View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Waterloo's Institutional Repository
manuscript version is made available under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/

Development of a new grading scale for tear ferning

Ali M Masmali[1,3,*], Christine Purslow[2,3], Paul J Murphy[4]

[1] *Cornea Research Chair (CRC), Optometry Department, College of Applied Medical Sciences, King Saud University, Riyadh, Saudi Arabia*

[2] *School of Health Professions, Plymouth University, Plymouth, UK*

[3] *Contact Lens and Anterior Eye Research (CLAER) Unit, School of Optometry and Vision Sciences, Cardiff University, Cardiff, UK*

[4] *School of Optometry and Vision Sciences, University of Waterloo, Waterloo, Canada*

Running title: New Grading Scale for Tear Ferning

<u>Corresponding author</u>

Dr. Ali M. Masmali

Cornea Research Chair (CRC), Optometry Department

College of Applied Medical Sciences

King Saud University

P.O. Box 10219, Riyadh 11433

Saudi Arabia

amasmali@ksu.edu.sa

Tel: +966 1 4693547

Fax: +966 1 4693536

None of the authors has any proprietary interest in this manuscript.

Abstract

Purpose: This paper reports on the development of a new tear ferning (TF) subjective grading scale, and compares it with the Rolando scale.

Method: TF patterns obtained from tear film samples collected from normal and dry eye subjects in previous studies were collated into a large image library. From this library, 60 images were selected, to represent the full range of possible TF patterns, and a further sub-set of 15 images was chosen for analysis. Twenty-five optometrists were asked to rank the images in increasing order between extreme anchors on a scale of TF patterns. Interim statistical analysis of this ranking found 7 homogeneous sub-sets, where the image rankings overlapped for a group of images. A representative image (typically the mean) from each group was then adopted as the grade standard. Using this new 7-point grading scale, 25 optometrists were asked to grade the entire 60 image library at two sessions: once using the 4-point Rolando scale and once using the new 7-point scale, applying 0.25 grade unit interpolation.

Results: Statistical analysis found that, for the larger image set, the Rolando scale produced 3 homogeneous sub-sets, and the 7-point scale produced 5 homogeneous sub-sets. With this refinement, a new 5-point TF scale (Grades 0–4) was obtained.

Conclusions: The Rolando grading scale lacks discrimination between its Type I and II grades, reducing its reliability. The new 5-point grading scale is able to differentiate between TF patterns, and may provide additional support for the use of TF for both researcher and clinician.

Keywords: tear ferning, dry eye, grading scale

Introduction

The chemical analysis of tear film composition is difficult due to the small volumes available, and to the transparent and dynamic nature of tears [1]. Clinicians and scientists recognise that biochemical analysis of osmolarity and other key components in a tear sample is the way forward, but the small volumes involved make biochemical analysis particularly challenging [2,3]. Techniques available are limited by the need for expensive equipment that is difficult to use under normal clinical conditions [4]. A simple, clinical tear film test, that is quick and inexpensive to perform, and can indicate the biochemical properties of the tear film, would be very useful.

One potential and clinically suitable test involves drying a tear sample on a glass microscope slide to produce a crystallisation pattern in the form of a fern [5–7]. This phenomenon occurs with many body fluids and follows a characteristic formation process. The first discovery of tear crystallisation was reported by Fourcroy and Vauquelin in 1791 [8], but remained unnoted until 1946, when observed by Papanicolaou during studying cervical mucus [9]. Ferning patterns have been used to test different body fluids, such as vaginal and cervical mucus as an indicator of the menstrual cycle [10], oestrogen activity and ovulation [11–14] and early pregnancy [13,15]. Ferning has also been used to test saliva [16], to consider the observation of salivary ferning as a new technique for determining the fertile period [17], and to correlate salivary ferning and the fertile period [18], and using of salivary ferning in ovulation detection in family planning [19].

Crystallisation begins with the formation of a nucleus, consisting of a regularly arranged number of ions. The nucleus is formed by aggregation when the solute evaporates and dissolved ions are concentrated until super-saturation of the tear film is reached [7]. The nucleation process begins at the peripheral edges of the drop, where the solution is thinnest and super-saturation is reached rapidly [7]. Each nucleus has the ability to grow into a large crystal unit with the addition of more ions, and, so long as the sample solute is able to diffuse into areas with a lower solute concentration area, normal crystals can form. This requires a slow growth rate, low solution viscosity and low impurity levels to permit free solute diffusion.

The absence of these conditions can lead to dendritic crystal growth [20]. In this situation the stems grow longer and branch at regular intervals along the main stem. The reason for this regularity is not understood [7], but it is known that fern-like dendritic growth can be promoted by increasing the evaporation rate of the drop, by reducing atmospheric humidity, by increasing the drying temperature, or when impurities are present in low concentration, which acts as additional nuclei for crystal deposition [7].

Since tears are a complex solution, with many organic and non-organic components, the tear fern pattern produced by drying a sample depends on the composition of the tear sample [4,7]. This variation in pattern has been suggested as a simple test for tear film quality at a gross biochemical level. This phenomenon gives tear ferning the potential, and the features, to be used as a diagnostic test in the clinic [5,21]. Previous studies have demonstrated it to show good repeatability [22], sensitivity and specificity [21,23,24]

Different scales for grading tear ferning patterns have been proposed [6,21,25], with the Rolando scale being adopted as the main method used in previous published work in this area. However, the Rolando scale was not originally developed to produce a repeatable, standardised grading instrument, rather it arose from Rolando’s observation that the Type I and II patterns were found in the majority of normal eyes, while Types III and IV were found in the majority of keratoconjuctivitis sicca (KCS) eyes [6].

