# Methods of Quality Assessment for Large Sample Sets

*Phil Green, Colour Imaging Group*
*London College of Communication London, United Kingdom*

*Siv Lindberg*
*Swedish Pulp and Paper Research Institute Stockholm, Sweden*

## Abstract

Psychophysical evaluation of large sample sets was studied with reference to the International Newspaper Colour Quality Club Jury Evaluation, in which 150 prints of the same image are assessed by category judgement under 10 different quality attributes.

In a series of experiments using sub-sets of the CQC prints, some psychophysical techniques which could affect the reliability and precision of the results were evaluated. It was also possible to gain insights into the relationship between the different psychophysical methods.

On the basis of these results a number of modifications to the category judgement task used in the Jury Evaluation are proposed. These include the adoption of an anchor image whose scores on the quality attributes are defined in a preliminary observer task; and a reduction in the number of attributes and judgement categories.

## Introduction

The International Newspaper Color Quality Club[1] is organised by the international newspaper research organisation IFRA, the Newspaper Association of America, and the Pacific Area Newspaper Publishers Association. Entry to the Club is based on the colour reproduction quality in a series of tests consisting of both colorimetric and visual assessments. The tests are repeated bi-annually, and membership of the Club is updated each time.

Entrants set great store by membership of the Color Quality Club, and the benefits in increased advertising revenues tend to offset the cost of entry. Hence there is considerable importance attached to the objectivity of the results.

The INCQC Jury Evaluation is the event in which the visual quality of the entries is judged by an international panel of experts.

The method of assessment used in the Jury Evaluation is a form of category judgement. In INCQC 2002, two test images were used, and judges were asked to score each reproduction on an 11-point scale. Initially an overall assessment of quality is made, and then scores are given in each of ten separate judgement criteria. The final scores for the jury evaluation are determined by simple summation of the scores for each reproduction, with a higher weighting given to the 'overall quality' category.

The final scores from the visual evaluation are aggregated with the data from a colorimetric evaluation of test prints to determine the ranking of the entrants. Only the top 50-scoring entrants are admitted into membership of the Color Quality Club each round. The individual category scores are also used by IFRA to provide an individualised report to each entrant on the strengths and weaknesses of their colour printing.

The number of entrants to the Color Quality Club is in excess of 150. With two test images, there are over 300 reproductions to be assessed, and this provides a unique opportunity to examine methods appropriate to the visual assessment of large numbers of samples.

The visual assessments should ideally result in precise and reproducible results. To achieve these aims, the design of the Jury Evaluation should as far as possible minimise the complexity and stress of the observer task; be defined in such a way that minimises inter-observer and intra-observer variability; and result in a score dispersion that minimises the uncertainty around the entrance threshold of the Color Quality Club

In examining the methods of assessment and analysis, a number of particular questions of interest arise.

### *Accuracy of the Results*

The accuracy of the scores obtained by the INCQC evaluation cannot be evaluated in an absolute sense, but the results can be compared to those obtained by other methods.

### *Repeatability of the Results*

How well do the scores stand up to judgements made by similar groups of expert observers? What is the repeatability at the level of the individual judge?

### *Category Scale*

How many judgement categories are necessary, in terms of the precision of the results and the just noticeable preferences[2]?

### *Number of Quality Dimensions*

How many independent quality attributes are necessary?

***Confidence Intervals***

Given the importance of the top-50 threshold, a means of expressing the confidence of the individual scores is desirable.

***Alternative Methods***

Are there other methods of assessment that will deliver equivalent results with less effort on the part of observers?

***Absolute versus Relative Scales***

To what degree are the results relative to the population of the sample set? Can more absolute scores be derived, which will, for example, permit comparison between different INCQC events?

These questions were addressed through analysis of previous INCQC result and through new experiments using samples from INCQC 2002.

## Analysis of INCQC 2000 and INCQC 2002 Results

Data from the 2000 and 2002 Jury Evaluations were analysed to consider patterns in the previous results and to evaluate the scale precision and repeatability of the final scores.

