Image Quality: Between Science and Fiction

Sergej Yendrikhovskij
IPO, Center for Research on User-System Interaction, Eindhoven, The Netherlands/ Colour & Imaging Institute, Derby, United Kingdom

Abstract

At present, most studies on Image Quality employ subjective assessment with only one goal - to avoid it in the future. Such an attitude is most likely caused by some methodological difficulties involved in psychophysical scaling, and the skepticism towards the prospect of modeling cognitive stages of human judgments. The following paper (1) reviews the main methodological problems, (2) discusses possible remedies, (3) considers a model, which specifies the major constraints imposed on Image Quality, and (4) provides a meaningful classification of different image categories.

Introduction

Image Quality (IQ) can be considered as fiction in three senses. First, such an abstract attribute as 'quality', defined by the Oxford Dictionary as 'degree of excellence', is fiction because it cannot be found in the material world and exists in *subjective realm* only. Second, an image of high quality is sometimes referred to 'ideal', which is a *figment of observer's imagination* and might not exist at all (even in subjective realm). Third, preference judgments on IQ are occasionally brought up in the literature as something, which is very difficult to be investigated at presence, but may be studied in the *distant future*.

The term 'fiction' in this article does not have a negative meaning. Instead of 'fiction', one can use more neutral terms 'psychological experience' or 'cognitive phenomena', which does not change its subjective nature. However, there always will be a pessimist who would question its existence and possibility of its scientific investigation. The research on IQ has to share the difficult fate of all researches on mental attributes (including its twin, another IQ, Intelligence Quotient), which involves the continual balance between satisfying the objectivity demanded by scientific approach (is it already science?) and preserving the subjectivity demanded by the ultimate nature of the phenomena under consideration (is it still fiction?). The aim of this paper is to analyze the current state of this balance for IQ research.

Is it Already Science?

Almost half century ago MacAdam[1] proposed the following program for IQ research: (1) to vary images in systematic manner; (2) to perform visual judgments; (3) to analyze the judgments and physical measurements for correlation; (4) to find the optima for different conditions; (5) to find exceptions; and (6) to improve the model of the optima. This is the program, which in general can be found in many present researches on IQ.

Let us analyze whether the study of IQ following this program has reached the level of being called 'science', i.e. body of knowledge that aims through observation, experiment and induction to provide a reliable explanation of phenomena with reference to physical world.[2] Further, science usually includes the main substantive theory, a set of often complex and problematic auxiliary theories, plus empirical conditions describing the experimental context of the research.[3]

Auxiliary Theories

It would not be an exaggeration to say that the present progress of the IQ research would not be possible without the progress of psychophysics and scaling techniques. The 'local' (concerning small stimulus differences) and 'global' (concerning the full range of stimulus differences) psychophysics, the laws of comparative and categorical judgments, multidimensional scaling, and the signal detection theory compose the core of modern psychophysical measurements (see Ref. 4 for review) and become the auxiliary basis for IQ measurement.

The possibility to measure, i.e. to quantify in a consistent and meaningful way, the attribute under investigation is a necessary condition for any scientific research. For an observer, in principle, there should not be much difference in saying 'image of a high quality' and 'image of a quality that equals 9 on the scale from 0 to 10' (at least, we assume so!). But for a scientist there is a fundamental distinction since the latter judgments include an origin and a unit that can be related to another origin and a unit of a physical scale. This relationship can be analyzed for consistency or correlation.

Early studies on IQ usually investigated what type of scale (ordinal, interval, or ratio) provides more stable results, and whether intervals of these scales are subjectively invariant. For example, Jones and Marks[5] found that judgments on the ratio scale are more consistent than on the ordinal scale and provides more meaningful intervals in the numerical responses.

