

# Interobserver and intraobserver reliability of a classification scheme for corneal topographic patterns

Karim Rasheed, Yaron S Rabinowitz, Diana Remba, Melvin J Remba

## Abstract

**Aims**—To determine the interobserver and the intraobserver reliability of a published classification scheme for corneal topography in normal subjects using the absolute scale.

**Method**—A prospective observational study was done in which 195 TMS-1 corneal topography maps in the absolute scale were independently classified twice by three classifiers—a cornea fellow, an ophthalmic technician, and an optometrist. From these observations the interobserver reliability for each category and the intraobserver reliability for each observer were determined in terms of the median weighted kappa statistic for each category and for each observer.

**Results**—For interobserver reliability, the median weighted kappa statistic for each category varied from 0.72 to 0.97 and for intraobserver reliability the range was 0.79 to 0.98.

**Conclusion**—This classification scheme is extremely robust and even in the hands of less experienced observers with minimal training it can be relied upon to provide consistent results.

(Br J Ophthalmol 1998;82:1401–1406)

Bogan *et al* described a classification scheme of the morphological patterns observed in corneal topography maps obtained from normal subjects; this classification scheme was further expanded by Rabinowitz *et al* who introduced five additional categories to enable classification of asymmetric corneal topographic patterns.<sup>1,2</sup> These studies provided much valuable information by describing the spectrum of variability in corneal topography patterns as observed in normal human corneas. A recent study has also applied the scheme proposed by Rabinowitz *et al* to studying the possible effects of contact lens warpage on corneal topography in patients with keratoconus.<sup>3</sup> It is likely that there will be many further applications where the classification of corneal topography will be used. However, in order for a classification scheme to be used as clinical or research tool, its reliability must be established. To achieve widespread use it must also be shown that the scheme proposed produces consistent results in the hands of less experienced observers and not just expert observers.

Since these classification schemes rely to a large extent on the subjective interpretation of observed topographical patterns it is inevitable

that there will be some degree of error as a result of observer variability. Errors may arise as a result of different interpretation by different observers (interobserver variability) and as a result of inconsistent interpretation of the same map on repeat observations by the same observer (intraobserver variability).<sup>4,5</sup> Measurement of such variability in classification schemes used in ophthalmology has been reported primarily for schemes classifying crystalline lens changes, visual field changes, and has been widely reported for a number of other grading systems in clinical use outside ophthalmology.<sup>5–9</sup> Such a determination of interobserver and intraobserver reliability of a classification scheme for corneal topographic map morphology has not previously been reported. The following is a description of the measurement of reliability of the classification scheme proposed by Rabinowitz *et al* in the hands of relatively inexperienced observers.<sup>2</sup>

## Materials and methods

A database of 195 normal subjects originally studied in the longitudinal study of genetic factors in keratoconus at the Cornea Genetic Medical Eye Clinic at the Cedars Sinai Medical Center was used. The selection of subjects, demographic characteristics, methods of obtaining corneal topographic maps, and the distribution of the patterns of topography have been described previously in detail.<sup>2</sup> In order to avoid bias due to intereye correlation only the maps of the subject's right eye were used. The TMS-1 corneal topography maps used were in the "absolute scale".<sup>10</sup> This is a standardised 26 colour scale where each colour has a designated dioptric range that does not vary. This use of a standard range of reference colour is thought to allow more consistent pattern recognition between different observers.<sup>2</sup>

The three classifiers were an optometrist, an ophthalmic technician, and a cornea fellow who were previously inexperienced in the classification of cornea topographic maps using this scheme. All three classifiers received training in the use of the classification scheme by an experienced classifier during one 3 hour teaching session. Maps used to train the observers were not used to test the reliability. The maps were assigned random numbers and all identifying marks and quantitative indices shown on the maps were obscured before the maps were presented to each observer. Each observer independently classified all maps in turn. After all three observers had completed the first set of observations the maps were shuffled so that they

Cornea-Genetic Eye Medical Clinic, Cedars Sinai Medical Centre, Los Angeles, California, USA  
K Rasheed  
Y S Rabinowitz

Division of Ophthalmology, Department of Surgery, Burns and Allen Research Institute, Cedars Sinai Medical Center, Los Angeles, California, USA  
K Rasheed  
Y S Rabinowitz

