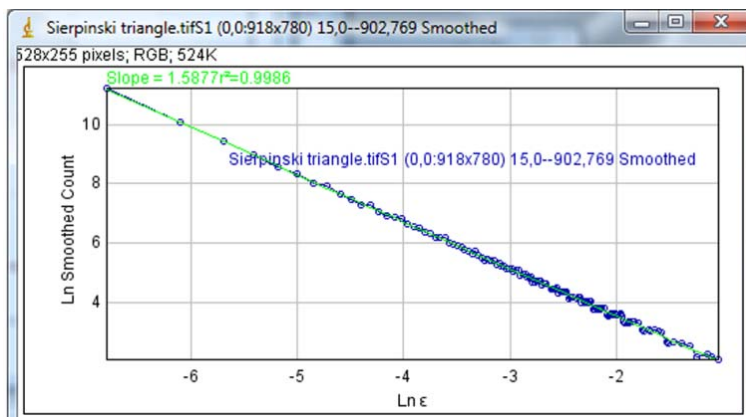


Figure 1-15: Log-log plot of count and scale (ϵ).

1.4.3.3 Application in ocular research

There are numerous shapes surrounding us in nature that have the attribute of being scale-invariant over a finite range of scale; that is, they can be modeled as a fractal, and their fractal dimension can be determined.¹²⁶⁻¹²⁸ The same can be said for the human body, where tree-like branching structures can be found in the bronchial tree or as cardiac muscle bundles, but particularly in a number of vascular branching systems such as the heart, the lungs or the kidney. Therefore, research in the medical and biological field has increasingly applied fractal analysis

to evaluate shapes and patterns of branching structures to derive their corresponding fractal dimension.^{128,129,134-139} The fractal dimensions were then used to investigate if changes to these structures could be detected, and to see if it could be applied for the discrimination of normal from pathological structures.^{128,137}

Fractal analysis has also been used to estimate the complexity of the retinal vasculature.^{128,129,134,137,139} After bifurcation at the optic disc, the arteries and veins of the inner layer of the retina show extended branching patterns, but normally arteries and veins do not cross themselves. The arteries and veins extend further into smaller branches (arterioles, venules, and capillaries) which form a vast vascular network throughout the retina.¹²⁸ This retinal vascular network in normal human beings was found to have a statistically self-similar structure corresponding to a fractal dimension of approximately $D=1.70$.^{128,129,137,139} Since this is almost identical to the fractal dimension of a computer simulated diffusion limited growth process, it has been suggested that the development of the human retinal vasculature may involve a diffusion process.^{137,139} However, the use of fractal analysis for the detection of pathological changes to the retinal vasculature did not provide sufficient sensitivity and specificity to being used as a diagnostic tool.^{128,129,137,139}

Branching structures can also be found in the anterior segment of the eye, for example corneal neovascularization¹³⁰ or in the rich vascular network of the conjunctiva.³³ There are numerous factors that can cause an irritation of the bulbar conjunctiva, including contact lens wear^{49,50,52,140,141}, hypoxia⁴⁷⁻⁵⁰, or foreign bodies.^{31,45} The ocular irritation is accompanied with an increased dilation of the conjunctival blood vessels that gives the eye a red appearance.^{43,51,62} Since a dilation of blood

vessels also goes along with changes to the pattern of the conjunctival vasculature, a part of this research investigated if fractal analysis was capable of quantifying redness in the conjunctival vasculature.

2 Rationale

The monitoring and management of changes to ocular tissues and structures is a basic requirement in clinical practice and research settings. One of the complications routinely assessed in contact lens wearers is bulbar hyperaemia, a dilation of the conjunctival blood vessels that gives the eye its red appearance.^{1,2} To estimate the severity of bulbar redness (and some other conditions), practitioners most commonly resort to the use of clinical grading scales. Typically, grading scales employ five reference levels³⁻⁵ that define increasing levels of severity by means of descriptions and/or illustrations.⁶⁻⁸

A problem commonly identified with the use of grading scales is the variability inherent to the subjective assessments when the scales are used by different observers or by the same observer over time.^{7,8} Aside from observer-induced variability, other criticism can be of the grading scales themselves, for example because of unequally spaced scale steps along the scale range. A number of bulbar redness grading scales currently exist and there have been reservations expressed about their interchangeable use because of different designs⁹, non-aligned scale steps¹⁰ or varying scale ranges.^{10,11} Despite being frequently used in clinical practice today, grading scales are poorly understood and have not been thoroughly tested⁷, with the consequence that only very little information is available about the grading scales themselves.^{9,12,13}

The use of automated objective techniques to quantify redness has been recommended to provide an alternative to grading scales^{7,14,15} and to determine the criteria that may be applied when the severity of a condition is assessed.^{7,8,13}

However, objective techniques have remained generally unused except in research settings. Because of their convenience and availability^{8,16}, it appears likely that grading scales will remain the preferred clinical tool for the assessment and management of patients.

Therefore, the global aim of this thesis was to use objective and subjective approaches to analyze bulbar redness grading scales in order to get a better understanding of the scales themselves and of the processes that are involved in clinical grading. A specific purpose of this research was to develop a technique that would allow for a cross-calibration of the grading scales so that grades obtained with different scales may be compared. Four bulbar redness grading scales were focus of this research:

- The McMonnies/Chapman-Davies (MC-D) bulbar redness grading scale, the first photographic scale that was developed for the assessment of bulbar redness, consisting of six reference levels ranging from 0 to 5.¹⁷
- The Institute for Eye Research (IER) scale for bulbar redness that was developed at the Cornea and Contact Lens Research Unit in Sidney, Australia, consisting of four photographic reference levels ranging from 1 to 4.^{18,19}
- The Efron scale for bulbar redness that differs from the other scales inasmuch as artist-rendered drawings are used (instead of photographs) to illustrate its five reference levels ranging from 0 to 4.^{6,20,21}
- The Validated Bulbar Redness (VBR) scale, a 100-point photographic scale that was developed at the School of Optometry in Waterloo, Canada.²² It is

the only scale that employs reference levels (10, 30, 50, 70, and 90) that have been objectively validated.

A number of objective and subjective techniques were used in the course of this PhD research to analyze the bulbar redness grading scales. Before the findings of the individual experiments are presented, chapter 3 provides an introduction on the procedures and their corresponding accuracy and repeatability.

In the first study (Chapter 4), fractal dimension (D) was introduced as new objective metric to quantify redness in the bulbar vasculature, and compared to two other physical redness attributes, % pixel coverage (% PC) and photometric chromaticity, u' . The resulting quantitative measures were then correlated to the nominal scale grades to determine the 'accuracy' of the scales and to investigate if a cross-calibration of the grading scales was possible.

In chapter 5, a psychophysical scaling method was used to estimate the perceived redness of the reference images of the grading scales. The images were to be scaled for a given 0 to 100 redness range, and their relative position was taken as their perceived redness. The perceived redness of the images was compared to the physical redness (chapter 4) in order to identify the criteria that may be used when subjectively scaling redness.

Chapter 6 represents a logical extension of the scaling experiment discussed in chapter 5. The experimental setup was slightly modified by providing the VBR scale reference images as additional anchors for redness scaling, and comparisons between non-anchored (chapter 5) and anchored scaling were made.

The perceived redness from anchored scaling was used for a cross-calibration of grades between scales.

The newly calibrated grading scales were used in chapter 7 to investigate the agreement of grading estimates across scales. The reference images for each scale were placed at the positions corresponding to their perceived redness as determined in chapter 6, one scale at a time. Sample images were to be placed relative to the unlabelled reference images in order to identify their redness for a 0 to 100 range. Physical redness attributes (D , % PC , and u') of the sample images were determined using image processing and a spectrophotometer, and compared to the perceived redness of the sample images.

Chapter 8 summarizes the results that were presented, and provides an outlook to potential future studies that may arise from the work presented in this thesis.

In the following chapter, the image processing, photometric, and psychophysical scaling procedures that were used in this research are introduced, and their 'accuracy' and 'precision' is examined. A number of statistical tests and terminology are discussed.

3 The Accuracy and Repeatability of Objective and Subjective Techniques to Estimate Bulbar Redness

3.1 Introduction

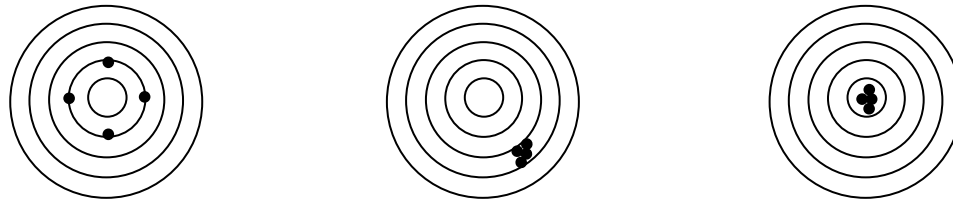
The performance of measurements obtained through new instruments or novel techniques is judged based upon both their reliability and validity, and is more specifically described by the measurements' accuracy, precision, repeatability, or reproducibility.^{1,2} Despite the importance of evaluating a measurement's performance, the interchangeable use of above terms creates some level of confusion in the scientific world.^{3,4} This is in part due to different definitions for the respective terms, as these may differ depending on the scientific discipline, or simply because the sometimes subtle differences between terms are not clear (e.g. repeatability and reproducibility).^{1,3} In addition, mismatching definitions between standardized vocabularies⁵⁻⁸, for example between the definitions for accuracy by the US Food and Drug Administration (FDA)⁷ and by the International Organization for Standardization (ISO)⁵, or the synonymous use of precision and accuracy in Merriam Webster's dictionary⁹, add to the confusion.¹ The terminology used in this section will closely follow the most recent definitions in the 'International Vocabulary of Metrology' (VIM)¹⁰, and is specified below.

The *accuracy* of a measurement relates to the "*closeness of agreement between the measured quantity value and a true quantity value of a measurand*".¹⁰ In other words, a measured value is compared to the accepted value of a reference that may be an established instrument with proven accuracy (such as a calibrated weighting scale) or a measurand with standard dimensions (e.g. the international

prototype kilogram). In the context of this thesis, the accuracy of the image processing software ImageJ (ver. 1.38x) and its plug-in FraCLac (ver. 2.5 Rel. 1b5i) that were used to derive the physical metrics was determined against references with known dimensions.

The precision of a measurement is defined as the “*closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions*”.¹⁰ In other words, the agreement of repeated measurements is assessed by how variable the data are, and is quantified by measures of imprecision such as standard deviation or variance.^{6,10}

The difference between accuracy and precision is frequently described with a target analogy⁴, where the bull’s eye represents the accepted reference value (or true value) that’s expected to be measured (Figure 3-1). Bearing the above definitions in mind, this means that shots (i.e. measurements) that are fairly evenly distributed and close to the bull’s eye are accurate, with shots closer to the bull’s eye being more accurate. Figure 3-1A shows this situation, however, as the shots are scattered on the target they are not very precise. Measurements are precise if their values are all the same, or, for the target analogy, the shots all hit the same spot, even if the spot is off the centre and thus the shots are inaccurate (Figure 3-1B). Because of the errors inherent in measurements, the implementation of new measurement techniques or instruments requires that these procedures are accurate and precise.¹ If repeated measurements are both accurate and precise, they consistently produce the same value that also matches the accepted reference value: We consistently hit the bull’s eye on the target (Figure 3-1C).



A) Imprecise but accurate

B) Inaccurate but precise

C) Accurate & precise

Figure 3-1: The target analogy for accuracy and precision.

The repeatability and reproducibility of measurements are specific kinds of precision, which differ with regard to their specified measurement conditions. The VIM provides definitions for repeatability and reproducibility, however, it should be mentioned that VIM actually uses the terms that are being defined within their definitions. Thus, the *repeatability* of measurements was defined as “*measurement precision under a set of repeatability conditions of measurement*”. These repeatability conditions require that the same operators perform the same measurement procedure with the same measuring system, under the same operating conditions and in the same location, on the same objects over a short period of time.^{3,10}

The *reproducibility* of measurements on the other hand is defined as “*measurement precision under reproducibility conditions of measurement*”.¹⁰ These reproducibility conditions require replicate measurements on the same measurand or object, however, other measurement conditions may be changed, for example by using different instruments or if different observers are involved.^{1,3,10}

The repeatability of the methods used in the individual experiments of this thesis was assessed by comparing the results of the test and the retest session (test-retest repeatability) using a number of statistical tests:

- *Coefficient of repeatability (COR)*^{11,12}

The COR describes the degree of scatter for repeated measurements on the same objects. The COR is the standard deviation of the differences (s_d) between test and retest session for all measurands multiplied by 1.96 (i.e. $COR=1.96*s_d$). It may be interpreted as the range of differences that 95% of pairs can be expected to fall within, with differences that are larger than the 95% confidence limits being statistically significant.² The smaller the COR, the better the repeatability of the measurements.

- *Limits of agreement (LOA)*¹³

The LOAs are a means to graphically display the differences between test and retest measurements that are quantified as the COR. The LOAs are the limits of the 95% confidence interval (i.e. COR) that are plotted with respect to the mean of the differences between test and retest for all measurands (\bar{d}); the upper and lower LOA are calculated by $\bar{d}\pm 1.96*s_d$, respectively. They may be interpreted as the range between which a repeated measurement can be expected to lie without representing a statistically significant change.²

- *Intraclass correlation coefficient (ICC)*¹⁴⁻¹⁶

The ICC estimates the variability of measurements between sessions to the overall variability between measurands, and is an indicator of the reliability of the data.³ It is a correlation coefficient that indicates the amount of variance that

can be attributed to differences between measurands. The ICC will be high (i.e. approach 1.00) if most of the variance is between measurands and will approach zero if most of the variance is between sessions.

- *Correlation coefficient of concordance (CCC)*¹⁷

CCC is a specific type of ICC³⁴ that describes the degree of deviation from perfect concordance. It is a correlation coefficient that defines the degree of concordance between sessions by calculating the perpendicular variation (i.e. the variation in both horizontal and vertical direction) of each pair of measurements from the 45°-line corresponding to perfect agreement between the two measurements. A CCC of 1.00 corresponds to perfect concordance, while a CCC approaching zero corresponds to poor concordance. To show the deviation, an orthogonal regression line is fitted to the data and plotted in comparison to the 45°-line of perfect concordance between session 1 and session 2.

The curve fitting for this thesis was carried out using Sigmaplot v10 (Systat Software Inc., San Jose, CA, USA). Except for the concordance plots for which orthogonal regression lines were fitted, an ordinary least squares regression was used to determine the best fit line for the data being compared. In least squares regression, only the variation in the vertical direction is considered when the curve is fitted to the data.³²

The purpose of this chapter is to describe the accuracy and precision of the instrumentation and the novel experimental procedures that were used to derive the physical and perceptual measures to estimate redness. Therefore, these

procedures will be briefly summarized at the beginning of each subsection, with experiment-specific descriptions of the respective methods for image processing, fractal analysis and photometry to follow in Chapter 4, and for psychophysical scaling in Chapters 5 and 6.

3.2 Methods

3.2.1 Image processing

The public domain Java image processing software ImageJ (v. 1.38x)¹⁸ was used to pre-process the reference images of the four grading scales. Pre-processing of the reference images was required to create binarized versions of the reference images. The original colour images were modified in a way so that pixels corresponding to vessels were displayed in black, and pixels corresponding to the background (i.e. the sclera beneath the transparent conjunctiva) were white. The binarized versions of the images were then analyzed with the ImageJ plug-in FraLac (ver. 2.5 Rel. 1b5i)¹⁹ to determine the area covered by the vessels (that we termed % pixel coverage, or % *PC*; Chapter 4) and the complexity of the vessels that was quantified by fractal dimension, *D*. To ensure that each reference image was treated identically during pre-processing, a macro was developed that systematically applied the same processing steps in identical order to derive a binarized version of each reference image.

Repeatability of image processing

The repeatability of the image processing macro was evaluated by pre-processing each reference image on two separate occasions about one week apart.

After both binarized versions of each reference image were obtained, each pair of images was opened in ImageJ, and the image calculator function was used to subtract image 2 from image 1. If the macro was consistent in binarization of the reference images, the subtraction operation was expected to result in a plain white image with all vessels (i.e. black pixels) removed. The histogram function in ImageJ was used to verify the pixel count after image subtraction.

Accuracy of image processing

To evaluate the accuracy of the image processing software, a base image with dimensions of 100 by 100 pixels (total 10,000 pixels) consisting of white pixels only was generated. Figure 3-2 shows modifications of this base image where black pixels were added to represent:

- a single vertical line (i.e. 100 black pixels);
- a single horizontal line (i.e. 100 black pixels);
- a cross (i.e. 199 black pixels);
- 5 vertical lines (i.e. $5 \times 100 = 500$ black pixels);
- $\frac{1}{2}$ of the image (i.e. $50 \times 100 = 5000$ pixels).

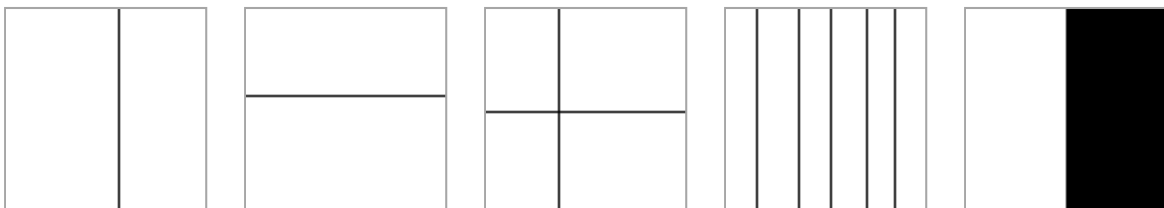


Figure 3-2: Generated line art images with known pixel count to evaluate the accuracy of ImageJ.

The accuracy of ImageJ was determined by comparison of the count of white and black pixels in each of these images (as derived by the histogram function in ImageJ) to the true number of white and black pixels.

3.2.2 Fractal analysis

The ImageJ plug-in FraCLac (ver. 2.5 Rel. 1b5i)^{18,19} was used to determine the fractal dimension, D , of the specified regions of interest (ROI) in the binarized grading scale reference images. FraCLac uses a box-counting algorithm that places a series of grids of boxes of decreasing size over the ROI, and the number of boxes including detail (i.e. having black pixels) is counted.¹⁹ As the size (or calibre) of the boxes decreases, the number of boxes containing detail increases. The ratio at which detail changes with changing scale is a measure of the complexity of the structure of interest.¹⁹ The slope of the regression line for the log-log relation between box count and scale is used to calculate the box counting fractal dimension, D_B , which quantifies the ratio at which detail and scale change.^{19,20} *Note: since box counting was the only method to calculate fractal dimensions in this work, it will be referred to as D only, neglecting the subscript.*

Box-counting algorithms also depend on the starting position of the grid, as the number of boxes containing detail for the same grid calibre can be quite different depending on the orientation (i.e. starting position) of the grid. Therefore, FraCLac allows the use of multiple orientations for the same set of grid calibres and delivers four different fractal dimensions that are calculated using different methods of data transformation.¹⁹ The four fractal dimensions provided by FraCLac are:

- *Averaged fractal dimension (D)*

The fractal dimension is derived by averaging over the number of global scan positions selected.

- *Slope-corrected fractal dimension (D_{sc})*

Identical to \bar{D} , but corrected for periods of no change for the log-log plot of box size and count. Since FraLac is set up to change from minimum to maximum box size in a linear fashion, there may be occasional plateaus where the number of boxes does not change. As these plateaus do not necessarily represent features of the structure of interest but affect the final slope of the regression line and thus the fractal dimension, they are removed.

- *Most-efficient covering fractal dimension (D_e)*

Identical to \bar{D} if only a single scan is used. Takes advantage of the possibility of using multiple scans at different starting grid positions, as only the box-count that required the lowest number of boxes at each grid size is used to calculate the fractal dimension.

- *Slope-corrected most-efficient covering D (D_{sce})*

Fractal dimension derived based on a combination of all of the above algorithms. If only a single scan position is used, it is equivalent to the slope-corrected fractal dimension.

Accuracy of the box-counting algorithm

Accuracy is defined as the closeness of agreement for a measured value to the true value of a reference. Therefore, references with known fractal dimension

were used to determine which of the four fractal dimensions provided by FraLac resulted in the closest agreement with the expected (true) fractal dimensions. These references included a single vertical line (Figure 3-2A; $D=1.00$), a single horizontal line (Figure 3-2B; $D=1.00$), the 'Koch Snowflake' (Figure 1-11; $D=1.26$), and the 'Sierpinski Triangle' (Figure 3-3; $D=1.58$). The latter two were generated for this purpose by software from www.efg2.com/Lab.

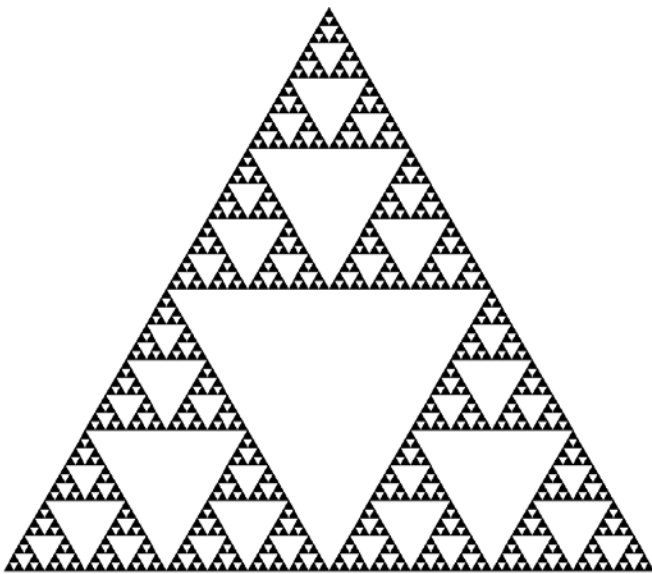


Figure 3-3: The Sierpinski Triangle.

Image created by software from www.efg2.com/Lab.

Standardized settings for FraLac were developed based on the recommendations of the FraLac user manual.¹⁹ The size of the series of grids was set to linearly decrease from a maximum box size of 45% of the image size to a minimum size of 1 pixel, and images were scanned either in a single global scan (with a fixed starting position) or with ten global scans (with ten randomly chosen starting grid positions). Based on these settings, the four fractal dimensions

provided by Fraclac were calculated and compared to the true fractal dimensions of the reference fractals. In addition, the generated line art images (Figure 3-2) were analyzed to estimate the accuracy of Fraclac's pixel count function.

3.2.3 Photometric chromaticity

The calibrated spectrophotometer SpectraScan PR650 (Photo Research Inc., Chatsworth, CA, USA) was used to measure the photometric chromaticity, u' , of the 20 reference images. The accuracy of the SpectraScan PR650 has been reported to be ± 0.0015 for CIE 1931 x , ± 0.001 for CIE 1931 y , and ± 0.006 for CIE 1931 xy for cathode ray tube (CRT) monitors.²⁴

The photometer was mounted on a tripod that was placed 30cm away from a flat screen liquid crystal display (LCD) monitor (LG Flatron L1511S; LG; Seoul, Korea), and was kept stationary throughout the experiment (Figure 4-4). Standardized experimental settings were photometer position, room illumination, monitor settings, and monitor running time (i.e. 3 hours before the measurement started; Chapter 4).

For each image within a scale, equally-sized ROIs were specified to cover the largest conjunctival area possible (Chapter 4; ROIs were identical to the ROIs as specified for scale version 1). Photometric chromaticity, u' , was measured for each of the ROIs and then averaged across ROIs to represent a global estimate of redness, taking the whole conjunctiva into account. Figure 3-4 shows a photograph taken through the eyepiece of the photometer. The yellow rectangle represents a sample ROI as displayed on the screen, and the black circular area corresponds to the measurement spot (approximately 19.63 mm²) of the photometer.²¹

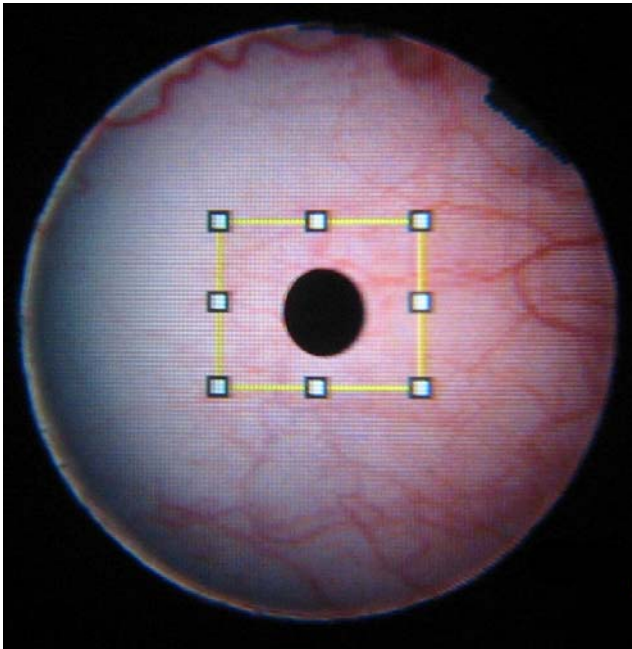


Figure 3-4: View through the eyepiece of the photometer.

The repeatability of the photometric measurements was evaluated by measuring u' of each reference image twice, separated by 14 days. The order of the images in the two sessions was exactly reversed, so that the image that was measured first in the test session was measured last in the retest session, and vice versa.

3.2.4 Psychophysical scaling

Psychophysical scaling was used to estimate the perceived redness of the reference images on a table top over a distance of 1.5m, for which only the start and end point were labelled by 0 and 100 to represent minimum and maximum redness, respectively. Scaling was done using three image sets that differed in the type of colour information they displayed (i.e. the vessels were shown in colour, greyscale, or binarized). In a subsequent experiment, scaling of the colour image

set was done for the same overall redness range, but using the VBR reference images as additional anchors to estimate redness (Chapters 5 & 6). Scaling was repeated for each of the image sets and for anchored scaling about four to six weeks after the first session had concluded.

The repeatability of psychophysical scaling was estimated based on the averaged perceived redness (across observers) for each reference image. For each image set, CORs, LOAs, ICCs, and CCCs were derived. To investigate if the variability between observers was different depending on severity, the standard deviation of redness estimates (sd) for each reference image was plotted vs. its averaged perceived redness.

Data analysis

Statistical analysis was performed using STATISTICA version 8 (StatSoft. Inc., Tulsa, OK, USA) and an alpha level of ≤ 0.05 was considered statistically significant. CCCs were calculated at <http://www.niwa.co.nz/our-services/online-services/statistical-calculators/lins-concordance>. Sigmaplot v10 (Systat Software Inc., San Jose, CA, USA) was used for plotting and curve fitting.

3.3 Results

3.3.1 Image processing

Repeatability

The macro developed for pre-processing of the reference images proved to be perfectly repeatable. For each pair of binarized images that were created from the same source reference image, the histogram function showed only white pixels

but no black pixels after subtraction of the images. Figure 3-5 shows this effect for the binarized versions of the VBR 50 reference image (Scale Version 2).

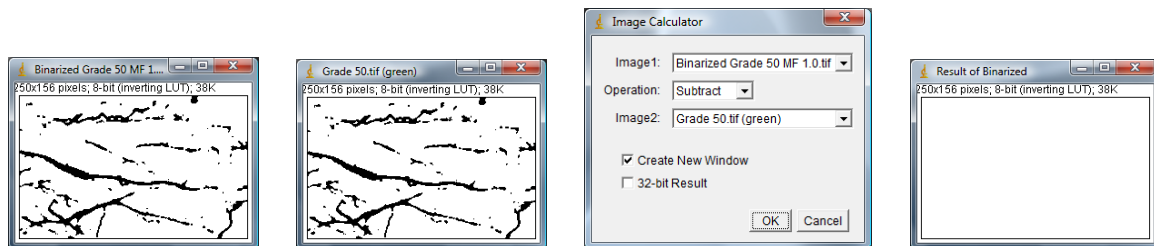


Figure 3-5: Result of image subtraction.

Accuracy

ImageJ determined the same number of pixels as expected from the (true) pixel count of black and white pixels for each of the line art images (Figure 3-2).

3.3.2 Fractal analysis





Accuracy of the box counting algorithm

Table 3-1 shows the true fractal dimensions for each of the references, the four fractal dimensions that were calculated by Fraclac, and the standard error (SE) for the regression line of the log-log plot of box count and scale. Fractal dimensions were derived for both a single scan only (i.e. a fixed starting grid position) and for ten global scans (i.e. ten random starting grid positions).

The pixel count for the five line art images (Figure 3-2) as calculated in Fraclac was in perfect agreement with ImageJ's histogram based count and with the true pixel count of the images.

Table 3-1 Accuracy of fractal analysis.

Shown are the fractal dimensions for the references based on their accepted (true) and the four calculated fractal dimensions (\bar{D} , D_{sc} , D_e , D_{sce}), for both a single and ten global scans. Fractal dimensions in bold indicate the most accurate representation of the true fractal dimension. SE represents the standard error of the regression line that was used to calculate the respective fractal dimension.

Reference		\bar{D}	SE	D_{sc}	SE	D_e	SE	D_{sce}	SE
 True: $D=1.00$	1 scan	1.00	0	1.00	0	1.00	0	1.00	0
	10 scans	1.00	0	1.00	0	1.00	0	1.00	0
 True: $D=1.00$	1 scan	1.00	0	1.00	0	1.00	0	1.00	0
	10 scans	1.00	0	1.00	0	1.00	0	1.00	0
 True: $D=1.26$	1 scan	1.29	0.11	1.24	0.11	1.29	0.11	1.24	0.11
	10 scans	1.22	0.05	1.22	0.06	1.29	0.09	1.25	0.08
 True: $D=1.58$	1 scan	1.56	0.08	1.59	0.06	1.56	0.08	1.59	0.06
	10 scans	1.52	0.06	1.53	0.05	1.57	0.05	1.58	0.04

3.3.3 Photometric chromaticity

Table 3-2 shows the CCCs, CORs, and the mean of the differences between test and retest (\bar{d}) for each of the scales. There was almost no variability between the photometric data obtained in test and retest session.

Table 3-2: Repeatability coefficients for photometric chromaticity, u' .

	CCC	COR	\bar{d}
MC-D	0.984	0.002	+0.0021
IER	0.996	0.001	+0.0018
Efron	0.990	0.001	+0.0024
VBR	0.996	0.001	+0.0026

Figure 3-6 shows the LOAs for the MC-D, IER, Efron, and VBR scale plotted vs. the mean of u' as determined in the test/retest session. The thin solid line indicates the mean of test/retest differences (\bar{d}) for u' . The upper and lower LOAs ($\bar{d} \pm 1.96 * s_d$) are shown as thick solid lines.

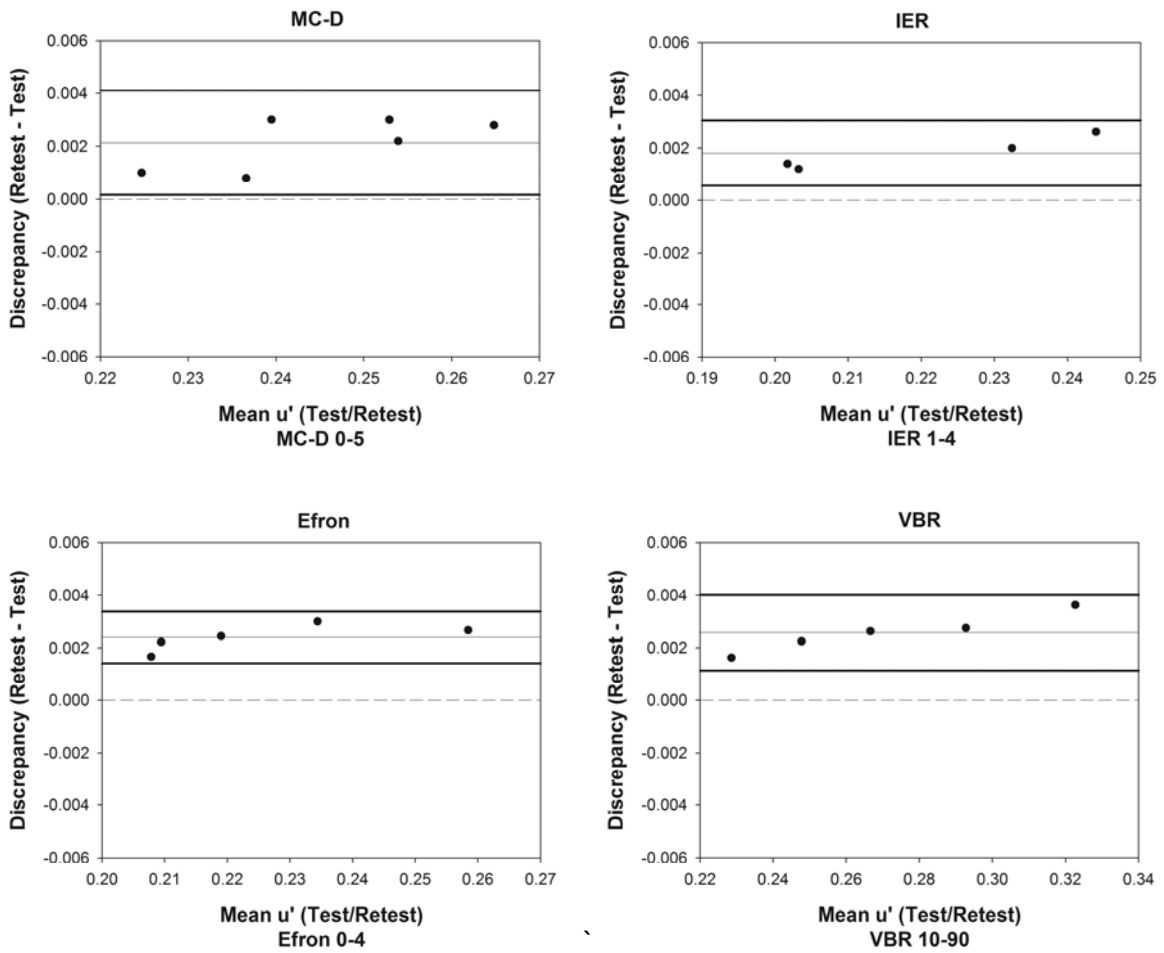


Figure 3-6: Limits of agreement for photometric chromaticity, u' .

The LOAs (ordinate) are plotted vs. the mean u' of test and retest (abscissa) for the MC-D, IER, VBR, and Efron scale (clockwise). The thin solid line indicates the mean of test/retest differences (\bar{d}) for u' . The upper and lower LOAs ($\bar{d} \pm 1.96 \cdot s_d$) are shown as thick solid lines.

3.3.4 Psychophysical scaling

Averaged psychophysical scaling data were highly repeatable for each of the image sets and for non-anchored and anchored scaling (Table 3-3), with almost perfect concordance ($CCC \geq 0.984$) for averaged perceived redness across observers. A number of intraclass correlation coefficients are possible, each appropriate for specific situations that depend on the experimental design of the study. In compliance with the classification by Shrout and Fleiss¹⁴, the participants that took part in this study were regarded as random sample from a large population, where each participant judged each target (in this case, each reference image): This corresponds to Shrout's and Fleiss' case 2. As the perceived redness in this study was determined to allow a comparison between reference images, the results based on the mean of k raters and not the results of a single rater were of interest. Based on this specification, the ICC data as presented in Table 3-3 are based on Shrout's and Fleiss classification ICC 2,k, where k corresponds to 10.¹⁴

Table 3-3: Repeatability coefficients for psychophysical scaling.

	COR	ICC 2,10	CCC
Colour Non-anchored	8.7	0.994	0.988
Greyscale	9.7	0.992	0.984
Binarized	4.4	0.999	0.997
Colour Anchored	6.0	0.995	0.990

Figure 3-7 shows the averaged perceived redness for the test session plotted vs. the retest session.

The test and retest differences for averaged perceived redness (ordinate) are plotted in Figure 3-8 vs. the test/retest mean perceived redness (abscissa). The thin solid line indicates the mean of test/retest differences in perceived redness. The upper and lower LOAs are shown as thick solid lines.

Figure 3-9 shows the variability between observers (expressed by sd) relative to the averaged perceived redness for each reference image. For each of the image sets, variability between observers was largest for images that were perceived approximately at the middle of the redness range. *Note: data are shown for the test session only; the variability in the retest session was very similar.*

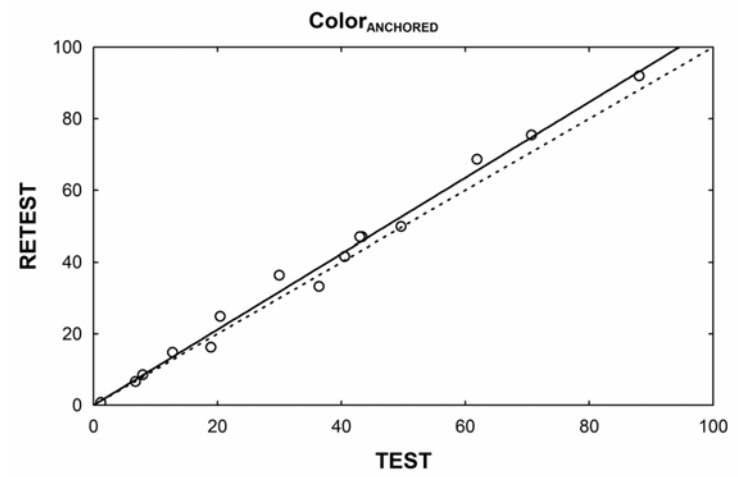
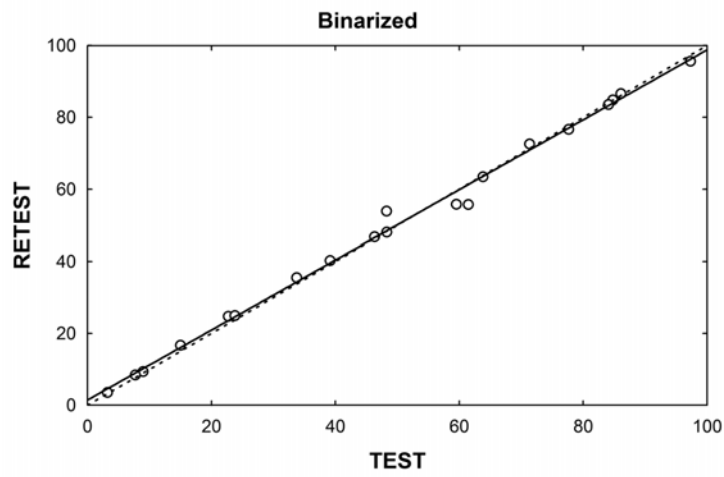
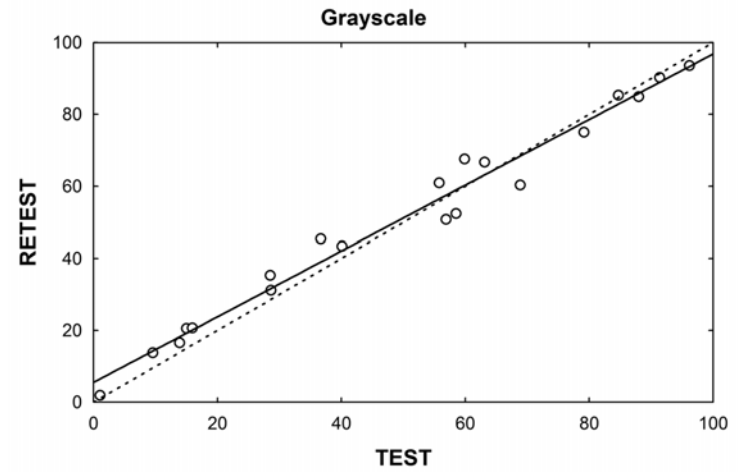
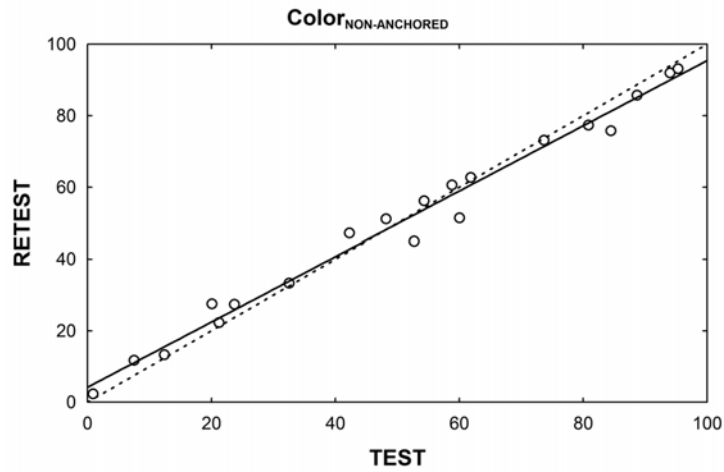


Figure 3-7: Averaged perceived redness plots for test vs. retest.