The focus of recent studies on IQ has shifted towards developing computational metrics that predict IQ in the form of a single number that correlates with observers judgments. The IQ metrics are generally derived from physical parameters of images and basic properties of the human visual system (HVS). Therefore, other important auxiliary theories of IQ research include models dealing with spatial, temporal and contrast sensitivity functions of HVS. Overviews of some modern IQ metrics and relevant HVS functions were given for example by Jacobson[6] and Farrell et al.[7]

Experimental Context

The major problem with all scaling methods is the impact of experimental context on the observer's response. The response is determined not by the presented stimulus only, but by the total context of the other stimuli in the experiment.[8] The detailed analyses of the response-bias can be found in Ref. 9, and its consequences for IQ research in Ref. 10. The context-dependency in the chain sensation-response presents a risk for internal validity of IQ experiments that are based on the psychometrical function stimulus-sensation. A possible remedy may come from studies that try to model the context effects[8,11] and a recently proposed 'integral' psychophysics,[12] which aims to investigate judgments of objects (e.g. people, flowers) rather then judgments of attributes (e.g. brightness, color). Another solution is to explore non-scaling techniques, for instance recording eye movements.[13] However, this approach might increase complexity of data interpretation.

Another problem of scaling methods, which is especially important for IQ research, is individual and cultural differences. There is evidence that magnitude estimation in general and subjective preferences in particular for the same stimulus might vary from one individual to another,[11] between different cultures,[14] and between naïve observers and experts.[13] These differences threatens the external validity of IQ research which is naturally interested in generalizing data found in a particular experiment on the whole population of potential customers. The appealing solution of this problem can be adopted from advanced multidimensional programs such as SINDSCAL[15] that provides an opportunity to represent a common stimulus space in which individuals (groups, cultures etc.) are permitted to have different weights for essentially the same underlying dimensions.

The Substantive Theory

It is important to realize that the auxiliary theories of visual psychophysics and physiology are not the substantive theory of IQ. Psychophysical scaling, for example, is just an instrument to relate the psychological impression to the numerical representation, and can be applied to the estimation of beauty, happiness, and 'attitude towards movies'[16] without reference to any physical magnitudes. A variable does not have scientific value just because it can be represented on a numerical scale.[17] Determining the underlying dimensions of the attribute is a critical first step in scientific progress.

The ADONIS research project aimed exactly to make this step and specify the primary perceptual factors that substantially contribute to IQ.[18] A panel of consumers was asked to list all words that they could think of in terms of IQ and trained to use these words in scaling judgments of displayed images. The results of these judgments were correlated well with some physical display's measurements.

The basic idea behind the ADONIS approach was to empirically define the dimensions of IQ and its subjective importance in order to develop a model, which would predict the overall impression of IQ without presence of an observer. This would be a very powerful model in practice, which can help engineers and designers to focus their effort on improvement the most significant aspects of IQ. However, such a model (if developed) would have a low scientific value, because it would not be able to elucidate *why* those were the primary dimensions, and *why* one dimension has a higher subjective weight than another. That is, the model would fail to provide a reliable explanation of the IQ phenomena. Despite a long history of IQ research no substantive IQ theory has yet become universally acceptable.

Is it Still Fiction?

There is a certain danger that the demand for objectivity of IQ research would 'throw the baby out with the bath water'. This danger comes from two directions. The phenomenon of IQ is sometimes replaced by the phenomenon of impairment, which although being closely related is different from IQ itself. Second, the modern IQ related indices do not go beyond the models of visual system, which are not the only models determining the subjective impression of IQ.

Quality and Impairment Phenomena

Due to (1) the absence of substantive IQ theory, (2) the close inverse relationship between quality and impairment judgments,[19] and (3) the evidence that the impairment terms produce more regular and context-independent scales than the quality terms,[5] it is appealing to think about quality as something opposite of impairment.

From this perspective, an ideal image is simply an image without noticeable distortions and its quality can be derived from IQ metrics based on visibility of annoying artifacts. The impairment metrics basically raise doubts about the existence of IQ phenomena. Moreover, such metrics would never be able to explain the fact that some artifacts *are* preferable by observers. For example, it was shown that subjects have a slight but significant preference for more colorful images, despite the fact that they realize that these images look somewhat unnatural.[20]

Visual and Cognitive Models

Most of modern IQ metrics operating with MTF, CSF, SQRI and sCIELAB are based on the properties of the visual system only, and do not take into account cognitive aspects involved in IQ judgments. Such metrics can be very useful to define the subjective tolerance to image errors, especially at the threshold level. They also can provide the answer on the question 'what do people *accept* to see in images'; but they are unable to respond on other important questions of the supra-threshold level 'what do people *expect* to see' and 'what do people *prefer* to see' in observed pictures.