Inter-observer variation for INCQC 2000 and 2002 was high, with median scores for a given attribute for a single observer ranging from approximately 4 to 8.5. A typical result is shown in Figure 1.

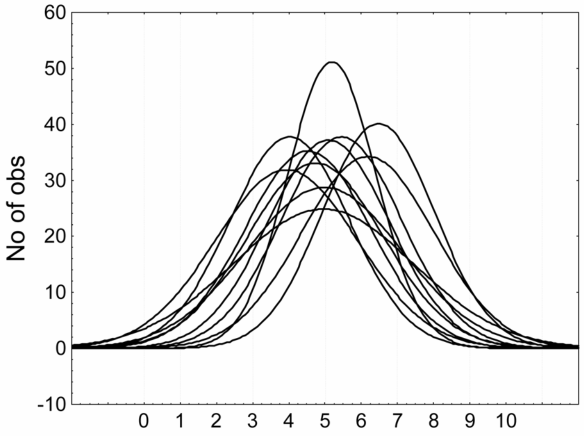

*Figure 1. The ratings of 'colour accuracy' for 11 jury members of the INCQC 2002. Each observer's histogram is fitted to the normal distribution.*

It can be seen that individual observers tend to distribute their ratings over different ranges of the 11-point scale, some being stricter and some more lenient in their criteria. These results indicated that the number of judgment categories might be too high. An analysis of the category data according to Thurstone's Law of Comparative Judgement[3] was performed. This approach derives interval scale values from rank order data and locates the boundaries for the numerical categories on the response continuum. The data are plotted in Figure 2. An inspection of the figure shows that the category sizes are unequal and that the upper categories are truncated. The categories representing the end points are populated by only a very small frequency of judgments.

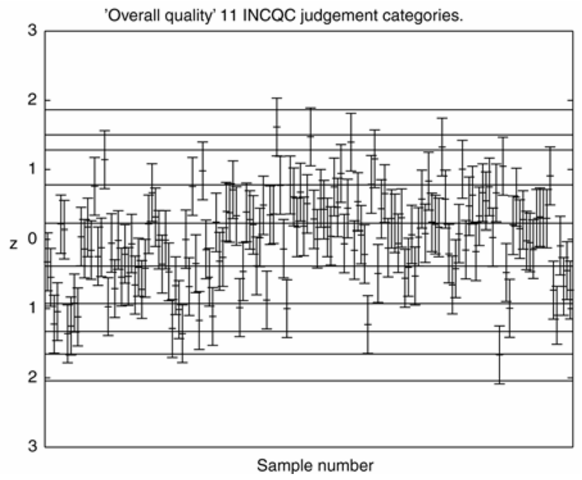

*Figure 2. Thurstonian scaling of the category data for 'overall quality of the Showtime image, showing mean and 95% confidence intervals. Category boundaries are illustrated as horizontal lines.*

Previous work has suggested that the number of fundamental dimensions of quality is small.[4] The correlation between the scores for individual attributes and the 'overall quality' in the INCQC 2000 and 2002 data is shown in Table 1, where S1, S2 and S2 refer to the different test images. It can be seen that the highest correlation with the overall score in 2002 was obtained for 'colour' judgement, which is in agreement with Montag.[4] However, when the criterion 'flesh tone' was present in 2002, this had the highest correlation with overall quality, and 'colour' had the second highest. In general, the correlation between the scores for individual attributes was high, and the high level of these correlations suggest that the attributes are not independent of each other and a smaller number would be sufficient to describe the quality dimensions.

An important question in image quality modelling is how well overall image quality can be predicted from the ratings of individual quality attributes (cf. sharpness, tone gradation etc). One approach is to use multiple regression to find which, if any, criteria, can explain most of the variance in the 'overall quality. For the INCQC 2000 it was found that that the perceived quality of the flesh tones adds substantially to the prediction of overall quality ($\beta$ = 0.563), together with the quality of the colour rendition ($\beta$ = 0,27). In the INCQC jury evaluation of 2002, the criterion 'flesh tones' was not included, and 'colour accuracy ($\beta$ = 0.584) and 'sharpness '($\beta$ = 0.204) were the largest contributing factors to the 'overall quality' judgement.