The main difficulty with using the Rolando scale lies with this gross categorisation of ferning patterns, restricting sensitivity – the variance around Types I and II is particularly large – and not all types of tear ferning patterns are represented by the scale [22]. If the tear ferning test is to become part of routine clinical examination of the tear film, it is important to have a grading scale that has been developed to meet the needs of the clinician, and to address the four fundamental design requirements of a grading scale [26].

The aim of this paper is to report on the development of an improved subjective grading scale for clinicians, and the comparison of the new subjective scale with the Rolando scale.

Methods

A digital image library was compiled from tear ferning patterns produced using a standardised protocol, all images were observed under digital microscope (Leica DMRA2) with 10X magnification, and all images were saved in JPEG file format [22]. In total, 560 images of tear ferning patterns were produced from tear samples collected from 157 subjects, and all images were graded to 0.25 increments of the Rolando scale, for increased sensitivity [26]. Sixty images were selected by the authors, according to Rolando's grading scale, to be representative of the full range of possible tear ferning patterns.

From the 60 image library, 15 images were further selected to represent the range of tear ferning patterns. Fifteen was judged to be a workable number for clinicians to rank at a single session in an experimental setting. Although the Rolando scale was used to assist in selecting an equal number of images across the range, this was a notional attribute used only to help in image selection.

Twenty-five experienced optometrists working in the School of Optometry and Vision Sciences at Cardiff University were presented with hard copies of the fifteen images and asked to rank the fifteen images in ascending order between two ‘anchors’ - Reference 1 (a densely branched Rolando Type I) to Reference 2 (a sparse Rolando Type IV). Each image had the same magnification (10X) and was printed to the same size (12 x 10 cm), then labelled with two random capital letters and laminated. Each volunteer was given a record sheet, with a numeric table from 1–15, on which they recorded the alpha-code of each image in the rank order they felt best matched the pattern progression between the two references images. There was no time limit given and each volunteer was reminded that there was no right or wrong ranking, only his or her opinion. A value (weighting) was assigned to each position in the ranking (*i.e*. position 1 was worth 1 point, position 2 worth 2 points, position 7 worth 7 points, *etc*.). This produced 25 weighted rankings for each image, and the average (and variance) weighting for each image was calculated (Table 1). The data was normally distributed (Kolmogorov-Smirnov; $p>0.05$). A one-way ANOVA was used to compare the score weightings attributed to each image, and a statistically significant difference ($p<0.0005$) was observed. Post-hoc Tukey HSD testing revealed seven homogeneous sub-sets, within which no statistically significant differences were found (Table 2).

The seven groups, representing the homogeneity amongst the 15 images, supported the strategy to use a single image from each group to represent the library: a new 7-item scale. The mean score of the images in each sub-set was used to select a representative image (Table 3), and the image score closest to the mean was chosen to be representative of the sub-set (Table 4). This produced seven images, selected to represent a new 7-point tear ferning grading scale (Figure 1).

This new scale was then validated against the larger sample of sixty images. Twenty-five optometrists, experienced in clinical grading attended the laboratory for two sessions. Each observer was asked to grade all sixty library images displayed *via* a random slide-show presentation (Microsoft PowerPoint). The images were displayed on the screen under identical luminance and resolution (screen size 13.3 inch, and resolution of 1280 x 800 pixels) at each session. Volunteers were provided with the Rolando scale at one visit, and the new 7-point scale at the other; with grading scale provision randomised for each observer between visits. Observers were asked to grade each image using each grading scale to 0.25 increments, rather than the preferred 0.1 increments, as interpolation of the Rolando scale to finer increments is problematic. Observers were not told which scale was a ‘new’ scale, in order to avoid bias. At the end of the session, each observer was given the option to write any comments on the ease of use of the grading scale. Furthermore, in order to assess the reproducibility of grading using the scales, five observers were asked to return for four more visits at which they repeated the grading, as above.

Data from both grading scales was not normally distributed (Kolmogorov-Smirnov; $p<0.05$), and the median grade for each image was calculated. While the appropriate statistical comparisons were made between the grades given by the 25 observers for each of the 60 library images (Kruskal-Wallis), the analysis was also repeated with ANOVA to facilitate post-hoc testing, which was used to detect/confirm homogeneous sub-sets. Reproducibility was assessed using paired testing between sessions, and mean differences (and their confidence intervals were calculated).

Results

1. Grading of image library using the Rolando Scale

The median grades for each Type were calculated (Table 5), indicating non-linearity across the scale, i.e. small difference between Types I and II, but large between Types III and IV. The variance around each grade also differed.

The non-parametric equivalent of ANOVA (Kruskal-Wallis Test) was used to compare the scores for the 60 images using the Rolando scale and a statistically significant difference was found between the grades ($p<0.001$; Figure 2). Post-hoc testing indicated that homogeneous sub-sets existed, but there was little distinction between Types I and II (Table 6).