To answer these questions, one has to think about cognitive models. Since reproduced in original images are seldom observed side by side, and image appraising is usually based on some 'mental recollection' of previously experienced sensations,[21] the critical role in analyzing quality judgments should be assigned to memory models.

In the case of color, a very useful description of 'mental recollections' was given within the concept of 'memory colors', i.e. colors that are recalled in association with familiar objects.[22] Several studies revealed a discrepancy between memory colors and average colors of actual objects.[21-24] There is evidence of a significant increase in saturation of memory colors for some object categories (e.g. grass, sky, and food items). However, other categories show no such shift (e.g. sand, skin), or produce it in the opposite direction (e.g. concrete).

Similar data have been obtained for preferred colors. Experimental results indicated that subjects not only remember but also prefer some object colors to be slightly more saturated in comparison with actual colors.[24] Again, the differences between preference and actual colors appear to be object-dependent.[25] For example, it was found that color coordinates of the preferred blue sky that were derived for reflection prints viewed in daylight are somewhat more saturated than real blue sky, while preferred green grass and Caucasian skin have about the same saturation as real grass and skin but were both a little yellow.[26]

To explain the object-dependency of memory colors, Newhall et al.[23] have proposed an interesting hypothesis that memory colors exhibit the shift in the direction of the *prototypical* colors associated with the actual objects. Since the prototypical color for sky is blue, than the actual color of sky might be shifted in memory towards 'more blue', while the prototypical color of concrete is gray, and therefore the memory color for concrete might be shifted towards 'more gray'. In short, the color shift in memory and preference judgments might be explained by the effect of prototypical colors.

These ideas bring into discussion the impact of generalization and categorization processes on the formation of memory colors and object appraising. Objects are seldom interpreted sui generis; usually they are associated with a particular group—what is called a 'natural kind' or category. Categories are extracted through the process of generalization from the population of apparent object seen in the past. The theory on generalization proposed by Shepard[27] can be considered an example of an explicit explanation for generalization principles that govern the organism's behavior.

The generalization process is assumed to result in construction of the category and its prototype, or the most typical category member. While the Classical Categorization model represented categories by sets of characteristic features, the Prototype Theory introduced the concept of probability and topology in the categorization process: a category is represented by a set of stimuli around the most frequently occurred items. The General Recognition Theory went further in explicit modeling the categorization process and assumed that the structure of natural categories can be effectively modeled by a multivariate normal (Gaussian) distribution. Basically, this is an extension of the Signal Detection Theory into multidimensional cognitive domain.

For overview of different models of categorization in perception and cognition see Ref. 28, and for their application to assessments of color reproductions see Ref. 29. At this point, it is necessary to emphasize that there *are* cognitive models that explicitly specify the basic psychological processes and, therefore can be utilized by IQ research as additional auxiliary theories.

Is There a Balance?

Let us summarize 'pros and cons' of whether the research on IQ has reached the balance between fulfilling the objectivity demanded by scientific approach and preserving the subjectivity demanded by the nature of IQ phenomena.

Pros:

1. a lot of experimental results on systematically varied images were amassed over the past 50 years
2. the results were obtained using rigorous measurements of psychophysical techniques
3. the results can be quantified and related to physical measurements
4. a number of computational IQ metrics has been developed trying to predict IQ in correspondence with observers' judgments
5. the IQ metrics are based on the sensitivity of the human visual system

Cons:

1. the observer's response obtained by psychophysical measurements is context-dependent
2. the observer's response may vary significantly from one individual to another, and between cultures
3. the possibility to replace IQ by the impairment, questions the existence of IQ phenomenon
4. the modern IQ metrics do not take into account the cognitive aspects involved in IQ judgments
5. the absence of universally acceptable IQ theory that provides a reliable explanation of IQ phenomena

This article has shown so far that the first four problems can be resolved in principle by 1) including context effect in the 'integral' psychophysical model, 2) representing a common stimulus space in which subjects are permitted to have different weights, 3) distinguishing impairment and quality judgments, and 4) comprising cognitive models in IQ metrics.