**Table 1. Correlations between the scores for individual attributes and the overall quality attribute in INCQC 2000 and INCQC 2002**

| | 2000 | 2002 | |
|---|---|---|---|
| Criterion | S1 | S2 | S3 |
| Overall | 1.00 | 1.00 | 1.00 |
| Colour | 0.92 | 0.95 | 0.98 |
| Neutrality | | 0.80 | 0.82 |
| Flesh tones | 0.94 | | |
| Highlight | 0.84 | 0.87 | 0.95 |
| Shadow | 0.62 | 0.88 | 0.77 |
| Gradation | | 0.92 | 0.94 |
| Smoothness | | 0.75 | 0.87 |
| Screening | 0.78 | 0.59 | 0.78 |
| Detail | | 0.90 | 0.93 |
| Sharpness | 0.83 | 0.86 | 0.88 |
| Register | 0.71 | | |
| Noise | 0.82 | | |
| Artifacts | | 0.42 | 0.63 |

Psychophysical experiments on image quality usually report average scale values, some exceptions being Cui's work on the repeatability of paired comparisons[5] and Keelan's studies of observer and scene variations.[2] It is important to know the variability of observers in order to understand the precision of the measurement. The INCQC results are interesting because of the large amount of samples that is evaluated by the judges. The agreement between judges in terms of correlation coefficients was investigated. The inter-individual correlations were found to be only moderate, 0.52 for the 'overall quality' criterion, and in the range of 0.45-0.5 for the different quality attributes. The least agreed upon criteria were in general the artefact criteria such as 'screening quality' or 'mis-register'. It should be remembered that each of these correlation coefficients are calculated over more than 150 samples, and we do not know what to expect with such large sets. Fatigue could be a contributing factor or the physical range of variations in different attributes might have been small, and although discernible, some of the characteristics might have been within the acceptability tolerances.

### What Psychophysical Methods Can Be Used With Large Sample Sets?

A characteristic of INCQC is that during the Jury Evaluation event all the judges are performing their assessments at the same time, and are making their assessments on multiple quality attributes. For an assessment which is both parallel and multi-dimensional, only a limited number of methods are suitable. Here we review the psychophysical methods in common use and consider which have potential for this application

Paired comparison is an indirect method designed to construct an interval scale from a matrix of data showing the proportion of times each of a number of stimuli is judged to be greater in magnitude with respect to some attribute than another stimulus. The method requires each stimulus to be compared with every other, and the type of scale generation is called Thurstonian Scaling.[6,3] A modern and detailed description of the method is found in Engeldrum.[7] Paired comparison is the best method available for assessing small differences between the samples, but is unsuitable for larger differences. This is because the method has a limited dynamic range, and if there are large gaps in magnitude between two samples so that the higher quality sample is never confused with the lower quality sample, this gap cannot be correctly quantified because proportions of 0 and 1.0 yield undeterminable z-value. This would cause a saturation of scale values in situations where differences between samples are unambiguous.

Category scales are designed to measure attributes on an equal-appearing interval scale. It is one of the most popular scaling techniques because it is simple to set up, easy to explain to the observers, and can be self-administered once started. Observers are asked to assign a number of samples to a specified number of categories.

The categories are usually specified as numbers or as adjectives such as poor, good excellent.[3] The psychological scale value is often taken to be the average (or median) value obtained from a large number of repetitions, as in the case of INCQC. Methods are available to convert category scale values to an approximately interval scale. Torgerson's Law of Categorical Judgment[3] parallels Thurstone's Law of Comparative Judgment and has a similar formulation. Like the paired comparison case, this method is also built on a confusion matrix and if all observers place the samples in the same category these methods will not yield a solution.