The solid line indicates the linear relationship between test and retest data, and the dashed line represents the 45°-line of equality (i.e. perfect concordance).

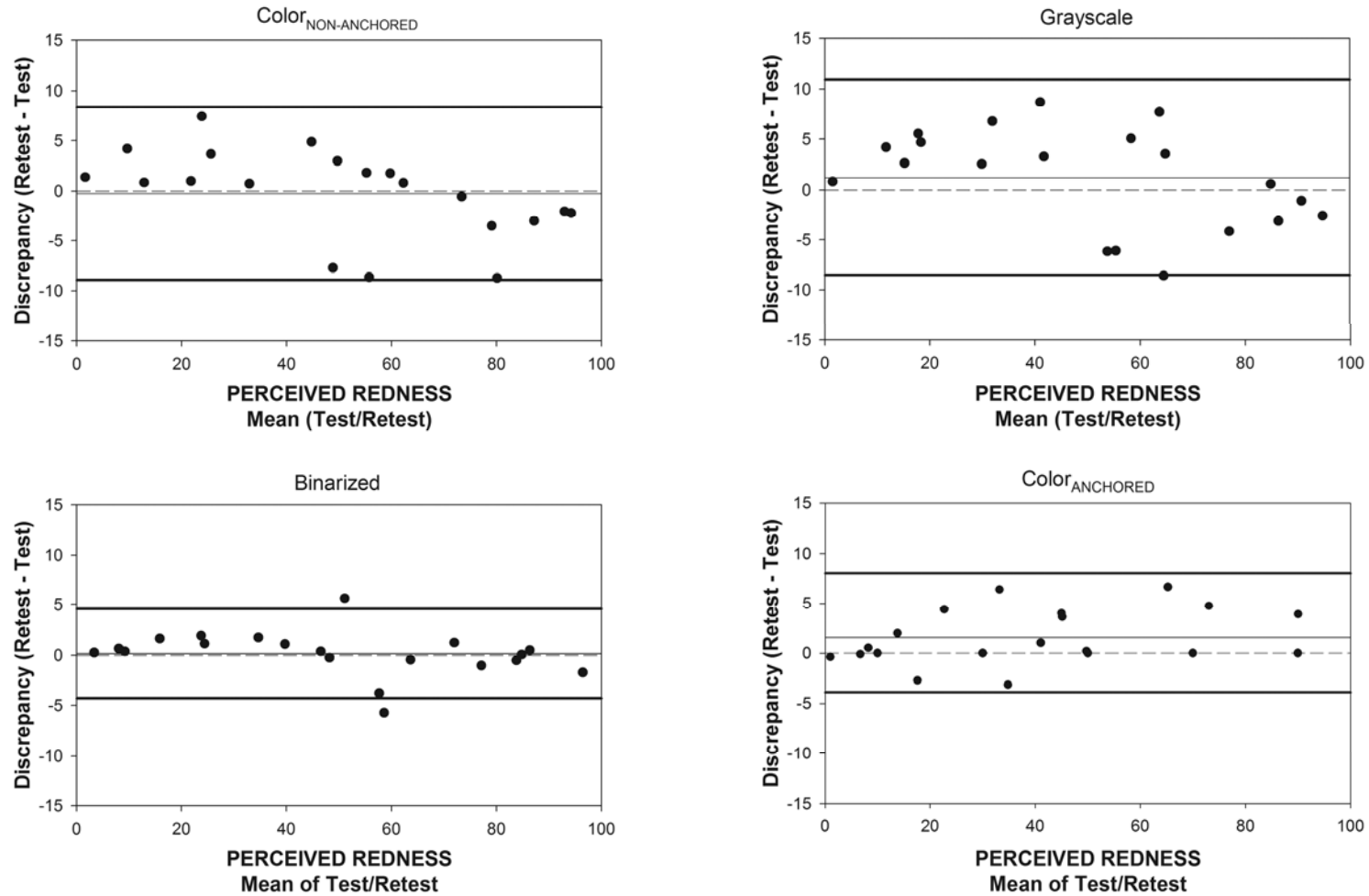


Figure 3-8: Limits of agreement for psychophysical scaling.

The LOAs (ordinate) are plotted vs. the mean perceived redness of test and retest for non-anchored colour, greyscale, anchored colour, and binarized scaling (clockwise). The thin solid line indicates the mean of test/retest differences for perceived redness. The upper and lower LOAs ($1.96 \cdot sd$) are shown as thick solid lines.

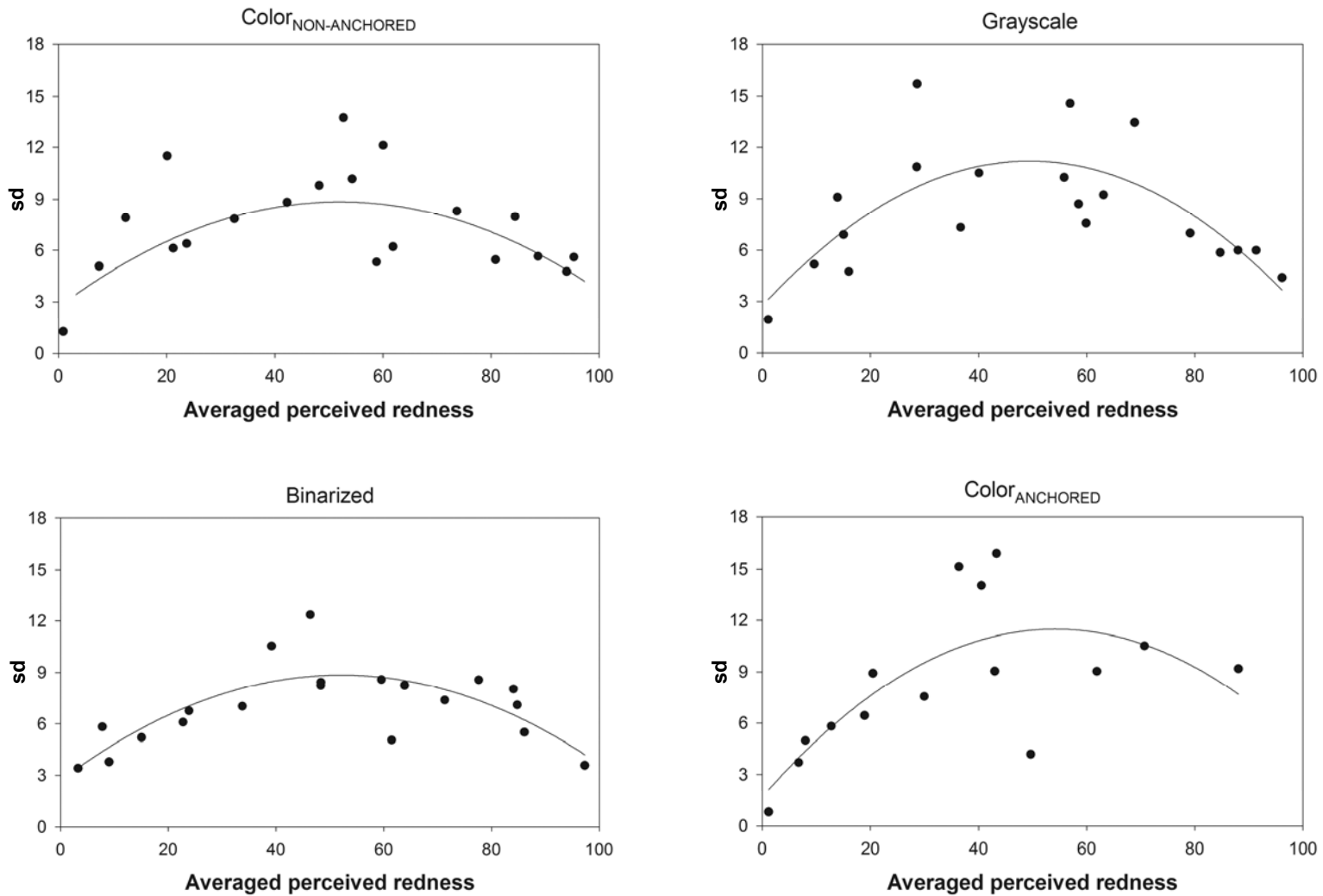


Figure 3-9: Variability between observers relative to averaged perceived redness.

The standard deviation (sd) of redness grades between observers (ordinate) is plotted vs. the averaged perceived redness for each reference image (abscissa). 2nd order polynomial estimates are shown using continuous lines.

3.4 Discussion

3.4.1 Image processing

The purpose of image processing is the application of a number of filtering and noise-reduction procedures to extract only the veridical information from the images of interest. The accomplishment of this task becomes difficult if images, as it was the case in this study, originate from different sources and vary with respect to image size and resolution. As the rationale of this work included the attempt to cross-calibrate the grading scales, it was important that the reference images were treated in the same way during image pre-processing. Therefore, the pre-processing steps for scale versions 1 and 2 (Chapter 4) were standardized by applying a custom image processing macro.

Processing of images at different time points did not have an effect on the resulting image. For each reference image and for both scale versions, the two binarized images that were derived based on repeated application of the image processing macro were absolutely identical, as shown in the resulting image in Figure 3-5. After subtraction of the images, all black pixels were removed, with only white pixels remaining. This finding, although somewhat expected, is important as it demonstrates that the macro that was developed was able to consistently process images at different times (weeks apart). The use of this processing macro also eliminated possible operator errors such as different filter settings or differences in the order of processing steps.

The accuracy of the pixel count is important if physical attributes of redness such as % *PC*, the area that is covered by vessels, will be used in the experiments.

Only if the pixel count can accurately be determined time after time, can appropriate conclusions about possible ocular changes be drawn. The histogram function in ImageJ gives the count of white and black pixels for the five generated line art images (Figure 3-2), and may be used for the objective quantification of the conjunctival vessels.

These results suggest that ImageJ and the described image processing macro can be used to objectively, accurately and repeatedly quantify bulbar conjunctival redness.

3.4.2 Fractal Analysis

Fractal dimensions quantify the complexity of a structure. In ocular research, they have been used as an objective metric to evaluate and detect changes of the retinal vasculature.^{22,23} In order to being able to use the fractal dimensions calculated in FracLac to quantify redness in the bulbar conjunctival vasculature, their accuracy had to be determined.

The accuracy of FracLac depends on the number of global scans used during box-counting.¹⁹ For a single scan, the starting position of the grid of boxes is placed by default in the top left corner of each image (coordinates $x=0$, $y=0$), whereas multiple scans have randomly set starting positions for the grid. Because the location of the starting grid affects the number of boxes that are required to cover the object of interest, the related fractal dimensions may change accordingly, so that multiple scans were recommended to improve the accuracy of the fractal dimensions that are derived by FracLac.¹⁹ The standard error (SE) that is shown in Table 3-1 is an indicator of how confident we can be in the calculated fractal

dimension.¹⁹ The use of ten multiple scans resulted in smaller standard errors than when only a single scan was used, independent of the reference analyzed and of the fractal dimension (Table 3-1). Therefore, ten global scans were used to calculate fractal dimensions throughout this research.

Using the data for ten global scans, the fractal dimension (of the four calculated using FraLac) that most closely matched the true fractal dimension of the test images was sought after. The data in Table 3-1 show that if simple structures such as the vertical and horizontal straight line were analyzed, each of the four calculated fractal dimensions exactly matched the expected fractal dimension of 1.00. If the structures became more complex, the agreement with the expected nominal fractal dimension of the Koch Snowflake and the Sierpinski Triangle depended on the algorithm that was used to calculate the fractal dimension. For both the Koch Snowflake and the Sierpinski Triangle, the slope-corrected most efficient covering fractal dimension (D_{sce}) most closely matched the nominal fractal dimension, with almost no deviation (≤ 0.01) from the expected nominal fractal dimension of the reference fractal images.

These findings suggest that FraLac is capable of accurately calculating fractal dimensions, and that it therefore can be used quantify redness in the bulbar vasculature. In the context of this work, the slope-corrected most-efficient covering fractal dimension (D_{sce}) based on ten global scans was used for analysis.

3.4.3 Photometric chromaticity

The repeatability of photometrically obtained data can be affected by a number of factors including the time of the measurement or inaccuracies in the

measured position. In the particular case of this study, the possibility of monitor inconsistencies (e.g. brightness fading over time) might have played a role as well.

The repeatability for test and retest session was very high, with almost perfect concordance between test and retest measures for each of the grading scales (Table 3-2). Inspection of Figure 3-6 and the data in Table 3-2 (COR) show that the discrepancies between test and retest were very small. The ROIs that were specified for this experiment were identical for both test and retest session. However, the positioning of the targeting spot at the centre of each ROI, taking the eight white squares of each yellow box (Figure 3-4) as references, was done based on the investigator's judgment only. Therefore, some variability in the positioning was to be expected. The high levels of repeatability between the two sessions suggest that these potential inaccuracies in the positioning of the targeting spot did not substantially affect the repeated measurements of u' .

However, there was a small but systematic bias towards higher measurements in the retest session (Table 3-2 [\bar{d}]) that appears to increase with higher nominal reference grades for each of the scales (Figure 3-6). The reasons for this finding are not entirely clear. Since the tripod and photometer had to be put back into place for the second session, small differences between the two sessions in photometer position and/or alignment to the screen, or a potentially slightly angled measurement off the screen might have contributed to this finding. Another explanation might be subtle differences in the ambient lighting in the room (due to differences in the power supply of the fluorescent light source in the ceiling).

A comparison of the chromaticity (CIE u') values for the VBR scale images between measurements off of the computer screen and the scale development³³ showed lower chromaticity for the measurements off of the computer screen. In particular, these differences were larger for the 'whiter' images, due to the higher luminance of the monitor compared to the printed versions of the images. However, in both cases, there was very little variability ($sd=0.001$) in chromaticity along the v' axis, indicating that bulbar redness measurements only vary along the u' -dimension (Figure 1-9, solid black line); it is worth noting that the invariant redness along the v' -axis for the redness measures - allowing an unidimensional description of redness - was the actual reason to prefer the u' and v' chromaticity coordinates over other systems such as the x,y system for scale development. Note that the position of the black line in Figure 1-9 represents an approximation of the redness range for the VBR scale only; however, closer inspection of this figure reveals that the redness range (black solid line) is quite distant from the white point of the chromaticity diagram, which is somewhat surprising considering that the scale was designed to include images from 'white' to 'red'. This finding might be due to the measurement setup for which the patient's eye (or here: the printed versions) was illuminated by a candescent light bulb, with all other room illumination turned off, thus potentially shifting the physical redness measures into more yellowish red hues and away from white.

The repeatable photometric measurements also provide important information for clinical research settings, where photometric setups such as this have been used for the objective assessment of redness in study participants.^{21,25} If the measurement spot on the patient's conjunctiva can be kept constant, for

example by having the participant look at a fixation spot, the repeatability of the photometer suggests that changes in chromaticity that are detected at different time points may be attributed to actual changes in bulbar redness.

3.4.4 Psychophysical scaling

It has been previously shown that subjective estimates of a condition often vary significantly between individual observers or over time.^{12,26-30} The plots in Figure 3-9 show that there was a similar trend for psychophysical scaling, as redness estimates varied significantly between individual observers, particularly for reference images that were perceived to be closer to the middle of the perceived redness range. This trend was observed for each of the image sets and for both non-anchored and anchored scaling. However, since the purpose of this study was to attempt a cross-calibration between scale levels within and between different grading scales, the variability between observers, although expected^{12,26-30}, might have masked similarities or differences between scales.¹² Therefore, it was decided to average perceived redness scores across observers, so that the performance of psychophysical scaling as measurement technique could be determined.

Psychophysical scaling provided highly repeatable averaged perceived redness, independent of the image set or scaling procedure (Table 3-3, Figure 3-7 and Figure 3-8). There was almost perfect concordance between test and retest session for averaged perceived redness across observers (CCCs all ≥ 0.984), with the (solid) best fit line of the test/retest plots being very close to the (dashed) 45°-line of equality corresponding to perfect concordance between sessions (Figure 3-7). The high levels for the ICCs indicate that the variability of psychophysical scaling

was very small relative to the overall variability between the subjects (i.e. the scaled reference images), and suggests that the measurements are highly reliable.³ The CORs were <10 (in perceived redness units) for each of the image sets and scaling procedures, with no systematic bias towards higher or lower averaged perceived redness between test and retest session (Figure 3-8).

These findings suggest that psychophysical scaling represents a robust methodical approach for the measurement of visual appearance³¹ that allows the comparison of scale levels on a common measurement scale. Since a period of four to six weeks elapsed before the retest scaling session was started, it appears unlikely that a recollection of redness estimates (i.e. scaled positions) from the test session might have triggered the high levels of agreement.

In summary, the objective and subjective techniques that were used for the estimation of redness in the reference images in this thesis were found to provide accurate and precise measurements of the parameters investigated.

In the following chapter, three physical redness attributes (D , % PC and u') will be quantified in order to determine the accuracy of bulbar redness grading scales. The resulting data will be compared across scales to investigate if a cross-calibration of the scales is feasible.

4 The Use of Fractal Analysis and Photometry to Estimate the Accuracy of Bulbar Redness Grading Scales

This chapter is published as follows:

Schulze MM, Hutchings N, Simpson TL. The use of fractal analysis and photometry to estimate the accuracy of bulbar redness grading Scales. *Invest Ophthalmol Vis Sci.* 2008;49(4):1398-1406.

Reprinted with permission. © Association for Research in Vision and Ophthalmology 2008

	Concept / Design	Recruitment	Acquisition of data	Analysis	Write-up / publication
Schulze	Y	n/a	Y	Y	Y
Hutchings		n/a		Y	Y
Simpson	Y	n/a		Y	Y

Table detailing role of each author in this publication (Y denotes significant contribution)

4.1 Overview

Purpose: To use physical attributes of redness to determine the accuracy of four bulbar redness grading scales, and to cross-calibrate the scales based on these physical measures.

Methods: Two image processing metrics, fractal dimension (D) and % pixel coverage (% PC), as well as photometric chromaticity were selected as physical measures to describe and compare redness in the McMonnies/Chapman-Davies, Institute for Eye Research, Efron, and a validated bulbar redness grading scale developed at the Centre for Contact Lens Research. Two sets of images were prepared using image processing: The first used multiple segments to cover the largest possible region of interest (ROI) within the bulbar conjunctiva in the original images; the second used modified scale images that were matched in size and resolution across scales, and a single, equally-sized ROI. To measure photometric chromaticity, the original scale images were displayed on a computer monitor, and multiple conjunctival segments were analyzed. Pearson correlation coefficients between each set of image metrics and the reference image grades were calculated to determine the accuracy of the scales.

Results: Correlations were high between reference image grades and all sets of objective metrics (all Pearson's r 's ≥ 0.88 , $p \leq 0.05$); each physical attribute pointed to a different scale as being most accurate. Independent of the physical attribute used, there were wide discrepancies between scale grades, with almost no overlap when cross-calibrating and comparing the scales.

Conclusions: Despite the generally strong linear associations between the physical characteristics of reference images in each scale, the scales themselves are not inherently accurate and too different to allow for cross-calibration.

4.2 Introduction

'Red eye', clinically known as bulbar hyperaemia, is an increased dilation of blood vessels in the bulbar conjunctiva that gives the eye its red appearance and is a prominent sign of ocular irritation. The recognition of change in redness is crucial for clinicians in management of the ocular surface, particularly in contact lens research and practice. Commonly, redness is estimated in a patient's eye by subjectively comparing it to references that represent different levels of severity for the condition, and changes over time can be monitored. The references are descriptive^{1,2}, illustrative³⁻⁶ or computer generated^{7,8}, and are presented in the form of grading scales. The subjectivity in grading is a criticism linked to the use of grading scales, and weak repeatability for inter- and intra-observer assessments is of particular concern for clinical practice.^{9,10} Aside from variability introduced by observer use, grading scales have been criticized for technical difficulties^{10,11} such as unequal steps, references not capable of covering the whole range of the scale, or biased depiction of references for different levels of severity. Hence it has been recommended that the different grading scales not be interchanged.^{8,12-14}

Repeatability has been the main focus of most research studies of traditional grading, either with respect to differences between observers^{3,11,12,15}, between grading scales^{8,12}, between levels of observer training^{6,16,17}, or as compared to novel objective techniques measuring the physical attributes of redness.^{9,14,18-23} The physical attributes to describe conjunctival redness have included various quantitative^{13,20-22,24} (e.g. number of vessels or % vessel coverage) and colorimetric variables^{9,18,19,23} (e.g. chromaticity levels or red intensity ratios) that were determined using either digital image processing or photometric techniques. However, only

three studies were found for which the physical attributes of the scales, per se, had been analyzed.^{13,14,19,23}

An interval or ratio scale level has been recommended for grading scales²² since it ensures uniform separation of reference images across the scale range; that is, a change from 10 to 20 on a 100-point scale represents the same difference as a change from 70 to 80 on the same scale. The extent of blood vessel coverage (% pixel coverage, an objective measure of redness) has been used to examine scales and compare them but not to specifically investigate the separation of the steps of the scales.^{13,14,19}

In this study we introduce fractal analysis, a new technique to analyze grading scales, and compare it to % pixel coverage^{13,14,19} and photometric chromaticity.²⁵ Fractal analysis has been shown to be a powerful objective technique to detect changes in various biological systems.^{26,27} It describes the complexity of the object or pattern by estimating the degree of branching of the vascular tree in the respective biological system.²⁸ Fractals found in nature are so-called random fractals, objects that are scale invariant over a finite range, which means that they look the same under different degrees of magnification or scale (e.g. the branches of a tree). They are quantized by a fractal dimension, D , describing the degree of branching. In a 2-dimensional photograph of vascular branches in the eye, the fractal dimension D can take on any decimalized value between 0 and 2.

Figure 4-1 shows simulated examples of vascular branching of the bulbar conjunctiva and the range of the expected fractal dimension, D . With respect to the eye, fractal analysis has been used to simulate corneal neo-vascularizations²⁹, and

has been successfully applied to investigate the vessel structure in normal and diseased retinæ.^{28,30,31}

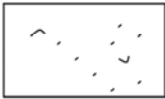

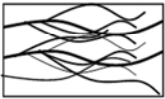

Degree of vascular branching	Scattered vascular artefacts	1 single, straight vessel	Common vascular branching	Complete vessel coverage
Conjunctiva shows:				
Fractal Dimension D	$0 < D < 1.0$	$D = 1.0$	$1.0 < D < 2.0$	$D = 2.0$

Figure 4-1: Simulated fractal dimensions (*D*) representing different degrees of vascular branching on the conjunctiva.

The purpose of this study was to estimate the accuracy of the grading scales by comparing the distribution and separation of the reference images of illustrative redness grading scales to objective physical attributes of redness defined by fractal analysis and photometric chromaticity, and to use these measures to cross-calibrate the scales.

4.3 Methods

4.3.1 Grading scale images

The bulbar redness reference images of four grading scales were analyzed: The McMonnies/Chapman-Davies scale³ (MC-D), the Institute for Eye Research scale⁴ (IER, previously known as CCLRU scale), the Efron scale⁵, and a validated bulbar redness scale⁶ (VBR) developed at the Centre for Contact Lens Research. The images in these four grading scales were generated using different procedures and

instrumentation, and differ in size, resolution, and the display of the area of interest (Figure 4-2).

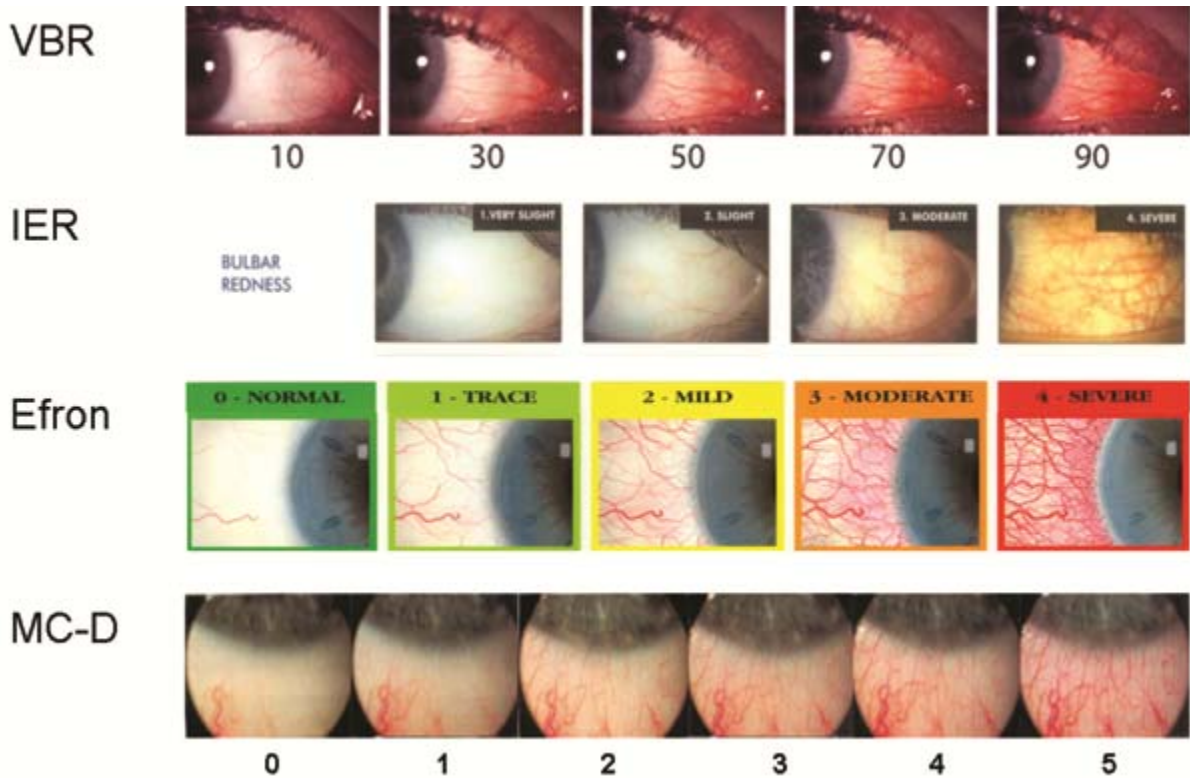


Figure 4-2: The bulbar redness grading scales analyzed.^{3,6}

The highest resolution reference images provided by the producers of each grading scale were used in the study. The MC-D scale images were scanned from the original hardcopy of the scale, and all others were digital in their original format. To perform fractal analysis it was required to save each image in the tagged image file format (TIFF); therefore the reference images for the Efron scale had to be converted from high resolution joint photographic experts group (JPEG) images into the TIFF format. Table 4-1 shows the original file types and the resolution and size of the original images.

Table 4-1: Image size and resolution.

Original file types for the grading scale reference images and their associated image size and resolution.

	Original File type	Size (<i>px x px</i>)	Resolution (<i>dpi</i>)
VBR	TIFF	1524 x 1012	300
IER	TIFF	700 x 525	100
Efron	JPEG	1628 x 1399	72
MC-D	TIFF	372 x 271	100

4.3.2 Image processing and fractal analysis

The public domain Java image processing program ImageJ 1.38x³² was used for the analysis of the color photographs or illustrations in the grading scales. Initially, the signal-to-noise ratio (SNR) was determined for each color channel to select the best channel for further analysis. The selected channel was then pre-processed to maximize the SNR and was then binarized prior to fractal analysis.^{33,34}

The image representing the lowest level of severity for each of the scales was used to determine the 8-bit color component (i.e. red, green or blue) that provided the highest SNR (in decibels). This was achieved by analyzing the SNR of each color component for the largest rectangle that included only the conjunctiva in each scale (see Methods: Scale Version 2 for details). An image has a high signal-to-noise ratio if the contrast of the object is large relative to the image noise; the noise in an image is characterized by the standard deviation of its brightness differences.³⁴ In this case, the numerator of the SNR ratio represents the conjunctival blood vessels and the denominator represents the noise in the background of the image. The

background noise is therefore determined for areas of the image that do not include vascular detail.

To determine the SNR, the procedure suggested by Young et al.³⁴ was followed. Since the signal (i.e. the blood vessels) is red, the red channel contained minimal target information¹⁹ and the whole image was used to determine the background noise. The signal-to-noise ratio for each color component of each scale was calculated using the following equation:

$$SNR = 20 \log_{10} * \frac{(a_{\max} - a_{\min})}{s_n},$$

where a_{\max} and a_{\min} represent the maximum and minimum brightness value within the whole image (brightness range) and s_n represents the standard deviation of the brightness values. Table 4-2 displays the data for each parameter and the SNR. For each of the scales, the green component provided the highest SNR and was consequently the image used for pre-processing and fractal analysis.

Table 4-2: Signal-to-noise ratio.

Grayscale brightness values, standard deviation and signal-to-noise ratio for each grading scale and each 8-bit RGB component.

	Red				Green				Blue			
	s_n	a_{\max}	a_{\min}	SNR	s_n	a_{\max}	a_{\min}	SNR	s_n	a_{\max}	a_{\min}	SNR
VBR	12	255	190	14.7	3	247	114	32.9	4	242	102	30.9
IER	13	255	117	20.5	2	251	103	37.4	3	253	76	35.4
Efron	6	255	147	25.1	1	255	20	47.4	2	255	45	40.4
MC-D	24	255	121	14.9	7	242	40	29.2	10	210	42	24.5

Due to the acquisition differences between the scales, two versions of the grading scales were generated for fractal analysis. The first version used the original reference images for each scale and used multiple segments of the region of interest (ROI) with the aim of covering the largest area possible of the conjunctiva (Scale Version 1); the size and number of segments required to achieve this varied between scales. The intention of this scale was to mimic clinical subjective grading, where a global estimate of the conjunctiva is commonly preferred to rating a single, prominent vessel.^{6,8} The second version modified the reference images prior to fractal analysis to match them in size and resolution across scales, so that for each reference image a single, rectangular and equally-sized ROI was analyzed (Scale Version 2). For both scale versions pre-processing settings were identical for all scales. The procedure to generate Scale Version 1 and Scale Version 2 is illustrated in Figure 4-3 for the Grade 50 reference image of the VBR scale.

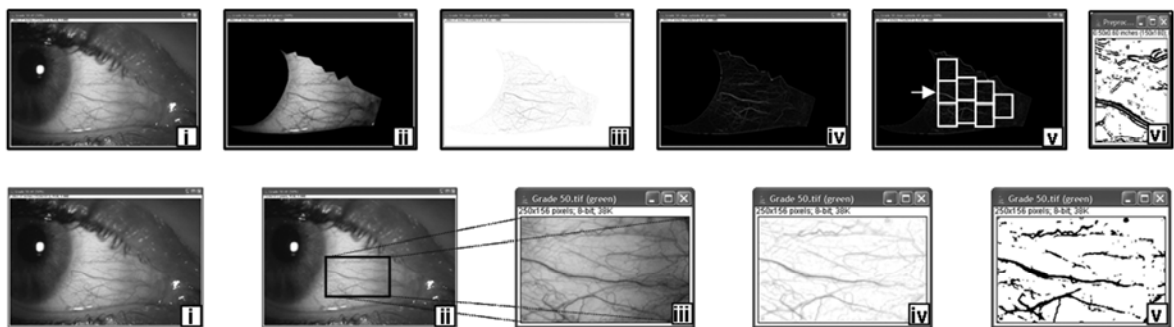


Figure 4-3: Image pre-processing steps.

Image pre-processing steps used for generating Scale Version 1 (A, top) and Scale Version 2 (B, bottom) are illustrated using the grade 50 reference image of the VBR scale. (i) original image; (ii) defined ROI; (A-iii) noise reduction & background subtraction; (A-iv) Sobel edge detection; (A-v) specification of multiple segments; (A-vi) Segment binarization; (B-iii) matched resolution & size for ROI; (B-iv) noise reduction & background subtraction; (B-v) ROI binarization.

Scale Version 1 - Greatest conjunctival area coverage

Cornea, eye-lids and lashes were excluded in the green component of the reference images to outline the overall region of interest (Figure 4-3A-i & Figure 4-3A-ii). A median filter was used to reduce the background noise. The filter settings were determined incrementally in order to obtain the highest correlation (for all scales) between scale grades and the objective fractal analysis measures. Next, the image background was subtracted to account for eyeball curvature (Figure 4-3A-iii).²⁵ Finally a Sobel edge detection algorithm³² was applied to highlight vessel edges as well as small capillaries (Figure 4-3A-iv). For each scale, eight to ten equally-sized segments of the ROI were selected to cover the largest area possible on the conjunctiva (Figure 4-3A-v). Although the size and number of the ROI segments was different between scales, the same segments were used across all steps within a scale. To complete pre-processing of the images, each segment was binarized using an automated thresholding procedure (Figure 4-3A-vi) where the pixel color (black or white) for foreground (blood vessels) and background was automatically assigned by ImageJ. For two of the images (Efron grades 3 and 4) the background and foreground pixel color was reversed, and the images were therefore inverted for consistency.

Scale Version 2 - Size-matching of reference images

To allow fractal analysis of a single, equally-sized ROI for each scale image, the resolution and size of the scale images were adjusted to account for the differences between scales shown in Table 4-1. First, the original images were matched in resolution with respect to the scale with the lowest resolution (Efron,

72dpi, see Table 4-1 and Figure 4-3B-i). Next, the largest possible ROI that could be fitted within the overall bulbar conjunctiva was determined for each scale image (Figure 4-3B-ii), and their sizes were compared to determine the ROI which was smallest for any of these images (MC-D, 250x156 pixels). This ROI was chosen as reference size, and proportionally scaled versions of this reference size (i.e. having the same ratio of horizontal to vertical pixels, approx. 1.6025 to 1) were fitted in the overall conjunctival outlines of the images in the three other scales. To allow the same x/y dimensions for all of the scales, the ROI was cropped out of the image in its original size (e.g. 400x250 for VBR grade 50) and down-scaled to the reference size of 250x156 pixels (Figure 4-3B-iii) and, in case of the Efron scale, was rotated by 90° counter-clockwise.

The background noise of the size and resolution matched green 8-bit image was removed using a median filter. The settings of the median filter were determined separately for Scale Version 2 (using the same method described in Scale Version 1), and the background was subtracted from each image (Figure 4-3B-iv). The Sobel edge detection filter was not applied to the image as it added significant noise and did not enhance the target. The reference images were then binarized as described for Scale Version 1 (Figure 4-3B-v).

Fractal Analysis

The ImageJ plug-in FracLac (v. 2.5 Release 1b5i)^{32,35} was used for fractal analysis. It employs box-counting algorithms to determine the fractal dimension of an object. During this procedure a series of grids of boxes with decreasing box size is placed over the ROI, and the number of boxes containing pixels with detail is

counted. The fractal dimension is then expressed as the slope of the regression line for the log-log plot of box size and count.³⁵

The following standardized settings in FracLac were developed according to the recommendations of the FracLac user manual.³⁵ The size of the series of grids was set to linearly decrease from a maximum box size of 45% of the horizontal ROI size to a minimum size of 1 pixel. Ten global scans were performed for each ROI, with randomly selected starting grid locations to improve the accuracy of the box-counting dimension. The following outcome measures were derived by FracLac and used to describe redness in terms of vascular branching (four different fractal dimensions) and area of vascular coverage (% pixel coverage):

- averaged D (D); *fractal dimension which is averaged over all 10 global scan locations*
- slope-corrected D (D_{sc}); *same as D , but corrected for periods of no change for the log-log plot of box size and count*
- most-efficient covering D (D_e); *same as D , but for each grid size the box-count that required the lowest number of boxes was used*
- slope-corrected most-efficient covering D (D_{sce}); *combination of all of the above*
- % Pixel Coverage (% PC); *the ratio of the number of black foreground pixels (representing vessels) to the overall number of pixels in the ROI.*

4.3.3 Photometric measurements

A spectrophotometer (SpectraScan PR650; Photoresearch Inc, Chatsworth, VA) was used to measure chromaticity (to estimate the amount of redness in the grading scale references). The photometer was mounted on a tripod and positioned at a fixed distance of 30cm from a LCD computer monitor (LG Flatron L1511S, Figure 4-4).



Figure 4-4: Standardized photometric setup.

The spectrophotometer was placed on a tripod 30cm away from the computer screen.

Experimental settings (room illumination, screen brightness and color, and photometer position) were standardized prior to the beginning of and controlled throughout the measurements. ImageJ v. 1.38x was used to display the unmodified reference images on the screen and to specify the associated, identical ROIs as established for Scale Version 1 (Figure 4-3A-v). To keep measurement settings stable, only the images on the screen and not the photometer itself were moved, and the position of each ROI was adjusted with respect to the fixation target in the eyepiece of the photometer. In a preliminary experiment it could be shown that

these measurements were highly repeatable and unaffected by factors such as brightness fading of the screen or inaccurate positioning of the ROI.²³

In a 30-second sequence five repeated measures of the same ROI were taken, and the chromaticity values u' and v' were subsequently averaged to estimate redness.⁶ The quantities u' and v' are described by the Commission International de L'Eclairage (CIE) in the CIE (u' , v') system and the CIE $L^*u^*v^*$ space.^{36,37} In the CIE (u' , v') system a chromaticity diagram with axes u' and v' is used to describe all possible colors, and the human perception of color differences in this diagram is approximately uniform across the whole diagram. While u' and v' determine the chromaticity of a color, the u^* and v^* of the CIE $L^*u^*v^*$ space includes a third dimension, lightness (L^*), the perception of luminance (L), to describe the color.^{6,37,38}

4.3.4 Data analysis

To determine fractal dimensions and % pixel coverage for the largest conjunctival area possible of each reference image, the results for the individual segmented ROIs of Scale Version 1 and the photometric measures were averaged to represent a global estimate of the conjunctival redness. The Pearson's product moment correlation coefficient (Pearson's r) was used to estimate the strength of linear association between scale steps and physical attributes of the scales (D , % PC , and photometric chromaticity).