While the balance between pros and cons presented above might be disputed, one can say with certainty that if the lack of balance is recognized then it mainly comes from the luck of the substantive IQ model. The Rubicon of induction, i.e. general inference from particular observations, has not been passed in IQ research yet. At present, there is no widespread model that combine all collected data into a coherent theory that can explain why subjects prefer one image to another, and how the subjective preferences can be predicted from objective parameters of images. The following section discusses a candidate for such a model.

A New Model of Image Quality

Any model that aims to explicitly describe a phenomenon should consider the ontological aspect (what *is* the phenomenon), the functionality aspect (what is the phenomenon *for*), and the causality aspect (what is the phenomenon *caused by*).

These aspects were partly discussed by Janssen and Blommaert[30] in a new model of IQ. In contrast to the traditional 'signal processing' approach towards IQ, they regarded the processing of images by the visuo-cognitive system as the processing of *visual information* about the outside world. From this perspective, IQ can be defined as 'the degree to which the image can be successfully exploited by the observer'.

They argued that an image with successful exploitation, or an image of 'high' quality should satisfy two constraints: usefulness and naturalness. The usefulness was described as the precision of the visual representation and related to discriminability of items in the image. The naturalness was described as the degree of correspondence between the visual representation and the memory representation, and related to recognizability of items in the image. Since the two constraints might conflict with each other, the quality of an image can be modeled as a compromise between them. Note that the outcomes of such a compromise may require exaggeration of certain features of the image (e.g. by means of increasing brightness or color contrast) resulting in a less natural but more useful appearance of the image. Indeed, for chroma and lightness variation of natural images in the CIELUV color space a small but systematic difference was found between quality and naturalness judgments.[20,30]

The model of Janssen and Blommaert is a refinement of an earlier model for the color domain by Yendrikhovskij et al.[29,31] who elaborated on under-standing, measuring and optimizing perceived color quality of natural images. The color IQ model assumes that when people look at images containing familiar categories of objects, two primary factors shape their subjective impression of how optimal colors are reproduced: colors should be perceived both realistically (naturalness constraint), and distinctly (colorfulness constraint). This hypothesis was tested in a number of experiments with CRT displayed images. One of the main results of these experiments is that the quality judgments can be almost perfectly modeled as a linear combination of the colorfulness and naturalness judgments.

They also specified the naturalness, colorfulness and quality indices on the basis of the analysis of the observer's judgments in relation to statistical parameters of the color point distribution across the images in the CIELUV color space. The naturalness index can be estimated locally within the 'skin', 'sky', and 'grass' segments in the CIE u'v' chromaticity diagram, and can be described by a probability density function. The colorfulness index can be defined as a sum of the average saturation value and its standard deviation. The color quality index can be calculated as a weighted sum of the colorfulness and naturalness indices. The proposed quality index was used to optimize the perceived quality of color reproduction of natural scenes.

A Space Called GUN

The classical (signal processing) approach and the new (information processing) approach have essentially different view on IQ. However, both approaches highlight important requirements that determine IQ. These requirements can be used to represent different categories of images in one common space. Because this space assumes the three major constraints determining IQ: Genuineness, Usefulness and Naturalness, it will be referred as the GUN space. These constraints are derived from the basic components involved in IQ judgments: (1) the original image&environment, (2) the imaging device, (3) the reproduced image & environment, and (4) the observer&task (see Figure 1).

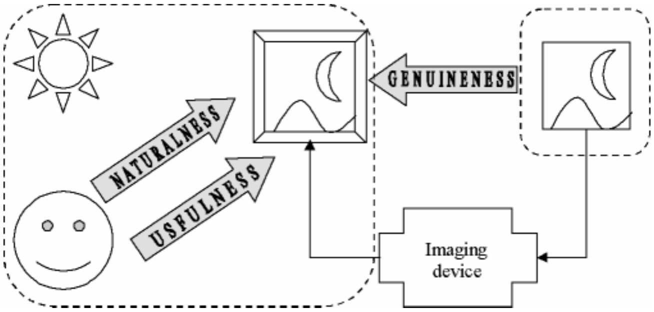

Figure 1. The diagram showing the GUN constraints.