2. Grading of image library using the new 7-point scale

The mean grade and standard deviation for each image (Figure 3) showed an overlap between Grades 2 and 3, and between Grades 6 and 7. A one-way ANOVA found a statistically significant difference between all grades ($p<0.001$), and Tukey's HDS test identified 5 homogeneous sub-sets within the 7-point scale by combining Grades 2 and 3 and Grades 6 and 7 into one grade each (Table 7). This analysis produced a final tear ferning grading scale with five images (Figure 4). When the grading scores for the over-lapping groups were combined in this new 5-point scale (Figure 5), a linear relationship between the homogeneous sub-sets was evident (Pearson, $r = 0.988$; $p<0.001$).

The new 5-point grading scale was classified from 0 to 4. The 0 grade was chosen to reflect lower limit of grading as being nothing less than zero and library image #**1** was used to represent this grade.

3. Subjective feedback on use of the 7-point scale

The observer's scoring sheet included a space for comments, and the following were written by the observers after they had used both scales:

<u>About the new 7-point scale</u>

"The current scales more accurate than the previous scales"

"More clear and easier to grade than Rolando's scales"

"I found it difficult to distinguish between grade 6 and 7 of the grading scales"

"Scales 1-7 are better than scales 1–4 as I can judge easily according to the given images as guideline"

"I like these scales much better than 4 scales (Rolando)"

<u>About Rolando's grading scales:</u>

"The Rolando's scales are harder to use than the 7 scales"

"I think the 7 scales give the examiner better tools of judgment"

"This set is more difficult to judge than the 7 scales"

"Harder than before, as had to decide what interpolation looks like. This could vary between practitioners"

"The first 7 scales are easier due to wide range of choices".

4. Reproducibility of scoring the image library

No statistically significant difference was found between sessions for grading of the image library when the 7-point scale was used (paired t-test, $p = 0.581$; coefficient of variation, 4%). In contrast, there was a significant difference in the grading of these images between the two sessions when the Rolando grading scale was used (Wilcoxon test, $p<0.001$; coefficient of variation, 6%).

Discussion

This series of studies has led to the development of a new tear ferning grading scale, which has improved discrimination and repeatability over the previous Rolando grading scale. The final 5-point grading scale demonstrated good linearity in grading score across the ferning image library, and significant differences were found between the mean scores of the 5 scales. Reproducibility between sessions was also better with the new scale compared to Rolando's scale, indicating improved reliability.

The availability of a reproducible and reliable tear ferning grading scale will help to support the evaluation and investigation of the tear film, and might contribute in the treatment of dry eye. This new grading scale offers exciting potential for both the researcher and the clinician.

The major weaknesses of the traditional Rolando grading scale are that scale has no protocol for sample preparation associated with it, the categorisation of ferning patterns is crude with large incremental steps, which restricts sensitivity, not all types of tear ferning patterns appear to be represented by the scale, and the variance around Types I and II is particularly large. Previous attempts have been made to try and improve the Rolando scale. Evans et al [27] adopted a refinement of the Rolando scale using 0.25 increments in line with Bailey et al [26], which increased the sensitivity in classification of TF patterns, but even with using these increments, classification was still restricted because there were no clear protocols in their use, and that may have produced inter- and intra-variation in examiner judgment.

Subjective grading scales come in many forms. Grading can be applied as numeric scales (*e.g.* 0–4) or as descriptive or qualitative terms (*e.g*. slight, moderate, severe) to describe the stage of development of any condition. Numeric scales are most often used and are quite widespread. Illustrative grading scales have the advantage of presenting the severity of a clinical condition as a series of photographs, paintings or drawings at various stages of severity [28]. The use of standard reference photographs and a numeric grading system have undeniably improved the reproducibility of clinical estimates, but the assumptions made in designing a clinical grading scale have important implications on the clinician's ability to detect change. Bailey et al. [26] suggested four assumptions to adopt when developing any grading scale, that: (1) the distribution of discrepancies (*i.e*. the variation in the condition) is normal, (2) there is no systematic bias (*i.e*. the mean discrepancy is zero), (3) variance is uniform across the range of the scale (*i.e*. the steps in the scale are evenly spread), and (4) no truncation effects are caused by restrictions at the end of the scale.

Some of these assumptions are not met by Rolando's grading scale; there should be no systematic bias, i.e. the mean discrepancy should be zero, but the Rolando scale has only four options which may cause bias between observers, especially when grading without the use of incremental units; on the other hand, the new developed grading scale has more options, helping to reduce this level of bias; variance should be uniform across the range of the scale, but with the Rolando scale there are many ferning patterns that do not seem to easily fit into any of the Rolando grades, particularly around Types I and II [22], in contrast, the new grading scale was based on an image library which contained a wide cross-section of ferning patterns that have been observed.

In contrast, by grading the image library using the initial 7 point scale, these limitations could be addressed. Although initial grading found an overlap across two grading standards (between Grades 2 and 3 and between Grades 6 and 7), the new 7-point scale showed a linear relationship across the library. Statistical analysis allowed the 7 point scale to be collapsed down to five grades, to create an acceptable working scale. An advantage of larger increment steps is that it promotes good repeatability [29] and reproducible classification [30], by making the test an easy and consistent method for TF pattern classification.