4.4 Results

Pre-processing of the grading scale reference images resulted in a sequence of binarized images that either consisted of individual segment ROIs (Scale Version 1) or single ROIs (Scale Version 2) for each of the scales. The changes in severity

across the scale range for these binarized images are shown in Figure 4-5A (Scale Version 1, one segment ROI per scale) and Figure 4-5B (Scale Version 2).

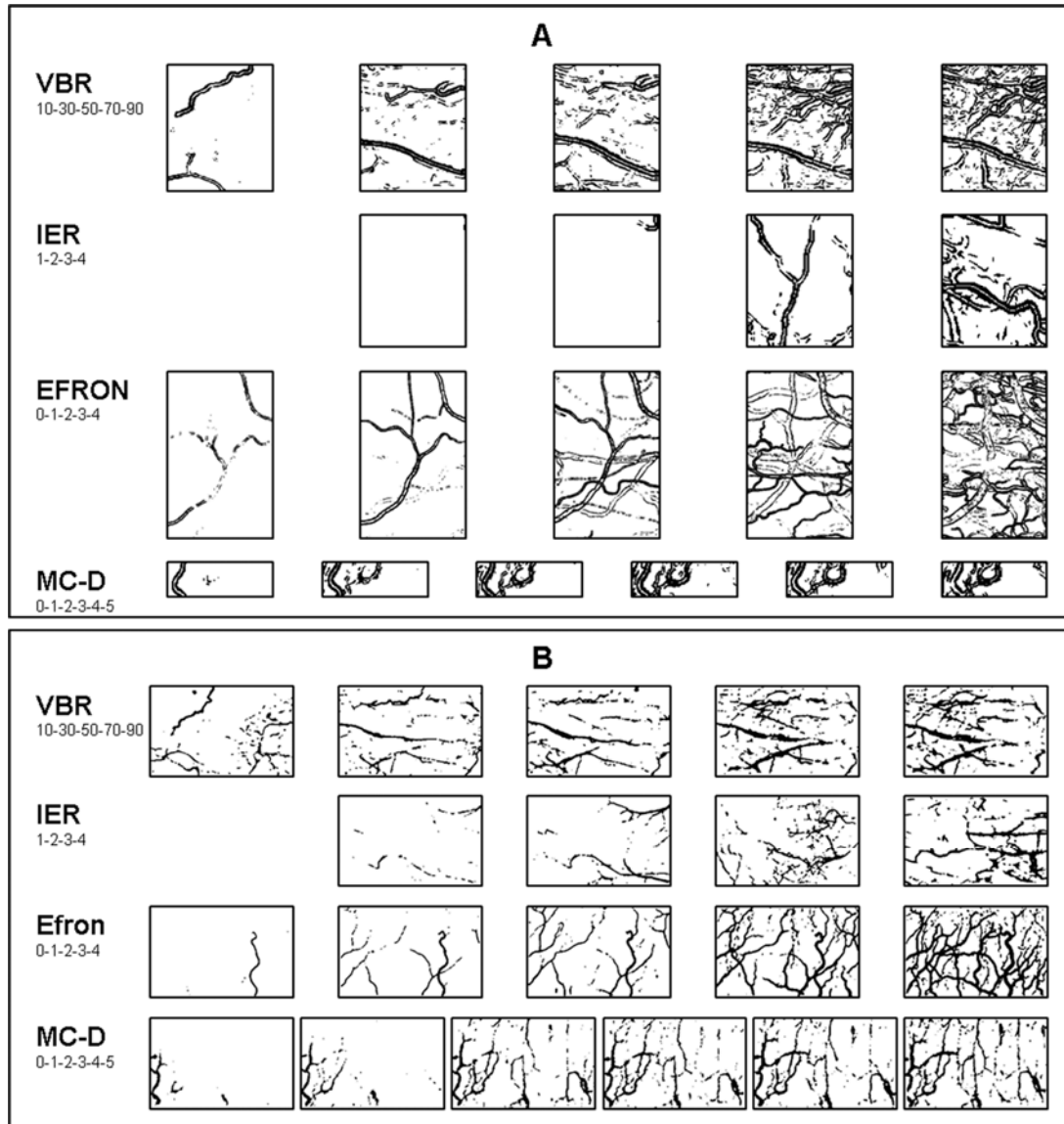


Figure 4-5: Resulting images from fractal analysis.

Resulting images from fractal analysis using Scale Version 1 (A, top) and Scale Version 2 (B, bottom). For Scale Version 1, a single segment of the ROI is illustrated across the reference levels for each scale. For Scale Version 2, the whole ROI (fixed size and resolution) is illustrated across the reference levels for each scale; the ROI in the Efron scale is rotated by 90° counter-clockwise to match the orientation of the other scale references. *Note: Within a scale, each image represents a scale step. The order of the scales is presented to be consistent with Figure 4-2.*

Independent of the measure (D , % PC , or photometric chromaticity), strong linear associations between grading scale steps and associated physical attributes were found, as expressed by Pearson's r 's of at least 0.88 (all $p \leq 0.05$) for any of the scales (Table 4-3). Correlation levels between scale steps and fractal dimensions as well as % pixel coverage for each grading scale are given - subdivided into Scale Version 1 and Scale Version 2 - in Table 4-3. Table 4-3 also shows the correlation levels for the combinations of scale steps and chromaticity measures u' and u^* . Pearson correlations for any combination of physical attributes were at least $r=0.89$ (all $p < 0.05$).

Table 4-3: Pearson correlation coefficients between scale grades and their associated physical attributes.

	VBR		IER		Efron		MC-D	
	1	2	1	2	1	2	1	2
<i>Scale Version</i>								
D	0.97	0.98	0.93	0.96	0.99	1.00	0.88	0.95
D_{sc}	0.97	0.98	0.93	0.96	0.99	1.00	0.88	0.95
D_e	0.98	0.98	0.93	0.95	0.99	0.99	0.88	0.94
D_{sce}	0.97	0.98	0.92	0.94	0.99	0.99	0.88	0.95
% PC	0.98	0.98	0.97	1.00	0.99	0.97	0.95	0.97
u'	0.99		0.95		0.94		0.97	
u^*	1.00		0.96		0.96		0.98	

Graphical display of the relation between scale grades and physical attributes of the scales are shown in Figure 4-6a-e. To allow display and comparison of all grading scales in the same graph the VBR scale was translated from its original 100-point format relative to the other scales. The results for D_{sce} are used to illustrate the relation between scale grade and fractal dimension for the images based on greatest conjunctival area covered (Scale Version 1, Figure 4-6a) and after size-matching of the scales (Scale Version 2, Figure 4-6b). The relation between scale grade and % pixel coverage for the greatest conjunctival area coverage and after size-matching of the images is shown in Figure 4-6c and Figure 4-6d, respectively. Figure 4-6e displays the relation between photometric chromaticity (shown as u') and associated scale grade for each grading scale.

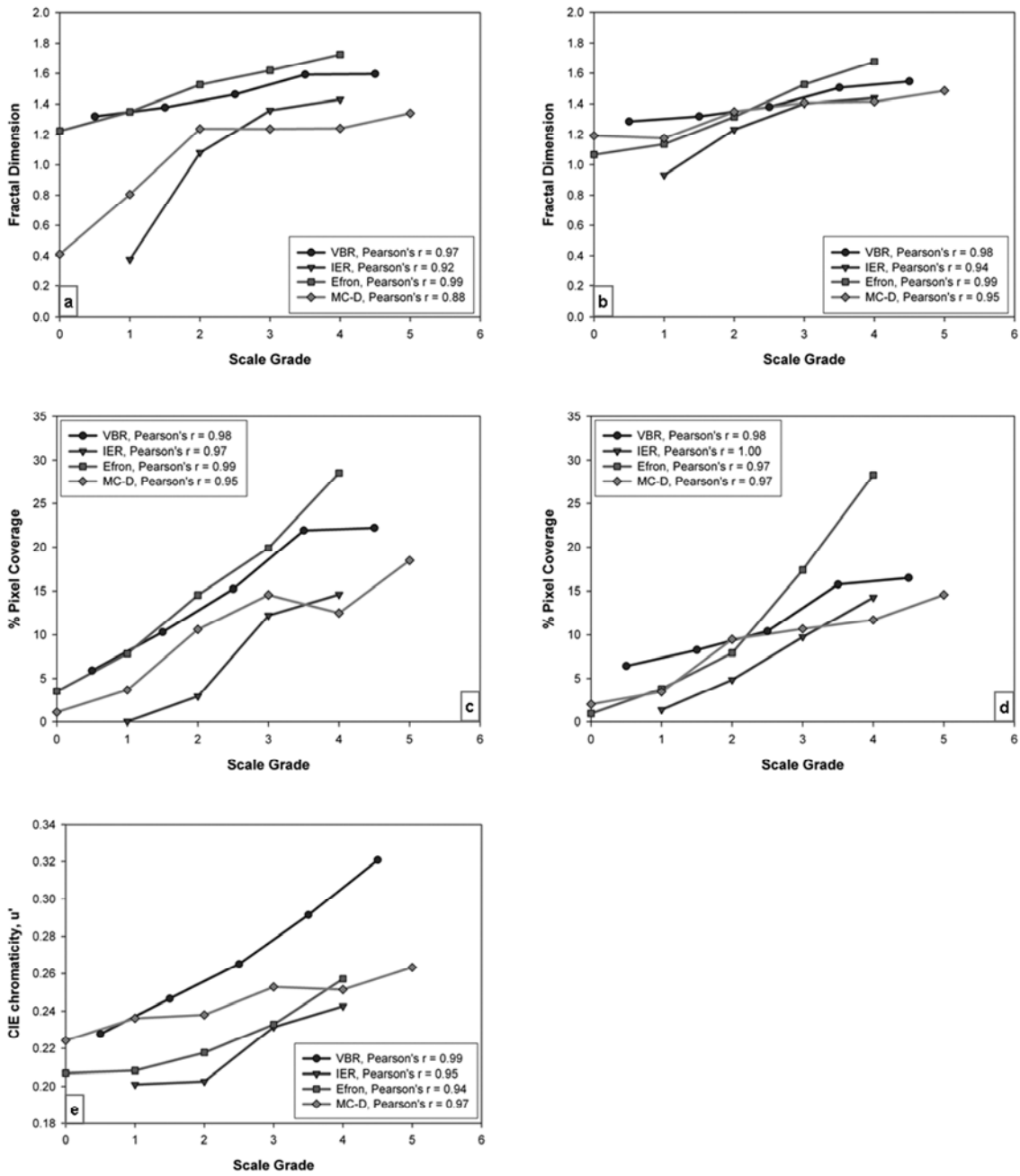


Figure 4-6: Scale grades vs. physical metrics.

Graphs showing the relationships between scale grades and D_{sce} for Scale Version 1 (6a) and Scale Version 2 (6b), between scale grades and % PC for Scale Version 1 (6c) and Scale Version 2 (6d), and between scale grades and CIE chromaticity, u' (6e).

4.5 Discussion

The purpose of this study was to estimate the accuracy of the grading scales by comparing the distribution and separation of the reference images of illustrative redness grading scales to objective physical attributes of redness defined by fractal analysis and photometric chromaticity, and to use these measures to cross-calibrate the scales.

One limitation of the study was that scale reference images were provided at different resolutions (72 – 300 dpi). As with all image processing techniques, the highest spatial resolution possible is advantageous particularly when small spatial details form the object of interest. However, even at the lowest resolution used in this experiment, there was a systematic relationship between grading scale steps and the estimated fractal dimension. If we were to attempt to glean recommendations from these resolution data, our results suggest that images acquired only at the resolution of common screen display (72 dpi) were sufficient to reasonably quantify redness based on fractal dimension. Most common clinical digital image acquisition instrumentation would provide much higher resolution than this.

4.5.1 Accuracy of the grading scales

The use of fractal analysis to analyze changes in vascular branching is an emerging strategy in clinical research.^{27,28,31} This is the first study in which fractal analysis has been used to evaluate vascular structures in the conjunctiva as well as to compare differences in the physical attributes between grading scale images. The strong correlations (all $r's \geq 0.89$; $p \leq 0.05$) between physical measures to

describe conjunctival redness (% pixel coverage^{13,14,19,22} and photometric chromaticity²⁵) and fractal dimensions indicate that fractal analysis is capable of describing changes in severity of bulbar redness.

The pre-processed scale images showing the changes in severity across the whole scale range are displayed in Figure 4-5a and Figure 4-5b. The physical attributes derived from these images (% *PC* and *D*) were highly correlated to the grading scale steps (Table 4-3). The types of fractal dimensions (*D*, *D_{sc}*, *D_e*, and *D_{sce}*) calculated by FraLac showed only minimal differences in the raw data for any scale for the same pre-processing procedure, which resulted in very small variations of the Pearson correlation coefficients (Table 4-3; differences within a scale ± 0.01). Therefore the slope-corrected most-efficient covering fractal dimension (*D_{sce}*), which eliminates periods of no change in the data by using the lowest number of boxes, was selected to illustrate the results of fractal analysis for Scale Version 1 (Figure 4-6a) and Scale Version 2 (Figure 4-6b).

The results of this study showed high levels of linear association between grading scale reference levels and physical attributes (Table 4-3, range of $0.88 \leq r \leq 1.00$) for all grading scales. A Pearson correlation of 1.00 represents a perfectly linear association between scale grades and physical attributes; grading scales that exhibit this feature may be characterized at least as interval. For the means of this study we decided to define the accuracy of a grading scale by the level of linear association it exhibited between scale steps and associated physical attribute (Pearson's *r*, Table 4-3). Thus, the most accurate grading scale would be the scale for which this correlation was the highest, and the least accurate scale the one with the lowest Pearson correlation level.

In our study, a Pearson correlation of $r=1.00$ between scale reference grades and one physical attribute was found for each grading scale except for the MC-D scale. However, each physical attribute extracted from the images pointed to a different scale as most accurate. If the accuracy of a scale were defined only on the amount of vessel coverage across the scale range, the IER scale would be the most accurate (Table 4-3, % PC). Based on fractal dimension, (Table 4-3, all types of D) the Efron and the VBR scales were more accurate. With the third physical attribute of redness in our study, photometric chromaticity, the VBR scale showed the highest linear association to the reference grades and this therefore might be described as the most accurate scale.

The high accuracy of the VBR scale with respect to chromaticity is not surprising since this scale was developed based on a combination of subjective estimates and photometric chromaticity measures.⁶ The reference images for the Efron scale were painted, perhaps focusing on highlighting certain features and simultaneously avoiding confounding artefacts.³⁹ Our study supports this intention as we determined consistently strong associations with degree of vascular branching and the amount of vessel coverage, whereas the linear association between photometric chromaticity changes and scale steps was lowest of all scales.

These results show that estimating the accuracy of a grading scale is closely related to the technique and the physical attributes used. Overall, fractal dimension, % pixel coverage, and photometric chromaticity all were capable of detecting changes in the severity of redness; the high linear associations between scale steps and physical attributes indicate that the anticipated change in the scales could be determined with the physical attributes used, and that all scales thus may

be considered accurate. Superior or inferior accuracy of a scale, however, should only be defined by indicating the physical attribute that was used. Based on our results it seems that each scale describes one characteristic of redness best; VBR and MC-D scale best describe redness in terms of photometric chromaticity, the Efron scale with respect to changes of vascular branching (D), and the IER scale with respect to changes in the area that is covered ($\% PC$). The consistently high correlation levels for the VBR scale (all $r \geq 0.97$), however, suggest that this scale is the least affected by the selection of the physical attributes or the pre-processing procedure.

4.5.2 Comparison and cross-calibration between the grading scales

In the past, various studies suggested that grading scales should not be used interchangeably.^{8,12-14} However, the physical attributes of each grading scale image could be used in an attempt to cross-calibrate the grading scales. An illustration of this is Figure 4-6, in which physical attributes of each image are plotted against their associated nominal scale grades.

Figure 4-6a-e shows that there are large differences in the physical attributes of each scale for equivalent grades highlighting that cross-calibration across all grades is impossible. The differences in size and resolution (among many others) between the scale images complicate the selection of the physical measure to cross-calibrate the scales.

Equivalency between grading scales and the physical attributes of the conjunctival images occurs at points where the scales coincide. For fractal dimension there is little convergence of the scales; indeed, at step 1 of the scales,

the fractal dimension ranges between 0.38 and 1.35. Percent pixel coverage also shows little convergence across scales as evidenced in Figure 4-6, graphing data for Scale Version 1. A contributory factor to the large spread of values might be the different ROI sizes. In this study, fractal dimensions appeared to be affected by the size of the ROI (the larger the ROI, the higher the number of box counts and the fractal dimension) and because each scale's images have inherently different sizes, this confounds the ability to cross-calibrate scales. This problem is ameliorated by using boxes that have the same size, as evidenced in the two panels at the right (Scale Version 2). The regions where there is greatest convergence are approximately at a grade of 2.2.

Figure 4-6e shows that three intersections between scales were found when chromaticity was used. Chromaticity was measured on the original, un-altered images when displayed on a computer screen under identical illumination and monitor settings. Except for adjusting magnification levels on the screen to account for the differences in size and resolution of the original images (Table 4-1), no further alterations of the images were required. Because of the least observer intervention involved in this procedure, the selection of photometric chromaticity would seem the most logical solution to cross-calibrate the scales. As can be seen in Figure 4-6e, however, there was a wide range of scale grades for the same photometric chromaticity measure. As an example, the image representing grade 4 in the IER scale corresponds to a value for u' of 0.24; for the other scales, this chromaticity value would correspond to interpolated grades of approximately 1.3 (VBR), 2.3 (MC-D), and 3.4 (Efron), showing that the same level of photometric chromaticity represents a range of about 3 scale steps for all scales. It seems that

chromaticity is very different between the scales, and that, except for the VBR scale, chromatic changes between scale steps are not linear across the full scale range.

An unfortunate conclusion that we are inevitably left with is that despite the generally strong linear associations between the physical characteristics of reference images in each scale, the scales themselves are not inherently accurate. What allows the switching between scales by clinicians is the huge non-linear compensations that are made by those using the various scales: We propose that this is accomplished by a novel mechanism that we refer to as clinical scale constancy. This is similar to perceptual mechanisms such as color constancy⁴⁰ which allow relatively invariant perceptions despite differences in physical attributes of the image. This is exactly what is occurring here: The images are physically different but, say, grade 0 is perceived to be low (not red) regardless of the scale used, similar to a dark object appearing the darkest of its surroundings regardless of strong or weak illumination.⁴⁰ This observation of our ability to rescale clinical attributes despite their physical content has profound implications for developing measuring tools and suggests that current measurement theories⁴¹⁻⁴³ are inadequate because certain measurement tenets are less important and soft/weak metrology⁴⁴ should take these looser constraints into account when defining what constitutes measurement. In addition, teaching clinical judgments is also influenced by understanding that under certain conditions we are able to ignore absolute physical attributes while using relative (within-scale) characteristics to reach appropriate clinical conclusions about redness (and presumably other aspects of ocular appearance).

4.6 Conclusion

The first goal of this study was to determine the accuracy of bulbar redness grading scales based on their correlation levels between scale grades and three physical measures (D , % PC , and photometric chromaticity). Based on our results all scales might be considered accurate based on the criteria we specified; however, it seems that each scale describes one characteristic of redness best: VBR and MC-D scale best describe redness in terms of photometric chromaticity, the Efron scale with respect to changes of vascular branching (D), and the IER scale with respect to changes in the area that is covered (% PC).

The second intent of this series of analyses was to cross-calibrate the scales so that they could be compared via the physical measures. We have shown that an objective cross-calibration between scales would be technically possible, but as is apparent because of the wide discrepancies and the relatively low overlap, it cannot be recommended. Differences between acquisition methods, image quality, and physically obtained measures show that the differences between grading scales are too severe to allow for cross-calibration. Among many things, this highlights the need for standardization in as much as scales are proposed with little physical similarities, bringing into question whether the numbers that represent the steps on the scale (and therefore the numbers derived using the scales) are measurements at all.

4.7 Acknowledgements

The authors would like to thank Dr. Charles McMonnies, Dr. Nathan Efron, and the International Association of Contact Lens Educators (IACLE) for providing us

with high resolution copies of the original reference images, and the Canada Foundation for Innovation (CFI) for an equipment grant.

The next chapter will determine the perceived redness of the reference images by means of psychophysical scaling, and compare the perceived redness to the physical redness metrics that were described in this chapter.

5 The Perceived Bulbar Redness of Clinical Grading Scales

This chapter is published as follows:

Schulze MM, Hutchings N, Simpson TL. The Perceived Bulbar Redness of Clinical Grading Scales. *Optom Vis Sci.* 2009;86(11): 1250-1258.

Reprinted with permission. © The American Academy of Optometry 2009

	Concept / Design	Recruitment	Acquisition of data	Analysis	Write-up / publication
Schulze	Y	Y	Y	Y	Y
Hutchings	Y			Y	Y
Simpson	Y			Y	Y

Table detailing role of each author in this publication (Y denotes significant contribution)

5.1 Overview

Purpose: To use a psychophysical scaling method to estimate the perceived redness of reference images of the McMonnies/Chapman-Davies (6 reference levels), IER (4), Efron (5), and VBR (5) bulbar redness grading scales.

Methods: Regions of interest were cropped out of the grading scale reference images; three separate image sets (color, grayscale, and binarized) were created for each scale, combining to a total of 20 images per image set. Ten naïve observers were asked to arrange printed copies of the 20 images per image set across a distance of 1.5m on a flat surface, so that separation reflected their perception of bulbar redness; only start and end point of this range were indicated. The position of each image was averaged across observers to represent the perceived redness for this image within the 0-100 range. Subjective data were compared to physical attributes (chromaticity & spatial metrics) of redness.

Results: For each image set, perceived redness of the reference images within each scale was ordered as expected, but not all consecutive within-scale levels were rated as having different redness. Perceived redness of the reference images varied between scales, with different ranges of severity being covered by the images. Perception of redness severity depended on the image set (RM ANOVA; all $p \leq 0.0002$). The perceived redness was strongly associated with the physical attributes of the reference images.

Conclusions: Subjective estimates of redness are based on a combination of chromaticity and vessel-based components. Psychophysical scaling of perceived redness lends itself to being used to cross calibrate these four clinical scales.

5.2 Introduction

The assessment of the ocular surface is a routine procedure in clinical practice, and of particular interest with respect to contact lens wear. A visible sign of ocular irritation is increased redness of the normally white conjunctiva, a condition clinically referred to as bulbar hyperaemia. In contact lens wear, bulbar hyperaemia is commonly observed as a consequence of mechanical irritation of the eye by the lens and its edges, or disaffected metabolism of the eye, or a chemical reaction of contact lens and cleaning solution. Any of these effects can result in increased vasodilation of the conjunctival blood vessels that gives the eye its red appearance.¹ For the clinician it is crucial to monitor conjunctival tissues over time, to be able to detect even small changes and to select an appropriate treatment, if required.²⁻⁵ To assess and monitor bulbar redness, grading scales have become increasingly popular with clinicians and researchers.⁶ They are used by comparing a patient's eye to a reference scale, which defines different levels of severity for a particular clinical sign. Ophthalmic scales typically employ five discrete reference levels based on descriptions and/or illustrations that correspond to increasing levels of severity between 0 and 4.⁷⁻¹² A number of illustrative grading scales have been developed in an effort to standardize clinical assessments (thus reducing inter- and intraobserver variability) and to assist in the detection of small, but possibly clinically significant changes of ocular structures.^{13,14} Interpolation between these steps has been recommended to increase the sensitivity of the assessments.^{7,8,10,15}

McMonnies and Chapman-Davies introduced the first photographic grading scale (MC-D) for the assessment of bulbar redness, and showed that their scale was

capable of detecting statistically significant changes between hard, soft, and non-contact lens wearers.² Two additional photographic scales for bulbar redness have since been introduced: the '*Institute for Eye Research*' grading scale¹⁶ (IER; previously known as CCLRU) and the '*Validated Bulbar Redness*' scale¹⁰ (VBR). The reference images in the latter scale were selected using psychophysical scaling, and the scale grades were subsequently validated by demonstrating their strong linear correlation to photometric chromaticity.¹⁰ Efron introduced a grading scale using artist-rendered illustrations, citing better standardization of the images with respect to eye, illumination, or angle of view as an advantage of drawn, rather than photographic images.^{6,7,17} The objective validation of the reference grades sets the VBR scale apart from the other three scales described above, for which reference image selection had been based on clinical experience and subjective judgments only.^{2,6,7} Figure 5-1 shows the four bulbar redness grading scales.

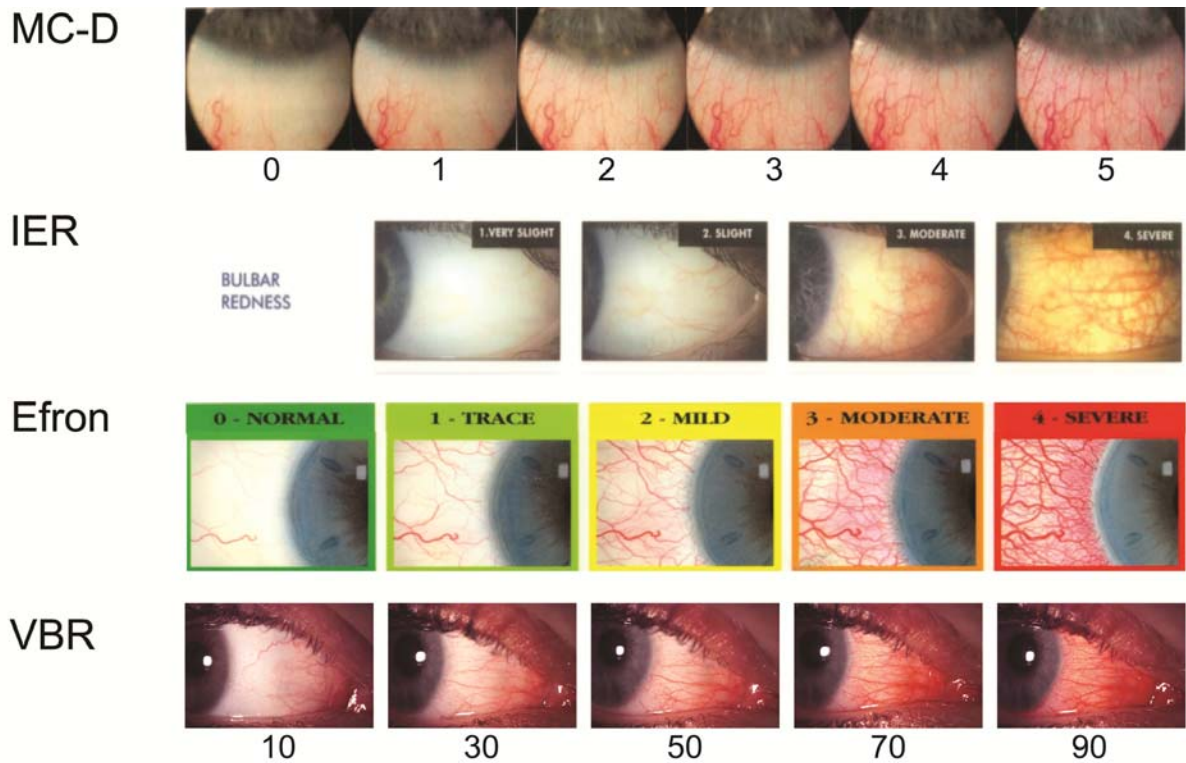


Figure 5-1: The McMonnies/Chapman-Davies (MC-D), Institute for Eye Research (IER), Efron, and Validated Bulbar Redness (VBR) grading scales.

The use of illustrations and the interpolation of grades have certainly contributed to a better standardization of clinical grading; however, the reliability of assessments remains a major criticism. Various studies have investigated the clinical performance of grading scales and shown that inter- and intraobserver repeatability was limited.^{2,9,18-21} Poor interobserver agreement has also been attributed to a lack of empiric information on how redness grading is performed. Redness assessments likely follow at least two general strategies that are either based on a color/luminance component (such as judging the chromaticity of the eye), based on spatial (vessel) criteria (such as appearance, complexity, or the area

covered by vessels), or on a combination of color and spatial (vessel) information.^{19,22}

Leaving reliability aside, technical difficulties of the scales such as unequal scale steps or the inability to cover the full range of severity have also been criticized.^{9,15} Therefore it is not surprising that the literature indicates that different grading scales not be interchanged.^{5,19,20,23,24} Despite depicting the same ocular sign, the scales differ in number of reference images used, the severity range covered, and the conjunctival region displayed in the images (Figure 5-1).

To overcome the vagaries associated with clinical grading, different objective techniques such as image analysis or photometric measurements have been introduced to estimate redness based on spatial criteria^{4,19,25,26} or colorimetric information.^{11,19,26-29} Due to their nature, by excluding or at least minimizing observer intervention, it is not surprising that objective measurements were found to be superior to subjective grading for repeatability¹⁹ and sensitivity to detect changes.⁴ Objective techniques represent valuable tools particularly in research settings, where the detection of even small conjunctival changes to highlight differences between contact lens materials or cleaning solutions is of interest. However, in clinical practice, it is questionable that objective techniques will replace subjective grading, since practicability, availability, and lower cost are strong reasons for the continued use of grading scales. In research settings, instruments that allow for objective estimates of redness may be available, but even here it appears likely that for studies focusing on multiple aspects of contact lens wear in particular, the routine subjective assessment of redness using grading scales will remain (at least) an additional standard assessment.

Given the likelihood of future utilization of grading scales in clinical settings, a better understanding of the available grading scales and the processes that underlie grading is desirable. In a previous study, three physical measures that describe redness based on spatial (fractal dimension and % pixel coverage) and colorimetric information (photometric chromaticity) were used to determine the accuracy of the MC-D, IER, Efron, and VBR bulbar redness grading scales.⁵ The current study used a psychophysical scaling method to estimate the perceptual relationship between the reference images of these scales. The perceptual estimates of redness were used to compare scale levels within and between scales, and to examine which general strategies are most likely applied in the subjective assessment of redness by comparing the perceived to the physical⁵ redness of the reference images.

5.3 Methods

5.3.1 Grading Scale Images

The reference images of the MC-D (6 images), IER (4), Efron (5), and VBR (5) bulbar redness grading scales (Figure 5-1) were modified according to the procedure described elsewhere.⁵ In brief, images were matched in size and resolution and modified so that they displayed conjunctival detail only, with confounding regions such as lids or lashes being excluded. To evaluate the effect of color on redness assessments, three sets of images were developed, showing vascular detail either in color, grayscale, or binarized. For each image set, the 20 reference images were printed in a size of 5x3cm and randomly coded on the reverse.

5.3.2 Participants

Ten participants (five males and five females) naïve to the use of grading scales were enrolled into the study; all participants were undergraduate or graduate students (mean age 25.5 years, range 23-31) at the School of Optometry, University of Waterloo, with no previous clinical experience. 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, ON, Canada; informed consent was obtained from each participant prior to starting the study.

5.3.3 Psychophysical scaling method

The psychophysical scaling method used to estimate the perceived redness of the reference images of the four grading scales was a combination of partitioning and magnitude estimation that might be referred to as relative magnitude scaling.^{30,31} The procedure was very similar to a previous psychophysical scaling experiment¹⁰: To allow redness severity scaling, a scale range of 1.5m was marked on a table top. The only indicators for redness severity were labels at the start and end point of this 1.5m range corresponding to a minimum redness score of 0 and a maximum redness score of 100, respectively. No intermediate severity levels were given. Minimum and maximum redness levels were not further defined and were intended to correspond to what each observer subjectively associated with minimal or maximal redness. The 20 images were presented randomly spread out on the table top, and participants were asked to position each image based on their perception of redness severity. Participants were asked to give a global estimate of redness, giving similar vessels the same weight. Images that were perceived as having a low amount of redness were to be placed closer to the start

point at 0, and images with a high degree of redness closer to the end point of the range at 100. The participants were instructed to place them based on estimating their perceived redness. The basic task of each observer was to look at the set of images and position them where they chose to, taking as long as needed. They were allowed to make as many adjustments as needed and were instructed that images could overlap. At the end of each scaling session, their arrangement included every single image, the position of which represented its “perceptual severity” for the 0 to 100 range. The position of each image was measured, scored between 0 and 100, and averaged across observers. For each observer, the order of image set presentation (color, grayscale, or binarized) was randomized, with a break of at least two days before scaling the next image set. Repeatability of the perceptual assessments was evaluated in a second session that was carried out approximately four to six weeks after the first session.

Except for the section illustrating the effect of color information on the perception of redness, the paper will focus on evaluating the averaged perceptual scores of the color image set, as these images are the ones used in clinical practice.

5.3.4 Perceived vs. physical redness

The perceived redness for the reference images was compared to previously determined physical redness for these references.⁵ The vessel related metrics fractal dimension (D) and % pixel coverage (% PC) were measured as previously.⁵ To describe redness in colorimetric terms, the CIE (Commission International d’Eclairage) u' component of the CIE (u' , v') system was used to define redness chromatically. Each reference image was placed in a stationary slide holder that was

attached to a modified slit lamp mount, and the SpectraScan PR650 spectrophotometer (Photo Research Inc, Chatsworth, CA, USA) was used to measure u' .¹⁰ Chromaticity was averaged across nine regularly spaced locations (3x3 grid) to represent a single, global estimate of redness for each reference image.

5.3.5 Data analysis

Statistical analysis was performed using STATISTICA version 8 (StatSoft. Inc., Tulsa, OK, USA); an alpha level of ≤ 0.05 was considered statistically significant. Test-retest correlation coefficients of concordance (CCC)³² and the coefficient of repeatability (COR; $1.96 * s_d$)^{18,33} were used to determine the repeatability of assessments of the first and second session. CCC describes the concordance of repeated measurements, with a CCC of 1.00 representing identical repeated scores.^{10,32} The Pearson product-moment correlation coefficient (Pearson's r) was used to evaluate the strength of linear association between both sessions. After verifying normality of the outcome variable (Kolmogorov-Smirnov test), repeated measures analysis of variance (RM ANOVA) with post-hoc Tukey Honestly Significant Differences (HSD) tests were used to determine perceptual differences for scale levels within each scale and to evaluate if redness perception was different between the three image sets. Pearson's r was used to quantify the strength of linear association between scale steps and perceptual reference image positions. The relationship between perceptual scale positions and objective metrics⁵ was analyzed using Pearson's r , partial correlation coefficients, and multiple regression analysis. Except for Table 5-1 (repeatability coefficients), all tables and figures in the results section of the paper show data obtained in the first scaling session.

5.4 Results

For each observer, scaling of perceived redness required about 5 to 10 minutes. In general, variability between observers was largest for images perceived to have redness scores close to the middle of the 0 to 100 range. Since the purpose of this study was to evaluate the perceptual relationship between scale levels within and between different grading scales, and not to compare the performance of individual observers, it was decided to average scores across observers. Averaging across observers reduced the impact of individual observers, which otherwise might have masked similarities or differences between scales.¹⁸ Therefore, the perceived redness of each reference image is represented by a single, averaged score in the following analyses.

Averaged scaling data were highly repeatable across observers for each image set, as expressed by COR levels (<10) and very high concordance (CCC>0.98) between sessions. Almost perfect linear associations between scaling sessions were found for each image set (Table 5-1).

Table 5-1: COR, CCC, and Pearson's r for the three image sets.

Coefficient of repeatability (COR), correlation coefficient of concordance (CCC), and Pearson's r for the three image sets.

	Color	Grayscale	Binarized
COR (1.96*sd)	8.7	9.7	4.4
CCC	0.99	0.98	1.00
Pearson's r	0.99	0.99	1.00

Figure 5-2 shows the relationship between averaged perceptual scores for session 1 and session 2 for the color image set; the plots for the grayscale and binarized image set showed very similar relationships. Because of the high levels of repeatability and agreement for the two sessions, the results of the first scaling session were used to compare perceived redness between scales.

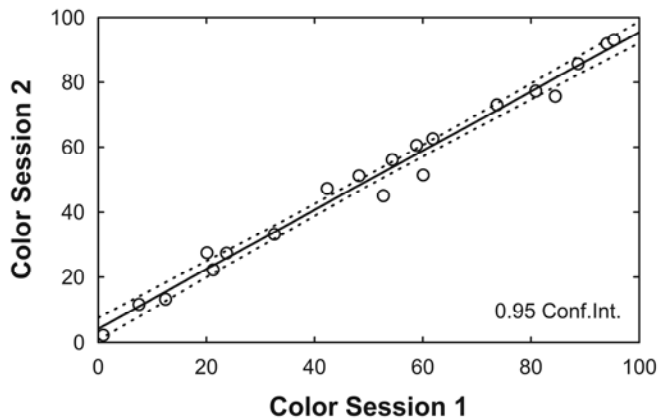


Figure 5-2: Averaged perceptual scores for session 1 vs. session 2

Averaged perceptual scores for session 1 vs. session 2 for the color image set. The dashed lines represent the 95% confidence intervals.

Figure 5-3 shows the reference images of the MC-D, IER, Efron, and VBR scale at the positions that correspond to their averaged perceptual redness severity within the 0 to 100 range. The perceptual order of the reference images within each scale was in agreement with their assigned reference grades.

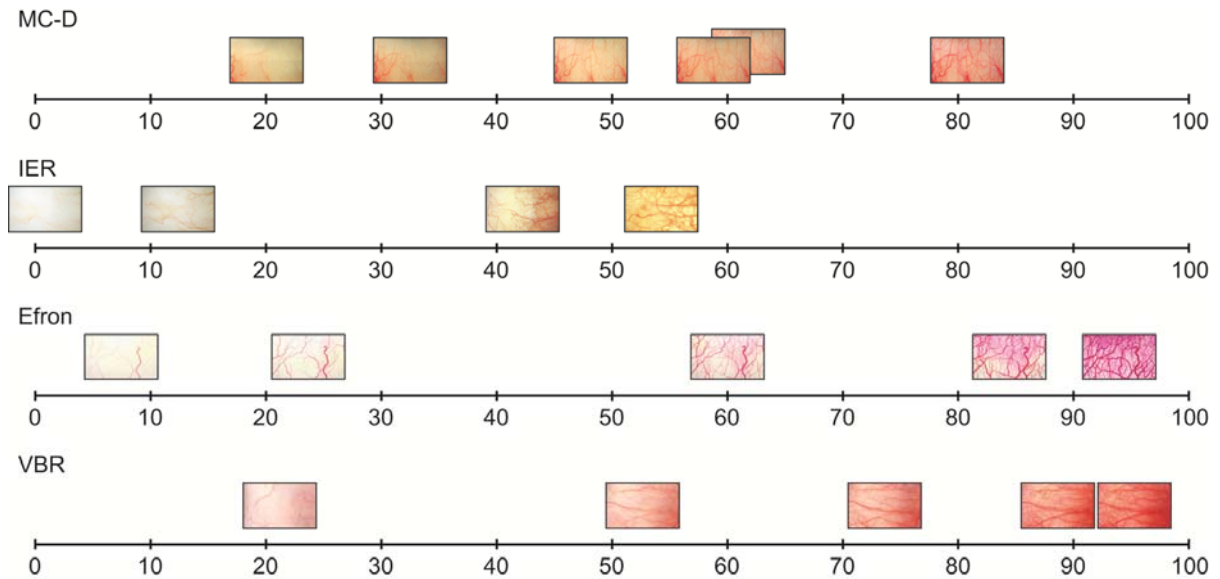


Figure 5-3: The perceptual scores for the reference images.