Genuineness will be referred to as the degree of apparent similarity of reproduced image&environment with the external reference, i.e. original image&environment. This is the attribute related to research on visibility of annoying artifacts, perceptual error measure, and fidelity metrics. The original image can be presented by unprocessed, unsampled, uncoded images etc. It is important to realize that the genuineness is an apparent attribute, which might be influenced by the image environment.

Images are usually viewed with different illumination and surroundings, under different conditions of adaptation and view angle than the original scene. All these factors influence the appearance of the image. The distinction made by Hunt[21] for different levels of color reproduction is especially important here. The recently[32] proposed CIE 1997 Color Appearance Model, CIECAM97s is an important step in going beyond the psychometrical spaces (e.g. CIELUV, CIELAB) and establishing a truly appearance color space.

Ideally, an image with highest degree of genuineness should give an impression of looking through the window, or mirror, which shows 'no lies'. This requirement is crucial for proofs, catalogues, fine art etc. (see Figure 2).

Usefulness will be referred to as the degree of apparent correspondence of the reproduced image&environment with the observer&task activity. The main criterion of usefulness– the maximum discriminability of the items that are represented—is under current research on optimal visual metrics.[33] One of the main argument of this research is that visual metrics are intrinsically flexible, and that this flexibility is exploited to optimize the discriminative power of the metrics. The tendency to increase discriminative power can be found not only in the well-known phenomenon of light-dark adaptation, but also among many our interactions with the environment in our daily life: illuminating a room, cleaning spectacles etc. Microscopes or telescopes would never be invented without the key-tendency of human beings 'to see more'.[34]

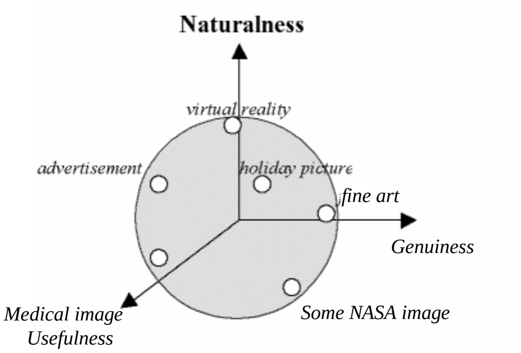

Figure2. A representation of image categories in the GUN space.

Interestingly, the 'exaggeration' of certain features demanded by the maximum discriminability can be found in completely different activities, such as using make-up. By analogy with the make-up idea, Judd[35] proposed to use the so-called Flattery Index, as a special case of the Color Rendering Index.[36] Illuminants that render object-color closer to the preferred rather than the actual color (e.g. by increasing its saturation) were proposed to have a high 'flattery index'. Many image enhancement algorithms actually follow the same inclination.

Ideally, an image with the highest degree of usefulness should represent the maximum distinguishable items. One can draw on the analogy with reading glasses, which are also task-dependent. The usefulness requirement is crucial for medical images, military night-vision images, for some NASA images etc. However, in other NASA images, e.g. pictures obtained from the Mars mission, one probably would like to keep the genuineness requirements at least partially valid (see Figure 2).

Naturalness will be referred to as the degree of apparent similarity between the reproduced image&environment and the internal references, i.e. memory prototypes. The influence of this attribute on the quality judgments becomes substantial when no external reference, i.e. the original, is available for observers: watching TV, looking at photos, browsing through the internet etc. Recently, the naturalness judgments were systematically studied,[37] and attempts were made to develop a Color Naturalness Index (CNI) that evaluates the perceived naturalness of reproduced colors in accordance with observer's judgments. Basically, it gives a measure of the similarity between the colors of objects presented in an image and the prototypical colors of the corresponding object categories. The following is a short overview of a procedure for determining the CNI.