Subjective grading relies upon the skill of the examiner to "subjectively" grade a particular condition, usually based on a fixed scale or standard. It has been used to monitor and quantify many ocular conditions, and different scales have been developed for subjective anterior ocular assessment, such as the Vistakon scales, which uses artist-rendered images for a large range of conditions [31]; the Cornea and Contact Lens Research Unit (CCLRU) scales, which have a 4-point scale for a range of conditions and use a series of photographs derived from clinical experience [32]; and the Efron scales [33] and Efron Millennium scales [34], which consist of a 5-point scale for a range of conditions, created from artist drawings. These different subjective scales are widely used because they are easy to use, cheap and portable. This means that a five-point grading system for tear ferning should be widely accepted by clinicians and easy for them to use, and to apply interpolation.

Tear film osmolarity is often assessed in the clinical setting using the TearLab (TearLab™ Corp., San Diego, California). This instrument has been shown to be effective at analysing osmolarity in the small sample sizes available from the tear film [35], but can be expensive to use, especially if the recommendation of Khanal and Millar [36] to take three repeat measurements is followed. Tear ferning offers an alternative method for practitioners to use, but full assessment of its clinical validity requires investigation of the ferning pattern obtained from a sample, with analysis of the same sample's osmolarity. However, in doing so, a grading scale which is able to consistently discriminate between ferning pattern is necessary.

This study has culminated in the production of a new grading scale for TF, which appears to be discriminating, linear and reliable. A new grading scale is necessary because of the limitations within the Rolando grading scale: the categorisation of ferning patterns lacks sensitivity, particularly with the overlap across Types I and II. The next stage of development is to examine the validity of grading scale in practice, for example by applying the new scale to normal and dry eyes, to examine the usefulness of the scale as a clinical and research measure.

References

[1] Pearce EI, Tomlinson A. Spatial location studies on the chemical composition of human tear ferns. Ophthal Physl Opt 2000; 20:306−13.

[2] Farris RL. Tear Osmolarity: a new gold standard? Adv Exp Med Biol 1984;350:495–503.

[3] Versura P, Profazio V, Campos EC. Performance of tear osmolarity compared to previous diagnostic tests for dry eye diseases. Curr Eye Res 2010;35:553–64.

[4] Kogbe O, Liotet S, Tiffany JM. Factors responsible for tear ferning. Cornea 1991;10:433–44.

[5] Tabbara KF, Okumoto M. Ocular ferning test. A qualitative test for mucus deficiency. Ophthalmology 1982; 89:712−14.

[6] Rolando M. Tear mucus ferning test in normal and keratoconjunctivitis sicca eyes. Chibret Int J Ophthalmol 1984;2:32–41.

[7] Golding TR, Brennan NA. The basis of tear ferning. Clin Exp Optom 1989;72:102–12.

[8] Murube J. Tear crystallization test: two centuries of history. Ocul Surf 2004;2:7–9.

[9] Papanicolaou G N. A general survey of the vaginal smear and its use in research and diagnosis. Am J Obstet Gynecol 1946;51:316–28.

[10] Rydberg E. Observation on the crystallization of the cervical mucus. Acta Obstet Gynecol Scand 1948;28:172–87.

[11] Rolando M. A simple test for the determination of ovulation, estrogen activity, and early pregnancy using the cervical mucus secretion. Am J Obstet Gynecol 1952;63:81–9.

[12] Zondek B. Some problems related to ovarian function, in Pincus G (ed). Recent Prog Horm Res 1954;10:391–423.

[13] Rolando M. The fern test. A critical analysis. Obstet Gynecol 1958;11:30–34.

[14] Zaneveld L J, Tauber P F, Port C, Propping D. Scanning electron microscopy of cervical mucus crystallization. Obstet Gynecol 1975;46:419–28.

[15] Abou-Shabanah E H, Plotz E J. A biochemical study of the cervical and nasal mucus fern phenomenon. Am J Obstet Gynecol 1957;74:559–68.

[16] Maragou M, Vaikousis E, Ntre A, Koronis N, Georgiou P, Hatzidimitriou E, et al. Tear and saliva ferning tests in Sjogren's syndrome (SS). Clin Rheumatol 1996;15:125–32.

[17] Guida M, Barbato M, Bruno P, Lauro G, Lampariello C. Salivary ferning and the menstrual cycle in women. Clin Exp Obstet Gynecol 1993;20:48–54.

[18] Barbato M, Pandolfi A, Guida M. A new diagnostic aid for natural family planning. Adv Contracept 1993;9:335–40.

[19] Guida M, Tommaselli GA, Palomba S, Pellicano M, Moccia G, Di Carlo C, et al. Efficacy of methods for determining ovulation in a natural family planning program. Fertil Steril 1999;72:900–904.

[20] Buckley HE. Crystal Growth. London: Chapman & Hall Ltd, 1951.

[21] Vaikoussis E, Georgiou P, Nomicarios D. Tear mucus ferning in patients with Sjogren's syndrome. Doc Ophthalmol 1994;87:145–51.

[22] Masmali AM. Development of a tear ferning test protocol and a new grading scale. PhD, Cardiff 2010.

[23] Norn M. Quantitative tear ferning. Clinical investigations. Acta Ophthalmol 1994; 72: 369–372.

[24] Ravazzoni L, Ghini C, Macri A, Rolando M. Forecasting of hydrophilic contact lens tolerance by means of tear ferning test. Graefes Arch Clin Exp Ophthalmol 1998; 236: 354–358.

[25] Norn M. Ferning in conjunctival-cytologic preparations. Crystallisation in stained semiquantitative pipette samples of conjunctival fluid. Acta Ophthalmol 1987;65:118–22.