Division of Optometry, Burns and Allen Research Institute, Cedars Sinai Medical Center, Los Angeles, California, USA  
M J Remba

Medical Genetics Birth Defects Center, Burns and Allen Research Institute, Cedars Sinai Medical Center, Los Angeles, California, USA  
Y S Rabinowitz  
D Remba  
M J Remba

Department of Pediatrics and Departments of Ophthalmology, UCLA School of Medicine, Los Angeles, California, USA  
Y S Rabinowitz

Correspondence to:  
Dr Yaron S Rabinowitz,  
Cornea-Genetic Eye Medical Clinic, Cedars-Sinai Medical Center, Mark Goodson Building, Suite 1102, 444 South San Vicente Boulevard, Los Angeles, California, CA 90048, USA.

Accepted for publication  
18 June 1998

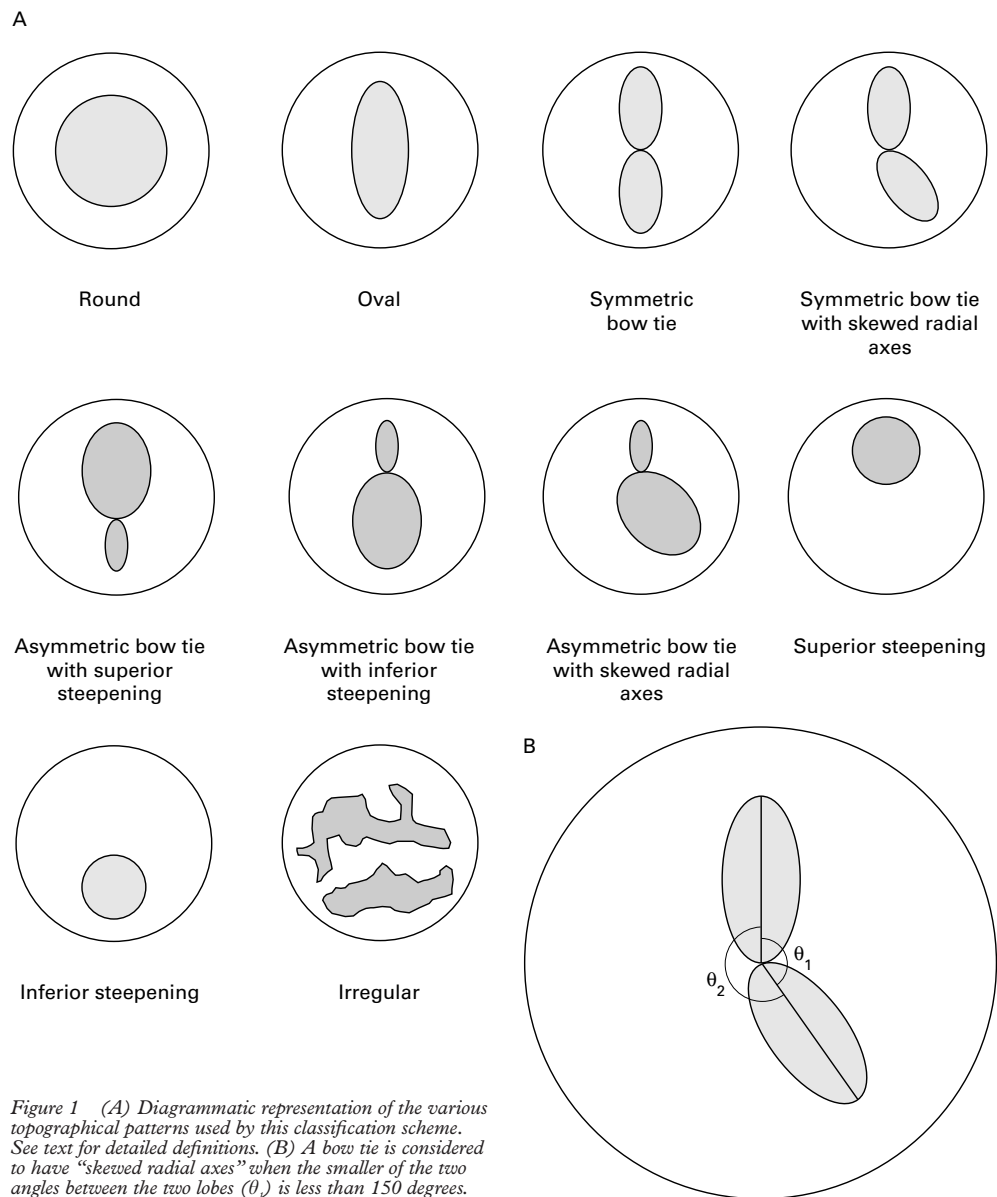


Figure 1 (A) Diagrammatic representation of the various topographical patterns used by this classification scheme. See text for detailed definitions. (B) A bow tie is considered to have "skewed radial axes" when the smaller of the two angles between the two lobes ( $\theta_1$ ) is less than 150 degrees.

were no longer in the same sequence when presented to the observers for the second time.

The classification scheme used was the one described by Rabinowitz *et al* with minor modifications<sup>2</sup> (Fig 1A and B). The descriptions of the categories are repeated here for the sake of clarity. The steepest colour in the central two thirds of the map was used to