The perceptual scores for the reference images of the MC-D, IER, Efron, and VBR scale (top to bottom); displayed are the results for the color image set (image size not to scale for better recognition of the images).

When evaluating perceptual differences between scale steps within each scale, there were at least two consecutive steps that were not perceived to be different for at least one of the image sets (color, grayscale, or binarized). For the example of the color image set, the reference images representing grade 3 and 4 for the MC-D scale and the reference images representing grade 70 and 90 for the VBR scale were perceived to have similar redness (Tukey; $p=0.92$ and $p=0.24$, respectively).

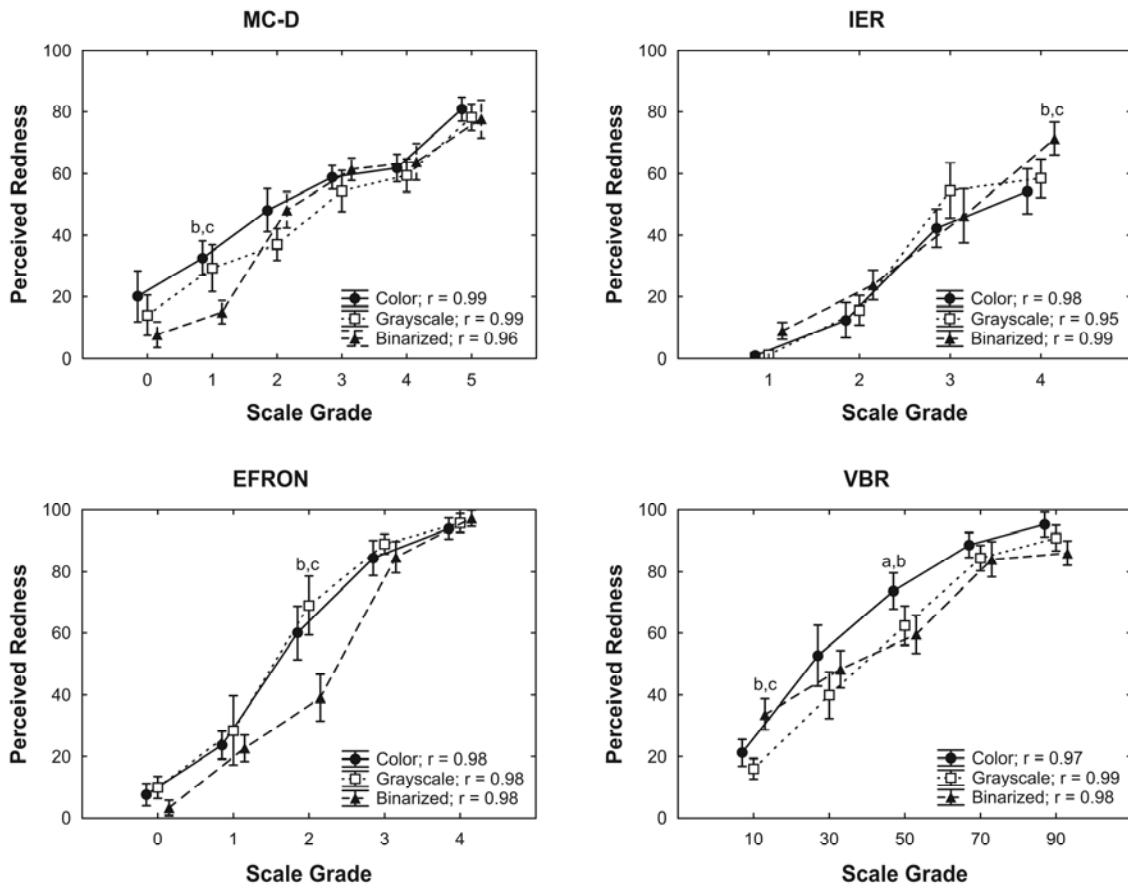


Figure 5-4: The effect of type of color information.

The effect of type of color information (color, grayscale, and binarized) on perception of redness severity for the MC-D, IER, VBR, and Efron scale (clockwise, top left to bottom left); error bars denote 95% confidence intervals. The scale levels that were perceptually different between image sets are indicated by the letters a (for significant differences between color and grayscale), b (color and binarized), and c (grayscale and binarized).

RM ANOVA and Tukey HSD were used to determine if the type of color information (color, grayscale, and binarized) affected the perception of redness in the reference images; for each scale, there was a statistically significant interaction between image set and scale levels (Figure 5-4; all $p \leq 0.0002$). The scale levels that were perceptually different between image sets are indicated by the letters *a* (for

significant differences between color and grayscale), *b* (color and binarized), and *c* (grayscale and binarized). In general, statistically significant differences were between the binarized image set and the color or grayscale image sets.

There were strong linear associations between each combination of physical redness attribute and perceptual scaling data (Table 5-2). These data suggest that spatial characteristics of redness (*D* and % *PC*, representing complexity and area, respectively) are slightly better correlated to redness perception than chromaticity (*u'*).

Table 5-2: Pearson correlation matrix between perceived and physical redness; all $p < 0.001$.

Pearson's r	<i>D</i>	% <i>PC</i>	<i>u'</i>
Color	0.92	0.90	0.90
Grayscale	0.92	0.90	0.83
Binarized	0.95	0.95	0.84

To investigate whether the physical redness attributes were related, partial correlation coefficients were calculated. Partial correlation coefficients are used to determine the individual contribution of a predictor variable (in this case, the objective metric corresponding to area, complexity, or color) to the correlation with the criterion variable (here: the perceptual data), while the other predictor variable(s) are controlled for.³⁴ As an example, by controlling for the spatial characteristics, *D* and % *PC*, the individual contribution of chromaticity, *u'*, to the correlation with the perceptual scores can be determined. Table 5-3 shows the partial correlation coefficients for the perceptual scaling data (color image set) and

each physical redness characteristic when controlling for one (3A) or two (3B) of the physical redness attributes. The first column indicates which predictor variable(s) was/were controlled for, and columns 2-4 indicate the partial correlation coefficients between the perceptual scores for the color image set and each of the three physical metrics.

Table 5-3: Partial correlation coefficients.

Partial correlation coefficients for the combination of perceptual redness scores and physical redness attributes (D = fractal dimension; % PC = % pixel coverage; u' = chromaticity) when one predictor variable is controlled for (A) and when two predictor variables (B) are controlled for. Bonferroni corrected significant correlations are marked by *.

A	Color vs. D	Color vs. % PC	Color vs. u'
u'	0.76*	0.70*	controlled
% PC	0.52	controlled	0.73*
D	controlled	0.23	0.70*
B	Color vs. D	Color vs. % PC	Color vs. u'
% PC, u'	0.44	controlled	controlled
D, u'	controlled	0.19	controlled
$D, \% PC$	controlled	controlled	0.69*

Multiple regression analysis was used to determine which combination of physical redness attributes provided the best correlate to subjective redness scaling for the color image set (Table 5-4).

Table 5-4: Multiple regression analysis.

Correlations between redness perception and combinations of physical redness attributes (Multiple R; all $p < 0.001$).

Color vs.	<i>D</i> & % <i>PC</i>	% <i>PC</i> & <i>u'</i>	<i>D</i> & <i>u'</i>	<i>D</i>, % <i>PC</i> & <i>u'</i>
Multiple R	0.925	0.952	0.960	0.962

5.5 Discussion

The purpose of this study was to evaluate the perceptual relationship between the reference images of four bulbar redness grading scales so that scale levels within and between scales could be compared, and to relate the perceptual estimates to physical redness characteristics⁵ in order to evaluate which strategies may likely be applied in the subjective assessment of redness.

Previous studies have shown that inter-observer agreement for subjective grading is often limited^{2,9,18-21}, and that variability between observers can extend over more than half of the available scale range for a single image.¹⁹ The variability between observers for this study followed a similar trend (Figure 5-4), although to a lesser extent, with maximum variation between any two observers for the same image ranging from 3.8 (IER 1, color image set) to 50.5 (Efron 2, grayscale image set).

Repeatability for averaged perceptual scores was very high (Table 5-1), independent of the image set. Levels for COR were < 10 for each image set, a variation that is fairly low considering the 100-point range of possible redness scores. There was almost perfect concordance for the averaged perceptual scores

obtained in the two sessions, as expressed by CCCs of at least 0.98 for each image set. Because a period of four to six weeks had elapsed between scaling sessions, it is rather unlikely that a recollection of image positions was reason for the high repeatability found in this study. This allows the assumption that, at least for this group of observers, another repetition of the scaling experiment is likely to reproduce almost identical averaged perceptual scores.

To evaluate differences between scale steps, averaged perceived redness for each reference image was compared within each scale, and between the scales. Figure 5-4 shows that the perceived redness associated to each reference image increased according to the assigned reference levels for each scale, with statistically significant differences between scale levels within each scale (RM ANOVA, all $p \leq 0.0001$). Some pairs of consecutive reference images were not perceived to be different, however. For the example of the color image set, this was found for the reference images representing grade 3 and 4 of the MC-D scale and grade 70 and 90 of the VBR scale. The finding that consecutive reference images were perceived to have similar content was not limited to the color image set only and was found within each scale at least once. There were strong linear associations, as expressed by Pearson correlation coefficients of at least $r=0.95$ ($p < 0.05$), between scale steps and perceptual reference image positions for each scale with each of the three image sets (Figure 5-4).

Figure 5-3 allows visualization of the relative position of the scale reference images on the 0 to 100 range and the qualitative assessment of relative redness between reference levels of different scales. Differences in the range of redness covered by the reference images of each scale are also obvious in this figure. The

reference images are fairly equally distributed along the full 0 to 100 range for the Efron and MC-D scale (although not extending to both extremes for the latter), but the IER scale appears to employ images that are more useful for lower levels of redness severity and does not provide reference images for more severe degrees of redness. Murphy et al⁹ have used the IER scale, interpolated to 0.1 increments, to investigate the prevalence of bulbar redness in non-contact lens wearers. In their sample of 121 eyes, they determined a range of redness severity between 1.2 and 2.9 (mean 1.93). They suggested that the dynamic range of the scale may be shifted towards the lower end, and that it may need to be extended to values greater than grade 4. In addition, they pointed out that the particularly white appearance of the reference image IER 1 may illustrate unusually low conjunctival redness, as no eye in their data was graded less than grade 1.⁹ In this study, the IER reference images were perceived to cover a range between 1 and 54, indicating that only reference images for the lower half of the 0 to 100 range are provided in this scale. The IER 1 reference image was perceived by all participants to have the lowest amount of redness out of all 20 reference images, resulting in an averaged perceptual score of 1 for the 0 to 100 scale (Figure 5-3). Bearing in mind that the recommendations for the use of the IER scale state that an eye rated greater than grade 2 has to be considered abnormal³⁵, and considering that IER 2 corresponded to an averaged perceptual severity of 12 for the 0 to 100 range, the findings of Murphy's and this study suggest that the IER scale employs images for which the reference grades overestimate the degree of redness severity that is actually shown in the images.

The reference images of the VBR scale were perceived to have higher redness levels than previously reported in a validation experiment.¹⁰ When the VBR scale was developed, the reference level for each image was determined in a similar psychophysical scaling experiment as was used in this study, and the severity levels were subsequently validated by comparison to the chromaticity (CIE u') of these images. Considering the almost perfect association between the validated scale levels and chromaticity (Pearson's $r=0.99$, $p<0.001$)^{5,10}, the higher perceptual scores found in this study were not expected. An explanation for this finding might be the (purposely) limited instructions that were given to the participants as only the minimum and maximum redness severity levels were indicated, or by an effort of the participants to use the complete 0 to 100 range. As Figure 5-3 shows, the reference images VBR 50, 70, and 90 were perceived to be more severely red than most of the images of the other scales, suggesting that this might have been an effect of the different severity ranges between the scales. Therefore, it appears that a combination of these factors might have caused the shift to higher scores for all of the VBR reference images, inducing a ceiling effect. This might partially account for why the reference images representing grade 70 and 90 were not being scaled to be different. A study is currently in progress to further investigate this finding.

Figure 5-4 shows that the type of color information had a significant impact on the assessments of redness severity, with post-hoc Tukey paired t-tests that showed significant differences in particular for the binarized image set when compared to the color or grayscale image set. The main difference between the binarized images and the color and grayscale image sets is based on the removal of information that is not directly attributable to vessel or background (sclera)

information during image processing. Pixel ‘color’ in the binarized images was assigned based on an automated thresholding procedure, where pixels that were part of blood vessels and which were found to be below the threshold level were assigned to be black, and all other pixels to be white, thus corresponding to the background.⁵ The differences between perceived redness for the three image sets imply that the additional redness information visible in the inter-vascular spaces of the color and grayscale images is likely used in the assessment of redness.

In clinical practice redness is assessed either while the patient is seen, or afterward using photographs that were captured during the eye examination. It appears that using vessel information only may have a positive impact on the reliability of assessments, as the removal of background information in the binarized image set resulted in the highest repeatability between assessments and the lowest variability between observers (Table 5-1 and Figure 5-4). It needs to be borne in mind, however, that good repeatability, although part of measurement, is not the only arbiter: Accuracy needs to be considered as well.

When the averaged perceived redness was compared for the color and grayscale image set, only one of the 20 reference images was found to be perceptually and statistically different (Figure 5-4), with the mean difference being 1.1 ± 6.8 (s_d) for the 100-point range. The finding that redness assessments using color or grayscale information yield very similar results is in agreement with Papas’ result²⁶ that for a group of experienced clinicians, redness grades for color and grayscale versions of the same eyes were highly correlated. This finding is interesting for clinical purposes, and may prove especially helpful for research studies. When photographs of the patient’s eyes are captured during the clinical

examination so that they can be used for later image analysis or reference, the similar redness estimates despite removing the color component suggest that these images may be stored in grayscale format. Since an equivalent color (RGB) image requires 24 bits (3 bytes) per pixel, whereas a grayscale image only uses 8 bits (1 byte) per pixel of computer memory, this will reduce storage space by a factor of 3.^{36,37}

To evaluate the processes that are involved in assessments of redness, the perceived redness for the reference images was compared to three physical attributes of redness: fractal dimension (D), % pixel coverage (% PC), and photometric chromaticity (u').⁵ D and % PC both are physical attributes that describe redness based on vessel characteristics: D is a measure of complexity, i.e. how the vessels fill up the space, and % PC can be used to describe how much area is covered by the vessels. Photometric chromaticity, u' , is a physical attribute of redness that relates to color/luminance differences. Chromaticity for the previously analyzed digital versions of the reference images⁵ was significantly lower than for the printed images (two-tailed paired t-test: $t=20.73$, $p<0.001$), but showed a very strong linear association (Pearson's $r=0.99$, $p<0.05$). Therefore, the perceived redness for the reference images was compared to u' of the printed images as those were the images used in the scaling experiment.

Independent of the type of color information used for the scaling experiment, all physical redness attributes showed high levels of linear association to the scaling data (Table 5-2). The highest correlation levels between subjective and objective redness estimates were found for the binarized image set in combination with D or % PC ; in general, spatial (vessel) criteria were found to be

more closely related to the averaged perceived redness for each of the image sets. Based on these data only, it appeared that vessel related criteria were mainly used in the assessment of redness, with chromaticity being less of a factor. For this reason we examined partial correlation coefficients for the relationship of perceived and physical redness. The partial correlation levels in Table 5-3A and Table 5-3B show that, as opposed to the results for simple Pearson correlation coefficients (Table 5-2), chromaticity has the better association with redness perception, with partial correlation levels of at least $r=0.69$ when either one or both spatial characteristics are controlled for. The spatial characteristics, on the other hand, appear to be closely related to each other, with a partial correlation coefficient of $r=0.84$ when controlling for u' . Their strength of linear association to redness scaling is significantly reduced when either of them is removed from the equation, however (Table 5-3). Multiple regression analysis showed that subjective scaling is best correlated to a combination of all three objective metrics (area, vessel complexity, and chromaticity; Table 5-4). However, leaving out the area component (% PC) does seem to only marginally affect subjective redness scaling; if the correlation levels are rounded to using only two decimals, no difference between two or all three objective metrics would be evident. Based on these results, redness assessments are likely based on a combination of two criteria: Chromaticity (color/luminance differences) on the one hand, and a vessel component based on complexity and area on the other hand.

These findings are in agreement with other studies where objective and subjective estimates of redness were compared. Fieguth and Simpson¹⁹ used a computer algorithm to automatically estimate bulbar redness, and suggested that a

combination of a color and an edge (vessel) feature best predicted clinical assessments of redness. Recently, Peterson and Wolffsohn²² used objective measures of redness that were based on edge detection and relative color extraction as correlates to subjective grades obtained by means of the IER and Efron scales. In agreement with Fieguth and Simpson's and the findings of this study, the best correlation between subjective grades and objective measures was found for the combination of both edge detection and relative color extraction features. When the edge detection and color feature were treated separately, the color component provided better agreement with the subjective grades, similar to the findings of this study. In a study where different objective parameters were derived to be compared to subjective grading when using the IER scale, Papas²⁶ found the opposite trend, suggesting that vessel area was better correlated to subjective redness assessments than any color parameter he investigated. A possible explanation for this might be that the images that were graded in Papas' study had been selected to represent a similar range as used in the IER scale, which, as the results of this study have shown, has a considerably shorter range and lower severity than the other scales. This would suggest that vessel features are more suitable for lower degrees of redness, which would be in agreement with Fieguth and Simpson's study who have reported a similar trend.¹⁹

It might be argued that the decision to select naïve participants as raters instead of experienced clinicians or contact lens specialists represented a potential disadvantage of this experiment. It has been shown that psychophysical scaling of redness was not different for groups of different levels of experience¹⁰, but in this experiment, explicitly images from (potentially) recognizable scales were used.

Since clinicians' ratings might be biased by this recognition, naïve participants were selected for redness scaling.

In summary, we have shown that a group of naïve participants is able to repeatedly scale perceived redness using the reference images of four clinical redness grading scales. The positioning of these reference images is not equal for each grading scale, even though the scales are generally designed to cover the same range of clinical redness. The perceived redness is strongly associated with the physical attributes of the images in each grading scale; based on these findings, redness is likely assessed by using both chromaticity *and* the vessel branching and coverage of the examined area of the conjunctiva. In addition, psychophysical scaling of perceived redness lends itself to being used to cross calibrate these four clinical scales.

5.6 Acknowledgments

The authors would like to thank Dr. Charles McMonnies, Dr. Nathan Efron, and the International Association of Contact Lens Educators (IACLE) for providing us with high resolution copies of the original reference images.

In the next chapter, the psychophysical scaling procedure will be slightly modified by providing the VBR reference images as additional anchors, and the perceived redness from anchored scaling will be used to cross-calibrate the grading scales.

6 The Conversion of Bulbar Redness Grades using Psychophysical Scaling

This chapter is published as follows:

Schulze MM, Hutchings N, Simpson TL. The Conversion of Bulbar Redness Grades using Psychophysical Scaling. *Optom Vis Sci.* 2010;87(3): in press.

Reprinted with permission. © The American Academy of Optometry 2010

	Concept / Design	Recruitment	Acquisition of data	Analysis	Write-up / publication
Schulze	Y	Y	Y	Y	Y
Hutchings	Y			Y	Y
Simpson	Y			Y	Y

Table detailing role of each author in this publication (Y denotes significant contribution)

6.1 Overview

Purpose: To use psychophysical scaling to investigate if the inclusion of reference anchors affected the perceived redness of the reference images of four bulbar redness grading scales, and to convert grades between scales.

Methods: Ten naïve participants were asked to arrange printed copies of the McMonnies/Chapman-Davies (6), IER (4), and Efron (5) grading scale images relative to each other, using the stationary but unlabeled 10, 30, 50, 70 & 90 reference images of the Validated Bulbar Redness (VBR) scale as additional anchors within a given 0 (minimum) to 100 (maximum) redness range (anchored scaling). The position of each image was averaged across observers to represent its perceived redness within this range. Anchored scaling data were then compared to data from a previous study, where the images of all four grading scales had been scaled for the same experimental setup, but with no reference anchors provided (non-anchored scaling). Averaged perceived redness as determined with anchored scaling was used to cross-calibrate grades between scales.

Results: Overall, perceived redness of the reference images was significantly different within each scale (RM ANOVA, all scales $p < 0.001$). There were differences in perceived redness range and when comparing reference levels between scales. Anchored scaling resulted in an apparent shift to lower perceived redness for all but one reference image compared to non-anchored scaling, with the rank order of the 20 images for both procedures remaining fairly constant (Spearman's $\rho = 0.99$).

Conclusions: The re-scaling of the reference images in the anchored scaling experiment suggests that redness was assessed based on within-scale

characteristics and not using absolute redness scores, a mechanism that can be referred to as clinical scale constancy. The perceived redness data allow practitioners to modify the grades of the scale they commonly use for comparison of their grading estimates to grades obtained with another calibrated scale.

6.2 Introduction

Grading scales are common tools in clinical practice and research settings, and are used to aid with the routine assessment of contact lens wear related changes to ocular structures like the cornea or conjunctiva.¹⁻¹⁴ A number of bulbar redness scales exist^{1,3,10,15-18}, and the selection to use a particular scale may depend on personal preference, but also on local availability or scale awareness. There seems to be some geographical association linked to the selection of a grading scale to be used, as a recent study¹⁹ has shown that Sickenberger's Kontaktlinsen Klassifizierungsschlüssel¹⁷ (i.e. contact lens classification system) is better known and more frequently used in Germany than the widely known Efron^{3,4} scale. The decision in favour of a particular scale might also be based on the severity range the scale is able to cover, if the scale has been validated^{10,20}, or if a grading system combines scales for different contact lens complications.^{3,18}

It has been recommended that grading scales are not interchanged^{6,12,20-22}, as it has been shown that the scales have unequal scale steps and differ with regard to the number of steps, the severity range displayed, or the way the condition is presented.^{12,14,20} However, practitioners generally prefer a particular scale, so that the use of different scales in the ophthalmic field is inevitable. In some situations the ability to convert between grades obtained with different grading scales would be desirable for the practitioner, for example if a paper contained results utilizing a different scale than an individual used, in multi-center research studies, or if a patient case requires the attention of multiple practitioners at different sites.

Independent of the scale being used in clinical practice, it is crucial for the practitioner to be able to detect small and possibly clinically significant changes during the assessment of ocular conditions.^{23,24} The ability to detect change is linked to the coarseness or fineness of the scale, and finer scales have been recommended because of their greater sensitivity to change.^{23,24} Since the illustrative scales used in clinical practice commonly consist of four to five reference images^{6,8,9,24-26}, interpolation of the grades associated to the images, either by decimalization for 0 to 4 or 0 to 5 scales or by the use of integers for 0 to 100 scales, has been recommended to translate them into finer scales.^{23,24} The interpolation (and extrapolation, if required) of grades has also been advised in the instructions for the application of grading scales^{3,10,27}, and is routinely applied in clinical research today.^{5,8,20,24,28}

The incentive for this study were results from a recent study¹⁴, in which the perceptual relationship between reference images of the MC-D (6), IER (4), Efron (5), and VBR (5) bulbar redness grading scales was investigated with the intention of attempting a conversion of grades between scales. In brief, ten naïve observers had used printed versions of the slightly modified reference images that only showed vascular detail but no lids or lashes (Figure 6-1) to psychophysically scale redness. Participants were not aware that the images were part of grading scales, and the only indicators for redness were labels of 0 and 100 at the start and end point of a 1.5m range, corresponding to minimum and maximum redness, respectively. After completion of the arrangement, the position of each image should reflect its perceived redness within the 0 to 100 range relative to the other reference images.¹⁴

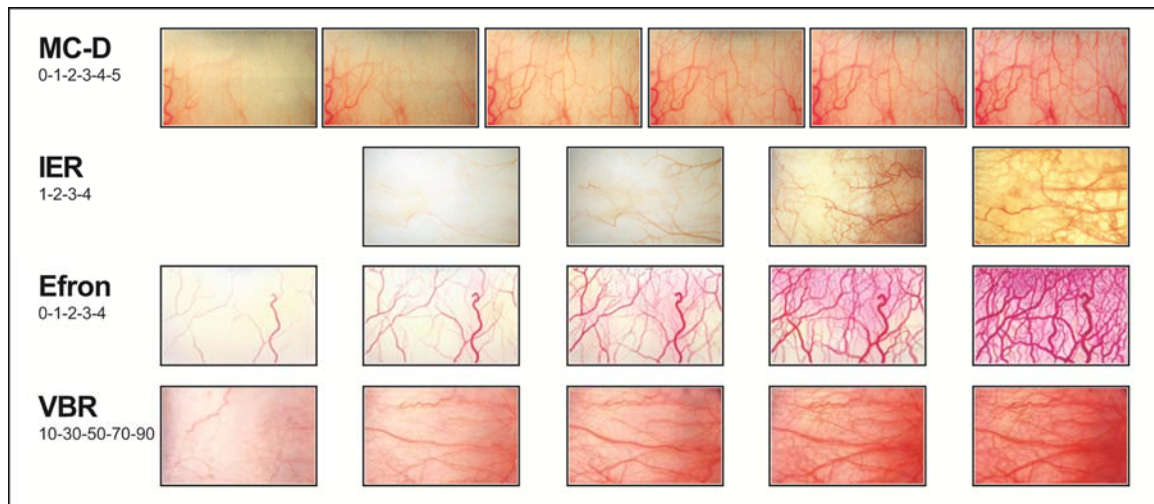


Figure 6-1: The modified reference images of the MC-D, IER, Efron, and VBR scales.

An interesting finding of this study¹⁴ was that the color versions of the reference images of the VBR scale were perceived to have higher redness severity levels than found in a previous validation experiment.¹⁰ This finding might be explained by a possible effort of the participants to use the full 0 to 100 range when scaling redness, with the effect that the generally less red reference images of the MC-D and IER scale were being placed towards the middle, and the more severely red reference images of the Efron and VBR scale towards the end of the redness range.

The purpose of this study was to further investigate this finding, and to evaluate if and how the inclusion of the VBR reference images as additional reference anchors would impact the perceptual relationship between the reference images of the four scales. Based on these findings, the perceived redness of the reference images was used to attempt a conversion of scale grades between the scales.

6.3 Methods

6.3.1 Psychophysical scaling

The methods of this study followed closely the previously described protocol.¹⁴ The identical experimental settings were used, with the only exception being the additional presentation of the VBR scale reference images as intermediate anchors. Scaling was done within the same 1.5m range on a table top, with the start and end point being labeled with 0 and 100 corresponding to minimum and maximum redness, respectively. The experiment took place in normal, full room illumination with cool white fluorescent lighting with a general color rendering index (CRI) of 84. Illumination settings were obtained with a light meter (DVM 1300; Velleman®, Gavere, Belgium), were consistent over time, and ranged from 350 to 390 lux across the table top on which scaling was performed. The same observers that had participated in the original study¹⁴ were asked to scale redness again; except for having scaled the reference images before, the participants were not experienced in the clinical application of grading scales, and were unaware that the images were part of grading 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, ON, Canada; informed consent was obtained from each participant prior to starting the study. To discriminate between the previous¹⁴ and current scaling procedure in this manuscript, they will be referred to as non-anchored and anchored redness scaling, respectively.

The anchored positions of the VBR reference images within the 0 to 100 range corresponded to their previously validated¹⁰ reference grades at 10, 30, 50,

70, and 90, but no numerical indicators were given to quantify their redness. The modified reference images of the MC-D, IER, and Efron scale were randomly spread out on the table top, and the participants were given verbal instructions asking them to place the images within the 0 to 100 range, so that their position represented their perceived redness relative to each other, but also to the stationary VBR reference images. To represent perceived redness, images could be placed separately, slightly overlapping (if they were perceived to be fairly similar), or at the same position (if they were perceived to have the same redness). Figure 6-2 shows the experimental setup before scaling was started (A), and a typical arrangement after completion of redness scaling (B). Perceived redness for each image was taken as its measured position with respect to 0, and averaged across observers. A second session was conducted about six weeks later to evaluate repeatability.

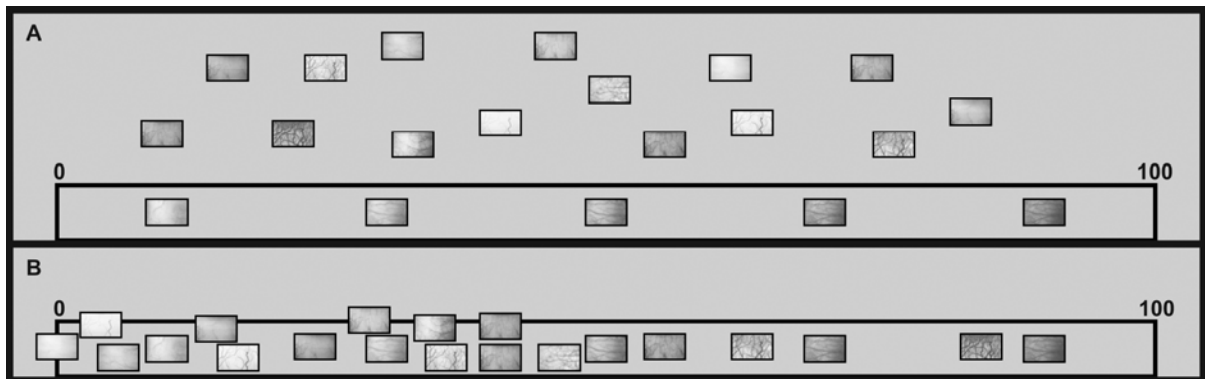


Figure 6-2: Experimental setup for anchored redness scaling.

Shown are the randomly spread images before arrangement and the VBR scale reference images as (unlabeled) intermediate anchors (A), and a typical arrangement after completion (B).

6.3.2 Conversion of grades between scales

For better sensitivity to detecting change, the interpolation of scale grades in clinical research settings is typically done by dividing the interval between scale grades into equal increments, e.g. into 0.1 steps.^{5,8,20,24,28} To comply with this strategy, the MC-D, IER, and Efron scale grades were divided into ten equal increments (i.e. to 1 decimal place). This was achieved by linear interpolation of the perceived redness interval (averaged across all observers) between consecutive scale grades and by dividing this distance (in perceived redness units) by 10 which corresponded to their respective decimalized scale grades.

To allow comparison of grading estimates across scales, base scale grades (SG_b) were converted to target scale grades (SG_t) using the following equation:

Equation 1

$$SG_t = \left(\frac{(PR_b - PR_{TL})}{(PR_{TH} - PR_{TL})} \right) + SG_{TL}$$

where PR_b is the perceived redness of the base scale image; PR_{TL} and PR_{TH} are the perceived redness of the lower and higher target scale image enclosing the base scale image within the 100-point perceived redness range, respectively; and SG_{TL} is the lower scale grade of the two target scale images enclosing the base scale image.

For each of the 100 integer scale increments of the VBR scale, the corresponding decimalized target scale grades (i.e. equivalent decimal step of MC-D, IER and Efron) were calculated, where applicable. To further illustrate the

conversion of scale grades between scales, a sample conversion - using the perceived redness data - is given in the results section.

6.3.3 Data Analysis

Statistical analysis and curve fits were done using STATISTICA v8 (StatSoft. Inc., Tulsa, OK, USA) and Sigmaplot v10 (Systat Software Inc., San Jose, CA, USA), respectively; an alpha level of ≤ 0.05 was considered statistically significant. After verifying normality of the outcome variable (Kolmogorov-Smirnov test), repeated measures analysis of variance (RM ANOVA) with post-hoc Tukey HSD's were used to evaluate differences between scale levels within each scale. Averaged perceived redness for non-anchored¹⁴ vs. anchored redness scaling was analyzed using RM ANOVA with post-hoc Tukey HSD's, and using Spearman's rank order coefficient (Spearman's ρ). The correlation coefficient of concordance (CCC)²⁹ and the coefficient of repeatability (COR; $1.96 * sd$)^{6,30} were used to evaluate repeatability between sessions.

6.4 Results

For each observer, scaling of perceived redness required about 5 to 8 minutes. Averaged scaling data were highly repeatable across observers, with almost perfect concordance (CCC=0.99) and a COR of 6 for the 100 point range.

Overall, there were statistically significant differences between the scale levels within each scale (RM ANOVA; all scales $F > 37$, $p < 0.0001$). However, not all consecutive scale levels were perceived to be different for the IER and the MC-D scales, as indicated by arrows on the abscissa between the scale grades in Figure 6-3 (Tukey HSD; $p > 0.05$).

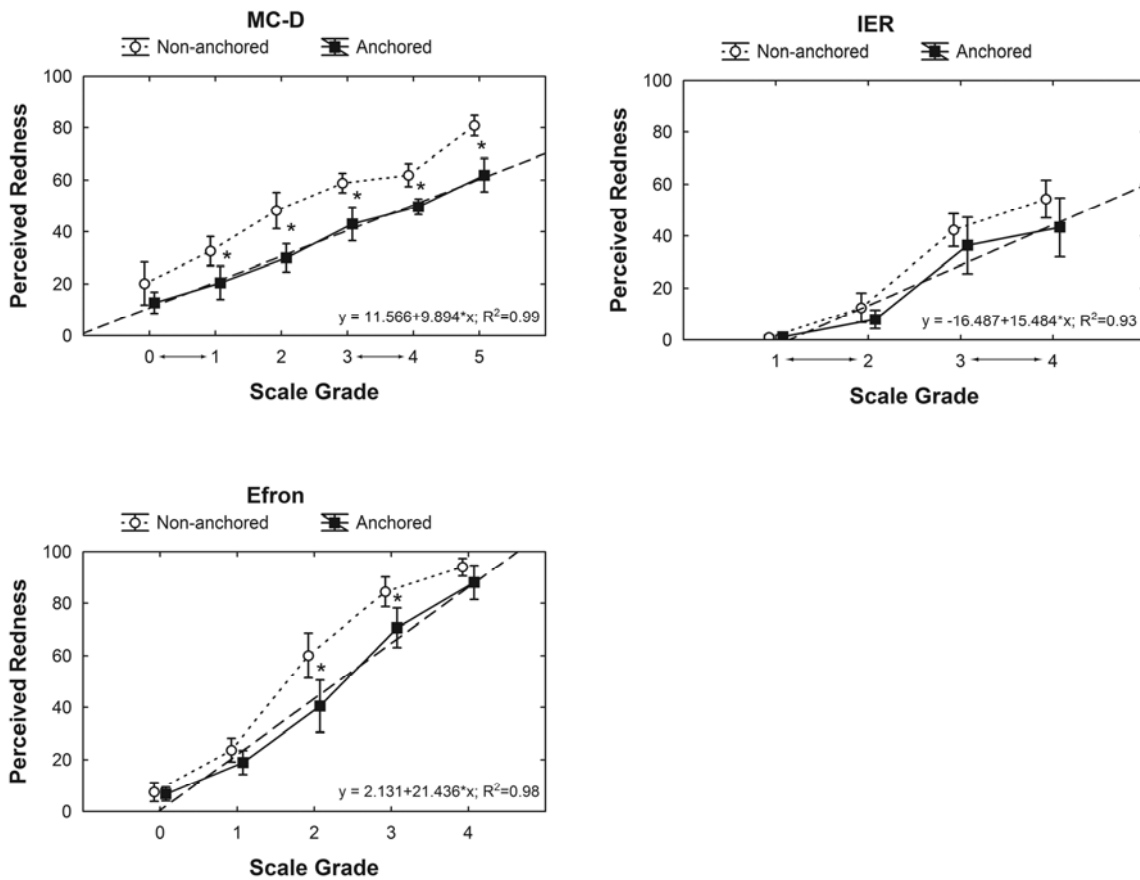


Figure 6-3: Shift to lower perceived redness with anchored (filled squares) compared to non-anchored scaling (open circles).

Vertical bars denote 95% confidence intervals. Images perceived to have significantly lower redness are indicated by *. Arrows between the scale grades (x-axis) indicate consecutive scale images that were perceived to have similar redness. The dashed line indicates the best linear fit for the anchored scaling data.

The use of the VBR reference images as intermediate anchors resulted in an apparent shift to lower perceptual scores (Table 6-1 and Figure 6-3). All images except IER 1 were perceived to be less red, with significantly lower redness for seven of these images (Figure 6-3; indicated by *). The differences between non-anchored and anchored scaling were only significant for the Efron scale ($p < 0.0001$),

but were not significant for the MC-D ($p=0.05$) and IER scale ($p=0.17$). Note: RM ANOVA was only applicable for the MC-D, IER, and Efron scale, as the VBR reference images as stationary anchors exhibited no variance.

Table 6-1: The perceived redness of each reference image for non-anchored (top) or anchored (bottom) psychophysical scaling.

Data are shown according to their order from the anchored scaling experiment; * indicates significant differences between anchored and non-anchored scaling.

	IER 1	EFRON 0	IER 2	VBR 10	MC-D 0	EFRON 1	MC-D 1 *	MC-D 2 *	VBR 30	IER 3	EFRON 2 *	MC-D 3 *	IER 4	MC-D 4 *	VBR 50	MC-D 5 *	VBR 70	EFRON 3 *	EFRON 4	VBR 90
Non-anchored	1	8	12	21	20	24	33	48	53	42	60	59	54	62	74	81	89	84	94	95
Anchored	1	7	8	10	13	19	20	30	30	36	41	43	43	50	50	62	70	71	88	90

The rank order of the images (based on increasing perceived redness) was very similar for the non-anchored¹⁴ and anchored scaling experiments, as expressed by Spearman's $\rho=0.99$ ($p<0.001$).

Because of the difference in perceived redness for the VBR reference images observed with non-anchored scaling¹⁴ compared to their chromatically validated scale grades (that were also determined by psychophysical scaling¹⁰), we evaluated the shift in perceived redness by plotting one against the other. This yielded an equation for the relationship between the objectively validated redness grades (i.e. 10, 30, 50, 70, and 90)¹⁰ and the perceived redness data from non-anchored scaling (Equation 2).