[26] Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects of scaling. Invest Ophth Vis Sci 1991;32:422–32.

[27] Evans KS, North RV, Purslow C. Tear ferning in contact lens wearers. Ophthal Physiol Opt 2009;29:199–204.

[28] Efron N. Contact Lens Complications. Oxford, UK: Butterworth-Heinemann-Optician, 1999.

[29] Pensyl CD, Dillehay SM. The repeatability of tear mucus ferning grading. Optom Vis Sci 1998;75:600–4.

[30] Felberg S, Cordeiro H, Sato EH, Martini Filho D, Nishiwaki-Dantas MC, Endo RM, Dantas PE. Reproducibility of the classification of ocular ferning patterns in Sjogren's syndrome patients. Arq Bras Oftalmol 2008;71:228–33.

[31] Andersen JS, Davies IP, Kruse A. Handbook of Contact Lens Management. Jacksonville: Vistakon, 1996.

[32] Terry R, Schnider C, Holden B, et al. CCLRU standards for success of daily and extended wear contact lens. Optom Vis Sci 1993;70:234–43.

[33] Efron N. Grading scales for contact lens complications. Ophthal Physiol Opt 1998;18:182–6.

[34] Efron N. Grading Scales for Contact Lens Complications. Famborough: Millennium Edition ed, Hydron, 2000.

[35] Tomlinson A, McCann LC, Pearce EI. Comparison of human tear film osmolarity measured by electrical impedance and freezing point depression techniques. Cornea 2010;29:1036–1041.

[36] Khanal S, Millar TJ. Barriers to clinical uptake of tear osmolarity measurments. Br J Ophthalmol 2012;96:341-344.

Figure Legends

Figure 1: Images of the new 7-point grading scale.

Figure 2: Mean grading score and standard deviation for each image using the Rolando grading scales, showing the overlap between Types I and II.

Figure 3: Mean grading score and standard deviation for each image using the 7-point scale, showing the overlaps between Grades 2 and 3, and between Grades 6 and 7.

Figure 4: Baseline images of the new 5-point grading scale.

Figure 5: Mean grading score and standard deviation for each image using the new 5-point scale.

Table 1: The average position score for each image.

Table 2: Seven homogeneous sub-sets were found using post-hoc Tukey HSD test; the table shows the mean weighting for the homogeneous sub-sets.

Table 3: The mean score of each homogeneous sub-set, and the chosen image mean score for each group.

Table 4: Selection of the 7 images of the new scale (mean score in bold and highlighted).

Table 5: Median score and inter-quartile range (IQR) for each Rolando Scale Type.

Table 6:Homogeneous sub-set mean scores for the Rolando Scale.

Table 7: Homogeneous sub-sets mean scores for the 7-point scale.

Tables

Image	Sum of Score	Mean Score	SD
1	107	4.28	2.98
2	101	4.04	2.94
3	97	3.88	1.96
4	86	3.44	2.22
5	140	5.6	2.10
6	117	4.68	1.70
7	124	4.96	2.17
8	159	6.36	1.89
9	221	8.84	1.25
10	247	9.88	1.72
11	263	10.52	1.50
12	287	11.48	2.20
13	329	13.16	0.37
14	349	13.96	0.54
15	372	14.88	0.33

Table 1: The average position score for each image.

Image	N	Sub-set for alpha = 0.05						
		1	2	3	4	5	6	7
4	25	3.44						
3	25	3.88	3.88					
2	25	4.04	4.04					
1	25	4.28	4.28					
6	25	4.68	4.68	4.68				
7	25	4.96	4.96	4.96				
5	25		5.60	5.60				
8	25			6.36				
9	25				8.84			
10	25				9.52			
11	25				10.52	10.52		
12	25					11.48	11.48	
13	25						13.16	13.16
14	25							13.96
15	25							14.88
Sig.		.278	.118	.143	.143	.920	.143	.118

Table 2: Seven homogeneous sub-sets were found using post-hoc Tukey HSD test; the table shows the mean weighting for the homogeneous sub-sets.

Group	Sub-set mean score	Chosen image number	Mean score of the image
1	4.21	1	4.28
2	4.57	6	4.68
3	5.40	5	5.60
4	9.62	10	9.52
5	11.00	12	11.48
6	12.32	13	13.16
7	14.00	14	13.96

Table 3: The mean score of each homogeneous sub-set, and the chosen image mean score for each group.

Image	N	Sub-set for alpha = 0.05						
		1	2	3	4	5	6	7
4	25	3.44						
3	25	3.88	3.88					
2	25	4.04	4.04					
1	25	**4.28**	4.28					
6	25	4.68	**4.68**	4.68				
7	25	4.96	4.96	4.96				
5	25		5.60	**5.60**				
8	25			6.36				
9	25				8.84			
10	25				**9.52**			
11	25				10.52	10.52		
12	25					**11.48**	11.48	
13	25						**13.16**	13.16
14	25							**13.96**
15	25							14.88
Sig.		**.278**	**.118**	**.143**	**.143**	**.920**	**.143**	**.118**

Table 4: Selection of the 7 images of the new scale (mean score in bold and highlighted).

Type	Median Score	IQR
I	1.15	0.36
II	1.46	0.36
III	2.81	0.36
IV	4	0.06

Table 5: Median score and inter-quartile range (IQR) for each Rolando Scale Type.