Step 1: Digitized color images are represented as color statistics, i.e. color point distributions, in some perceptual and approximately uniform color space.

Step 2: The color statistics of images of natural scenes are divided into three privilege segments, for convenience referred to as 'skin', 'grass', and 'sky' segments.

Step 3: The local CNI is calculated by means of a Gaussian function of the differences between the coordinates of the average and 'prototypical' colors for a privileged segment.

Step 4: The global CNI is calculated as the arithmetic mean weighted of the 3 local CNI.

The current calculations of CNI are based on the three internal references only. The same concept can be extended for a larger number of categories. For example, the CNI with six references (e.g. red, orange, yellow, green, blue, and purple) can be developed using basic colors found by Boynton and Oslon.[38] Figure 3 schematically illustrates a possible segmentation the color statistics of an image into six regions, and comparison of the segments' mean values with the six prototypical colors. Note that this approach is essentially similar to calculating the CIE Color-Rendering Index and the Flattery Index.

The application of prototypical colors in the calculation of the naturalness index might be considered as a first step towards incorporating memory aspects in quality models. It can also form the basis for developing a uniform cognitive color space. The equality of distances between categories or prototypes (not between percepts) can be the criterion of uniformity for such a space. Following the large impact of the categorization process on perception and cognition, one can speculate that some image processing techniques, e.g. gamut mapping, color quantization, segmentation, coding etc. might be more appropriate to perform in the cognitively uniform color space.

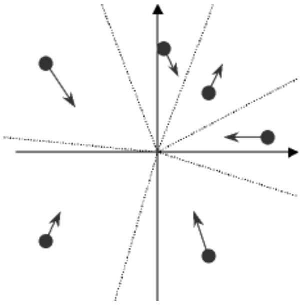

Figure 3. A depiction of calculating the CNI with six average (tips of arrows) and prototypical (solid circles) colors.

Ideally, an image with the highest degree of naturalness should represent only prototypical properties of objects in an image (one of the close analogy are sketches and caricatures). The naturalness requirement is expressed for examples in cartoons and partially in virtual reality images, i.e. simplified and somehow 'idealized' representation of the real world, which fulfils observer's expectation about it.

To summarize, consider an over-saturated original picture that has to be reproduced by three color devices. The Genuineness-devise ('no lies' strategy) will reproduced this picture perceptually the same; the Usefulness-devise ('see more' strategy) will try to increase the saturation further; the Naturalness-devise ('no surprise' strategy) will decrease the saturation down to the prototypical level.

Conclusion

The phenomenon of Image Quality is sometimes illusive, and its measurement techniques are not always reliable. Because of this, most of the popular Image Quality metrics are based on the low-level visual processing only. However, the analysis provided in this article shows that there is *no* real reason why Image Quality research should stay at this level any longer. There are many exciting perspectives ahead.