A triplet comparison[8] is a two-stage assessment where samples are first assigned category values, and in the second stage, subsets of three samples are compared at the time. In an aggregated rank order, a large sample set is first divided into sub-sets of eight, each sub-set placed in rank order by the observers, and finally two samples from each sub-set of eight is assessed in a category judgement. The category scores are then used as reference points to scale the samples within the subset from which they were taken, making it possible to aggregate the results from each rank ordering.

Many of these alternate methods have advantages of low observer stress when performed on a single quality attribute. However, in practice only category judgement could be scaled up to the large judgement task of INCQC, since the sample sorting and re-presentation required become logistically intractable in a very large sample set, particularly where judges perform assessments in parallel on multiple quality attributes.

## Experimental

A series of experiments were conducted using the INCQC 2002 test prints, which allowed us to explore some of the issues described above. Sub-sets of between 8 and 70 of the prints were used, depending on the experiment design.

Two different scenes were evaluated in the INCQC 2002 sessions. The images are referred to as 'Showtime' and 'Autumn Leaves' respectively. In the INCQC Jury Evaluation, a photographic print (reflection copy) of 'Showtime' was present and used as a reference for the scaling task. 'Autumn Leaves' however was not available as reflection copy. The jury was instructed that the 'Showtime' image should be reproduced as close as possible to the photographic print and the 'Autumn Leaves' picture, should be reproduced to give a 'pleasing' reproduction that was close to what most people would call 'realistic'.

Subsets of print samples were selected on the basis of the results from the INCQC 2002 evaluation. An overview of samples, methods, image and observers, can be found in Table 2.

**Table 2. Overview of Studies Carried Out**

| Study | Method | Prints | Anchor/ control | Observers Naïve | Experts |
|---|---|---|---|---|---|
| I A | Pair | 10 | N | 15 | 0 |
| I B | Pair | 8 | Y | 6 | 7 |
| II A | Category | 16 | N | 11 | |
| II B | Category | 18 | Y | | 13 |
| II C | Category | 18 | Y | | 12 |
| III | Triplet | 20 | Y | 5 | 5 |

For the studies focusing on the repeatability of category scaling, using experienced observers (II B, C:1 and C:2), a set of 20 samples was selected that the reflected the variation in the total set evaluated by the INCQC jury. For the pair comparison tasks, two subsets were selected, one set representing a small variation in the original scores and one set representing a wider range of scores. All of the results reported here, with one exception, are based on the 'Showtime' image.

In the category scaling experiments, only six criteria were used. These were 'colour reproduction', (CR) 'colour balance' (CB), 'highlight reproduction' (HR), 'shadow reproduction' (SR), 'sharpness', and 'detail rendering' (detail). Experiment II B, C: 1 and C: 2 all investigated the same subset of samples using the exact same criteria and settings and are thus replicates of each other. An anchor sample was used selected on the basis of previous INCQC judgments.

Observations were made under ISO 3664 P1 viewing conditions.[9] This specifies an illuminance of 2000 lux with a chromaticity of D50, with a surround that is neutral and matt and has a reflectance of less than 60%. Most observers were experienced in colour judgement, and included 20 people working in colour reproduction at UK national newspapers.

## Results

In the results are discussed below under the headings of inter-observer agreement, the precision of the category scale, and the repeatability of the results.

### Agreement Among Observers

The average inter-observer correlation for the different criteria adopted in the INCQC judgments are in the range of approx 0.3-0.5. The same pattern holds for both 2000 and 2002. Although the correlation coefficients appear small, they are statistically significant ($p<0.05$) given this large amount of samples ($n=155$), which suggests there are relationships in the data between the judges. For the repeated studies with subsets of samples rated by groups with similar expertise as the INCQC jury, the correlation coefficients are only marginally higher, indicating that the disagreements among judges may not be an effect of the sample size only. However, on average, judges tend to agree, i.e., the average scale values correlate to a much higher degree than individuals agree on average. This is illustrated in Figures 3 and 4 where the scale values for 'highlight reproduction' obtained in study II B is plotted against those of II C:1, and the 'colour accuracy' scale values from study II C:1 are plotted against those of II C:2. It can be seen that the average scales agree well, with correlations of 0.93 and 0.88 respectively.