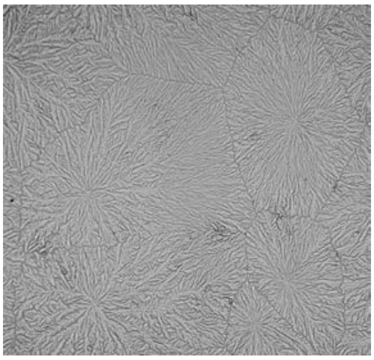
Type	N	Subset for alpha = 0.05		
		1	2	3
1	15	1.2693		
2	15	1.4067		
3	15		2.7240	
4	15			3.9860
Sig.		.100	1.000	1.000

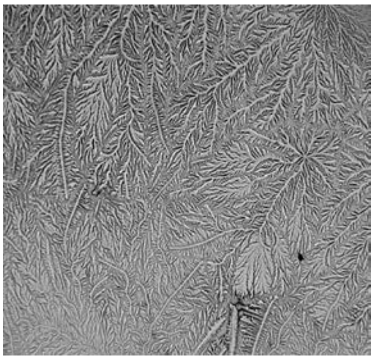
Table 6:Homogeneous sub-set mean scores for the Rolando Scale.

Type	N	Subset for alpha = 0.05				
		1	2	3	4	5
1	14	1.5279				
2	10		2.5620			
3	6		2.8400			
4	10			4.3620		
5	5				4.9020	
6	4					6.4695
7	11					6.6982
Sig.		1.000	.437	1.000	1.000	.663

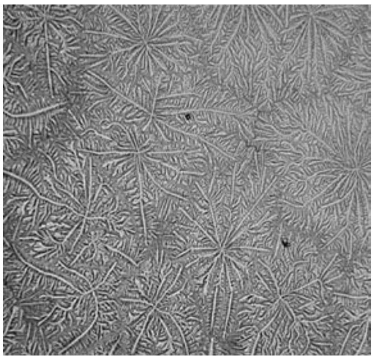
Table 7: Homogeneous sub-sets mean scores for the 7-point scale.

Figures

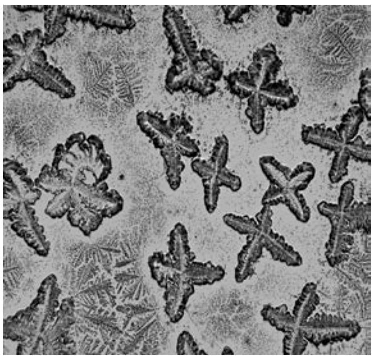


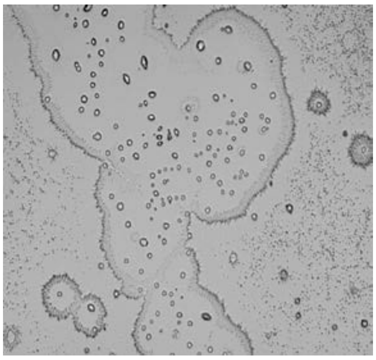


1 **2**

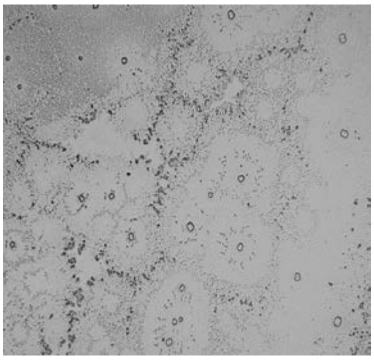


3 **4**





5 **6**



7

Figure 1: Images of the new 7-point grading scale.