Equation 2

$$VBR \text{ scale grade} = 5.823 e^{0.029x}$$

where x is the perceived redness for the VBR reference images as experimentally obtained from non-anchored scaling of all 20 reference images from the four scales.¹⁴ *VBR scale grade* is the validated redness grade obtained from the development¹⁰ of the VBR scale. Note that this relationship is derived only using the VBR scale images.

Figure 6-4 shows the relationship between experimentally obtained perceived redness for non-anchored¹⁴ and anchored scaling for the MC-D, IER, and Efron scales. A reference line, representing the shift in perceived redness observed for the VBR scale, is shown (dotted line) for comparison to the shift in perceived redness between anchored and non-anchored scaling for each of the other scales (filled squares with solid fit line). If the shift in perceived redness for the MC-D, IER, and Efron scale is similar to the shift observed for the VBR scale, the solid and dotted line should closely match and be parallel.

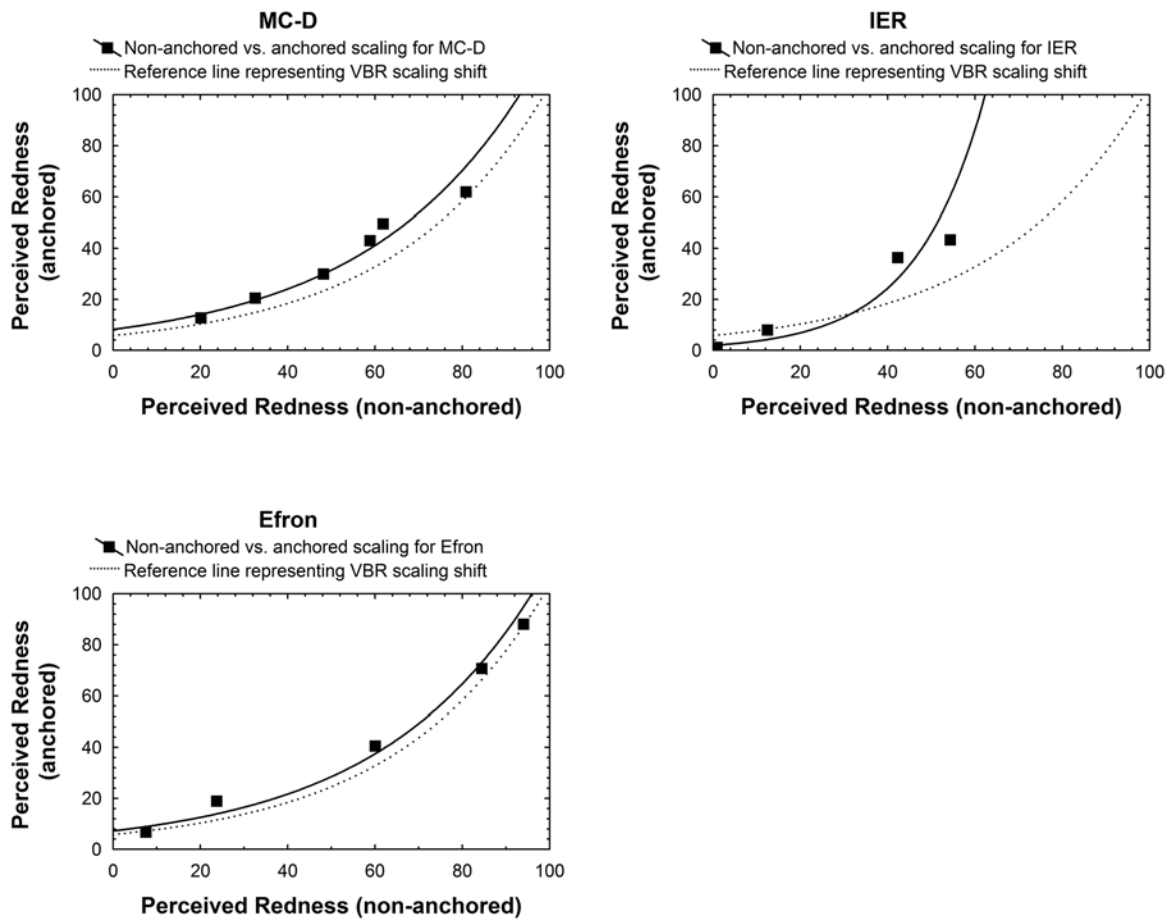


Figure 6-4: Non-anchored vs. anchored redness scaling (solid fit line) for the MC-D, IER, and Efron scales.

The dotted line represents the shift in perceived redness observed for the VBR scale and is shown for reference only. If the perceived redness for the MC-D, IER, and Efron scale shifted as expected (i.e. as observed for the VBR scale), both lines should closely match and be parallel.

Figure 6-5 shows the results of the anchored scaling experiment for comparison between scales. The image positions correspond to their averaged perceived redness within the 0 to 100 range. Since scaling of the reference images was done relative to the VBR anchors within the 0 to 100 redness range, the averaged perceived redness for each reference image is equivalent to its VBR scale

grade, and therefore demonstrates the relative perceived redness between scales. The perceptual order of the reference images within each scale was in agreement with the respective scale grades.

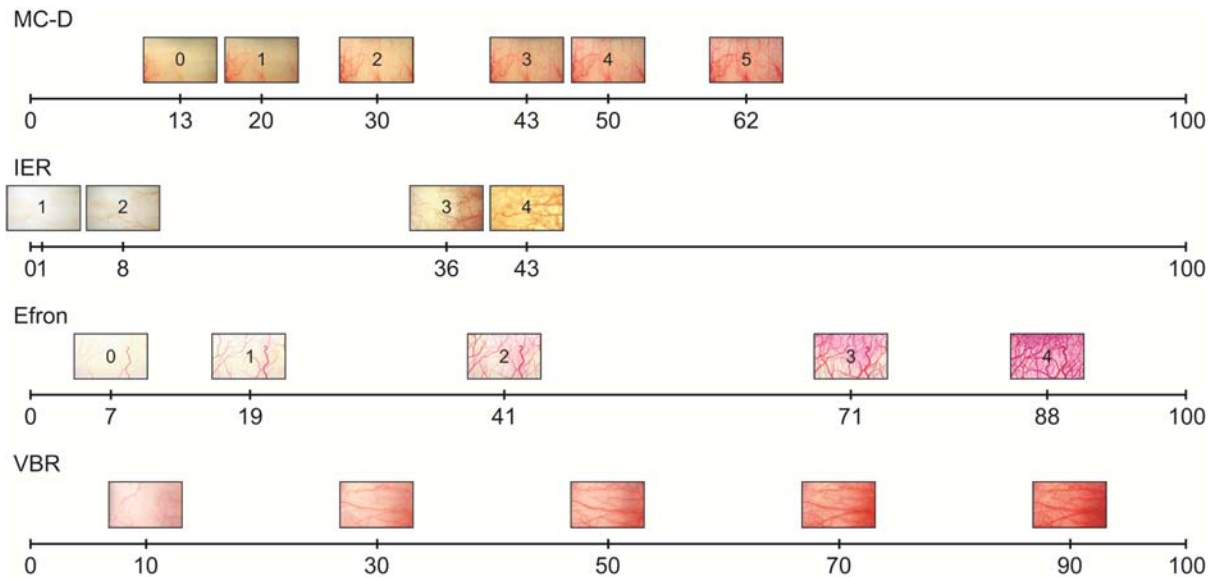


Figure 6-5: Perceptually scaled reference images within the 0 to 100 range for the anchored scaling experiment.

Perceptually scaled reference images of the MC-D, IER, Efron, and VBR scale (top to bottom) within the 0 to 100 range for the anchored scaling experiment. The image positions correspond to the averaged perceived redness for each image. Original scale grades are indicated at the centre of each reference image.

Table 6-2 shows the base scale grades for the reference images (gray boxes) and their corresponding target scale grades after conversion (Equation 1). Bold signifies the decimalized scale steps within the MC-D, IER and Efron scale that were used as base scale increments for conversion to target scale grades.

Table 6-2: Conversion table between scales.

Conversion table showing the integer (gray boxes) and decimalized (bold) base scale grades and their corresponding target scale grades after conversion based on the perceived redness of each scale's reference images (anchored scaling data). Bold signifies the decimalized scale steps within the MC-D, IER and Efron scale that were used as base scale steps for conversion.

MC-D	IER	Efron	VBR	MC-D	IER	Efron	VBR
--	<1	--	0	4.0	>4	2.3	50
--	1.0	--	1	4.1	>4	2.3	51
--	1.1	--	2	4.2	>4	2.4	52
--	1.3	--	3	4.3	>4	2.4	53
--	1.4	--	4	4.3	>4	2.4	54
--	1.6	--	5	4.4	>4	2.5	55
--	1.7	--	6	4.5	>4	2.5	56
--	1.9	0.0	7	4.6	>4	2.5	57
--	2.0	0.1	8	4.7	>4	2.6	58
--	2.0	0.2	9	4.8	>4	2.6	59
--	2.1	0.3	10	4.8	>4	2.6	60
--	2.1	0.3	11	4.9	>4	2.7	61
--	2.1	0.4	12	5.0	>4	2.7	62
0.0	2.2	0.5	13	>5	>4	2.7	63
0.1	2.2	0.6	14	>5	>4	2.8	64
0.3	2.3	0.7	15	>5	>4	2.8	65
0.4	2.3	0.8	16	>5	>4	2.8	66
0.6	2.3	0.8	17	>5	>4	2.9	67
0.7	2.4	0.9	18	>5	>4	2.9	68
0.9	2.4	1.0	19	>5	>4	2.9	69
1.0	2.4	1.0	20	>5	>4	3.0	70
1.1	2.5	1.1	21	>5	>4	3.0	71
1.2	2.5	1.1	22	>5	>4	3.1	72
1.3	2.5	1.2	23	>5	>4	3.1	73
1.4	2.6	1.2	24	>5	>4	3.2	74
1.5	2.6	1.3	25	>5	>4	3.2	75
1.6	2.6	1.3	26	>5	>4	3.3	76
1.7	2.7	1.4	27	>5	>4	3.4	77
1.8	2.7	1.4	28	>5	>4	3.4	78
1.9	2.8	1.5	29	>5	>4	3.5	79
2.0	2.8	1.5	30	>5	>4	3.5	80
2.1	2.8	1.5	31	>5	>4	3.6	81
2.2	2.9	1.6	32	>5	>4	3.6	82
2.2	2.9	1.6	33	>5	>4	3.7	83
2.3	2.9	1.7	34	>5	>4	3.8	84
2.4	3.0	1.7	35	>5	>4	3.8	85
2.5	3.0	1.8	36	>5	>4	3.9	86
2.5	3.1	1.8	37	>5	>4	3.9	87
2.6	3.3	1.9	38	>5	>4	4.0	88
2.7	3.4	1.9	39	>5	>4	>4	89
2.8	3.6	2.0	40	>5	>4	>4	90
2.8	3.7	2.0	41	>5	>4	>4	91
2.9	3.9	2.0	42	>5	>4	>4	92
3.0	4.0	2.1	43	>5	>4	>4	93
3.1	>4	2.1	44	>5	>4	>4	94
3.3	>4	2.1	45	>5	>4	>4	95
3.4	>4	2.2	46	>5	>4	>4	96
3.6	>4	2.2	47	>5	>4	>4	97
3.7	>4	2.2	48	>5	>4	>4	98
3.9	>4	2.3	49	>5	>4	>4	99
4.0	>4	2.3	50	>5	>4	>4	100

As an example, the base scale grade IER 2.5 was linearly interpolated to obtain its perceived redness of 22, and then converted to an Efron (target) scale grade of 1.1:

$$SG_{Efron} = \left(\frac{(PR_B - PR_{TL})}{(PR_{TH} - PR_{TL})} \right) + SG_{TL} = \left(\frac{(PR_{IER\ 2.5} - PR_{Efron\ 1})}{(PR_{Efron\ 2} - PR_{Efron\ 1})} \right) + SG_{Efron\ 1} = \left(\frac{(22 - 19)}{(41 - 19)} \right) + 1 = 1.1$$

6.5 Discussion

The purpose of this study was to investigate if the inclusion of reference anchors in a psychophysical scaling experiment affected the perceptual relationship between reference images. Based on these findings, a method could then be developed to convert scale grades between grading scales based on the averaged perceived redness of the reference images.

The selection of the VBR images as reference images for psychophysical scaling was based on two factors. First, their scale grades had been determined by estimating their perceived redness in a similar psychophysical scaling experiment¹⁰ for the same 0 to 100 range as used in this experiment. Second, the validity of the VBR scale grades had been demonstrated by their high linear association to an objective metric, photometric chromaticity¹⁰, as opposed to the other scales for which image selection had been based on subjective estimation and clinical experience only.^{1,4,14} The high linear association indicates equality of scale steps along the full scale range, which is a characteristic of at least interval scale level measurement.^{5,12,31} Therefore, the scaling of the reference images relative to the VBR

images, and to each other, provided perceived redness as common measurement unit for each of the reference images and allowed measurement and conversion of scale grades at the interval level.

Repeatability of the averaged psychophysical scaling data was very high, with anchored scaling having marginally superior repeatability than non-anchored scaling.¹⁴ The high levels of linear association between perceived data and objective metrics¹⁴ and the high levels of repeatability for psychophysical scaling found in both experiments suggest that psychophysical scaling represents a robust methodical approach for the measurement of visual appearance³² that allows the comparison of scale levels on a common measurement scale.

Overall, there were statistically significant differences between scale levels within each scale (RM ANOVA; all scales $p < 0.0001$), however, both the IER and MC-D scale were found to employ pairs of consecutive reference images that were not perceived to be different (Figure 6-3; arrows). This has particular impact for the IER scale, as this suggests that the four reference images/levels appear to correspond perceptually to two levels only. There were also differences regarding the perceived redness range covered by each scale (Figure 6-5). While the VBR and Efron scale have reference images that cover almost the complete 0 to 100 redness range, the reference images of the MC-D and the IER scale appear to better assist in the assessment of complications of lower severity, as their severity range is capped at 62 and 43, respectively. There were also differences in perceived redness when comparing reference levels across scales, as the perceived redness e.g. for reference level 2 differs largely between the MC-D (30), IER (8), and Efron (41) scale. The IER scale is intended to be used for the management of the severity range

typically seen with contact lens complications⁵, which might partially account for the rather low levels of redness displayed in the scale. However, this study corroborates that the scale grades of the IER scale overestimate the degree of redness that is actually depicted in its images (the perceived redness of IER 1 and IER 2 is 1 and 8, respectively)^{8,14}, and explains why Efron²⁰ found that bulbar redness grades assessed using the IER scale were 0.6 higher on average than with the Efron scale.

6.5.1 Non-anchored vs. anchored scaling

The primary purpose of this study was to investigate if the inclusion of the VBR reference images as anchors would have an impact on the perception of redness compared to scaling when no anchors were provided. There was an apparent shift to lower perceived redness for the anchored scaling experiment (Table 6-1 and Figure 6-3), which supports the hypothesis that the participants in the non-anchored scaling experiment¹⁴ made an attempt to use the full 0 to 100 range when estimating the redness of the reference images. By doing so, they shifted scale images that were perceived to be more red (Efron and VBR) to the top end of the range closer to 100 (representing maximum redness), while most reference images of the IER and, in part, of the MC-D scale were perceived to be comparatively less red and thus placed towards the low end of the range at 0 (representing minimum redness). The inclusion of the VBR images as stationary reference anchors resulted in a re-scaling of redness severity, with the consequence that the perceived redness associated to the reference images decreased (Figure 6-3 and Table 6-1), while the rank order of the images was found to only be marginally affected (Spearman's $\rho=0.99$). By providing stationary anchors it

appeared that the participants were able to re-scale redness relative to within (VBR) scale redness characteristics, despite large physical differences between the images of the scales.¹²

This re-scaling of redness is similar to another perceptual mechanism known as color constancy.³³ Human observers are able to recognize the color of objects irrespective of the light used to illuminate them, as their visual system is able to balance out these illumination differences. In the context of grading scales, this balancing effect occurred as a re-scaling of redness severity according to the additional information provided by the stationary anchors, a mechanism that we have referred to as clinical scale constancy.¹² Independent of the scale being used and despite the physical differences in the reference images of different scales, an eye that displays a low amount of redness will be assigned to a low grade relative to the redness characteristics within the particular scale being used.

Despite the apparent shift to lower perceived redness for all reference images except IER 1, the differences between non-anchored and anchored scaling (Figure 6-3) were only significant for the Efron scale, showed a trend towards significance for the MC-D Scale (due to a floor effect for MC-D 0 with anchored scaling), and were not significant for the IER scale. A closer look at the graph for the IER scale and Table 6-1 shows that there was almost no change in perceived redness independent if anchors were provided or not, and that in both cases the dynamic range was the smallest of all scales. The Efron scale images on the other hand were placed towards the high end of the redness range when no anchors were provided (similar to the VBR scale images), as their images were perceived to being redder than most images in the IER and MC-D scales. When the VBR scale images

were placed at their validated scale positions for anchored scaling, the Efron images shifted in an almost identical way relative to the VBR scale, as expressed by the almost perfect match between the dotted reference line (VBR) and the experimentally obtained perceived redness shift for the Efron scale (Figure 6-4). A very similar trend was found for the MC-D scale. However, the shift for the IER scale differed strongly from the shift in perceived redness for the VBR scale (reference line), likely because of its short dynamic range with only two perceptually different levels contributing to this model.

6.5.2 Conversion of grades between scales

Increasing the incremental steps of grading scales in the hope of improving the ability to detect small but significant change is usually done by interpolation between scale grades.^{23,24} In addition to using finer and/or more comprehensive scales, it would also be beneficial to compare or convert grades obtained with different grading scales. In this study, the averaged perceived redness for the reference images was used to convert reference grades and their respective scale increments between scales.

Table 6-2 shows the interpolated scale grades for the MC-D, IER, and Efron scales (10 decimalized steps between grades) and the VBR scale (100 integer steps). All scales were referenced to the 100-point perceived redness range of the VBR scale because it represents the only scale with sufficient increments to completely cover the redness range of all scales in addition to having a known relationship¹⁰ between increments. The duplicate representation of some decimalized scale steps is because these decimalized steps cover multiple integer steps, and also due to the

inequality of perceived redness intervals between consecutive scale steps within the MC-D, IER, and Efron scales (Figure 6-3). This inequality of perceived redness intervals also explains the absence of a few decimalized scale steps; if the perceived redness interval between consecutive scale steps was too small (e.g. 7 between IER 1 and 2), not all ten possible decimalized scale steps could be represented. However, the absence of decimalized scale steps only occurred between consecutive scale grades that were not perceptually distinguishable (Figure 6-3; arrows).

The inequality of perceived redness intervals associated to the scale steps was also the reason why linear regression equations were not used for the conversion of scale grades. While the association between scale grades and perceived redness was almost perfectly linear for the MC-D scale, the IER scale was better described by a sigmoidal rather than a linear fit for the full scale range (Figure 6-3). For completeness, the sigmoidal best fit line would have the equation $y=43.783/(1+\exp(-(x-2.484)/0.526))$ [$R^2=1.00$; $p<0.0001$]. A similar, although smaller trend towards a better fit by using a sigmoidal function was found for the Efron scale. For both the Efron and IER scale, the application of a linear fit would result in a misrepresentation of the experimentally determined perceived redness data, as the linear fit line progresses outside the 95% confidence interval for the IER 1, IER 2, Efron 0, and Efron 1 reference images. Thus, the use of linear regression equations is not applicable for the conversion of scale grades.

The approach described here uses subjective estimation of redness, as opposed to physical metrics, to convert between scale grades. Two reasons support this decision. First, a conversion of scale grades based on physical metrics of

redness, although technically feasible, cannot be reasonably achieved as differences between the physical measures of each scale are too severe to allow for such a conversion.¹² With these differences in mind, an even more fundamental reason for our decision can be found in clinical practice itself: As grading scales are used for the subjective assessment of a patient's eye, a conversion based on perceptual rather than physical data also appears the more appropriate and clinically relevant approach.

The perceived redness data in Table 6-1 (anchored) and Table 6-2 lend themselves to being used in the clinical application of grading scales and in clinical research, particularly in multi-center studies. By assigning the experimentally derived perceived redness to the reference images instead of using the actual scale grades, practitioners are able to improve the sensitivity and range of their preferred scale beyond the scale grades available, and allow comparison between grading estimates obtained with a different calibrated scale.

In summary, we have shown that the scaling of redness when using reference images as stationary anchors results in a shift in redness perception and in a re-scaling of redness severity. The re-scaling of redness severity had an impact on the actual (perceived redness) grade associated to the reference image, but the order and the perception of relative differences remained fairly constant. This suggests that despite the physical differences between the grading scales, and independent of the scale being used, it appears that practitioners are able to ignore absolute redness characteristics while using relative (within-scale) information to come to an appropriate clinical conclusion. This ability to re-scale redness based on within-scale characteristics is what we refer to as clinical scale constancy.¹²

The approach taken in this paper provides clinicians with a novel method to modify their scale of choice by applying the perceived redness to the reference images instead of using the original scale grades. This serves not only the purpose of increasing the sensitivity to detecting change, but also assists in the comparison of grades between the available redness scales. As grading estimates are done relative to within-scale redness characteristics (clinical scale constancy), it seems likely that the use of scales with calibrated grades might provide less variable redness estimates if different scales are used in multi-center studies.

6.6 Acknowledgements

The authors would like to thank Dr. Charles McMonnies, Dr. Nathan Efron, and the International Association of Contact Lens Educators (IACLE) for providing us with high resolution copies of the original reference images.

The between-scale agreement of grading estimates will be evaluated for the newly calibrated grading scales in chapter 7. Fractal dimension, % pixel coverage and chromaticity will be determined as physical attributes of redness, and compared to the subjective redness estimates.

7 Grading Bulbar Redness Using Cross-Calibrated Clinical Grading Scales

7.1 Overview

Purpose: To grade bulbar redness using cross-calibrated versions of the McMonnies/Chapman-Davies (MC-D), Institute for Eye Research (IER), Efron, and Validated Bulbar Redness (VBR) grading scales.

Methods: Modified reference images (5x3cm, showing only vascular detail) of each grading scale were distributed on a desk, one grading scale at a time. The positions of the reference images within each scale was determined in a previous psychophysical scaling experiment and corresponded to their perceived redness on a 0 to 100 range; the upper limit of the dynamic range for each scale was 43 (IER), 62 (MC-D), 88 (Efron), and 90 (VBR). 10 naïve 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. The order of scale presentation was randomized. 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 (RM ANOVA; $p=0.0008$); 6 of the 16 images had a perceived redness that was significantly different between at least two of the scales. Between-scale correlation coefficients of concordance (CCC) ranged from 0.93 (IER vs. Efron) to 0.98 (VBR vs. Efron). Between-scale coefficients of repeatability (COR) ranged

from 5 units (IER vs. VBR) to 8 units (IER vs. Efron) of the 0 to 100 range. There was a trend towards higher grades for less red images and lower grades for more red images when using MC-D and IER scales than for the Efron and VBR scales.

Conclusions: The use of cross-calibrated reference grades for bulbar redness grading scales allows comparison between scales. Perceived redness is dependent upon the dynamic range of the reference images of the scale.

7.2 Introduction

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) scale², the Efron scale that uses artist-rendered illustrations^{3,4}, and the validated bulbar redness (VBR) scale with objectively validated reference levels.⁵ Despite the advantages that illustrative grading scales possess over verbal descriptions or the use of purely descriptive scales^{1,6-8}, their use is still problematic. Typically, variability of grading estimates has been attributed to the subjectivity associated with the clinical application of grading scales, with differences between observers or for the same observer over time.⁸⁻¹⁴

Grading estimates have also been 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 supported by Peterson and Wolffsohn.¹⁵ There are apparent differences between the four scales in the number of reference images, the scale levels and range, and the conjunctival region displayed (Figure 4-2). Objective techniques have been used to quantify these differences in the scale images for various physical redness characteristics¹⁶⁻¹⁸, and have supported 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 exhibit the same degree of redness in another scale^{19,20}).

Because of these reasons it has been suggested that scales not be interchanged, and that grading estimates not be compared across scales.^{9,12,17} The ability to convert redness grades obtained with different grading scales would be valuable for clinical practice, 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 model^{21,22} that allowed the perceived redness of the 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 metrics¹⁶⁻¹⁸, the perceived redness data also indicated a misalignment of the reference images between scales, and pointed to different dynamic ranges (i.e. the range that is covered by the reference images) of the scales. Table 7-1 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

Table 7-1: Calibrated scale grades from psychophysical scaling experiment.²²

MC-D		IER		Efron		VBR
Original	Calibrated	Original	Calibrated	Original	Calibrated	Original
0	13			0	7	10
1	20	1	1	1	19	30
2	30	2	8	2	41	50
3	43	3	36	3	71	70
4	50	4	43	4	88	90
5	62					

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, one purpose of this study was to determine the between-scale agreement of the newly calibrated MC-D, IER, Efron, and VBR scales in order to estimate the reproducibility of redness estimates for a set of 16 sample images.

The inconsistency in grading has also been attributed to a lack of information about the criteria that are used for clinical decision making.^{10,13} To address this question, various studies have been conducted in order to establish objectively measurable parameters of redness that correlate well with subjective grading estimates.^{10,13,15,24-28} In general, because morphometric and/or chromatic information were used to objectively quantify redness in these studies, this suggests that at least two strategies may be involved in the clinical assessment of redness: One being based on colour/luminance differences, and one by judging the appearance of the vessels.¹³

Papas¹⁰ used image processing techniques to investigate which individual, objectively measurable characteristics were best associated to clinical estimates of bulbar redness. Subjective estimates were best described by two morphometric parameters, number of vessels and percentage area covered by vessels, while all colour-based parameters showed only poor to moderate linear associations to redness grading. Therefore he concluded that redness estimates appeared to be essentially based on vascular information.¹⁰ Fieguth and Simpson¹³ and Peterson

and Wolffsohn¹⁵ on the other hand have reported that a combination of colour and vessel information was best suitable to describe subjective redness estimates. A comparison of the results from psychophysical scaling²¹ and physical redness¹⁸ also suggests the latter, with the perceived redness of the reference images being best represented by chromaticity (CIE u') and a combination of vessel-based characteristics.²¹ In addition, it was shown that assessments of vessel appearance were not solely based on estimating the area covered by vessels (% PC), but also by judging the degree of vessel-branching.²¹ Fractal analysis has been used to quantify the degree of conjunctival vascular branching in terms of their fractal dimension (D), a measure that has been previously used to estimate the complexity of the retinal vasculature.²⁹⁻³²

The clinical applicability of fractal dimension as an objective metric to describe conjunctival redness has not been investigated yet, and was the second focus of this study. Thus, the subjective grading estimates for the 16 sample images were compared to D , % PC , and CIE u' , in order to determine which of these metrics was best associated with the clinical assessments of redness.

7.3 Methods

7.3.1 Sample images

Sample images were selected from a database of photographs available at the Centre for Contact Lens Research (CCLR). Previously collected redness data (subjective grading estimates and photometric chromaticity, CIE u') were analyzed in order to select sample images that represented a contact lens population that

might typically be seen in clinical practice; 16 sample images were selected for which chromaticity data and redness grades were normally distributed.

In order to allow subjective and objective redness estimates from the same images, regions of interest in a size of 250x156 pixels were cropped out of the sample images, so that only conjunctival vascular detail (without lids or lashes) was visible.¹⁸ This corresponded to an area of approximately 1.1x0.69cm 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 perceived to be the least red of all sample images (independent of the scale and for all physical metrics), can be seen in Figure 7-1.

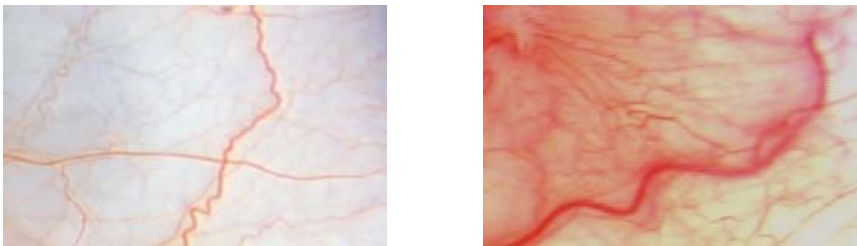


Figure 7-1: The sample images perceived to be the least (left) and most red (right).

7.3.2 Subjective redness assessments

Redness was subjectively estimated on a table top within the same 1.5m range that had been used for the psychophysical scaling^{21,22} and the subsequent calibration²² of the reference images of the MC-D, IER, Efron, and VBR scale. The start and end point of this 1.5m range corresponded to the minimum and

maximum redness level, and were labelled by 0 and 100, respectively. Within this range, the modified reference images of each scale were presented so that their position matched their calibrated reference grades (Table 7-1).²²

The same ten participants who had participated in the previous redness scaling experiments^{21,22} were asked to estimate perceived bulbar redness of printed colour copies of the sample images (5x3cm) by placing them relative to the unlabelled reference images of one of the four scales. After the placement of each image, its position was measured, and the image was 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, ON, Canada; informed consent was obtained from each participant prior to starting the study.

7.3.3 Objective redness measurements

The bulbar redness in the 16 sample images was objectively quantified by image processing and spectrophotometry. Image processing was used to compute two vessel-based characteristics of redness, % pixel coverage (% *PC*) and fractal dimension (*D*), and spectrophotometry was used to derive a colourimetric quantity, photometric chromaticity (CIE u').

7.3.3.1 Image processing and fractal analysis

A pre-processing macro was written for the public domain Java image processing program ImageJ 1.38x³³ to produce binarized versions of the sample images that consisted of black and white pixels only, representing vessels or background, respectively. First, the RGB sample images were split into their three colour channels red, green, and blue, and the channel that provided the highest vessel signal-to-noise ratio (green) was used for further pre-processing (Figure 7-2i & Figure 7-2ii). A median filter with a 3x3 kernel was used to reduce the background noise in the sample images before the contrast between vessels and the white scleral background was enhanced using histogram equalization (Figure 7-2iii), a non-linear mapping technique that re-assigns the intensity values of pixels in the original image in order to achieve a more uniform distribution of gray level intensities across the full range of the histogram.^{33,34} Therefore, images with a low dynamic range (gray level intensities that are clustered in a certain range of the histogram) such as pixels corresponding to a low contrast blood vessel against the scleral background, would have a wider dynamic range (improved contrast) and better visibility of detail (the vessel) after stretching. After histogram equalization, a Sobel edge detection algorithm³³ was applied to highlight vessel edges and small capillaries^{28,35}, and the background was subtracted to account for eyeball curvature (Figure 7-2iv & Figure 7-2v).²⁷ Image pre-processing was completed by binarizing each sample image using an automated (and thus observer-independent) thresholding procedure that assigned pixels corresponding to the scleral background to a gray level of 1, and pixels corresponding to vessels to 0 (Figure 7-2vi).

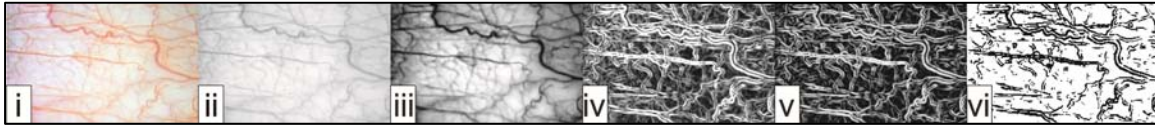


Figure 7-2: Image pre-processing steps for the sample images.

(i) original image; (ii) green channel; (iii) median filtering and contrast enhancement; (iv) edge detection; (v) background subtraction); (vi) final, binarized image.

Previously established standardized settings¹⁸ for the ImageJ plug-in FracLac (v. 2.5 Release 1b5i)^{33,36} were used to calculate % *PC* and *D* for the binarized sample images. % *PC* is a measure of the relative area covered by the vessels, and is computed as the ratio of the number of black pixels (representing vessels) to the overall number of all pixels in the image.^{10,17,18,24,25} FracLac calculates a number of fractal dimensions, *D*, of which the slope-corrected most-efficient covering fractal dimension was selected to quantify the complexity of the conjunctival vasculature in the sample images.^{18,36}

7.3.3.2 Photometric chromaticity measurements

The SpectraScan PR650 spectrophotometer (Photo Research Inc, Chatsworth, VA, USA) was used to measure photometric chromaticity, CIE u' , under controlled illumination settings. The spectrophotometer was mounted on a modified slit lamp stand, and a slide holder was attached to the chin rest to keep the sample images stationary during the measurements (Figure 7-3). The printed colour copies of the sample images that were also used for the subjective estimates of redness were placed in the slide holder, and u' was measured for nine regularly spaced locations (3x3 grid) and averaged to represent a single, global estimate of redness.

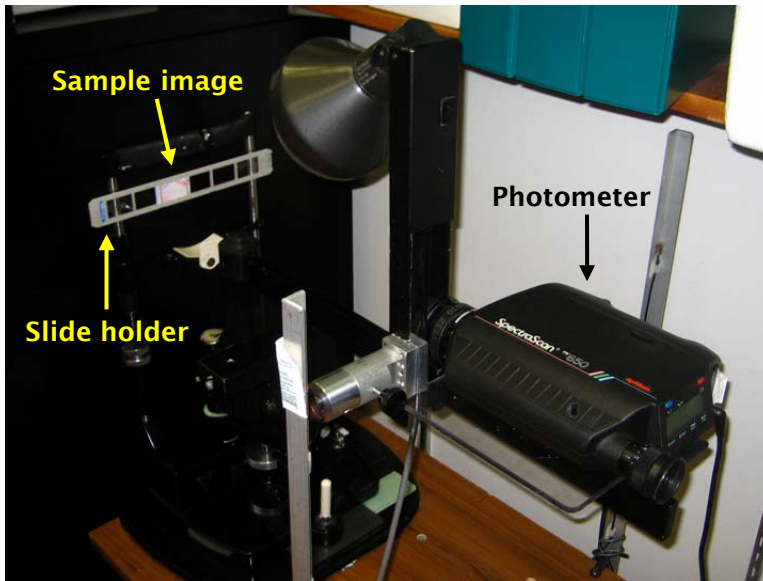


Figure 7-3: Setup for photometric measurement of the sample images.

Chromaticity was measured for a regular 3x3 grid in each sample image. The images were placed in a stationary slide holder that was attached to the chin rest of a modified slit lamp mount.

7.3.4 Data Analysis

Statistical analysis was performed using STATISTICA version 8 (StatSoft Inc., Tulsa, OK, USA); an alpha level of ≤ 0.05 was considered statistically significant. Repeated measures analysis of variance (RM ANOVA) was used to determine if the redness estimates for the sample images depended on the grading scale being used. Bonferroni corrected post hoc tests were used 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 redness agreement was evaluated with the intraclass correlation coefficient (ICC)³⁷⁻³⁹, the correlation coefficient of concordance (CCC)⁴⁰, the coefficient of repeatability (COR; $1.96 \cdot s_d$)^{6,9} and Bland-Altman's limits of agreement (LOA; $\bar{d} \pm \text{COR}$).⁴¹ The relationship between perceived redness and physical redness

was analyzed using Pearson's r , partial correlation coefficients and multiple regression analysis.

7.4 Results

Figure 7-4 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; $F=1.88$, $p<0.001$).

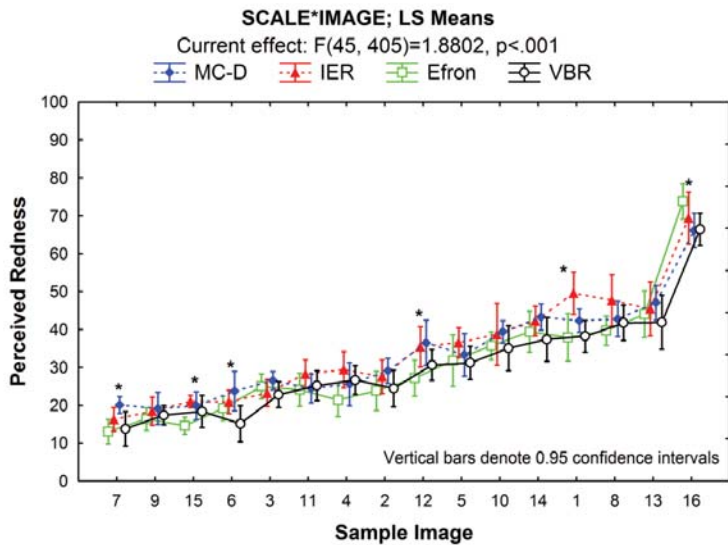


Figure 7-4: Grading estimates compared between scales.

* denotes significant differences between scales.

The mean redness estimates with each scale were 33.7 (MC-D), 34.4 (IER), and 30.4 for both Efron and VBR. There was a statistically significant difference between scales ($F=4.14$; $p=0.015$), with no significant difference between the MC-D and the IER scale and between the VBR and the Efron scale (Fisher LSD test, $p=0.64$ and $p=0.98$, respectively). Testing for simple effects, RM ANOVA showed statistically significant differences for 6 of the 16 sample images between at least

two of the scales (Figure 7-5; significant differences after Bonferroni corrected multiple comparisons are indicated by the respective scale names).

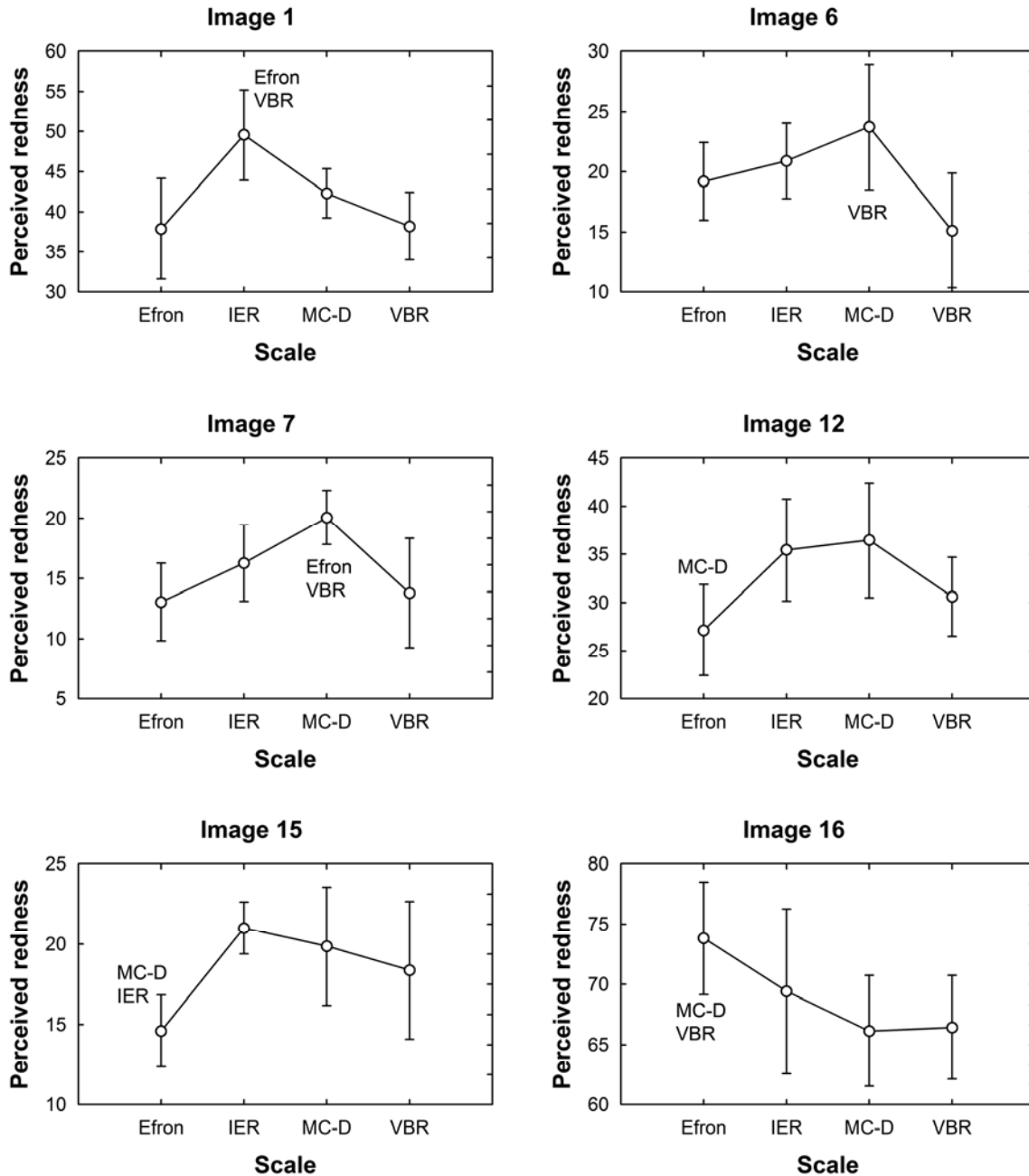


Figure 7-5: Images with significantly different grading estimates between scales.