determine the classification type provided that the pattern occupied at least 10% of the central two thirds area. The three observers were provided with a transparent overlay with the 10% and two thirds areas marked on it, which allowed a more consistent appreciation of the proportions. The overlay can be seen in use in Figures 2A and B.

A round pattern implies a predominantly circular pattern occupying the centre of the map such that the smallest diameter is not less than two thirds of the greatest diameter. Oval implies a predominantly oval pattern occupying the centre of the map where the smallest diameter of the oval is less than two thirds and greater than one third of the greatest diameter. Bow tie implies an oval pattern with a central constriction such that the width of the central constriction is less than one third of the width of either of the two lobes. This pattern is further classified as symmetric bow tie, when the smaller of the two lobes has not less than

Table 1 Number of maps for which there was complete or partial agreement between observers in each classification category

	Complete agreement between all three observers	Agreement between two observers	Total
Round	36	32	68
Oval	11	16	27
Superior steepening	4	4	8
Inferior steepening	12	8	20
Symmetric bow tie	27	5	32
Symmetric bow tie with skewed radial axes	0	0	0
Asymmetric bow tie with superior steepening	7	5	12
Asymmetric bow tie with inferior steepening	9	6	15
Asymmetric bow tie with skewed radial axes	0	0	0
Irregular	6	7	13
All groups combined	112	83	195