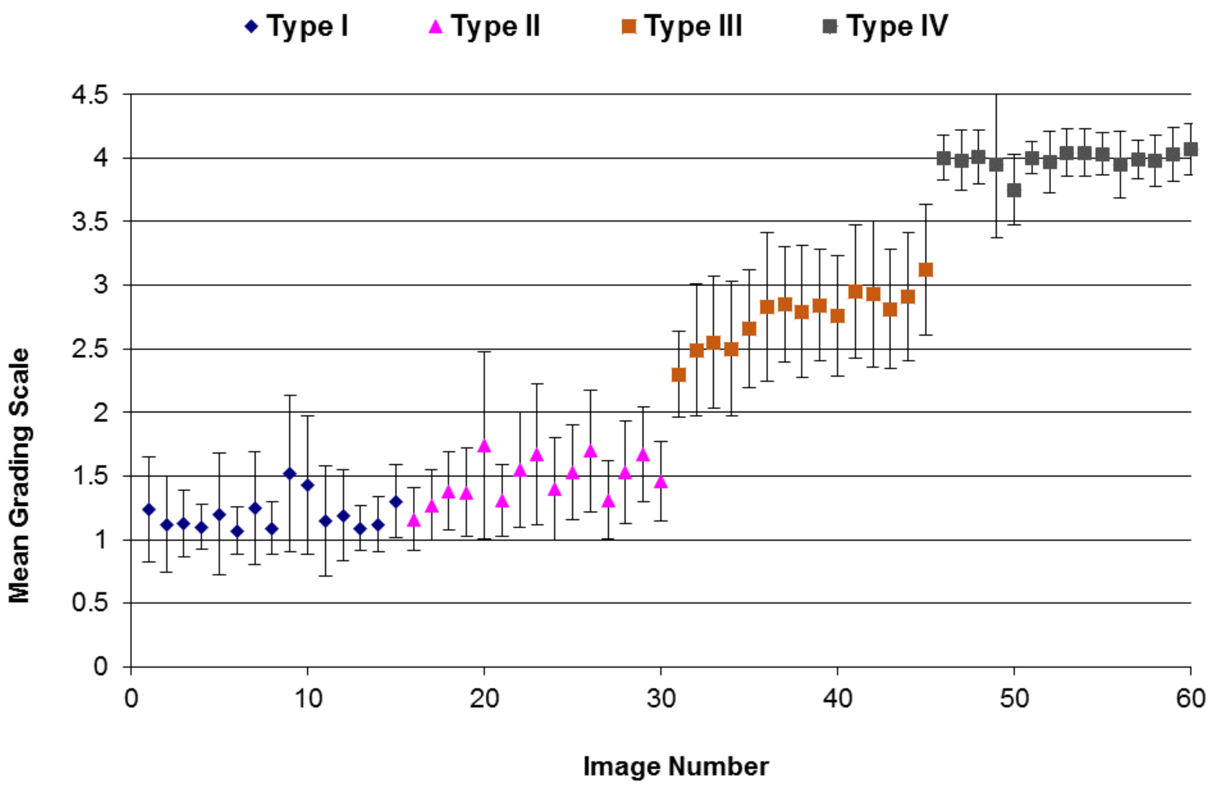


Figure 2: Mean grading score and standard deviation for each image using the Rolando grading scales, showing the overlap between Types I and II.

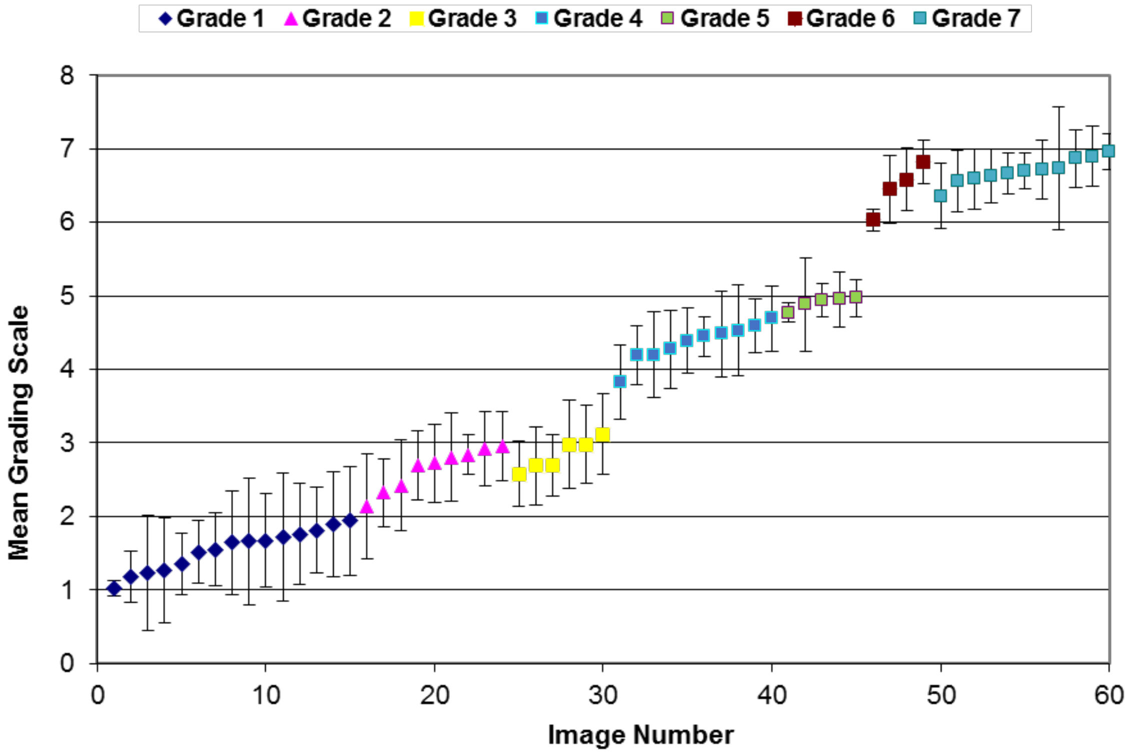
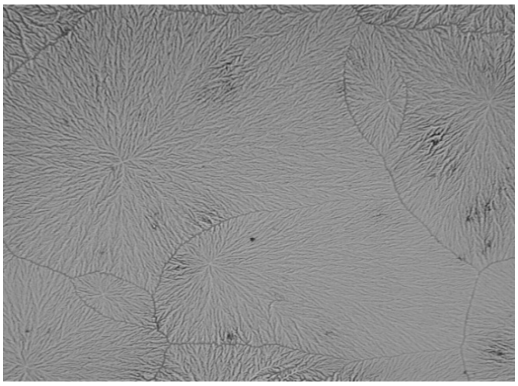
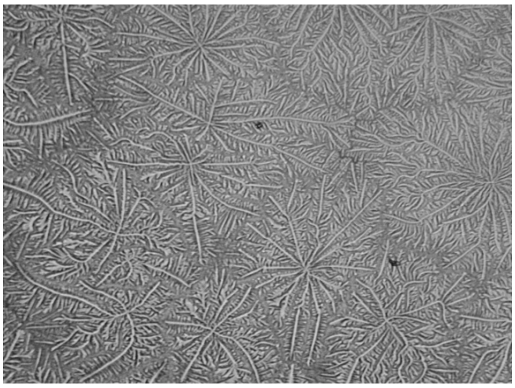
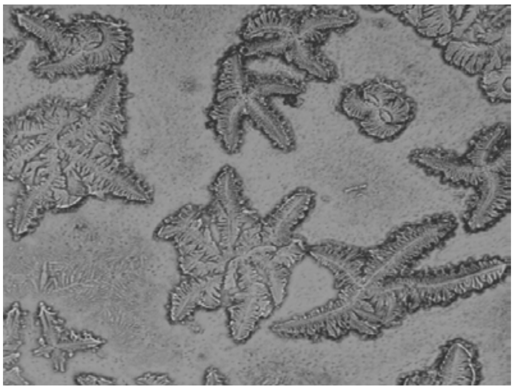