Bonferroni corrected significant differences between scales are indicated by the respective scale names.

Table 7-2 shows the between-scale ICCs, CCCs, and the CORs for each pair of scales. The last column shows the mean of the differences (\bar{d}) between each pair of scales; the scale that produced higher grades is indicated next to the associated mean difference. There was a very strong linear association between grading estimates for each pair of scales (all Pearson's r 's=0.98 except IER vs. Efron [$r=0.96$]).

Table 7-2: ICC, CCC, COR, and mean of the differences for each pair of scales.

	ICC (2,10)	CCC	COR	\bar{d}
IER vs. MC-D	0.99	0.97	6.4	+0.7 (IER)
IER vs. Efron	0.96	0.93	7.8	+4.0 (IER)
IER vs. VBR	0.97	0.94	5.0	+4.0 (IER)
MC-D vs. Efron	0.97	0.94	7.1	+3.3 (MC-D)
MC-D vs. VBR	0.97	0.94	5.6	+3.3 (MC-D)
Efron vs. VBR	0.99	0.98	5.9	0.0

Figure 7-6 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.

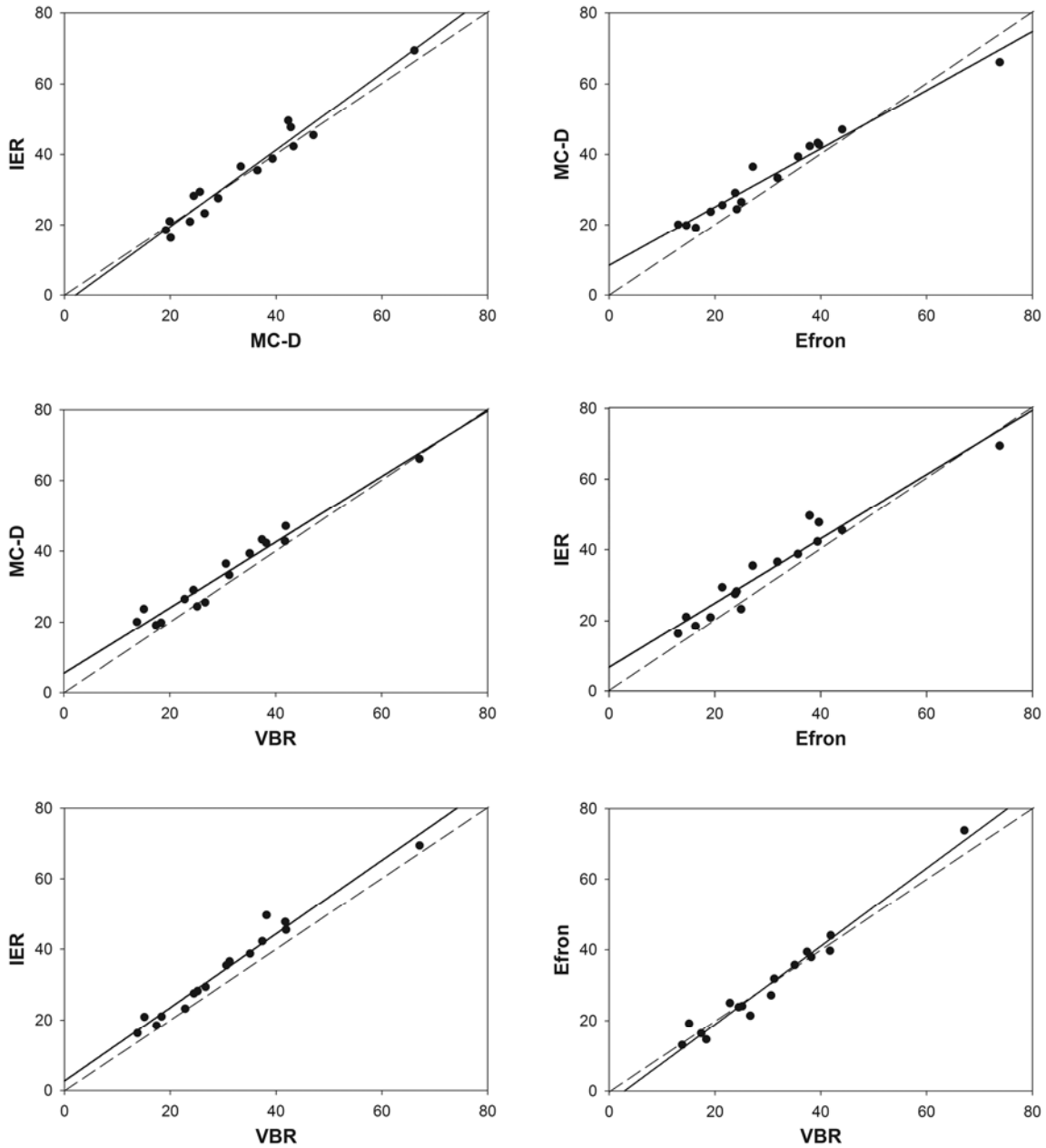


Figure 7-6: Between-scales concordance of grading estimates.

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 ($\bar{d} \pm \text{COR}$) are shown in Figure 7-7 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 ($\bar{d} \pm \text{COR}$).

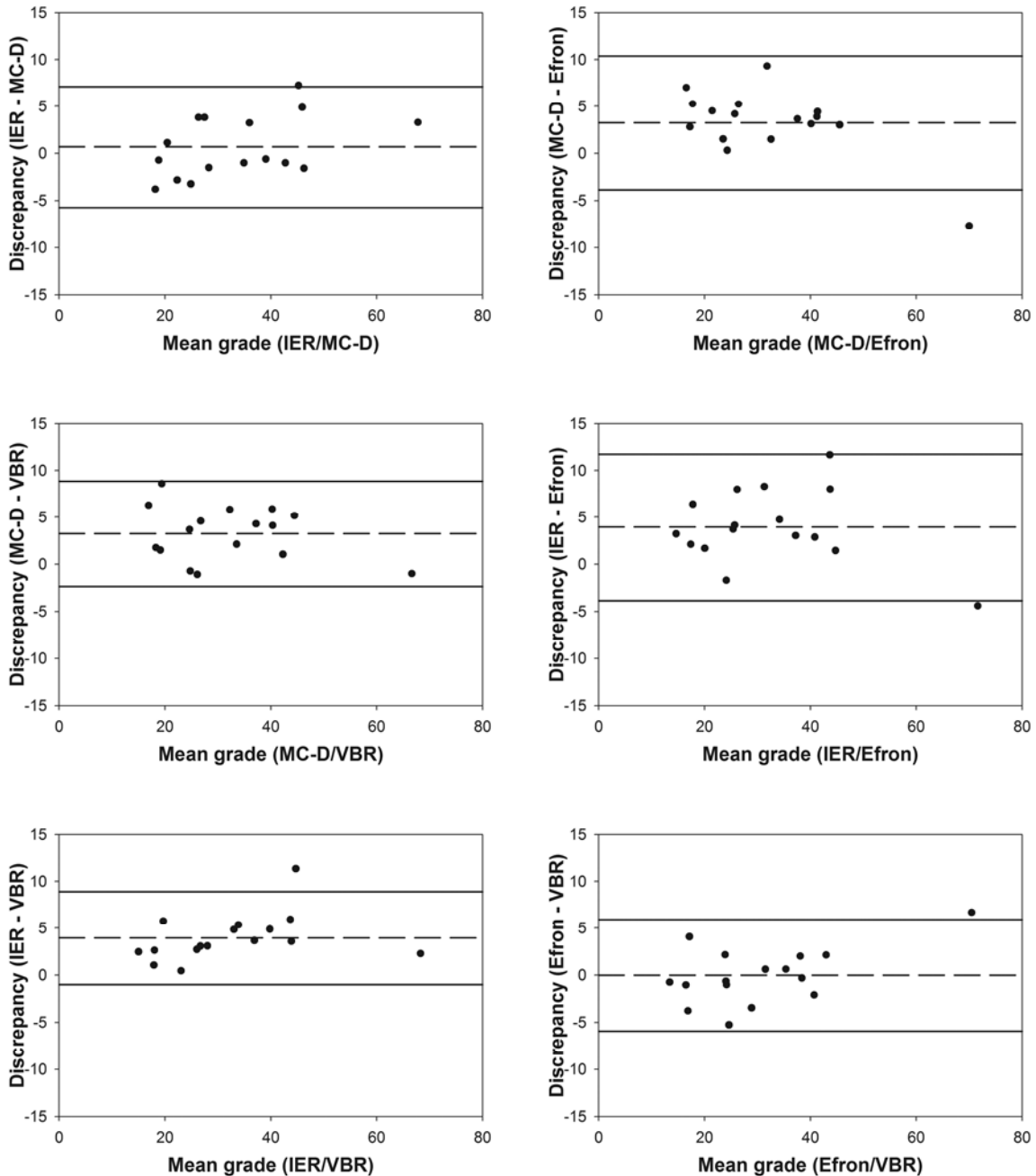


Figure 7-7: Between-scales limits of agreement (LOA).

The dashed line near zero represents the mean of the differences between the two scales; the solid lines show the limits of agreement ($\bar{d} \pm \text{COR}$).

For each of the scales, there were strong linear associations for each combination of grading estimates and physical redness attributes (Table 7-3). In general, grading estimates were best correlated to chromaticity, u' .

Table 7-3: Pearson correlation matrix.

Pearson correlation matrix for subjective grading estimates and physical redness characteristics; all $p < 0.05$.

	<i>D</i>	<i>% PC</i>	<i>u'</i>
MC-D	0.88	0.91	0.90
IER	0.86	0.87	0.90
Efron	0.87	0.90	0.95
VBR	0.86	0.88	0.94

Partial correlation coefficients were used to determine the individual contribution of the predictor variables (i.e. the physical redness attributes corresponding to pixel area, complexity, and colour) to the correlation with the criterion variable (graded redness with each scale), while the other predictor variable(s) are controlled for.⁴² Table 7-4 shows the partial correlation coefficients for the grading estimates and each physical redness attribute, separated for each scale. The top part of each table shows the partial correlation coefficients when one predictor variable (i.e. PRA) is controlled for, and the bottom part when the two other predictor variables are controlled for.

Table 7-4: Partial correlation coefficients.

Partial correlation coefficients for the combination of grading estimates and physical redness attributes (PRA). For each scale, the top part of the table shows the partial correlation coefficients when one predictor variable (i.e. PRA) is controlled for, and the bottom part when the two other predictor variables are controlled for.

Contr. for		MC-D vs.	
1 PRA	<i>D</i>	% <i>PC</i>	<i>u'</i>
	0.66*	0.71*	controlled
	-0.12	controlled	0.67*
	controlled	0.50	0.72*
2 PRAs	<i>D</i>	% <i>PC</i>	<i>u'</i>
	-0.05	0.35	0.67*

Contr. for		IER vs.	
1 PRA	<i>D</i>	% <i>PC</i>	<i>u'</i>
	0.59*	0.55*	controlled
	0.09	controlled	0.68*
	controlled	0.27	0.73*
2 PRAs	<i>D</i>	% <i>PC</i>	<i>u'</i>
	0.25	-0.02	0.70

Contr. for		Efron vs.	
1 PRA	<i>D</i>	% <i>PC</i>	<i>u'</i>
	0.71*	0.73*	controlled
	0.09	controlled	0.89*
	controlled	0.27	0.90*
2 PRAs	<i>D</i>	% <i>PC</i>	<i>u'</i>
	0.09	0.26	0.89*

Contr. for		VBR vs.	
1 PRA	<i>D</i>	% <i>PC</i>	<i>u'</i>
	0.64*	0.6*	controlled
	0.07	controlled	0.81*
	controlled	0.31	0.84*
2 PRAs	<i>D</i>	% <i>PC</i>	<i>u'</i>
	0.30	0.04	0.83*

PRA = Physical redness attribute; *D* = fractal dimension; % *PC* = % pixel coverage; *u'* = chromaticity.
 * = Bonferroni corrected significant correlations.

Stepwise multiple regression analysis was used to determine which combination of physical redness attributes provided the best correlate to the grading estimates for each of the scales (Table 7-5). Independent of the scale, the subjective grading estimates were best correlated with a combination of chromaticity and one of the two spatial attributes; adding the third physical redness attribute did not increase the correlation to subjective grading with either of the scales.

Table 7-5: Stepwise multiple regression analysis.

Correlations between physical redness attributes (PRA) and subjective grading estimates with each scale (Multiple R; all $p \leq 0.001$).

Grading estimates vs.	1 PRA	R	2 PRAs	R
MC-D	$\% PC$	0.91	$\% PC + u'$	0.95
IER	u'	0.89	$u' + D$	0.94
Efron	u'	0.95	$u' + \% PC$	0.98
VBR	u'	0.94	$u' + D$	0.96

7.5 Discussion

7.5.1 Agreement between scales

The first purpose of this study was to determine the between-scale agreement of 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 (Figure 7-4; RM ANOVA, $p < 0.001$). The perceived redness of six of the 16 images was significantly different between at least two of the four grading scales (Figure 7-5). In general, these sample images were found to

have different redness if assessed with the IER or MC-D scale compared to the Efron or VBR scale. Only image 16, the eye that was perceived to be the most red of all images, deviated from this trend. This finding suggests that redness estimates depend on the dynamic range of the scale being used (Table 7-1), as the scales having shorter dynamic ranges (IER and MC-D) generally resulted in higher redness estimates than the scales with wider dynamic ranges (Efron and VBR).

Despite these differences between single images, there was close agreement for the grading estimates between all scales (Table 7-2; Figure 7-6 and Figure 7-7). There were very high levels of linear association for each combination of scales (all Pearson's r 's ≥ 0.96). In this experiment, the ICC was used to quantify the reproducibility of grading estimates obtained with different scales. ICC (2,10) was selected since it estimates the agreement between assessments of a random sample of k raters (10) that can be generalized to other raters within some population, and represents an indicator of the interchangeability of the grading scales.³⁸ 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 1.00 representing perfect concordance.^{5,40} There was high concordance between grading estimates for each combination of scales, with levels of CCCs of at least 0.93. Figure 7-6 provides a pictorial representation of these relationships, 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 to the Efron and VBR scale appears to subside with increasing redness, as indicated by

the converging solid fit line towards 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.

The variability of grading estimates between any pair of scales was very low, as indicated by the between-scale CORs (Table 7-2) and LOAs (Figure 7-7). The COR describes the degree of scatter for repeated measurements on the same sample images. The COR is the standard deviation of the differences (s_d) between test and retest session for all measurands multiplied by 1.96 (i.e. $COR=1.96*s_d$). The LOAs are the limits of the 95% confidence interval (i.e. COR) that are plotted with respect to the mean of the differences between test and retest for all measurands (\bar{d}); the upper and lower LOA are calculated by $\bar{d}\pm 1.96*s_d$, respectively. The mean of the differences (\bar{d}) indicates if there is systematic bias in the grading estimates between scales. There was a small but systematic bias towards higher grades for scales with shorter dynamic range (MC-D & IER), while scales with similar dynamic range showed no such trend (Table 7-2 and Figure 7-7, dashed horizontal line). Overall, the between-scale CORs were small (indicating low variability and good repeatability) and ranged from 5 (IER vs. VBR) to 8 grading units (IER vs. 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 to 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 redness using different scales. Efron et al.¹² reported that the mean bulbar redness (across all observers) was about 0.6 grading units higher (for a 0 to 4 range) with the IER scale compared to the Efron scale for the same set of sample images. Proportionally, this means that grades were about 15% higher on average when the IER scale was used, whereas mean redness grades were only up to 4% different between any pair of the newly calibrated scales (Table 7-2). In general, CORs are often used to quantify the variability of grades for test/retest settings with a single scale^{5,9,10,12,43}, while for this study CORs were calculated to estimate the differences of grading estimates between scales. This complicates a direct comparison to other studies, however, it allows an estimation of how the variability between scales compares to the test/retest variability that has been reported. In this study, the between-scale CORs (Table 7-2) were found to be similar or even smaller than within-scale test/retest CORs that have been previously reported.^{9,10,12} Therefore, the calibration of the grading scales produces *closer* agreement between grading estimates using *different* scales than previously reported when the *same* scales were used^{12,15}; this implies that the newly calibrated grading scales may be used interchangeably. 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.

7.5.2 Physical redness attributes vs. subjective grading estimates

The second purpose of this study was to evaluate the relationship between the subjective grades and a number of physical attributes of redness, in particular fractal dimension.

Overall, there were strong linear associations between all physical attributes and the subjective redness estimates with each of the scales (Table 7-3; all Pearson's $r's \geq 0.86$). Subjective redness estimates were most closely related to chromaticity, u' , while the complexity of the conjunctival vasculature, as quantified by D , showed the weakest correlation to the subjective grades. Since an interrelation of the two spatial redness characteristics in the assessment of redness was suspected²¹, partial correlation coefficients were used to examine the individual contribution of each physical redness attribute to the correlation with the subjective grading estimates for each of the scales. When two of the three predictor variables were controlled for (Table 7-4; *contr. for 2 PRAs*), D and % PC were poorly correlated to subjective grading (all $r's \leq 0.35$), while u' was found to have the greatest association with the redness assessments (all $r's \geq 0.67$; all Bonferroni-corrected $p's < 0.05$). When either of the two spatial characteristics (D or % PC) was controlled for, the linear association of the other spatial metric to subjective grading was significantly reduced, as indicated by the partial correlation coefficients of $r \leq 0.50$ (Table 7-4, *contr. for 1 PRA*). A subsequent analysis of the objective metrics alone showed that % PC and D were highly associated, with a significant partial correlation coefficient of $r = 0.95$ when u' was controlled for.

Stepwise multiple regression showed that the grading estimates were best described by a combination of chromatic and vascular information (Table 7-5), which is in agreement with previous reports.^{13,15,21} However, the grading estimates with each of the scales were best described by combining u' with either D or $\% PC$, while the respective other spatial characteristic did not provide additional information. If we recall that controlling for one spatial characteristic reduced the association to the subjective grades for the other one (Table 7-4, contr. for 1 PRA), these findings seem to be somewhat contradictory. A possible explanation for this effect may be that redness assessments are mainly based on chromatic information, while judgments on the conjunctival vasculature appear to be more supplemental rather and might thus be sufficiently represented by one spatial characteristic only. In turn, the close agreement between D and $\% PC$ ($r=0.95$, see above) may also suggest that, despite describing different attributes of redness, the complexity and area of the conjunctival vasculature may be difficult to discern perceptually.

7.6 Conclusion

The newly calibrated grading scales were capable of producing highly reproducible redness estimates across scales. There were differences in 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 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 of the scales and it appears that using the newly calibrated grading scales might reduce the between-scale variability when subjectively estimating redness.

Independent of the scale being used, the redness estimates were highly associated to all physical attributes of redness. Chromaticity appears to be the main factor when redness is assessed, while *D* and % *PC* as physical analogues of vessel complexity and area were found to be closely related, but appeared to provide supplemental information only. Overall, however, a combination of chromatic *and* vessel-based information was found to best predict the subjective redness estimates for the observers in this study.

8 Conclusions and Future Work

Grading scales are the most commonly used assessment tool in clinical practice to monitor and manage changes to ocular structures and tissues. Despite being frequently used, grading scales are only poorly understood¹, and differences in scale design and the severity range covered by the reference images have limited an interchangeable use of the scales.^{2,3}

The global aim of this thesis was to gain a better understanding of grading scales and of the processes that are involved in clinical grading. In order to achieve this, objective and subjective techniques were used to analyze the MC-D, IER, Efron, and VBR bulbar redness grading scales, with the more specific goal in mind to establish a technique that would allow cross-calibration between scales.

In the first experiment (chapter 4), the accuracy of the bulbar redness scales was determined by correlating their nominal scale grades with three *physical* attributes of redness (D , % PC, and u). This study was the first to successfully use fractal dimension, D , as a measure of the complexity of the conjunctival vasculature, which prior to this had only been used for the analysis of retinal vessels⁴⁻⁸ and had briefly been suggested to be applicable in characterising the bulbar conjunctiva.⁹ There were strong linear associations between scale grades and all physical redness attributes, so that each scale individually might be considered accurate. However, there were discrepancies between the physical measures when comparing *across* the scales, so that a cross-calibration of the scales based on any of the physical redness attributes was not feasible. In addition, the accuracy of each scale depended on the physical attribute used for the analysis,

and each scale appeared to best describe one *particular* attribute of redness. The VBR and MC-D scale were found to best represent redness in terms of photometric chromaticity, the Efron scale vascular branching (*D*), and the IER scale area covered by the vessels (% *PC*).

After having demonstrated these physical differences, a different approach was taken in chapter 5 by investigating the *perceptual* relationship between the reference images of the MC-D, IER, Efron and VBR scales. Ten observers with no previous clinical experience or exposure to grading scales were asked to participate in the experiment. Psychophysical scaling was used to determine the perceived redness of each reference image based on its position within a given redness range for which only the start (0) and end point (100) were indicated. There were differences in the dynamic ranges of the scales and in reference levels across scales, even though the scales are generally designed to cover the same range of clinical redness.

To evaluate the criteria that may be used for clinical assessments of redness, the three physical attributes of redness (chapter 4) were compared to the perceived redness of the reference images. There were strong linear associations between perceived redness and the physical redness attributes corresponding to the reference images of each grading scale. Corroborating findings of some studies^{1,3,10} it was demonstrated in chapter 5 that redness is likely assessed by using both chromaticity *and* vessel-based information.

The findings of this study indicated that the perceived redness may be used for a cross-calibration of the grading scales. However, the increased perceived

redness of the VBR reference images compared to their previously validated scale grades¹⁹ necessitated a modification of the psychophysical scaling experiment before such a cross-calibration could be attempted. Therefore, chapter 6 represents a logical extension of the experiment conducted in chapter 5 (non-anchored scaling) inasmuch as the VBR scale images were now provided as unlabelled reference anchors at their respective scale levels at 10, 30, 50, 70, and 90 (chapter 6) within the same given 0 to 100 range (anchored scaling).

As could be demonstrated, the scaling of redness when using reference images as stationary anchors resulted in a shift in perceived redness and in a re-scaling of redness severity. The re-scaling of redness severity had an impact on the actual (perceived redness) grade associated to the reference image, but the order and the perception of relative differences remained fairly constant. This suggests that despite the physical differences between the grading scales, and independent of the scale being used, observers are able to ignore absolute redness characteristics (chapter 4) while using relative information to come to an appropriate clinical conclusion.

In chapter 6, we have referred to this mechanism as clinical *scale* constancy. That being said, it appears that this re-scaling of all reference images relative to the provided anchors might be more appropriately referred to as clinical *scaling* constancy. Clinical *scale* constancy would be the perceptual resetting of each scale (for example zero and dynamic range) so that it is applied the same way as any of the other (physically different) scales. For example, I perceptually reset the IER 1 to 4 scale or the MC-D 0 to 5 scale into a 'white eye' to 'red eye' scale for which the

'reddest' image of either scale corresponds to maximum red, despite physical and perceptual differences in the images.

These differences between the images of the four scales was one of the hypothesized potential causes (chapters 5 and 6) for the observation that the VBR reference images were placed at positions corresponding to higher redness than found when the scale was developed.¹⁹ However, further factors might have contributed to this finding, mainly because of somewhat different experimental settings – the current experiments were conducted eight years after the scale had been developed – so that observer cohorts, printer calibrations, or lighting were inevitably different.

A different observer cohort might have impacted the observed shift in redness for the VBR scale images due to a number of reasons. The observer cohort when the scale was developed consisted of optometrists and undergraduate students (mean age 29.9 years; range 21-39), and was slightly older than the one consisting of students for the current study (mean age 25.5 years, range 23-31). The results of both experiments indicated very high linear associations (Pearson's $r's \geq 0.92$) between any pair of observers within each individual scaling experiment, independent of their age, which suggests that age, at least for the fairly small difference for the two cohorts, was not a factor. The distribution of gender in both studies was quite different, with seven female and two male participants when the scale was developed¹⁹ compared to five observers for each gender in the current study. However, gender was found not to affect redness scaling, as RM ANOVA with gender as categorical factor showed no significant difference between sexes for

non-anchored scaling (RM ANOVA; $F=0.89$, $p=0.37$) or scaling of any other image set (grayscale, binarized, or anchored scaling).

It might be argued that observer cohorts with different distributions of colour vision might have contributed to the observed shift in redness for the VBR scale images. It is unknown if colour-defective observers participated in the scale-development experiment, however, the current research included one participant with a deuteranomaly. There were no significant differences for perceived redness for non-anchored redness scaling between the colour-defective participant and the colour-normal participants; the range of redness estimates was 0 to 100 for the colour-normal observers (average range: 3 to 97), and 0 to 99 for the colour-defective participant. Without exact knowledge of the distribution of colour vision in both cohorts it is difficult to estimate the potential impact this may have had, however, the almost identical perceived redness for colour-defective and colour-normal participants in this study suggests that a potential impact might have only been small.

The observed shift in redness might also be attributed to different distributions of handedness (i.e. the preference for using one hand more than another²⁰) for the observers of the two cohorts. It has been suggested that observers tend towards that end of a horizontal scale that is associated to their individual hand dominance.²¹ The distribution of handedness for the scale development cohort is not known, however, for the current experiment, three of ten participants were left-handed, which is considerably more than the percentage typically found in the population (10% left-handers²²). Therefore it seems rather

unlikely that the observed shift to the right (i.e. to higher redness) can be attributed to a more right-handed observer cohort in the current study.

Despite the use of different printers to print the colour images in the two experiments, there was no significant difference between the images' chromaticity when measured off of the printed copies using the SpectraScan PR650 spectrophotometer in both occasions (two-tailed paired t-test; $t=0.90$, $p=0.41$). This suggests that the shift in perceived redness of the VBR scale images was probably not caused by (potentially) different calibrations of the printers.

Lastly, another possible reason for the observed shift to higher redness of the VBR scale images in the current experiment might be attributed to an inevitable change of the general experimental setup. Although the physical distance to be used for scaling remained the same, the current experiment was conducted in a different laboratory, as the original room had been transformed into an examination room in the meantime. Therefore, the illumination settings in the two rooms might have been different. Both experiments were conducted with full room illumination using cool white fluorescent lighting typical for laboratory settings; however, no illuminance measurements of the scale development experiments are available to allow for a direct comparison of the illumination settings. That being said, it appears that differences in illumination are not likely to have triggered the observed shift in redness, as it is generally accepted that observers have fairly invariant perceptions of colour despite differences in illumination, a mechanism that is known as colour constancy.²³

To conclude, the reasons for the observed shift in redness are still not completely clear, and it appears that a combination of (some of) the discussed potential factors have triggered this finding.

Returning to the general discussion, the physical (chapter 4) and perceived redness (chapters 5 and 6) was found to be materially different between the reference images, which indicates that the images of the various scales are not aligned. In order to allow a comparison between scale grades despite these differences, the perceived redness from anchored scaling was used to cross-calibrate the scales and to develop a conversion table for the comparison of grading estimates between scales if the scales are used with their originally provided reference grades. In addition, the approach taken in chapter 6 also provides the option to modify the grading scales by a relative renumbering of the reference images (as shown in Figure 6-5) instead of using the original scale grades. This serves not only the purpose of extending the scale range (and thus theoretically increases the sensitivity to detecting change^{11,12}), but may also allow a better comparison of grades between the available redness scales. Therefore, the next logical step was to investigate if the calibration of the grading scale reference images based on perceived redness data resulted in less variable (and thus more comparable) redness estimates between different scales than previously reported.^{2,10}

In chapter 7, the reference images of the newly calibrated grading scales were presented at the positions within the 0 to 100 range that corresponded to their perceived redness from anchored scaling (chapter 6); each scale was shown alone, with only the start and end point of the range, but not the reference images, being identified by their redness grades. The same 10 participants who had

participated in the scaling experiments (chapters 5 & 6) were asked to represent redness in 16 sample images by placing each image, one image at a time, relative to the unlabelled reference images of each of the four scales.

Overall, there was very high agreement between the grading estimates of all of the scales. The bias and variability between grading estimates across scales was very low, and was similar to or lower than commonly observed within-scale results in repeated experimental assessments of the same eye.^{2,13-15} In addition, the agreement between scales seems to be connected to their dynamic range, as closer agreement of redness estimates was found for scales with similar dynamic ranges (MC-D/IER and Efron/VBR). A comparison between the subjective redness estimates obtained with each of the scales to physical redness attributes (D , % PC and u') showed high associations. Similar to the findings when scaling the reference images (chapter 5), the subjective redness estimates were best described by a combination of both chromaticity *and* vessel-based information, but the larger impact on the assessments appeared to be due to colour rather than vessel information.

In conclusion, I hope that the findings of my PhD research have contributed to a better understanding of the currently available grading scales. However, the research on grading scales is by no means complete yet, and a number of potential studies may arise from the experiments that were conducted during this PhD.

8.1 Future work

As stated before, illustrative grading scales are and will be used in clinical practice and research settings, probably primarily due to their availability and

practicability. For specific research settings, however, the objective quantification of small changes may be required. In chapter 4, the capability of fractal analysis to characterise redness in the reference images of the bulbar redness grading scales was demonstrated. Nevertheless, the close agreement between D and % PC , as demonstrated in chapters 5 and 7, leaves room for further investigation, as it is not clear yet if *either*, or *which*, or *both* of these physical attributes should be used in order to achieve the best spatial representation of redness. It, of course, may also be possible that these are simply mathematical transformations of the same aspect of images.

The results from chapter 7 showed that the previously reported bias between scales^{2,10} and the variability between grading estimates² was reduced when using the newly calibrated scales, which suggests that the cross-calibration of the grading scales based on perceived redness was successful. However, in chapter 7, redness was assessed by placing the sample images relative to the reference images and determining the associated grade by reading off of a meter stick. Despite generating a redness grade, this procedure was fairly different to typical clinical grading, when an eye is compared to reference images of a scale and a numerical grade is assigned presumably by matching. Therefore, a potential future study could investigate the agreement of the newly calibrated scales when redness is directly estimated in a set of sample images or a group of patients.

The experiments described in chapters 4 to 7 all included modified versions of the reference images only of conjunctival detail, without lids, lashes or the cornea. This was a necessary prerequisite for image processing, but to allow a direct comparison to the physical redness attributes, these modified versions were

also used for redness scaling (chapters 5 & 6) and for the grading experiment in chapter 7. In the light of this, it would be interesting to investigate how grading using the original reference images compares to grading with the modified versions with conjunctival detail only, as the removal of potential confounders^{16,17} such as lids and lashes might have partly contributed to the close agreement between scales as demonstrated in chapter 7. This would provide novel and potentially valuable information on the design of bulbar redness grading scales.

Despite putative advantages of a 100-point grading system in being more sensitive to detecting change^{11,12,18}, practitioners may be hesitant to use a 100-point grading system and might prefer to use a more familiar 0 to 4 grading scale. The data listed in Table 6-2 provides practitioners with a simple option to compare grades between scales. However, further research is required to validate the converted grades in Table 6-2, for example by using the scales with their original grades and subsequent comparison of the resulting redness estimates to the converted grades as shown in Table 6-2.

Despite the contributions that this research on grading scales may have provided, clinicians would still benefit from a non-subjective and thus non-observer dependent technique to quantify redness that would eliminate the inter- and intra-observer variability inherent to subjective clinical assessments. Therefore, the development of a usable, reliable and affordable objective technique to objectively quantify redness remains the ultimate goal to be accomplished. The use of fractal analysis to quantify the complexity of, and the area covered by, the bulbar vasculature might represent an affordable option for both private practice and research settings.

REFERENCES

CHAPTER 1

1. Merton RK, Sills DL, Stigler SM. The Kelvin dictum and social science: an excursion into the history of an idea. *J Hist Behav Sci* 1984;20 (4): 319-31.
2. BIPM - Bureau international des poids et mesures. Metrology. 2004. Available at: <http://www.bipm.org/en/convention/wmd/2004/>. Accessed: 2009, 01/28.
3. Merriam Webster. Dictionary. 2009. Available at: <http://www.merriam-webster.com/dictionary/>. Accessed: 2009, 07/28.
4. Microsoft Encarta Online Encyclopedia. Number (mathematics). Microsoft Corporation. Updated 2009. Available at: http://encarta.msn.com/encyclopedia_761557367/Number_%28mathematics%29.html. Accessed: 2009, 07/30.
5. Russell B. Introduction to mathematical philosophy, 2nd. ed. London: G. Allen & Unwin; 1956.
6. Martinez JL. The nature of fractals. 2006. Updated 2008/03/24/. Available at: http://www.fractovia.org/art/what/what_ing1.shtml Accessed: 2009, 03/17.
7. Massof RW. The measurement of vision disability. *Optom Vis Sci* 2002;79 (8): 516-52.
8. Finkelstein L. Widely, strongly and weakly defined measurement. *Measurement* 2003;34 (1): 39-48.
9. von Helmholtz HLF. Counting and Measuring. New York: Van Norstrand; 1930.
10. Kisch B. Scales and Weights: a historical outline. New Haven: Yale University Press; 1965.
11. Central Intelligence Agency. The World Factbook. 2009. Appendix G. Available at: <https://www.cia.gov/library/publications/the-world-factbook/appendix/appendix-g.html>. Last accessed: 02/08, 2009.
12. Rossi GB. Measurability. *Measurement* 2007;40 (6): 545-62.
13. Michels E. Measurement in physical therapy. On the rules for assigning numerals to observations. *Phys Ther* 1983;63 (2): 209-15.
14. Stevens SS. On the Theory of Scales of Measurement. *Science* 1946;103 (2684): 677-80.

15. Stevens SS. On the averaging of data. *Science* 1955;121 (3135): 113-6.
16. Stevens SS. Measurement and man. *Science* 1958;127 (3295): 383-9.
17. Pfanzagl J. *Theory of Measurement*. Wuerzburg / Wien: Physica - Verlag; 1971.
18. Suppes P, Zinnes JL, R.D.Luce, R.Bush, E.Galanter. Basic measurement theory. In: *Handbook of Mathematical Psychology*. New York Wiley, 1965. 3-76.
19. Finkelstein L. Problems of measurement in soft systems. *Measurement* 2005;38 (4): 267-74.
20. Finkelstein L, Leaning MS. A review of the fundamental concepts of measurement. *Measurement* 1984;2 (1): 25-34.
21. European Commission. *Measuring the Impossible*. 2004. Available at: ftp://ftp.cordis.europa.eu/pub/nest/docs/path_meas_refdoc_1204.pdf
Accessed: 2009, 03/17.
22. Pointer MR. *Measuring visual appearance - a framework for the future*; 2003. Report No.: NPL COAM 19.
23. Hunter RS, Harold RW. *The Measurement of Appearance*. New York: John Wiley & Sons, Inc.; 1987.
24. Stevens SS. On the psychophysical law. *Psychol Rev* 1957;64 (3): 153-81.
25. Schulze M, Jones D, Simpson T. The development of validated bulbar redness grading scales. *Optom Vis Sci* 2007;84 (10): 976-83.
26. Joint Committee for Guides in Metrology. *International vocabulary of metrology - Basic and general concepts and associated terms (VIM)*; 2008. Report No.: 200.
27. Torgerson WS. *Theory and Methods of Scaling*. New York: John Wiley & Sons, Inc.; 1958.
28. Efron N. Bulbar hyperaemia. In: *Contact Lens Complications Vol. 1st*. Oxford; Boston: Butterworth-Heinemann, 1999. Chapter 5: 33-9.
29. Jones LW, Jones DA. *Common contact lens complications : their recognition and management*. Oxford; Boston; Butterworth-Heinemann; 2000.
30. Munro F, Covey M. Differential diagnosis in contact lens aftercare: Part I - ocular redness. *Optician* 1999;217 (5683): 24-34.
31. Lens A, Langley T, Coyne Nemeth S, Shea C. *Ocular Anatomy and Physiology*, 1st. ed. Thorofare, NJ 08086: Slack Inc.; 1999.

32. Fatt I. Physiology of the eye - an introduction to the vegetative functions, 1st. ed. Boston: Butterworths; 1978.
33. Heath G. The episclera, sclera and conjunctiva - An overview of relevant ocular anatomy. Optometry Today 2006: 36-42.
34. Oyster CW. The human eye: structure and function, 1st. ed. Sunderland, MA: Sinauer; 1999.
35. Andersen JS, Davies IP, Kruse A, Lofstrom T, Ringman LA. Handbook of Contact Lens Management: Vistakon; 1996.
36. Leitman M. Manual for eye examination and diagnosis, 6th ed. Malden, MS: Blackwell Publishing Inc.; 2004.
37. Csillag A. Atlas of the sensory organs, 1st ed. Totowa, NJ: Humana Press Inc.; 2005.
38. Rapuano JR, Luchs JI, Kim T. Anterior Segment. St. Louis, Missouri: Mosby-Year Book, Inc.; 2000.
39. Papas EB. The limbal vasculature. Cont Lens Anterior Eye 2003;26 (2): 71-6.
40. Snell R, Lemp M. Clinical anatomy of the eye. Cambridge, MA: Blackwell Scientific Publications; 1989.
41. Lin L. Overview of agreement statistics for medical devices. J Biopharm Stat 2008;18 (1): 126-44.
42. Guillon M, Shah D. Objective measurement of contact lens-induced conjunctival redness. Optom Vis Sci 1996;73 (9): 595-605.
43. Murphy PJ, Lau JS, Sim MM, Woods RL. How red is a white eye? Clinical grading of normal conjunctival hyperaemia. Eye 2006;21: 633-8.
44. Pult H, Murphy PJ, Purslow C, Nyman J, Woods RL. Limbal and bulbar hyperaemia in normal eyes. Ophthalmic Physiol Opt 2008;28 (1): 13-20.
45. Efron N, Brennan NA, Hore J, Rieper K. Temperature of the hyperemic bulbar conjunctiva. Curr Eye Res 1988;7 (6): 615-8.
46. Lemp ME. The definition and classification of dry eye disease: report of the Definition and Classification Subcommittee of the International Dry Eye WorkShop (2007). Ocul Surf 2007;5 (2): 75-92.
47. Holden BA, Sweeney D, Swarbrick H, Vannas A, Nilsson KT, Efron N. The vascular response to long-term extended contact lens wear. Clinical and Experimental Optometry 1986;69 (3): 112-9.