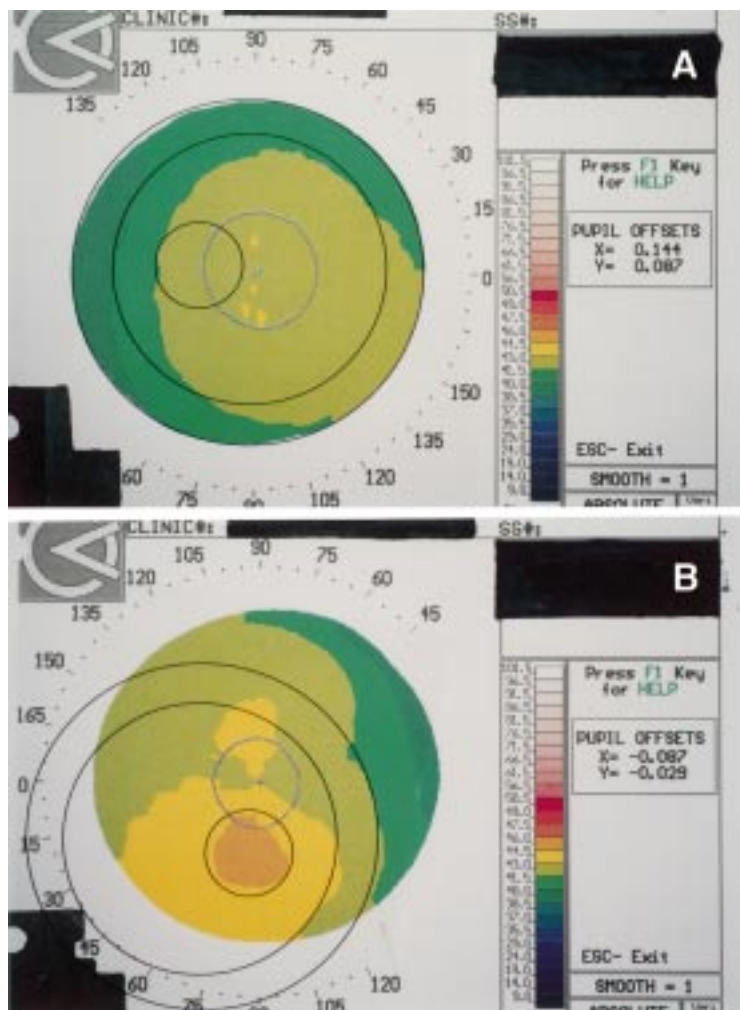


Figure 2 (A) Transparent overlay placed over a TMS 1 map. The outermost circle occupies the area that a TMS 1 map would occupy when all points on all 25 rings are digitised. The intermediate circle occupies an area equal to two thirds of the entire map. This area is used to determine the classification of a map. The innermost circle occupies an area equal to 10% of the central two thirds area of the entire map. (B) The overlay being used to assess whether a colour occupies 10% of the central two thirds area of the map.

two thirds of the area of the larger lobe and the smaller of the two angular differences between the steepest radial axes of the two lobes is not less than 150 degrees (Fig 1B); symmetric bow tie with skewed radial axis, when the smaller of the two lobes has not less than two thirds of the area of the larger lobe and the smaller of the two angular differences between the steepest radial axes of the two lobes is less than 150 degrees; asymmetric bow tie with superior steepening when the inferior lobe has less than two thirds of the area of the superior lobe and

the smaller of the two angular differences between the steepest radial axes of the two lobes is not less than 150 degrees; asymmetric bow tie with inferior steepening, when the inferior lobe has more than two thirds of the area of the superior lobe and the smaller of the two angular differences between the steepest radial axes of the two lobes is not less than 150 degrees and asymmetric bow tie with skewed radial axis, when the smaller of the two lobes has less than two thirds of the area of the larger lobe and the smaller of the two angular differences between the steepest radial axes of the two lobe is less than 150 degrees. Superior steepening implies that there is an area which occupies at least 10% of the central two thirds area, of increased power above the horizontal meridian and inferior steepening implies that there is an area which occupies at least 10% of the central two thirds area, of increased power below the horizontal meridian.

The maps were classified twice by each observer to assess intraobserver reliability. Maps were assigned to a specific category when there was agreement between at least two observers. The intervals between the first and second observations were sufficient to minimise the effect of memory on the results. All data were entered into a computer spreadsheet for statistical analysis.

STATISTICAL METHODS

Each map was classified twice by each of the three observers. The first set of observations was used to construct pairs for the assessment of interobserver variability (observers 1 v 2, 1 v 3, and 2 v 3). The first set of observations was compared with the second set to assess intraobserver variability for each observer (observers 1a v 1b, 2a v 2b, and 3a v 3b where “a” denotes the first observation set and “b” denotes the second observation set for each eye).

The index used to evaluate reliability was the weighted kappa statistic.<sup>4 11-13</sup> This is a chance corrected index of agreement which penalises interobserver disagreement and is appropriate for testing a classification scheme which produces categorical or ordinal data. The scale weights should vary in linear increments based upon the number of divisions in the scale between these limits. It has been proposed by Bogan *et al* that round, oval, symmetric bow tie, asymmetric bow tie, and irregular are gradations in a continuous spectrum.<sup>1</sup> It may therefore have been possible to assign weights based on this possible continuous spectrum. However, this hypothesis is unproved and we therefore assigned a weight of 1 for agreement and a weight of 0 for disagreement. A 2 x 2 matrix was constructed for each pair of observation. Each category was tested with the scale reduced to binary form—that is, round and not round, oval and not oval, etc, with the assigned weights of 0 and 1 as described. The system used was based on weighted agreement as proposed by Hall.<sup>4</sup> The equation used to calculate weighted kappa( $\kappa_w$ ) was:

$$\kappa_w = (P_o - P_d) / (1 - P_d)$$

Table 2 Number of maps that were consistently classified in each classification category on repeat observation

	No of consistent pairs of observations		
	Observer 1	Observer 2	Observer 3
Round (n=68)	52	49	57
Oval (n=27)	17	14	22
Superior steepening (n=8)	7	7	6
Inferior steepening (n=20)	18	17	19
Symmetric bow tie (n=32)	29	31	30
Asymmetric bow tie with superior steepening (n=12)	7	9	12
Asymmetric bow tie with inferior steepening (n=15)	11	11	13
Irregular (n=13)	10	9	12
All groups combined (n=195)	151	147	171