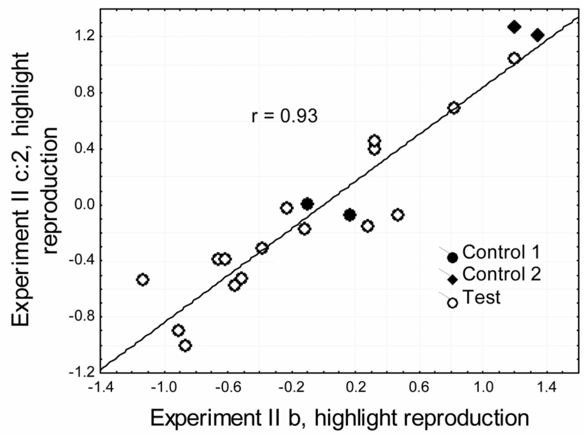

*Figure 3. Interval scale values obtained from two different groups for the 'highlight reproduction' attribute. Filled circles and squares are single-observer replicates.*

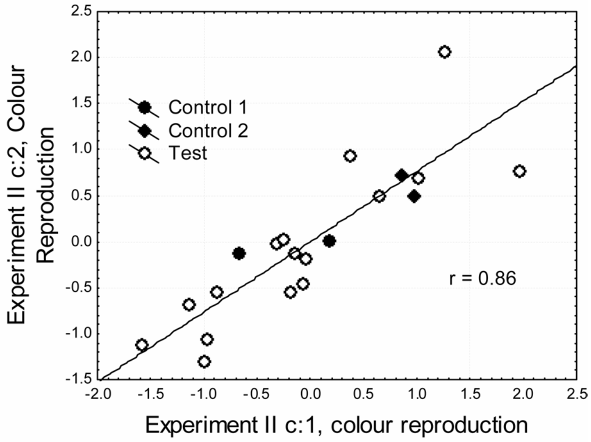

*Figure 4. Interval scale values obtained from two different groups for the 'colour reproduction' attribute.*

### Number of Categories and Just-Noticeable Preferences

Scale values from the two pair comparison experiments, study I A and I B (see Table 2), were calculated according to Thurstone's model. In addition, the just-noticeable preferences were derived from the pair comparison data as suggested by Keelan.[2] For the JND calculations, an arcsine distribution was used in order to better compensate for extreme proportions (0 and 1.0). The angular deviates were plotted against the mean score differences for the corresponding INCQC sample sets. The

JND was taken to be the point at which 75% of the observers reported that they preferred one print above the other. In experiment I A, the JND was found to correspond to 1.4 of the INCQC categories.

In study I B, the confusion between the samples was too high, resulting in a proportion matrix where very few proportions approached the 75% level, and the resulting angular deviates were too small and scattered to obtain a reliable slope of the psychometric function. These samples also obtained very similar scores from the INCQC jury, where the maximum score difference in the set was 0.85.

If the size of the JND, 1.4, obtained in study I A is taken as an indication of the 'true' JND of preference, this indicates that all 11 INCQC categories might not have been discernable. It should however be noted that the present result is based only on one image (Autumn Leaves) and the JND should be measured for both images before any firm conclusions can be made.

### Repeatability of the Results

The repeatability of the category scaling method was investigated by comparing the INCQC results with scales obtained from three different groups of observers scaling the same subset of samples. Interval scales were derived according to Torgerson's law of categorical judgment.[3] The scales were regressed against each other and against the INCQC results.

The results from INCQC evaluation are reported as sums of the individual criteria, weighted for the importance of the 'overall quality' criterion. Of the category judgment experiments reported here, all had six criteria and no 'overall' criterion. Therefore, only the sum of individual criteria without the weighting is reported here. The six criteria from the INCQC data that corresponded to the ones used in studies II B, C:1 and C:2 were summed together and correlated to the three scales from study II. The repeatability was found to be good for the summed categories. The correlations with the INCQC results were 0.86, 0.87, and 0.88 for the three replicate studies respectively. An example fit is shown in Figure 5.