48. 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 (12): 911-21.
49. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. Part II. *Am J Optom Physiol Opt* 1987;64 (4): 251-5.
50. Papas E, Willcox M. Reducing the Consequences of Hypoxia: The Ocular Redness Response. *Contact Lens Spectrum*. 2006;21 2. Available at: <http://www.clspectrum.com/article.aspx?article=12954>. Last accessed: 02/08, 2009.
51. Duench S, Simpson TL, Jones LW, Flanagan JG, Fonn D. Assessment of Variation in Bulbar Conjunctival Redness, Temperature, and Blood Flow. *Optom Vis Sci* 2007;84 (6): 511-6.
52. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. Part I. *Am J Optom Physiol Opt* 1987;64 (4): 246-50.
53. Woods R. Quantitative slit lamp observations in contact lens practice. *J Brit Contact Lens Assoc* 1989;12: 42-5.
54. Bundesrepublik Deutschland. Gesetz über Medizinprodukte. 2002. Available at: <http://bundesrecht.juris.de/bundesrecht/mpg/gesamt.pdf>.
55. Mandell RB. Slit lamp classification system. *J Am Optom Assoc* 1987;58 (3): 198-201.
56. Efron N. Clinical application of grading scales for contact lens complications. *Optician* 1997;213 (5604): 26-35.
57. Efron N. Grading scales for contact lens complications. In: *Contact Lens Complications*. Oxford; Boston: Butterworth-Heinemann, 1999. Appendix A: 171-9.
58. Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects of scaling. *Invest Ophthalmol Vis Sci* 1991;32 (2): 422-32.
59. Lloyd M. Lies, statistics, and clinical significance. *J Brit Contact Lens Assoc* 1992;15 (2): 67-70.
60. Efron N. Grading scales for contact lens complications. *Ophthalmic Physiol Opt* 1998;18 (2): 182-6.
61. Chong T, Simpson T, Fonn D. The repeatability of discrete and continuous anterior segment grading scales. *Optom Vis Sci* 2000;77 (5): 244-51.

62. Schulze M, Hutchings N, Simpson T. The use of fractal analysis and photometry to estimate the accuracy of bulbar redness grading Scales. *Invest Ophthalmol Vis Sci* 2008;49 (4): 1398-406.
63. Efron N, Chaudry A. Grading static versus dynamic images of contact lens complications. *Clin Exp Optom* 2007;90 (5): 361-6.
64. Efron N, McCubbin S. Grading contact lens complications under time constraints. *Optom Vis Sci* 2007;84 (12): 1082-6.
65. Efron N, Morgan PB, Jagpal R. The combined influence of knowledge, training and experience when grading contact lens complications. *Ophthalmic Physiol Opt* 2003;23 (1): 79-85.
66. MacKinven J, McGuinness CL, Pascal E, Woods RL. Clinical grading of the upper palpebral conjunctiva of non-contact lens wearers. *Optom Vis Sci* 2001;78 (1): 13-8.
67. Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci* 2000;41 (3): 687-91.
68. Peterson RC, Wolffsohn JS. Sensitivity and reliability of objective image analysis compared to subjective grading of bulbar hyperaemia. *Br J Ophthalmol* 2007;91 (11): 1464-6.
69. Peterson RC, Wolffsohn JS. Objective grading of the anterior eye. *Optom Vis Sci* 2009;86 (3): 273-8.
70. Sorbara L, Simpson T, Duench S, Schulze M, Fonn D. Comparison of an objective method of measuring bulbar redness to the use of traditional grading scales. *Cont Lens Anterior Eye* 2007;30 (1): 53-9.
71. Terry RL, Schnider CM, Holden BA, Cornish R, Grant T, Sweeney D, et al. CCLRU standards for success of daily and extended wear contact lenses. *Optom Vis Sci* 1993;70 (3): 234-43.
72. FDA. Slit Lamp Findings Classification Scale. In. Washington, DC: U.S. Food and Drug Administration; 1984.
73. Kahn HA, Leibowitz H, Ganley JP, Kini M, Colton T, Nickerson R, et al. Standardizing diagnostic procedures. *Am J Ophthalmol* 1975;79 (5): 768-75.
74. Lofstrom T, Anderson JS, Kruse A. Tarsal Abnormalities: A New Grading System. *CLAO J* 1998;24 (4): 210-5.
75. Terry R, Sweeney D, Wong R, Papas E. Variability of clinical investigators in contact lens research. *Optom Vis Sci* 1995;72 (suppl 12): 16.

76. CCLRU. CCLRU grading scales. In: Contact Lenses Vol. 4th. London: Elsevier Butterworth-Heinemann, 1997. 863-7.
77. IER Grading Scales. Institute for Eye Research, Sydney, Australia. Available at: <http://www.siliconehydrogels.org/resources/index.asp>. Last accessed: 05/19/2009.
78. Efron N. Grading scales. *Optician* 2000;219 (5733): 44-5.
79. Schulze M. The production of an enhanced grading scale for determination of ocular hyperaemia: University of Waterloo, Ontario, Canada; 2000.
80. Efron N, Morgan PB, Jagpal R. Validation of computer morphs for grading contact lens complications. *Ophthalmic Physiol Opt* 2002;22 (4): 341-9.
81. Efron N. Grading scales and morphs. In: Contact Lens Complications. Oxford; Boston: Butterworth-Heinemann, 1999. Chapter 19: 161-7.
82. Chong T, Simpson T, Pritchard N, Dumbleton K, Richter D, Fonn D. Repeatability of discrete and continuous clinical grading scales. *Optom Vis Sci* 1996;73 (suppl 12): 232.
83. Simpson T, Chong T, Pritchard N. Continuous clinical grading scales using morphing software. *J Brit Contact Lens Assoc* 1996;19: 151.
84. Simpson T, Pritchard N. Continuous grading scales of ocular redness, papillae and corneal staining. *Optom Vis Sci* 1995;72 (suppl 12): 77.
85. du Toit R, Pritchard N, Heffernan S, Simpson TL, Fonn D. A comparison of three different scales for rating contact lens handling. *Optom Vis Sci* 2002;79 (5): 313-20.
86. 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 (2): 119-24.
87. Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. *Ophthalmic Physiol Opt* 2001;21 (1): 17-29.
88. Wolffsohn JS. Incremental nature of anterior eye grading scales determined by objective image analysis. *Br J Ophthalmol* 2004;88 (11): 1434-8.
89. Miller G. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol Rev* 1956;63: 81-97.
90. Long B. Comparison of Subjective Grading of Limbal Redness By Eyecare Practitioners in Three Countries. *Eye Contact Lens* 2009.

91. Dundas M, Walker A, Woods R. Clinical grading of corneal staining of non-contact lens wearers. *Ophthalmic Physiol Opt* 2001;21 (1): 30-5.
92. Brennan NA, Coles ML, Ang JH. An evaluation of silicone-hydrogel lenses worn on a daily wear basis. *Clin Exp Optom* 2006;89 (1): 18-25.
93. Dumbleton KA, Chalmers RL, Richter DB, Fonn D. Vascular response to extended wear of hydrogel lenses with high and low oxygen permeability. *Optom Vis Sci* 2001;78 (3): 147-51.
94. Fieguth P, Simpson T. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci* 2002;43 (2): 340-7.
95. Norcross JC, Karg RS, Prochaska JO. Clinical psychologists in the 1990's: II. *The Clinical Psychologist* 1997;50 (3): 4-11.
96. Perez-Cabre E, Millan MS, Abril HC, Otxoa E. Image processing of standard grading scales for objective assessment of contact lens wear complications. In: *Proc Soc Photo Opt Instrum Eng*; 2004; 2004. p. 107-12.
97. Willingham FF, Cohen KL, Coggins JM, Tripoli NK, Ogle JW, Goldstein GM. Automatic quantitative measurement of ocular hyperemia. *Curr Eye Res* 1995;14: 1101-8.
98. Conrad KJ, Smith EV. International conference on objective measurement - Applications of Rasch analysis in health care. *Med Care* 2004;42 (1): 1-6.
99. Court H, Greenland K, Margrain TH. Content development of the Optometric Patient Anxiety Scale. *Optom Vis Sci* 2007;84 (8): 729-37.
100. Cavanagh RF, Romanoski JT. Rating scale instruments and measurement. *Learning Environments Research* 2006;9 (3): 273-89.
101. Bron AJ, Evans VE, Smith JA. Grading of corneal and conjunctival staining in the context of other dry eye tests. *Cornea* 2003;22 (7): 640-50.
102. McMonnies CW, Ho A. Conjunctival hyperaemia in non-contact lens wearers. *Acta Ophthalmol (Copenh)* 1991;69 (6): 799-801.
103. O'Donnell C, Wolffsohn JS. Grading of corneal transparency. *Cont Lens Anterior Eye* 2004;27 (4): 161-70.
104. Chen PCY, Kovalcheck SW, Zweifach BW. Analysis of Microvascular Network in Bulbar Conjunctiva by Image-Processing. *Int J Microcirc Clin Exp* 1987;6 (3): 245-55.
105. Owen CG, Fitzke FW, Woodward EG. A new computer assisted objective method for quantifying vascular changes of the bulbar conjunctivae. *Ophthalmic Physiol Opt* 1996;16 (5): 430-7.

106. Palmer JR, Owen CG, Ford AM, Jacobson RE, Woodward EG. Optimal photographic imaging of the bulbar conjunctival vasculature. *Ophthalmic Physiol Opt* 1996;16 (2): 144-9.
107. Simpson T, Chan A, Fonn D. Measuring ocular redness: first order (luminance & chromaticity) measurements provide more information than second order (spatial structure) measurements. *Optom Vis Sci* 1998;75 (suppl 12): 279.
108. Villumsen J, Ringquist J, Alm A. Image-analysis of conjunctival hyperemia - a personal-computer based system. *Acta Ophthalmol (Copenh)* 1991;69 (4): 536-9.
109. Wolffsohn JS, Purslow C. Clinical monitoring of ocular physiology using digital image analysis. *Cont Lens Anterior Eye* 2003;26 (1): 27-35.
110. Schulze M, Hutchings N, Simpson T. The perceived bulbar redness of clinical grading scales. *Optom Vis Sci* 2009;86 (11): 1250-8.
111. Palmer JM. Radiometry and Photometry FAQ. 2003. Updated 2003/10/26. Available at: <http://www.optics.arizona.edu/Palmer/rpfaq/rpfaq.pdf> Accessed: 2009, 03/17.
112. Matkovic K. Tone Mapping Techniques and Color Image Difference in Global Illumination: Technische Universitaet Wien; Institut fuer Computergraphik; 1997.
113. Kuehni RG. Color Space and Its Divisions: color order from antiquity to the present. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2003.
114. Poynton CA. Frequently Asked Questions about Color. 1997. Updated 2006/11/28/. Available at: <http://www.poynton.com/PDFs/ColorFAQ.pdf> Accessed: 2009, 03/17.
115. Carpenter RH, Robson JG. Vision research : a practical guide to laboratory methods: Oxford; New York; Oxford University Press; 1999.
116. Schaeffers S. Comparison of spectroradiometric measures and clinical grading of interpalpebral bulbar vascularity of the human eye - a new method of evaluating objective data: University of Waterloo, School of Optometry, Centre for Contact Lens Research; 2000.
117. Situ P, Simpson TL, Fonn D. Objective measure of ocular redness: the repeatability and the association with subjective scale. *Invest Ophthalmol Vis Sci* 2001;42 (4): s597#3212.
118. Baxes GA. What is image processing. In: Digital image processing - principles and applications. New York: John Wiley & Sons, Inc., 1994. Chapter 1: 1-10.

119. Baxes GA. The digital image. In: Digital image processing - principles and applications. New York: John Wiley & Sons, Inc., 1994. Chapter 3: 37-67.
120. Castellano G, Bonilha L, Li LM, Cendes F. Texture analysis of medical images. Clin Radiol 2004;59 (12): 1061-9.
121. Eperjesi F, Wolffsohn JS, Phillips AJ, Speedwell L. Clinical instrumentation in contact lens practice. In: Contact Lenses Vol. 5th. London: Elsevier Butterworth-Heinemann, 2007. 159-71.
122. Owen CG, Ellis TJ, Rudnicka AR, Woodward EG. Optimal green (red-free) digital imaging of conjunctival vasculature. Ophthalmic Physiol Opt 2002;22 (3): 234-43.
123. Morgan P, Frankish C. Image quality, compression and segmentation in medicine. J Vis Commun Med 2002;25 (4): 149-54.
124. Peterson RC, Wolffsohn JS. The effect of digital image resolution and compression on anterior eye imaging. Br J Ophthalmol 2005;89 (7): 828-30.
125. Guillon JP, Godfrey A, Phillips AJ, Speedwell L. Tears and contact lenses. In: Contact Lenses Vol. 5th. London: Elsevier Butterworth-Heinemann, 2007. 111-27.
126. Lesmoir-Gordon N, Rood W, Edney R. Introducing fractal analysis: Allen & Unwyn Pty. Ltd., PO Box 8500, 9 Atchison Street, St. Leonards, NSW, 2065; 2000.
127. Mandelbrot BB. The fractal geometry of nature. New York: W.H.Freeman and Company; 1982.
128. Masters BR. Fractal analysis of the vascular tree in the human retina. Annu Rev Biomed Eng 2004;6 (1): 427-52.
129. Masters BR. Fractal analysis of human retinal vessels. In: Proc Soc Photo Opt Instrum Eng; 1990; 1990. p. 250-6.
130. Landini G, Misson G. Simulation of corneal neovascularization by inverted diffusion limited aggregation. Invest Ophthalmol Vis Sci 1993;34 (5): 1872-5.
131. Mandelbrot B. How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. Science 1967;156 (3775): 636-8.
132. Rasband WS. ImageJ Version 1.38x (07/13/2007). Available at: <http://rsb.info.nih.gov/ij/>. U. S. National Institutes of Health, Bethesda, Maryland, USA.

133. Karperien A. FracLac for ImageJ - FracLac Advanced User's Manual v.2.5. 2007. Available at: <http://rsbweb.nih.gov/ij/plugins/fracLac/FLHelp/Introduction.htm>. Accessed: 2009, 07/29.
134. Avakian A, Kalina RE, Sage EH, Rambhia AH, Elliott KE, Chuang EL, et al. Fractal analysis of region-based vascular change in the normal and non-proliferative diabetic retina. *Curr Eye Res* 2002;24 (4): 274-80.
135. Baish JW, Jain RK. Cancer, angiogenesis and fractals. *Nat Med* 1998;4 (9): 984.
136. Cross SS. Fractals in pathology. *J Pathol Bacteriol* 1997;182 (1): 1-8.
137. Family F, Masters BR, Platt DE. Fractal pattern formation in human retinal vessels. *Physica D* 1989;38 (1-3): 98-103.
138. Kamiya A, Takahashi T. Quantitative assessments of morphological and functional properties of biological trees based on their fractal nature. *J Appl Physiol* 2007;102 (6): 2315-23.
139. Masters BR. Fractal analysis of human retinal blood vessel patterns: developmental and diagnostic aspects. In: Masters BR, editor. *Noninvasive diagnostic techniques in ophthalmology*. New York: Springer-Verlag, 1990. 515-27.
140. 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 (10): 758-68.
141. McMonnies CW, Chapman-Davies A, Holden BA. The vascular response to contact lens wear. *Am J Optom Physiol Opt* 1982;59 (10): 795-9.

CHAPTER 2

1. Efron N. Bulbar hyperaemia. In: Contact Lens Complications Vol. 1st. Oxford; Boston: Butterworth-Heinemann, 1999. Chapter 5: 33-9.
2. Guillon M, Shah D. Objective measurement of contact lens-induced conjunctival redness. *Optom Vis Sci* 1996;73 (9): 595-605.
3. Woods R. Quantitative slit lamp observations in contact lens practice. *J Brit Contact Lens Assoc* 1989;12: 42-5.
4. Lloyd M. Lies, statistics, and clinical significance. *J Brit Contact Lens Assoc* 1992;15 (2): 67-70.
5. Efron N. Grading scales for contact lens complications. *Ophthalmic Physiol Opt* 1998;18 (2): 182-6.
6. Efron N. Clinical application of grading scales for contact lens complications. *Optician* 1997;213 (5604): 26-35.
7. Fieguth P, Simpson T. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci* 2002;43 (2): 340-7.
8. Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci* 2000;41 (3): 687-91.
9. Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. *Ophthalmic Physiol Opt* 2001;21 (1): 17-29.
10. Pult H, Murphy PJ, Purslow C, Nyman J, Woods RL. Limbal and bulbar hyperaemia in normal eyes. *Ophthalmic Physiol Opt* 2008;28 (1): 13-20.
11. Murphy PJ, Lau JS, Sim MM, Woods RL. How red is a white eye? Clinical grading of normal conjunctival hyperaemia. *Eye* 2006;21: 633-8.
12. Perez-Cabre E, Millan MS, Abril HC, Otxoa E. Image processing of standard grading scales for objective assessment of contact lens wear complications. In: *Proc Soc Photo Opt Instrum Eng*; 2004; 2004. p. 107-12.
13. Wolffsohn JS. Incremental nature of anterior eye grading scales determined by objective image analysis. *Br J Ophthalmol* 2004;88 (11): 1434-8.
14. Willingham FF, Cohen KL, Coggins JM, Tripoli NK, Ogle JW, Goldstein GM. Automatic quantitative measurement of ocular hyperemia. *Curr Eye Res* 1995;14: 1101-8.
15. Wolffsohn JS, Purslow C. Clinical monitoring of ocular physiology using digital image analysis. *Cont Lens Anterior Eye* 2003;26 (1): 27-35.

16. Schulze M, Hutchings N, Simpson T. The perceived bulbar redness of clinical grading scales. *Optom Vis Sci* 2009;86 (11): 1250-8.
17. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. Part I. *Am J Optom Physiol Opt* 1987;64 (4): 246-50.
18. CCLRU. CCLRU grading scales. In: *Contact Lenses Vol. 4th*. London: Elsevier Butterworth-Heinemann, 1997. 863-7.
19. IER Grading Scales. Institute for Eye Research, Sydney, Australia. Available at: <http://www.siliconehydrogels.org/resources/index.asp>. Last accessed: 05/19/2009.
20. Efron N. Grading scales for contact lens complications. In: *Contact Lens Complications*. Oxford; Boston: Butterworth-Heinemann, 1999. Appendix A: 171-9.
21. Efron N. Grading scales. *Optician* 2000;219 (5733): 44-5.
22. Schulze M, Jones D, Simpson T. The development of validated bulbar redness grading scales. *Optom Vis Sci* 2007;84 (10): 976-83.

CHAPTER 3

1. Barnhart HX, Haber MJ, Lin LI. An Overview on Assessing Agreement with Continuous Measurements. *J Biopharm Stat* 2007;17 (4): 529-69.
2. White SA, van den Broek NR. Methods for assessing reliability and validity for a measurement tool: a case study and critique using the WHO haemoglobin colour scale. *Stat Med* 2004;23 (10): 1603-19.
3. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol* 2008;31 (4): 466-75.
4. Lin L. Overview of agreement statistics for medical devices. *J Biopharm Stat* 2008;18 (1): 126-44.
5. ISO. Accuracy (Trueness and Precision) of Measurement Methods and Results - Part 1: General Principles and Definitions. Geneva, Switzerland: International Organization for Standardization; 1994.
6. Taylor BN, Kuyatt CE. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, Appendix D.1: Terminology. Gaithersburg, MD: National Institute of Standards and Technology; 1994.
7. FDA. Guidance for Industry: Bioanalytical Method Validation. In: Food and Drug Administration; 2001.
8. BIPM - Bureau international des poids et mesures. Metrology. 2004. Available at: <http://www.bipm.org/en/convention/wmd/2004/>. Accessed: 2009, 01/28.
9. Merriam Webster. Dictionary. 2009. Available at: <http://www.merriam-webster.com/dictionary/>. Accessed: 2009, 07/28.
10. Joint Committee for Guides in Metrology. International vocabulary of metrology - Basic and general concepts and associated terms (VIM); 2008. Report No.: 200.
11. Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects of scaling. *Invest Ophthalmol Vis Sci* 1991;32 (2): 422-32.
12. Chong T, Simpson T, Fonn D. The repeatability of discrete and continuous anterior segment grading scales. *Optom Vis Sci* 2000;77 (5): 244-51.
13. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1 (8476): 307-10.
14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86 (2): 420-8.

15. Bartko JJ. Measures of Agreement - A Single Procedure. *Stat Med* 1994;13 (5-7): 737-45.
16. Streiner DL, Norman GR. *Health Measurement Scales - A practical guide to their development and use*. New York: Oxford University Press Inc.; 1995.
17. Lin LI. A concordance correlation-coefficient to evaluate reproducibility. *Biometrics* 1989;45 (1): 255-68.
18. Rasband WS. ImageJ Version 1.38x (07/13/2007). Available at: <http://rsb.info.nih.gov/ij/>. U. S. National Institutes of Health, Bethesda, Maryland, USA.
19. Karperien A. FraCLac for ImageJ - FraCLac Advanced User's Manual v.2.5. 2007. Available at: <http://rsbweb.nih.gov/ij/plugins/fraclac/FLHelp/Introduction.htm>. Accessed: 2009, 07/29.
20. Lesmoir-Gordon N, Rood W, Edney R. *Introducing fractal analysis*: Allen & Unwyn Pty. Ltd., PO Box 8500, 9 Atchison Street, St. Leonards, NSW, 2065; 2000.
21. Duench S, Simpson TL, Jones LW, Flanagan JG, Fonn D. Assessment of Variation in Bulbar Conjunctival Redness, Temperature, and Blood Flow. *Optom Vis Sci* 2007;84 (6): 511-6.
22. Masters BR. Fractal analysis of human retinal blood vessel patterns: developmental and diagnostic aspects. In: Masters BR, editor. *Noninvasive diagnostic techniques in ophthalmology*. New York: Springer-Verlag, 1990. 515-27.
23. Masters BR. Fractal analysis of the vascular tree in the human retina. *Annu Rev Biomed Eng* 2004;6 (1): 427-52.
24. Photo Research® Inc. PR® -650 Brochure. Chatsworth, CA; 1999.
25. Sorbara L, Simpson T, Duench S, Schulze M, Fonn D. Comparison of an objective method of measuring bulbar redness to the use of traditional grading scales. *Cont Lens Anterior Eye* 2007;30 (1): 53-9.
26. Efron N, Morgan PB, Jagpal R. Validation of computer morphs for grading contact lens complications. *Ophthalmic Physiol Opt* 2002;22 (4): 341-9.
27. Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. *Ophthalmic Physiol Opt* 2001;21 (1): 17-29.
28. Fieguth P, Simpson T. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci* 2002;43 (2): 340-7.
29. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. Part I. *Am J Optom Physiol Opt* 1987;64 (4): 246-50.

30. Murphy PJ, Lau JS, Sim MM, Woods RL. How red is a white eye? Clinical grading of normal conjunctival hyperaemia. *Eye* 2006;21: 633-8.
31. Pointer MR. Measuring visual appearance - a framework for the future; 2003. Report No.: NPL COAM 19.
32. Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*, 3rd ed. Hamilton, ON: BC Decker; 2008.
33. Schulze M, Jones D, Simpson T. The development of validated bulbar redness grading scales. *Optom Vis Sci* 2007;84 (10): 976-83.
34. Nickerson CA. A note on "A concordance correlation coefficient to evaluate reproducibility". *Biometrics* 1997;53: 1503-7.

CHAPTER 4

1. FDA. Slit Lamp Findings Classification Scale. In. Washington, DC: U.S. Food and Drug Administration; 1984.
2. Mandell RB. Slit lamp classification system. *J Am Optom Assoc* 1987;58 (3):198-201.
3. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. Part I. *Am J Optom Physiol Opt* 1987;64 (4):246-50.
4. IER. IER Grading Scales. Institute for Eye Research, Sydney, Australia. Available at: <http://www.siliconehydrogels.org/resources/index.asp>. Last accessed: 05/19/2009.
5. Efron N. Clinical application of grading scales for contact lens complications. *Optician* 1997;213 (5604):26-35.
6. Schulze M, Jones D, Simpson T. The development of validated bulbar redness grading scales. *Optom Vis Sci* 2007;84 (10):976-83.
7. Simpson T, Pritchard N. Continuous grading scales of ocular redness, papillae and corneal staining. *Optom Vis Sci* 1995;72 (suppl 12) (77).
8. Chong T, Simpson T, Fonn D. The repeatability of discrete and continuous anterior segment grading scales. *Optom Vis Sci* 2000;77 (5):244-51.
9. Fieguth P, Simpson T. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci* 2002;43 (2):340-7.
10. Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects of scaling. *Invest Ophthalmol Vis Sci* 1991;32 (2):422-32.
11. Murphy PJ, Lau JS, Sim MM, Woods RL. How red is a white eye? Clinical grading of normal conjunctival hyperaemia. *Eye* 2006;21 (6):33-8.
12. Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. *Ophthalmic Physiol Opt* 2001;21 (1):17-29.
13. Perez-Cabre E, Millan MS, Abril HC, Otxoa E. Image processing of standard grading scales for objective assessment of contact lens wear complications. In: *Proc Soc Photo Opt Instrum Eng*; 2004; 2004. p. 107-12.
14. Wolffsohn JS. Incremental nature of anterior eye grading scales determined by objective image analysis. *Br J Ophthalmol* 2004;88 (11):1434-8.
15. Efron N. Grading scales for contact lens complications. *Ophthalmic Physiol Opt* 1998;18 (2):182-6.

16. 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 (2):119-24.
17. Efron N, Morgan PB, Jagpal R. The combined influence of knowledge, training and experience when grading contact lens complications. *Ophthalmic Physiol Opt* 2003;23 (1):79-85.
18. Villumsen J, Ringquist J, Alm A. Image-analysis of conjunctival hyperemia - a personal-computer based system. *Acta Ophthalmol (Copenh)* 1991;69 (4):536-9.
19. Willingham FF, Cohen KL, Coggins JM, Tripoli NK, Ogle JW, Goldstein GM. Automatic quantitative measurement of ocular hyperemia. *Curr Eye Res* 1995;14 (11):1-8.
20. Owen CG, Fitzke FW, Woodward EG. A new computer assisted objective method for quantifying vascular changes of the bulbar conjunctivae. *Ophthalmic Physiol Opt* 1996;16 (5):430-7.
21. Guillon M, Shah D. Objective measurement of contact lens-induced conjunctival redness. *Optom Vis Sci* 1996;73 (9):595-605.
22. Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci* 2000;41 (3):687-91.
23. Sorbara L, Simpson T, Duench S, Schulze M, Fonn D. Comparison of an objective method of measuring bulbar redness to the use of traditional grading scales. *Cont Lens Anterior Eye* 2007;30 (1):53-9.
24. Wolffsohn JS, Purslow C. Clinical monitoring of ocular physiology using digital image analysis. *Cont Lens Anterior Eye* 2003;26 (1):27-35.
25. Simpson T, Chan A, Fonn D. Measuring ocular redness: first order (luminance & chromaticity) measurements provide more information than second order (spatial structure) measurements. *Optom Vis Sci* 1998;75 (suppl 12) (279).
26. Mandelbrot BB. *The fractal geometry of nature*. New York: W.H. Freeman and Company; 1982.
27. Masters BR. Fractal analysis of human retinal blood vessel patterns: developmental and diagnostic aspects. In: Masters BR, editor. *Noninvasive diagnostic techniques in ophthalmology*. New York: Springer-Verlag, 1990: 515-27.
28. Masters BR. Fractal analysis of human retinal vessels. In: *Proc Soc Photo Opt Instrum Eng*; 1990; 1990. p. 250-6.
29. Landini G, Misson G. Simulation of corneal neovascularization by inverted diffusion limited aggregation. *Invest Ophthalmol Vis Sci* 1993;34 (5):1872-5.

30. Avakian A, Kalina RE, Sage EH, Rambhia AH, Elliott KE, Chuang EL, et al. Fractal analysis of region-based vascular change in the normal and non-proliferative diabetic retina. *Curr Eye Res* 2002;24 (4):274-80.
31. Masters BR. Fractal analysis of the vascular tree in the human retina. *Annu Rev Biomed Eng* 2004;6 (1):427-52.
32. Rasband WS. ImageJ Version 1.38x (07/13/2007). Available at: <http://rsb.info.nih.gov/ij/>. U. S. National Institutes of Health, Bethesda, Maryland, USA.
33. Smith SW. *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing; 1997. Available at: <http://www.dspguide.com> Accessed: 2009, 03/17.
34. Young IT, Gerbrands JJ, van Vliet LJ. *Image Processing Fundamentals*. 1998. Updated 2007/06/07/. Available at: <http://www.ph.tn.tudelft.nl/Courses/FIP/noframes/fip-Statisti.html#Heading24> Accessed: 2009, 03/17.
35. Karperien A. *FraCLac for ImageJ - FraCLac Advanced User's Manual v.2.5*. 2007. Available at: <http://rsbweb.nih.gov/ij/plugins/fraclac/FLHelp/Introduction.htm>. Accessed: 2009, 07/29.
36. CIE - Commission International de L'Eclairage. *Recommendations on Uniform Color Spaces, Color-Difference Equations, Psychometric Color Terms*; 1978. Report No.: Supplement No. 2 of CIE Publ. No 15 (E-1.3.1) 1971.
37. Wyszecki G, Stiles WS. *Color Science - Concepts and Methods, Quantitative Data and Formulae*, 2nd ed. New York: John Wiley & Sons; 1982.
38. Poynton CA. *Gamma FAQ - Frequently Asked Questions about Gamma*. 1998. Updated 2002/12/16/. Available at: <http://www.poynton.com/PDFs/GammaFAQ.pdf> Accessed: 2009, 03/17.
39. Efron N. Bulbar hyperaemia. In: *Contact Lens Complications Vol. 1st*. Oxford; Boston: Butterworth-Heinemann, 1999: 33-9.
40. Land EH, McCann JJ. Lightness and Retinex Theory. *J Opt Soc Am* 1971;61 (1):1-11.
41. Finkelstein L. Widely, strongly and weakly defined measurement. *Measurement* 2003;34 (1):39-48.
42. Finkelstein L. Problems of measurement in soft systems. *Measurement* 2005;38 (4):267-74.
43. Rossi GB. Measurability. *Measurement* 2007;40 (6):545-62.
44. Pointer MR. *Measuring visual appearance - a framework for the future*; 2003. Report No.: NPL COAM 19.

CHAPTER 5

1. Efron N. Bulbar hyperaemia. In: Contact Lens Complications Vol. 1st. Oxford; Boston: Butterworth-Heinemann, 1999: 33-9.
2. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. Part I. Am J Optom Physiol Opt 1987;64 (4):246-50.
3. Efron N. Grading scales and morphs. In: Contact Lens Complications. Oxford; Boston: Butterworth-Heinemann, 1999: 161-7.
4. Peterson RC, Wolffsohn JS. Sensitivity and reliability of objective image analysis compared to subjective grading of bulbar hyperaemia. Br J Ophthalmol 2007;91 (11):1464-6.
5. Schulze M, Hutchings N, Simpson T. The use of fractal analysis and photometry to estimate the accuracy of bulbar redness grading Scales. Invest Ophthalmol Vis Sci 2008;49 (4):1398-406.
6. Efron N. Clinical application of grading scales for contact lens complications. Optician 1997;213 (5604):26-35.
7. Efron N. Grading scales for contact lens complications. Ophthalmic Physiol Opt 1998;18 (2):182-6.
8. Lloyd M. Lies, statistics, and clinical significance. J Brit Contact Lens Assoc 1992;15 (2):67-70.
9. Murphy PJ, Lau JS, Sim MM, Woods RL. How red is a white eye? Clinical grading of normal conjunctival hyperaemia. Eye 2006;21 (6):33-8.
10. Schulze M, Jones D, Simpson T. The development of validated bulbar redness grading scales. Optom Vis Sci 2007;84 (10):976-83.
11. Sorbara L, Simpson T, Duench S, Schulze M, Fonn D. Comparison of an objective method of measuring bulbar redness to the use of traditional grading scales. Cont Lens Anterior Eye 2007;30 (1):53-9.
12. Woods R. Quantitative slit lamp observations in contact lens practice. J Brit Contact Lens Assoc 1989;12 (42):5.
13. Kahn HA, Leibowitz H, Ganley JP, Kini M, Colton T, Nickerson R, et al. Standardizing diagnostic procedures. Am J Ophthalmol 1975;79 (5):768-75.
14. Terry R, Sweeney D, Wong R, Papas E. Variability of clinical investigators in contact lens research. Optom Vis Sci 1995;72 (suppl 12) (16).
15. Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects of scaling. Invest Ophthalmol Vis Sci 1991;32 (2):422-32.

16. IER. IER Grading Scales. Institute for Eye Research, Sydney, Australia. Available at: <http://www.siliconehydrogels.org/resources/index.asp>. Last accessed: 05/19/2009.
17. Efron N. Grading scales. *Optician* 2000;219 (5733):44-5.
18. Chong T, Simpson T, Fonn D. The repeatability of discrete and continuous anterior segment grading scales. *Optom Vis Sci* 2000;77 (5):244-51.
19. Fieguth P, Simpson T. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci* 2002;43 (2):340-7.
20. Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. *Ophthalmic Physiol Opt* 2001;21 (1):17-29.
21. Efron N, Morgan PB, Jagpal R. Validation of computer morphs for grading contact lens complications. *Ophthalmic Physiol Opt* 2002;22 (4):341-9.
22. Peterson RC, Wolffsohn JS. Objective grading of the anterior eye. *Optom Vis Sci* 2009;86 (3):273-8.
23. Perez-Cabre E, Millan MS, Abril HC, Otxoa E. Image processing of standard grading scales for objective assessment of contact lens wear complications. In: *Proc Soc Photo Opt Instrum Eng*; 2004; 2004. p. 107-12.
24. Wolffsohn JS. Incremental nature of anterior eye grading scales determined by objective image analysis. *Br J Ophthalmol* 2004;88 (11):1434-8.
25. Owen CG, Fitzke FW, Woodward EG. A new computer assisted objective method for quantifying vascular changes of the bulbar conjunctivae. *Ophthalmic Physiol Opt* 1996;16 (5):430-7.
26. Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci* 2000;41 (3):687-91.
27. Villumsen J, Ringquist J, Alm A. Image-analysis of conjunctival hyperemia - a personal-computer based system. *Acta Ophthalmol (Copenh)* 1991;69 (4):536-9.
28. Willingham FF, Cohen KL, Coggins JM, Tripoli NK, Ogle JW, Goldstein GM. Automatic quantitative measurement of ocular hyperemia. *Curr Eye Res* 1995;14 (11):1-8.
29. Horak F, Berger U, Menapace R, Schuster N. Quantification of conjunctival vascular reaction by digital imaging. *J Allergy Clin Immunol* 1996;98 (3):495-500.
30. Gescheider GA. *Psychophysics: Method, Theory, and Application*. Hillsdale, New Jersey: Lawrence Earlbaum Associates, Inc.; 1985.

31. Kuehni RG. Color Space and Its Divisions: color order from antiquity to the present. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2003.
32. Lin LI. A concordance correlation-coefficient to evaluate reproducibility. *Biometrics* 1989;45 (1):255-68.
33. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1 (8476):307-10.
34. Norman GR, Streiner DL. PDQ Statistics, 3rd ed. Hamilton, ON: BC Decker Inc.; 2003.
35. CCLRU. CCLRU grading scales. In: Contact Lenses Vol. 4th. London: Elsevier Butterworth-Heinemann, 1997: 863-7.
36. Morgan P, Frankish C. Image quality, compression and segmentation in medicine. *J Vis Commun Med* 2002;25 (4):149-54.
37. Baxes GA. The digital image. In: Digital image processing - principles and applications. New York: John Wiley & Sons, Inc., 1994: 37-67.

CHAPTER 6

1. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. Part I. *Am J Optom Physiol Opt* 1987;64 (4): 246-50.
2. Terry RL, Schnider CM, Holden BA, Cornish R, Grant T, Sweeney D, et al. CCLRU standards for success of daily and extended wear contact lenses. *Optom Vis Sci* 1993;70 (3): 234-43.
3. Efron N. Clinical application of grading scales for contact lens complications. *Optician* 1997;213 (5604): 26-35.
4. Efron N. Grading scales for contact lens complications. In: *Contact Lens Complications*. Oxford; Boston: Butterworth-Heinemann, 1999. Appendix A: 171-9.
5. Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci* 2000;41 (3): 687-91.
6. Chong T, Simpson T, Fonn D. The repeatability of discrete and continuous anterior segment grading scales. *Optom Vis Sci* 2000;77 (5): 244-51.
7. MacKinven J, McGuinness CL, Pascal E, Woods RL. Clinical grading of the upper palpebral conjunctiva of non-contact lens wearers. *Optom Vis Sci* 2001;78 (1): 13-8.
8. Murphy PJ, Lau JS, Sim MM, Woods RL. How red is a white eye? Clinical grading of normal conjunctival hyperaemia. *Eye* 2006;21: 633-8.
9. Sorbara L, Simpson T, Duench S, Schulze M, Fonn D. Comparison of an objective method of measuring bulbar redness to the use of traditional grading scales. *Cont Lens Anterior Eye* 2007;30 (1): 53-9.
10. Schulze M, Jones D, Simpson T. The development of validated bulbar redness grading scales. *Optom Vis Sci* 2007;84 (10): 976-83.
11. Peterson RC, Wolffsohn JS. Sensitivity and reliability of objective image analysis compared to subjective grading of bulbar hyperaemia. *Br J Ophthalmol* 2007;91 (11): 1464-6.
12. Schulze M, Hutchings N, Simpson T. The use of fractal analysis and photometry to estimate the accuracy of bulbar redness grading Scales. *Invest Ophthalmol Vis Sci* 2008;49 (4): 1398-406.
13. Peterson RC, Wolffsohn JS. Objective grading of the anterior eye. *Optom Vis Sci* 2009;86 (3): 273-8.
14. Schulze M, Hutchings N, Simpson T. The perceived bulbar redness of clinical grading scales. *Optom Vis Sci* 2009;86 (11): 1250-8.