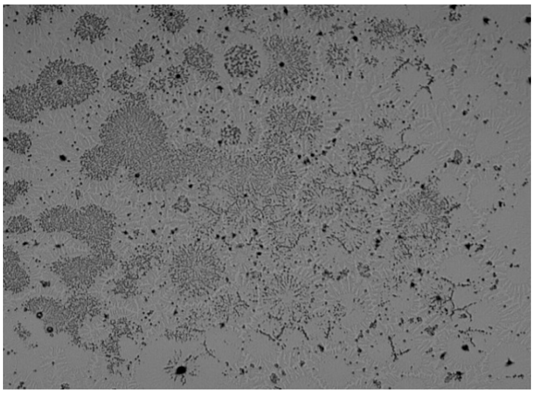
Figure 3: Mean grading score and standard deviation for each image using the 7-point scale, showing the overlaps between Grades 2 and 3, and between Grades 6 and 7.

Grade 0 Grade 1

Grade 2 Grade 3



Grade 4

Figure 4: Baseline images of the new 5-point grading scale.

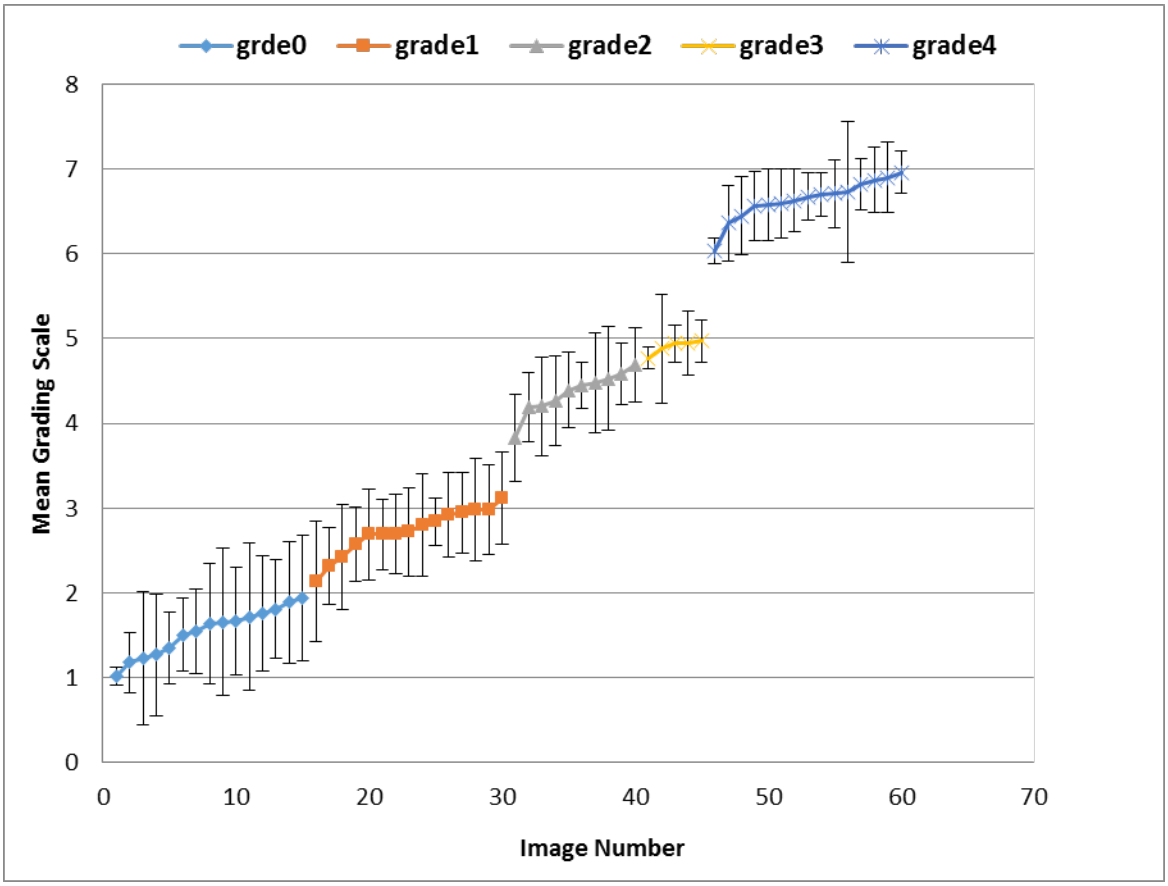


Figure 5: Mean grading score and standard deviation for each image using the new 5-point scale.