Table 3 Interobserver reliability expressed as the weighted kappa statistic for each pair of observers

	Weighted kappa statistic			Median weighted kappa statistic for category
	For pair 1 v 2 (SE)	For pair 1 v 3 (SE)	For pair 2 v 3 (SE)	
Round	0.67 (0.0099)	0.76 (0.0097)	0.72 (0.0104)	0.72
Oval	0.79 (0.0052)	0.86 (0.0067)	0.84 (0.0064)	0.84
Superior steepening	0.99 (0.0057)	0.98 (0.0053)	0.98 (0.0053)	0.98
Inferior steepening	0.95 (0.006)	0.95 (0.0058)	0.96 (0.0058)	0.95
Symmetric bow tie	0.89 (0.0073)	0.93 (0.0067)	0.94 (0.0068)	0.93
Symmetric bow tie with skewed radial axes*	1.0 (0.0051)	0.99 (0.005)	0.99 (0.0051)	0.99
Asymmetric bow tie with inferior steepening	0.94 (0.0059)	0.96 (0.0056)	0.96 (0.0056)	0.96
Asymmetric bow tie with superior steepening	0.97 (0.0055)	0.97 (0.0055)	0.99 (0.0057)	0.97
Asymmetric bow tie with skewed radial axes*	1.0 (0.0051)	1.0 (0.0051)	0.99 (0.0051)	1.0
Irregular	0.96 (0.0054)	0.97 (0.0059)	0.95 (0.0057)	0.96
Median weighted kappa statistic for whole group	0.95	0.96	0.96	0.96

\*There were no maps classified by at least two observers in this category. The weighted kappa statistic is shown here to illustrate the low false positive rate for this category but it has not been used in determining the median weighted kappa statistic for the whole group.

Table 4 Intraobserver reliability expressed as the weighted kappa statistic for each pair of observers

	Weighted kappa statistic			Mean weighted kappa statistic for category
	For pair 1a v 1b (SE)	For pair 1a v 2b (SE)	For pair 3a v 3b (SE)	
Round	0.79 (0.0093)	0.75 (0.0097)	0.86 (0.0101)	0.79
Oval	0.86 (0.0065)	0.86 (0.0066)	0.90 (0.0068)	0.86
Superior steepening	0.99 (0.0055)	0.99 (0.0055)	0.98 (0.0054)	0.99
Inferior steepening	0.98 (0.0062)	0.97 (0.0061)	0.99 (0.0061)	0.98
Symmetric bow tie	0.96 (0.0069)	0.92 (0.0072)	0.96 (0.007)	0.96
Symmetric bow tie with skewed radial axes*	0.99 (0.0051)	1.0 (0.0051)	0.99 (0.0051)	0.99
Asymmetric bow tie with inferior steepening	0.97 (0.0057)	0.95 (0.0054)	0.99 (0.0058)	0.97
Asymmetric bow tie with superior steepening	0.97 (0.0054)	0.98 (0.0055)	1.0 (0.0058)	0.98
Asymmetric bow tie with skewed radial axes*	1.0 (0.0051)	1.0 (0.0052)	1.0 (0.0051)	1.0
Irregular	0.94 (0.0057)	0.95 (0.0056)	0.99 (0.0059)	0.95
Median weighted kappa statistic for whole group	0.97	0.95	0.99	0.97

\*There were no maps classified by at least two observers in this category. The weighted kappa statistic is shown here to illustrate the low false positive rate for this category but it has not been used in determining the median weighted kappa statistic for the whole group.

where  $P_o$  is the total proportion of weighted observer agreement,  $P_c$  is the total proportion of weighted chance agreement.

Variance of weighted kappa was calculated by the formula:

$$\text{Variance}(\kappa_w) = \frac{1}{N} (1 - P_o)^4 \{ \sum P_{ij} [\omega_i (1 - P_o) - (\omega_i + \omega_j) (1 - P_o)]^2 - (P_o P_c - 2P_c + P_o)^2 \}$$

where  $P_{ij}$  is the observed proportion of observations for each cell,  $\omega_i$  is the weighted average of the weights in each row;  $\omega_j$  is the weighted average of the weights in each column;  $\omega_{ij}$  is the weight assigned to each cell in the matrix;  $P_o$  is the total proportion of weighted observed agreement;  $P_c$  is the total proportion of weighted chance agreement; and  $N$  is the total number of observations made by each observer.