The average scale values for each separate criterion were correlated against the INCQC results in order to study the repeatability of the single attributes. Table 3 shows the resulting correlations between the three different group scales and the corresponding INCQC subset scales. (Note that the scales are coded in opposite direction, resulting in the relationship having a negative sign.)

The agreement between scale values shown in Table 3 ranges from fair to good. An example is shown in Figure 6 where the scale values for 'highlight reproduction', obtained from the INCQC jury, are plotted against those obtained from a group of 12 observers, working mainly at UK newspapers.

**Table 3. Correlation between average scale values, obtained from three different groups, and the corresponding sub-scale from the INCQC evaluation.**

| Study | Colour | Hilights | Shadows | Sharpness | Detail |
|---|---|---|---|---|---|
| II B | -0.65 | -0.79 | -0.75 | -0.71 | -0.77 |
| II C:1 | -0.69 | -0.80 | -0.69 | -0.69 | -0.80 |
| II C:2 | -0.74 | -0.79 | -0.89 | -0.73 | -0.85 |

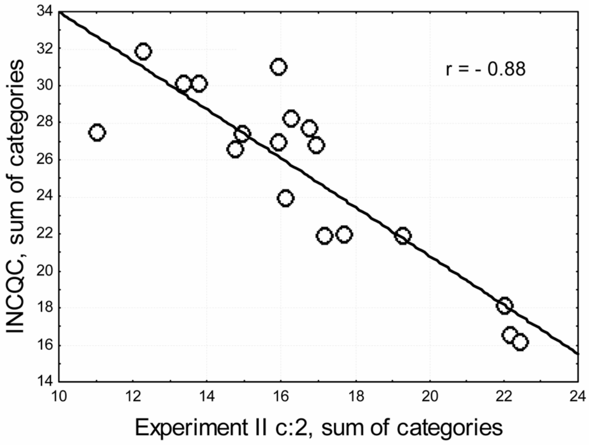

*Figure 5. Repeatability of the finalized scores. Five criteria: 'colour', 'highlight', 'shadow', 'detail' and 'sharpness', are summed and compared to the INCQC results summed over five corresponding categories.*

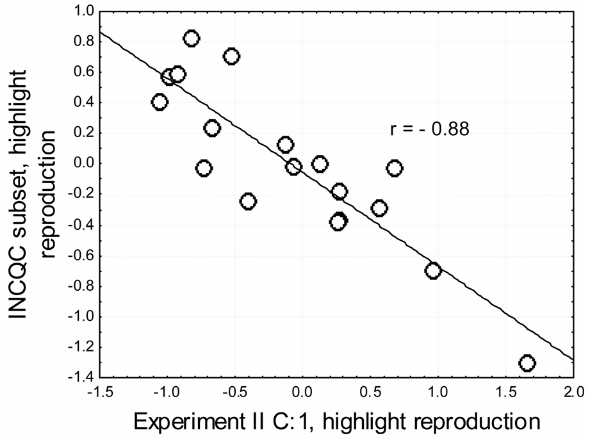

*Figure 6. Repeatability of individual criteria. Scale values for 'highlight' are plotted against the INCQC scale values. The category scales are converted to interval scales according to Thurstone.*

One set of reproductions was evaluated by the method described in ISO/CD 20462,8 which is a combination of category judgement and triplet comparison methods. If a constant scaling is applied to the scores arrived at by this method, there is good agreement with the scores obtained from a category judgement experiment performed with the same samples.

### Discussion

Psychophysical experiments on image quality usually report average scale values. It is however important to know the variability and repeatability in these scales. The relatively low correlation between individual judges in all experiments may be due to samples varying in more than one dimension, thus giving rise to inter-judge disagreement because they cannot effectively be ordered along one single perceptual continuum. If this were the case however, it could be expected that the correlation would be smallest for the 'overall quality' criterion, and a larger agreement would be found for specific criteria such as ‘sharpness’, highlight reproduction’, etc. However, the opposite occurs: correlation coefficients are in general smaller for the specific criterion than for the overall quality criteria. The reason for this can be twofold: The specific criterion may not be ambiguous, and/or it may in fact be multidimensional. Second, the range of physical variation of the criterion within the set of samples might be too small to be reliably discriminated among and thus resulting in ‘random scales’.