15. Andersen JS, Davies IP, Kruse A, Lofstrom T, Ringman LA. Handbook of Contact Lens Management: Vistakon; 1996.
16. Lofstrom T, Anderson JS, Kruse A. Tarsal Abnormalities: A New Grading System. CLAO J 1998;24 (4): 210-5.
17. Sickenberger W. Klassifikation von Spaltlampenbefunden - Ein praxisnahes Handbuch für Kontaktlinsenanpasser, 1st ed. Großostheim: Ciba Vision Vertriebs GmbH; 2001.
18. IER Grading Scales. Institute for Eye Research, Sydney, Australia. Available at: <http://www.siliconehydrogels.org/resources/index.asp>. Last accessed: 05/19/2009.
19. Wiegleb M, Sickenberger W. Optimization of Grading Scales to Classify Slit Lamp Findings. In: BCLA Conference Manual; 2009; Manchester; 2009. p. 100.
20. Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. Ophthalmic Physiol Opt 2001;21 (1): 17-29.
21. Perez-Cabre E, Millan MS, Abril HC, Otxoa E. Image processing of standard grading scales for objective assessment of contact lens wear complications. In: Proc Soc Photo Opt Instrum Eng; 2004; 2004. p. 107-12.
22. Wolffsohn JS. Incremental nature of anterior eye grading scales determined by objective image analysis. Br J Ophthalmol 2004;88 (11): 1434-8.
23. Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects of scaling. Invest Ophthalmol Vis Sci 1991;32 (2): 422-32.
24. Lloyd M. Lies, statistics, and clinical significance. J Brit Contact Lens Assoc 1992;15 (2): 67-70.
25. Efron N. Grading scales for contact lens complications. Ophthalmic Physiol Opt 1998;18 (2): 182-6.
26. Woods R. Quantitative slit lamp observations in contact lens practice. J Brit Contact Lens Assoc 1989;12: 42-5.
27. Efron N. Grading scales. Optician 2000;219 (5733): 44-5.
28. Fieguth P, Simpson T. Automated measurement of bulbar redness. Invest Ophthalmol Vis Sci 2002;43 (2): 340-7.
29. Lin LI. A concordance correlation-coefficient to evaluate reproducibility. Biometrics 1989;45 (1): 255-68.
30. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1 (8476): 307-10.

31. Stevens SS. On the Theory of Scales of Measurement. Science 1946;103 (2684): 677-80.
32. Pointer MR. Measuring visual appearance - a framework for the future; 2003. Report No.: NPL COAM 19.
33. Land EH, McCann JJ. Lightness and Retinex Theory. J Opt Soc Am 1971;61 (1): 1-11.

CHAPTER 7

1. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. Part I. *Am J Optom Physiol Opt* 1987;64 (4): 246-50.
2. IER Grading Scales. Institute for Eye Research, Sydney, Australia. Available at: <http://www.siliconehydrogels.org/resources/index.asp>. Last accessed: 05/19/2009.
3. Efron N. Clinical application of grading scales for contact lens complications. *Optician* 1997;213 (5604): 26-35.
4. Efron N. Grading scales. *Optician* 2000;219 (5733): 44-5.
5. Schulze M, Jones D, Simpson T. The development of validated bulbar redness grading scales. *Optom Vis Sci* 2007;84 (10): 976-83.
6. Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects of scaling. *Invest Ophthalmol Vis Sci* 1991;32 (2): 422-32.
7. Kahn HA, Leibowitz H, Ganley JP, Kini M, Colton T, Nickerson R, et al. Standardizing diagnostic procedures. *Am J Ophthalmol* 1975;79 (5): 768-75.
8. Terry R, Sweeney D, Wong R, Papas E. Variability of clinical investigators in contact lens research. *Optom Vis Sci* 1995;72 (suppl 12): 16.
9. Chong T, Simpson T, Fonn D. The repeatability of discrete and continuous anterior segment grading scales. *Optom Vis Sci* 2000;77 (5): 244-51.
10. Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci* 2000;41 (3): 687-91.
11. Peterson RC, Wolffsohn JS. Sensitivity and reliability of objective image analysis compared to subjective grading of bulbar hyperaemia. *Br J Ophthalmol* 2007;91 (11): 1464-6.
12. Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. *Ophthalmic Physiol Opt* 2001;21 (1): 17-29.
13. Fieguth P, Simpson T. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci* 2002;43 (2): 340-7.
14. Poynton CA. Frequently Asked Questions about Color. 1997. Updated 2006/11/28/. Available at: <http://www.poynton.com/PDFs/ColorFAQ.pdf> Accessed: 2009, 03/17.
15. Peterson RC, Wolffsohn JS. Objective grading of the anterior eye. *Optom Vis Sci* 2009;86 (3): 273-8.

16. Perez-Cabre E, Millan MS, Abril HC, Otxoa E. Image processing of standard grading scales for objective assessment of contact lens wear complications. In: Proc Soc Photo Opt Instrum Eng; 2004; 2004. p. 107-12.
17. Wolffsohn JS. Incremental nature of anterior eye grading scales determined by objective image analysis. *Br J Ophthalmol* 2004;88 (11): 1434-8.
18. Schulze M, Hutchings N, Simpson T. The use of fractal analysis and photometry to estimate the accuracy of bulbar redness grading Scales. *Invest Ophthalmol Vis Sci* 2008;49 (4): 1398-406.
19. Pult H, Murphy PJ, Purslow C, Nyman J, Woods RL. Limbal and bulbar hyperaemia in normal eyes. *Ophthalmic Physiol Opt* 2008;28 (1): 13-20.
20. Woods R. Quantitative slit lamp observations in contact lens practice. *J Brit Contact Lens Assoc* 1989;12: 42-5.
21. Schulze M, Hutchings N, Simpson T. The perceived bulbar redness of clinical grading scales. *Optom Vis Sci* 2009;86 (11): 1250-8.
22. Schulze M, Hutchings N, Simpson T. The conversion of bulbar redness grades using psychophysical scaling. *Optom Vis Sci* in press.
23. Taylor BN, Kuyatt CE. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, Appendix D.1: Terminology. Gaithersburg, MD: National Institute of Standards and Technology; 1994.
24. Willingham FF, Cohen KL, Coggins JM, Tripoli NK, Ogle JW, Goldstein GM. Automatic quantitative measurement of ocular hyperemia. *Curr Eye Res* 1995;14: 1101-8.
25. Guillon M, Shah D. Objective measurement of contact lens-induced conjunctival redness. *Optom Vis Sci* 1996;73 (9): 595-605.
26. Owen CG, Fitzke FW, Woodward EG. A new computer assisted objective method for quantifying vascular changes of the bulbar conjunctivae. *Ophthalmic Physiol Opt* 1996;16 (5): 430-7.
27. Simpson T, Chan A, Fonn D. Measuring ocular redness: first order (luminance & chromaticity) measurements provide more information than second order (spatial structure) measurements. *Optom Vis Sci* 1998;75 (suppl 12): 279.
28. Wolffsohn JS, Purslow C. Clinical monitoring of ocular physiology using digital image analysis. *Cont Lens Anterior Eye* 2003;26 (1): 27-35.
29. Avakian A, Kalina RE, Sage EH, Rambhia AH, Elliott KE, Chuang EL, et al. Fractal analysis of region-based vascular change in the normal and non-proliferative diabetic retina. *Curr Eye Res* 2002;24 (4): 274-80.
30. Family F, Masters BR, Platt DE. Fractal pattern formation in human retinal vessels. *Physica D* 1989;38 (1-3): 98-103.

31. Masters BR. Fractal analysis of human retinal vessels. In: Proc Soc Photo Opt Instrum Eng; 1990; 1990. p. 250-6.
32. Masters BR. Fractal analysis of the vascular tree in the human retina. Annu Rev Biomed Eng 2004;6 (1): 427-52.
33. Rasband WS. ImageJ Version 1.38x (07/13/2007). Available at: <http://rsb.info.nih.gov/ij/>. U. S. National Institutes of Health, Bethesda, Maryland, USA.
34. Fisher R, Perkins S, Walker A, Wolfart E. HIPR (Hypermedia Image Processing Reference). Edinburgh, UK; 2004. Available at: <http://homepages.inf.ed.ac.uk/rbf/HIPR2/histeq.htm>. Accessed: 2009, 09/14.
35. Villumsen J, Ringquist J, Alm A. Image-analysis of conjunctival hyperemia - a personal-computer based system. Acta Ophthalmol (Copenh) 1991;69 (4): 536-9.
36. Karperien A. FraCLac for ImageJ - FraCLac Advanced User's Manual v.2.5. 2007. Available at: <http://rsbweb.nih.gov/ij/plugins/fraclac/FLHelp/Introduction.htm>. Accessed: 2009, 07/29.
37. Bartko JJ. Measures of Agreement - A Single Procedure. Stat Med 1994;13 (5-7): 737-45.
38. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86 (2): 420-8.
39. Streiner DL, Norman GR. Health Measurement Scales - A practical guide to their development and use. New York: Oxford University Press Inc.; 1995.
40. Lin LI. A concordance correlation-coefficient to evaluate reproducibility. Biometrics 1989;45 (1): 255-68.
41. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1 (8476): 307-10.
42. Norman GR, Streiner DL. PDQ Statistics, 3rd ed. Hamilton, ON: BC Decker Inc.; 2003.
43. 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 (2): 119-24.

CHAPTER 8

1. Fieguth P, Simpson T. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci* 2002;43 (2): 340-7.
2. Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. *Ophthalmic Physiol Opt* 2001;21 (1): 17-29.
3. Wolffsohn JS. Incremental nature of anterior eye grading scales determined by objective image analysis. *Br J Ophthalmol* 2004;88 (11): 1434-8.
4. Masters BR. Fractal analysis of human retinal vessels. In: *Proc Soc Photo Opt Instrum Eng*; 1990; 1990. p. 250-6.
5. Masters BR. Fractal analysis of the vascular tree in the human retina. *Annu Rev Biomed Eng* 2004;6 (1): 427-52.
6. Avakian A, Kalina RE, Sage EH, Rambhia AH, Elliott KE, Chuang EL, et al. Fractal analysis of region-based vascular change in the normal and non-proliferative diabetic retina. *Curr Eye Res* 2002;24 (4): 274-80.
7. Family F, Masters BR, Platt DE. Fractal pattern formation in human retinal vessels. *Physica D* 1989;38 (1-3): 98-103.
8. Masters BR. Fractal analysis of human retinal blood vessel patterns: developmental and diagnostic aspects. In: Masters BR, editor. *Noninvasive diagnostic techniques in ophthalmology*. New York: Springer-Verlag, 1990. 515-27.
9. Simpson T, Chan A, Fonn D. Measuring ocular redness: first order (luminance & chromaticity) measurements provide more information than second order (spatial structure) measurements. *Optom Vis Sci* 1998;75 (suppl 12): 279.
10. Peterson RC, Wolffsohn JS. Objective grading of the anterior eye. *Optom Vis Sci* 2009;86 (3): 273-8.
11. Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects of scaling. *Invest Ophthalmol Vis Sci* 1991;32 (2): 422-32.
12. Lloyd M. Lies, statistics, and clinical significance. *J Brit Contact Lens Assoc* 1992;15 (2): 67-70.
13. Chong T, Simpson T, Pritchard N, Dumbleton K, Richter D, Fonn D. Repeatability of discrete and continuous clinical grading scales. *Optom Vis Sci* 1996;73 (suppl 12): 232.
14. Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci* 2000;41 (3): 687-91.

15. Peterson RC, Wolffsohn JS. Sensitivity and reliability of objective image analysis compared to subjective grading of bulbar hyperaemia. *Br J Ophthalmol* 2007;91 (11): 1464-6.
16. Efron N. Clinical application of grading scales for contact lens complications. *Optician* 1997;213 (5604): 26-35.
17. Efron N. Grading scales for contact lens complications. *Ophthalmic Physiol Opt* 1998;18 (2): 182-6.
18. Guillon M, Shah D. Objective measurement of contact lens-induced conjunctival redness. *Optom Vis Sci* 1996;73 (9): 595-605.
19. Schulze M, Jones D, Simpson T. The development of validated bulbar redness grading scales. *Optom Vis Sci* 2007;84 (10):976-83.
20. Merriam Webster. Dictionary. 2009. Available at: <http://www.merriam-webster.com/dictionary/>. Accessed: 2009, 07/28.
21. Cardona G, Seres C. Grading contact lens complications: the effect of knowledge on grading accuracy. *Curr Eye Res* 2009;34 (12): 1074-81.
22. Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 1971;9 (1): 97-113.
23. Land EH. Retinex Theory of Color-Vision. *Sci Am* 1977;237 (6): 108-28.

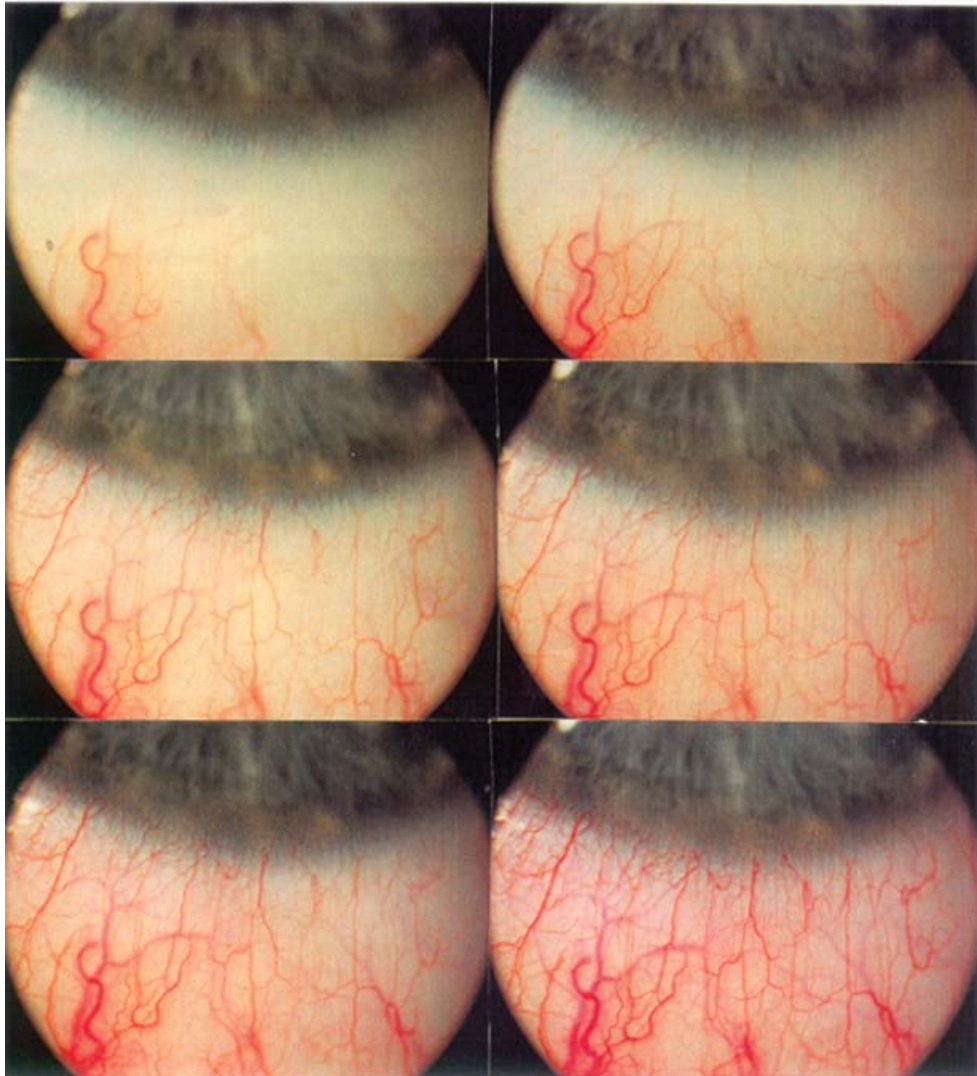
Appendix

Appendix A: Original Grading Scales

McMonnies/Chapman-Davies

HYPERAEMIA SCALE

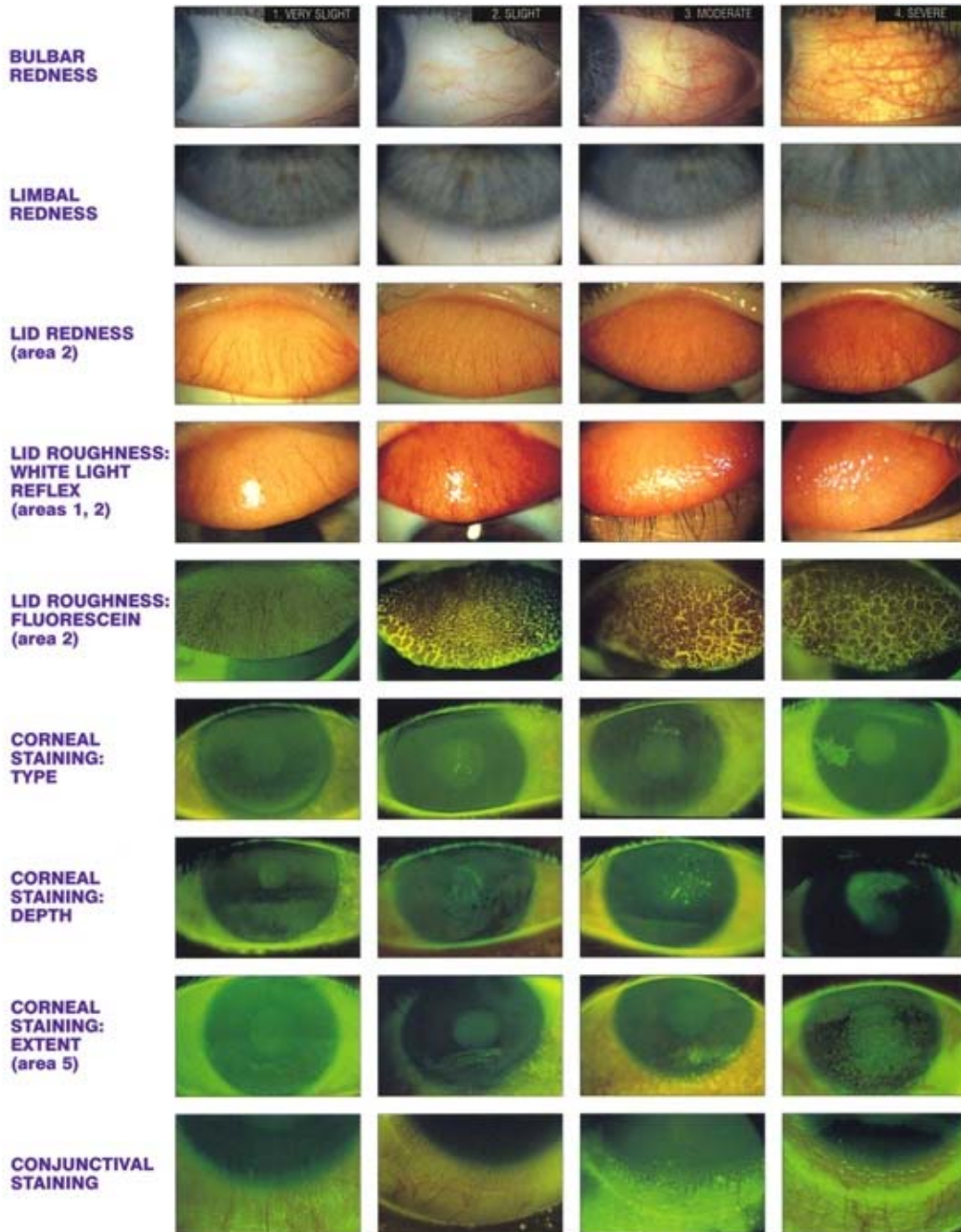
CHARLES W. McMONNIES & ANTHONY CHAPMAN-DAVIES



IER scale (front)



GRADING SCALES



Provided by THE VISION CARE INSTITUTE
A Division of Johnson & Johnson Vision Care, Inc.

AS-07-98-00



GRADING SCALES

APPLICATION OF GRADING SCALES

- Patient management is based on how much the normal ocular appearance has changed.
- In general, a rating of slight (grade 2) or less is considered within normal limits (except staining).
- A change of one grade or more at follow up visits is considered clinically significant.

PALPEBRAL CONJUNCTIVAL GRADES



- The palpebral conjunctiva is divided into five areas to grade redness and roughness.
- Areas 1, 2 and 3 are most relevant in contact lens wear.

ADVERSE EFFECTS WITH CONTACT LENSES

CLPC CONTACT LENS PAPILLARY CONJUNCTIVITIS

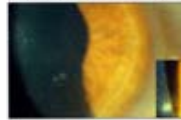
Inflammation of the upper palpebral conjunctiva



- Signs**
- Redness
 - Enlarged papillae
 - Excess mucus
- Symptoms**
- Itchiness
 - Mucus strands
 - Lens mislocation
 - Intolerance to lenses

INFILTRATES

Accumulation of inflammatory cells in corneal sub-epithelial stroma.
Inset: high magnification view



- Signs**
- Whitish opacity (focal) or grey haze (diffuse)
 - Usually confined to 2-3mm from limbus
 - Localized redness
- Symptoms**
- Asymptomatic or scratchy, foreign body sensation
 - Redness, tearing and photophobia possible

CLARE CONTACT LENS REDUCED RED EYE

An acute corneal inflammatory episode associated with sleeping in soft contact lenses



- Signs**
- Unilateral
 - Intense redness
 - Infiltrates
 - No epithelial break
- Symptoms**
- Wakes with irritation or pain
 - Photophobia
 - Lacrimation

POLYMEGATHISM



VASCULARIZATION



Vessel extension beyond translucent limbal zone is recorded (mm)

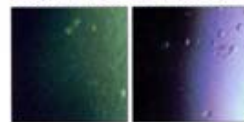
STROMAL STRIAE and FOLDS



One striae = 5% edema
One fold = 8% edema
(each additional striae or fold indicates 1% more edema)

Record number observed

MICROCYSTS and VACUOLES



Located in epithelium. Identified by side showing brightness.

Record number observed

CORNEAL STAINING GRADES

- Staining assessed immediately after single instillation of fluorescein using cobalt blue light and wratten 12 (yellow) filter over the slit lamp objective.
- The cornea is divided into five areas. The type, extent and depth of staining are graded in each area.



- Type**
- 1 Micropunctate
 - 2 Macropunctate
 - 3 Coalescent macropunctate
 - 4 Patch

Extent: Surface area

- 1 1 - 15%
- 2 16 - 30%
- 3 31 - 45%
- 4 > 45%

Depth*

- 1 Superficial epithelium
- 2 Deep epithelium, delayed stromal glow
- 3 Immediate localized stromal glow
- 4 Immediate diffuse stromal glow

*Based on penetration of fluorescein and slit lamp optic section

EROSION

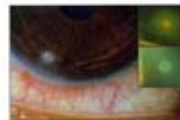
Full thickness epithelial loss over a discrete area



- Signs**
- No stromal inflammation
 - Immediate spread of fluorescein into stroma
- Symptoms**
- Can be painful
 - Photophobia
 - Lacrimation

CLPU CONTACT LENS PERIPHERAL ULCER

Round, full thickness epithelial loss with inflamed base, typically in the corneal periphery which results in a scar. Inset: with fluorescein, scar



- Signs**
- Unilateral, "white spot"
 - Localized redness
 - Infiltrates
 - Post healing scar
- Symptoms**
- Varies from foreign body sensation to pain
 - Lacrimation and photophobia may occur

INFECTED ULCER

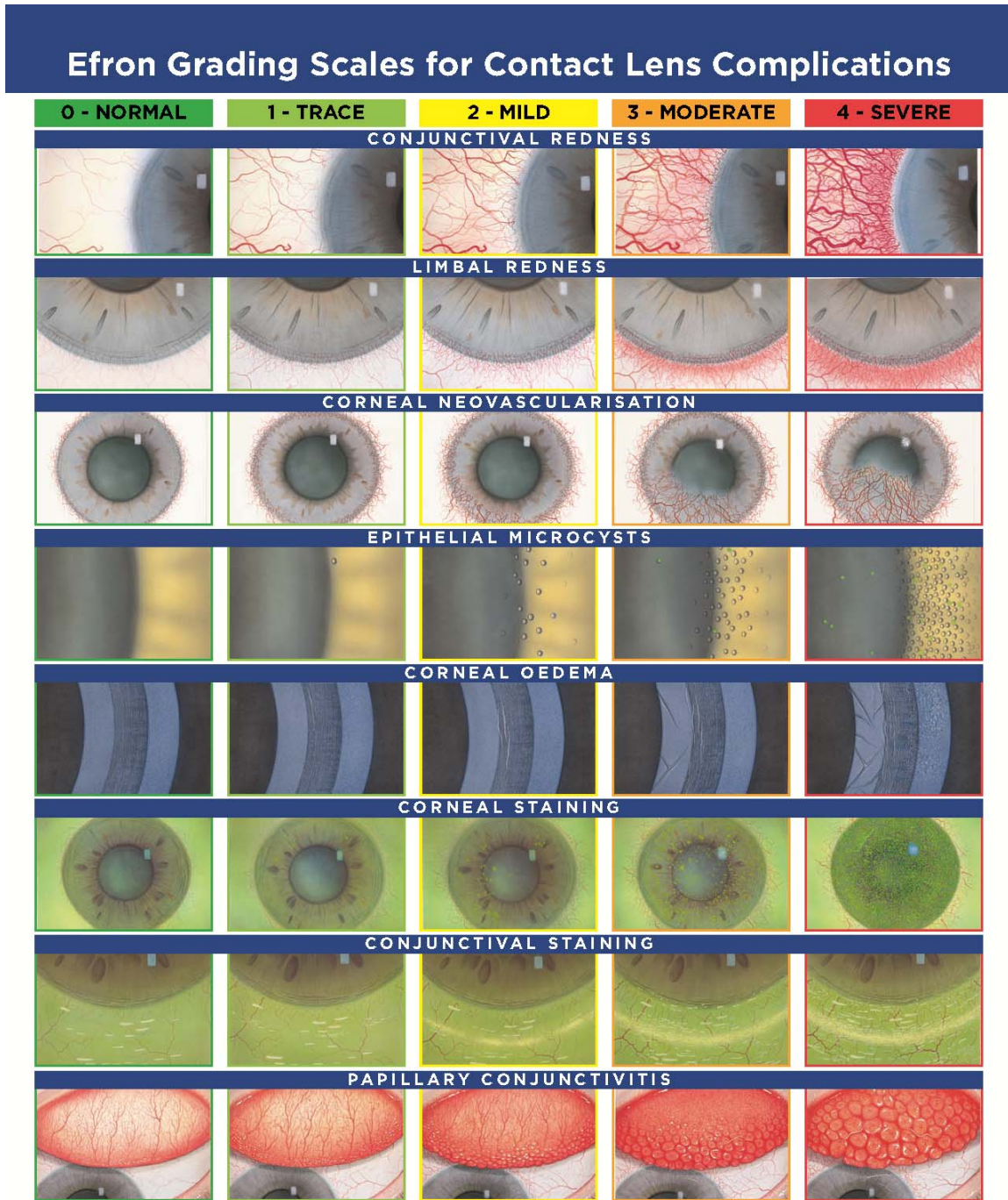
Full thickness epithelial loss with stromal necrosis and inflammation, typically central or paracentral



- Signs**
- Intense redness
 - "White patch" (raised edges)
 - Infiltrates
 - Epithelial and stromal loss
 - Anterior chamber flare
 - Conjunctival and lid edema
- Symptoms**
- Pain, photophobia
 - Redness, mucoid discharge
 - VA (if over pupil)



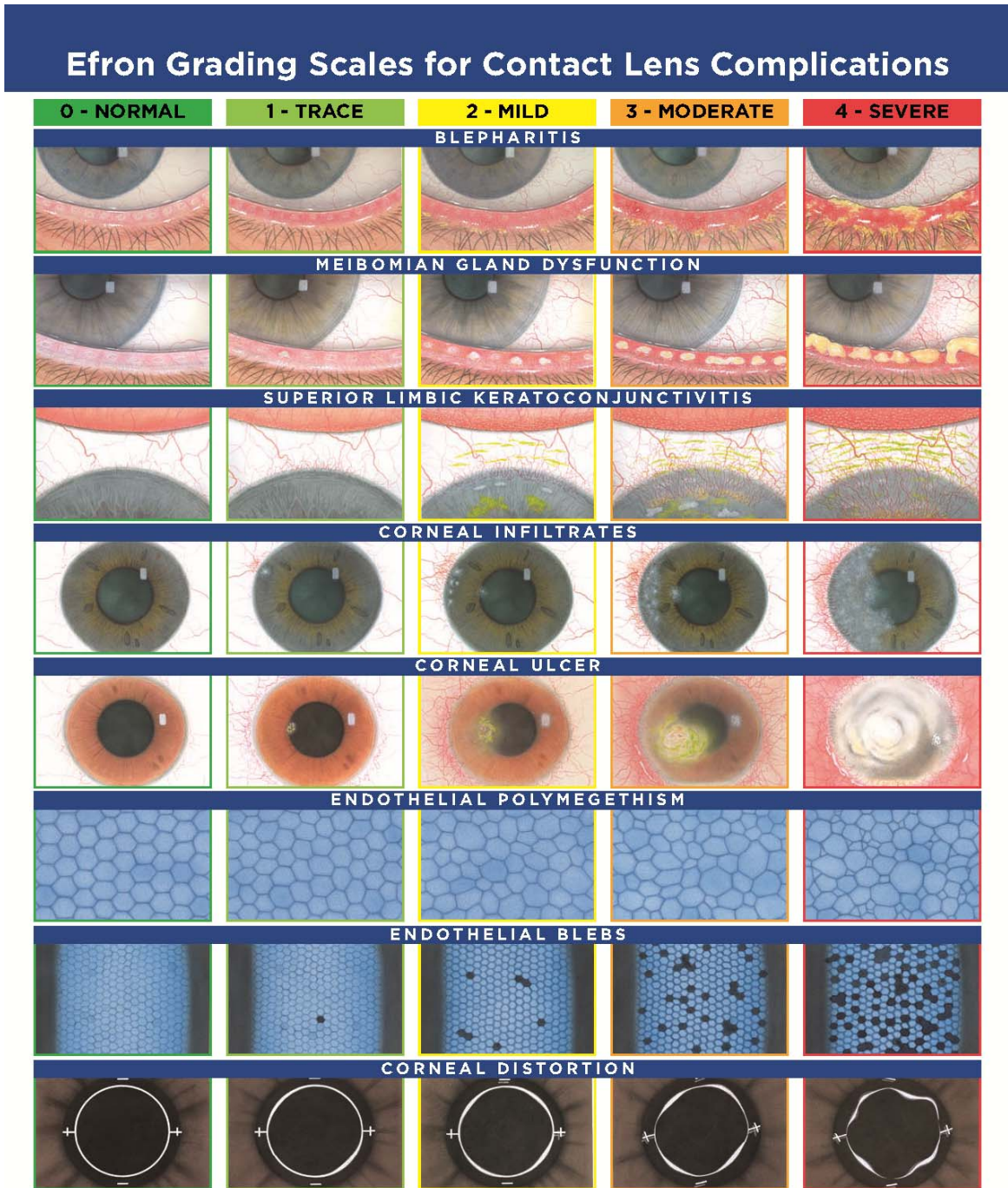
Efron scale (front)



IMPORTANT NOTE: This grading scale, along with the instructions for use, were developed by Professor Nathan Efron. The grading scale is offered as an educational tool that you may choose to use as part of your patient evaluations. These materials are not intended as, and do not constitute, medical or optometric advice.



Efron scale (back)



"Used by permission from Nathan Efron and Butterworth-Heinemann/Elsevier"
 Supplement to the book *Contact Lens Practice*, 2nd edition by Nathan Efron,
 published by Butterworth-Heinemann, 2010, ISBN 978-0-7506-8869-7



Appendix B: List of Abbreviations

$\% PC$	% pixel coverage
BIPM	Bureau International des Poids et Mesures (English: International Bureau of Weights and Measures)
CCC	Correlation coefficient of concordance
CCLRU	Cornea and Contact Lens Research Unit
CIE	Commission Internationdale d'Eclairage (English: International Commission on Illumination)
COR	Coefficient of repeatability
CRT	cathode ray tube (monitor)
d	Euclidean dimension (integer)
\bar{d}	mean of the differences (between test and retest)
D	Fractal (or fractional) dimension
\bar{D}	Averaged fractal dimension
D_B	Box-counting fractal dimension
D_e	Most-efficient covering fractal dimension
D_{sc}	Slope-corrected fractal dimension
D_{sce}	Slope-corrected most-efficient covering fractal dimension
DIP	digital image processing
dpi	dots per inch
ICC	Intraclass correlation coefficient
HSD	Honestly significant differences (Tukey test)
IER	Institute for Eye Research
ISO	International Organization for Standardization

LCD	liquid crystal display (monitor)
LOA	Limits of agreement ($\bar{d} \pm 1.96 * s_d$)
LSD	Least significant differences (Fisher test)
MC-D	McMonnies/Chapman-Davies (scale)
Pearson's r	Pearson's product moment correlation coefficient
ppi	pixels per inch
PR	perceived redness
PR _B	perceived redness of the base scale image
PR _{TH}	perceived redness of the higher target scale image enclosing the base scale image
PR _{TL}	perceived redness of the lower target scale image enclosing the base scale image
PRA	physical redness attribute
ρ	Spearman's rank-order coefficient; Spearman's rho
sd	standard deviation
s_d	standard deviation of the differences (between test and retest)
SG _B	Base scale grade
SG _T	Target scale grade
SG _{TL}	Lower scale grade of the 2 target scale images enclosing the base scale image
VBR	Validated Bulbar Redness (scale)
VIM	Vocabulary of Metrology

Appendix C: Copyright Permissions

Investigative Ophthalmology and Vision Science (Chapter 4)

Subject: RE: Request for permission to use article in doctoral thesis
From: "Debbie Chin" <dchin@arvo.org>
Date: Fri, 17 Apr 2009 17:45:55 -0400
To: "Marc Schulze" <m3schulz@sciborg.uwaterloo.ca>

Dear Dr. Schulze,

Permission is hereby granted to reprint the following article in your doctoral thesis:

Schulze MM, Hutchings N, Simpson TL. The use of fractal analysis and photometry to estimate the accuracy of bulbar redness grading scales. *Invest Ophthalmol Vis Sci.* 2008;49:1398-1406.

A reprint of this material must include a full article citation and acknowledge the Association for Research in Vision and Ophthalmology as the copyright holder.

Regards,
Debbie Chin

IOVS Editorial Office
12300 Twinbrook Parkway, Suite 250
Rockville, MD 20852 U.S.A.
Direct: +1.240.221.2926 | Fax: +1.240.221.0355
www.iovs.org

Reducing Disparities
In Eye Disease and Treatment
ARVO 2009 Annual Meeting
Fort Lauderdale, Florida
May 3-7

From: Marc Schulze [mailto:m3schulz@sciborg.uwaterloo.ca]
Sent: Thursday, April 16, 2009 4:54 PM
To: IOVS Email Account
Cc: Marc Schulze
Subject: Request for permission to use article in doctoral thesis

To whom it may concern,

I would like permission to use a published paper in IOVS of which I am the main author in my doctoral thesis. Here is the reference:

Schulze M, Hutchings N, Simpson T. The Use of Fractal Analysis and Photometry to Estimate the Accuracy of Bulbar Redness Grading Scales. *Invest Ophthalmol Vis Sci.* 2008;49(4):1398-1406.

Please let me know if you would charge for this request.

Thank you very much,

Regards,

Marc Schulze

Optometry and Vision Science (Chapter 5)

RE: Permission to use published work in thesis

Subject: RE: Permission to use published work in thesis
From: "Zadnik, Kurt" <ovs@osu.edu>
Date: Wed, 5 Aug 2009 09:58:33 -0400
To: Marc Schulze <m3schulz@sciborg.uwaterloo.ca>

-->

HI Marc,

Permission is granted, but not until your accepted manuscript is formally published (after November 4, 2009).
Please use the following credit line with the manuscript:

R20;Schulze MM, Hutchings N, Simpson TL. The perceived bulbar redness of clinical grading scales. *Optom Vis Sci* 2009;86;add page numbers. Reprinted with permission. ©The American Academy of Optometry 2009."

Please let me know if you have any questions.

Best regards,

Kurt

Optometry and Vision Science
Kurt A. Zadnik, Managing Editor
The Ohio State University, College of Optometry
338 West 10th Avenue
Columbus, OH 43210
Tel: (614) 292-4942; Fax: (614) 292-4949;
E-mail: ovs@osu.edu
<http://ovs.edmgr.com>

From: Marc Schulze [mailto:m3schulz@sciborg.uwaterloo.ca]
Sent: Tuesday, August 04, 2009 11:35 AM
To: Zadnik, Kurt
Subject: Permission to use published work in thesis

Dear Sir / Madam,

I would like to seek your permission to use my article which is currently "In Press" in *Optometry and Vision Science* for my Doctoral thesis. The reference is OVS9075: MM Schulze, N Hutchings, TL Simpson. The Perceived Bulbar Redness of Clinical Grading Scales. *Optom Vis Sci*. In press (scheduled for Nov. issue). The thesis will be completed late November, early December 2009.

Optometry and Vision Science (Chapter 6)

RE: Request for permission to use article in Doctoral thesis

Subject: RE: Request for permission to use article in Doctoral thesis
From: "Zadnik, Kurt" <ovs@osu.edu>
Date: Mon, 21 Dec 2009 14:25:23 -0500
To: Marc Schulze <m3schulz@sciborg.uwaterloo.ca>

-->

Hi Marc,

Permission is granted. Please use the following credit line with each figure:

Schulze MM, Hutchings N, Simpson TL. The conversion of bulbar redness grades using psychophysical scaling. *Optom Vis Sci* 2008;87:in press. Reprinted with permission. ©The American Academy of Optometry 2010."

Page numbers may be available before your thesis is completed, so keep in touch. Let me know if you have any questions.

Happy Holidays,

Kurt

Optometry and Vision Science
Kurt A. Zadnik, Managing Editor
The Ohio State University, College of Optometry
338 West 10th Avenue
Columbus, OH 43210
Tel: (614) 292-4942; Fax: (614) 292-4949;
E-mail: ovs@osu.edu
<http://ovs.edmgr.com>

From: Marc Schulze [mailto:m3schulz@sciborg.uwaterloo.ca]
Sent: Monday, December 21, 2009 11:43 AM
To: Zadnik, Kurt
Subject: Request for permission to use article in Doctoral thesis

Dear Kurt,

I would like to seek your permission to use my article which is currently "In Press" in *Optometry and Vision Science* for my Doctoral thesis. The reference is OVS9131: MM Schulze, N Hutchings, TL Simpson. The Conversion of Bulbar Redness Grades using Psychophysical Scaling. *Optom Vis Sci*. In press (scheduled for March 2010 issue). The thesis will be completed by mid January, but will only be available online 4 months after that date (mid May).