Standard error of weighted kappa was calculated by the formula:

$$SE(\kappa_w) = \{ \text{Variance}(\kappa_w) / N \}^{0.5}$$

A computer program was written to automate the solution of these equations. The computer program was tested for accuracy by solving data sets with known solutions before using it in this study. The median of these measurements was used to determine the median weighted kappa statistic for each of the classification categories. The median of all the categories was calculated to estimate interobserver and intraobserver reliability of the whole system.

**Results**

Each observer classified the test set of 195 maps over a 2 day period. The minimum

period between the two sets of observation used to test for intraobserver reliability was 13 days and the maximum was 26 days. Table 1 shows the results of the interobserver analysis and Table 2 shows the results of the intraobserver analysis in terms of percentages of agreement between the observers. In this set of maps we found no maps that were classifiable as either symmetric bow tie with skewed radial axes or as asymmetric bow tie with skewed radial axes. Of the 195 maps there was complete agreement between all three observers on the classification of 112 maps and there was agreement between two observers on the classification of 83 maps. There were no maps where all three observers disagreed. Tables 3 and 4 present the results of the interobserver and the intraobserver analyses in terms of the weighted kappa statistic. The weighted kappa statistic shown for symmetric bow tie with skewed radial axes and asymmetric bow tie with skewed radial axes are shown to illustrate the low false positive rate for these categories, but have not been used in the calculation of the median weighted kappa statistic for the whole group.

**Discussion**

The classification scheme described by Bogan et al for corneal topography patterns was based on the computerised modelling system maps using sagittal corneal topography and the normalised scale.<sup>1</sup> There were five classification categories—round, oval, symmetric bow tie, asymmetric bow tie, and irregular. The scheme

introduced by Rabinowitz *et al* added five new categories; superior steepening, inferior steepening, and the bow tie categories were expanded to include symmetric bow tie with skewed radial axes, asymmetric bow tie with skewed radial axes, asymmetric bow tie with superior steepening, and asymmetric bow tie with inferior steepening<sup>2</sup>(Fig 1). These new categories were meant to allow for the classification of more complex patterns seen in normal maps. In the studies reported by Bogan *et al* and by Rabinowitz *et al*, the inventors of the classification scheme were included in the observers who classified the corneal topographic maps. This is a common practice when a classification scheme is introduced for the first time. In the first study, Bogan *et al* reported that the first independent observation by their three observers resulted in a complete agreement (between all three observers) for 82.7% and partial agreement (where two of the three observers agree) for 16.3% and complete disagreement for 1% of the maps studied; on the second series of observations of the same maps, the rates improved to 90%, 10%, and 0% respectively.<sup>1</sup> Rabinowitz *et al* using their more complex classification scheme reported very similar results.<sup>2</sup> Neither of these two studies assessed intraobserver reliability and neither study assessed the reliability of the individual classification categories. Also, the method of using percentage of total agreement between observers provides an incomplete assessment of the reliability as partial agreement between observers is not included in the assessment, incorrect classifications are not penalised and no compensation is made for possible chance agreement.<sup>4 11-13</sup> The weighted kappa statistic overcomes most of the difficulties and enables a comprehensive expression of reliability in one single index.<sup>4 11-13</sup> This index (similar to the Pearson's product moment correlation) varies between -1.0 and +1.0 and enables meaningful comparison of reliability between different categories and even between different classification schemes. Landis and Koch suggest that a weighted kappa statistic of 0.81-1.0 implies "almost perfect" agreement.<sup>11</sup>

In this study, where the classifications were carried out by less experienced observers, the complete agreement rate (all three observers agree) was less than that reported by either Bogan *et al* or by Rabinowitz *et al*.<sup>1 2</sup> It is therefore obvious that there is a higher degree of disagreement in the hands of non-expert observers. However, it still appears that the classification scheme is extremely robust providing extremely good overall interobserver and intraobserver reliability as is reflected by the high weighted kappa statistic for all the analyses. There were two categories under which no maps were classified. This to some extent limited our assessment of these categories in that we were not able to assess the false negative rates for these categories, but there were very few false positives as can be seen by the very high weighted kappa statistics for these categories.