## Conclusions

From the results we obtained, we conclude that the category judgement method supports the assessment of a large sample set better than the other psychophysical techniques considered.

We also conclude that the repeatability of assessments is good. Sub-scales obtained from similar groups of observes agree well with the INCQC results. Although individual observers may agree only to a moderate degree, the average scale values show good agreement.

The category judgement method of visual assessment as used in INCQC can be improved. The results suggest:

1. The inclusion of an anchor image could reduce inter-observer and intra-observer variability. The quality of this anchor image on the attributes evaluated could be determined in a preliminary observer task
2. The use of 11 judgement categories is too fine a scale, and has the effect of increasing observer stress and setting the category boundaries too close to just-noticeable difference thresholds. Analysis of the results for INCQC 2000 and 2002 indicates this number could be reduced to seven.
3. The use of 11 quality attributes increases observer stress for no apparent benefit, since the number of independent quality attributes is considerably lower. Hence the number of quality attributes in INCQC could be reduced to approximately six.
4. The data available for INCQC is insufficient to determine intra-observer variation, and more repeats should be made during an assessment in order to better quantify this parameter.

We also suggest that further work is needed on the methods of assessing large sample sets. As an example, just-noticeable preferences for criteria such as those included in the INCQC evaluation cannot be determined directly from category scaling observations but could be assessed by selected paired comparison and related to the category scales.[2]

Some of these conclusions have already been recognized by IFRA and will be implemented in INCQC 2004.[10] In particular the number of quality attributes to be judged has been reduced to five: ‘colour quality’, ‘reproduction of detail’, ‘tone gradation’, ‘sharpness and screening’ and ‘overall image quality’.

## Acknowledgements

The authors would like to express their gratitude to IFRA for providing material support for this work, including access to data and prints from previous INCQC events. We would also like to thank all the observers who gave up their time to participate in the experiments.

## References

1. Ifra (2003) *The International Newspaper Color Quality Club* – http://www.colorqualityclub.org
2. Keelan, B.W. (2002) *Handbook of image quality: characterization and prediction*. Optical Engineering series Vol. 75. New York: Marcel Dekker Inc.
3. Torgerson, W. S. (1958). *Theory and Methods of Scaling*. New York: Wiley.
4. Montag, E. (2001) Multidimensional analysis reveals importance of color for image quality. *Proc. 9th IS&T/SID Color Imaging Conf* 17-21
5. Cui, C. (2001) On the repeatability of paired comparison based scaling methods. *Proc IS&Ts 2001 PICS Conference* 113-118
6. Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.
7. Engeldrum, P. (2000) *Psychometric Scaling*, Winchester, MA: Imcotek Press
8. ISO (2003) *ISO /CD 20462 Photography — Psychophysical experimental method to estimate image quality — Part 2: Triplet comparison method* ISO, Geneva
9. ISO (2000). *ISO 3664:2000 Viewing conditions - Prints, transparencies and substrates for graphic arts technology and photography.* Geneva: ISO.
10. Ifra (2004) *INCQC 2004-2006 Instructions for participants* Darmstadt: Ifra

## Biography

**Phil Green** is a member of the Colour Imaging Group at London College of Printing, and Course Director of the college's postgraduate programme in Digital Colour Imaging.

He worked in the printing industry from 1975, and joined the London College of Printing in 1986. He received an MSc from the University of Surrey in 1995, and a PhD from the University of Derby in 2003. He has authored a number of graphic arts textbooks, and recently edited Colour Engineering (Wiley) with Lindsay MacDonald.

He is active in CIE TC8-03 Gamut Mapping and his research interests are in graphic arts colour reproduction.