In most clinical grading systems, it is usual to find that the intraobserver reliability is higher than the interobserver reliability.<sup>3 9</sup> In

this study as well, the intraobserver kappa statistic for most categories was better than the interobserver kappa statistic. This may have been because of the confounding effect of memory. The duration between the two sets of observations may have played a role in causing this effect. In previous studies classifying crystalline lens changes Sparrow *et al* waited from 7 to 148 days and King *et al* in their study classifying Goldmann visual fields waited 2 months between the two sets of observations used to test for intraobserver reliability.<sup>5 9</sup> In these and another paper where intraobserver reliability has been tested, the time interval between observations was selected arbitrarily. In the absence of historical data to guide us, we took the following measures that we hoped would minimise the confounding effect of memory. We presented each observer with a large set of maps (195 in each set), we waited at least 13 days before presenting the same set of maps again to the same observer, all identifying marks, were obscured on each map and we shuffled the maps after the first set of observations was made so that the sequence of maps was not the same when the second observations were made. In spite of doing so we have no proof that the confounding effect of memory had been completely abolished.

One significant difference between our study and that of Rabinowitz *et al* is that we used a transparent overlay (Fig 2A and B).<sup>2</sup> This allowed us to accurately judge whether a colour occupied at least 10% of the central two thirds area and also demarcated the central two thirds from the outer one third area on the maps. Although we did not formally test the reliability of the scheme with and without the use of this overlay it is our impression that this overlay is extremely useful in improving consistency.

We conclude therefore that the 10 category classification scheme using the absolute scale is extremely robust and can, with minimal training, be relied upon to provide excellent reproducibility even in the hand of less experienced observers. This makes it a reliable research tool for determining subtle deviations from normal corneal curvature in longitudinal studies of corneal topography of patients with familial keratoconus and for detecting subclinical keratoconus as it develops in these patients based on changes in their corneal topography pattern as seen over time.

Supported in part by NIH grant EY09052, the Cedars-Sinai Young Investigator's award, and the Eye Birth Defects Research Foundation, Inc.

The authors have no financial interest in any device described in this article. Yaron S Rabinowitz, MD, has received a grant in the form of travel support from Tomey Technology Inc, the distributor of the topographic modelling system.

- 1 Bogan SJ, Waring GO, Ibrahim O, *et al*. Classification of normal corneal topography based on computer assisted videokeratography. *Arch Ophthalmol* 1990;108:945-9.
- 2 Rabinowitz YS, Yang H, Brickman Y, *et al*. Videokeratography database of normal human corneas. *Br J Ophthalmol* 1996;80:610-16.
- 3 Szczotka LB, Rabinowitz YS, Yang H. Influence of contact lens wear on the corneal topography of keratoconus. *CLAO J* 1996;22:1-4.
- 4 Hall JN. Inter-rater reliability of ward rating scales. *Br J Psychiatry* 1974;125:248-55.
- 5 Sparrow JM, Ayliffe W, Bron A, *et al*. Inter-observer and intra-observer variability of the Oxford clinical cataract classification and grading system. *Int Ophthalmol* 1988;11:151-7.

- 6 Gibson RA, Sanderson HF. Observer variations in ophthalmology. *Br J Ophthalmol* 1980;**64**:457-60.
- 7 Cicchetti LT, Leske MC, Cotlier E. Assessment of observer variability in the classification of human cataract. *Yale J Biol Med* 1982;**55**:81-8.
- 8 Chylack LT, Leske MC. Validity and reliability of photo derived human cataract classification. *Invest Ophthalmol Vis Sci* 1986;**27**:44.
- 9 King AJ, Farnworth D, Thompson JR. Inter-observer and intra-observer agreement in the interpretation of visual fields in glaucoma. *Eye* 1997;**11**:687-91.
- 10 Gormley DJ, Gersten M, Koplin RS, Lubkin V. Corneal modeling. *Cornea* 1988;**7**:30-5.
- 11 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159-74.
- 12 Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Mental Defic* 1981;**86**:127-37.
- 13 Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: John Wiley, 1981:212-35.