References

1. D. L. MacAdam, "Quality of color reproduction," *Proceedings of the Institute of Radio Engineers*, **36**, 468, (1951).
2. "The Hutchinson encyclopedia of science," Helicon, Oxford, 1998.
3. K. Popper, "The logic of scientific discovery," Basic Books, New York, 1959.
4. R. D. Luce, and C. L. Krumhansl, "Measurement, scaling, and psychophysics," in: *Stevens's handbook of experimental psychology*, 2nd Ed., Ed. R. Atkinson, R. Herrnstein, G. Lindzey and R. D. Luce, New York, Wiley Interscience, 3, (1988).
5. B. L. Jones, and L. E. Marks, "Picture quality assessment: a comparison of ratio and ordinal scale," *SMPTE Journal,* **94**, 1244, (1985).
6. R. E. Jacobson, "An evaluation of image quality metrics," *The Journal of Photographic Science*, **43,** 7, (1995).
7. J. E. Farrell, X. Zhang, van den Branden Lambrecht, and D. A. Silverstein, "Image quality metrics based on single and multichannel models of visual processing," *Proceedings of the IEEE COMPCON*, (1997).
8. A. Parducci, and D. H. Wedell, "The category effect with rating scales: number of categories, number of stimuli, and method of presentation," *Journal of Experimental Psychology: Human Perception and Performance*, **12,** 496, (1986).
9. E. Poulton, "Bias in quantifying judgments," Hove and London, London, 1989.
10. H. de Ridder, "Psychophysical evaluation of image quality," *Proceedings of SPIE Conference*, **3299**, 252, (1998).
11. S. N. Yendrikhovskij, H. de Ridder, E. A. Fedorovskaya, and F. J. J. Blommaert, "Colourfulness judgements of natural scenes," *Acta Psychologica*, **97,** 73, (1997).
12. G. R. Lockhead, "Psychophysical scaling: judgements of attributes or objects?," *Behav. and Brain Scien.*, **15,** 543, (1992).
13. G. Deffner, M. Yuasa, M. McKeon, and D. Arndt, "Evaluation of display-image quality: experts vs. non-experts," *SID 94 DIGEST,* 475, (1994).
14. M. Saito, "Comparative studies on color preference in Japan and other Asian regions, with special emphasis on the preference for white," *Color Res. and Applic.*, **21,** 35, (1996).
15. S.S. Schiffman, M. L. Reynolds, and F. W. Young, "Introduction to Multidimensional Scaling. Theory, Methods, and Applications," Academic Press, New York, 1981.
16. L. L. Thurstone, "A scale for measuring attitude towards the movies," *Journal of Educational Research***, 22**, 89, (1930).
17. R. M. Dawes, "Psychological measurement," *Psychological Review***, 101**, 278, (1994).
18. S. Bech, R. Hamberg, M. R.M. Nijenhuis, C. Teunissen, H. Looren de Jong, P. Houben, and S. K. Pramanik, "The RaPID Perceptual Image Description Method (RaPID)," *Proceedings of the SPIE Conference*, **2657,** 317, (1996).
19. M. R. M. Nijenhuis, and F. J. J. Blommaert, "Perceptual error measure for sampled and interpolated images," *Journal of Imaging Science and Technology***, 41**, 249, (1997).
20. E. A. Fedorovskaya, H. de Ridder, and F. J. J. Blommaert, "Chroma variations and perceived quality of color images of natural scenes," *Color Research and Applications***, 22**, 96, (1997).
21. R. W. G. Hunt, "The reproduction of Color in Photography, Printing and Television," Tolworth, Fountain, 1987.
22. C. J. Bartleson, "Memory colors of familiar objects," *Journal of the Optical Society of America***, 50**, 73, (1960).
23. S. M. Newhall, R.W. Burnham, and J.R. Clark, "Comparison of successive with simultaneous color matching," *Journal of the Optical Society of America***, 47**, 43, (1957).
24. P. Siple, and R. M. Springer, "Memory and preference for the colors of objects," *Perception and Psychophysics***, 34**, 363, (1983).
25. S. Sanders, "Color preference for natural objects," *Illuminating Engineering***, 54**, 452, (1959).
26. R. W. G. Hunt, I. T. Pitt, and L. M. Winter, "The preferred reproduction of blue sky, green grass and Caucasian skin in colour photography," *Journal of Photographical Science***, 22**, 144, (1974).
27. R. N. Shepard, "Toward a universal law of generalization for psychological science," *Science,* **273**, 1317, (1987).
28. F. Ashby, "Multidimensional models of categorization," in: *Multidimensional models of perception and cognition*, Ed. F. Ashby, Erlbaum, Hillsdale, 1992.
29. S. N. Yendrikhovskij, "Color reproduction and the naturalness constraint," PhD Thesis, ISBN: 90-386-0719-9, 1998.
30. T. J. W. M. Janssen, and F. J. J. Blommaert, "Image quality semantics," *Journal of Imaging Science and Technology***, 41**, 555, (1997).
31. S. N. Yendrikhovskij, F. J. J. Blommaert, and H. de Ridder, "Perceptually optimal color reproduction," *Proceedings of the SPIE Conference*, **3299**, 274, (1998).
32. M. R. Luo, and R. W. G. Hunt, "The structure of the CIE 1997 Colour Appearance Model (CIECAM97s)," *Color Research and Application*, **23,** 138, (1998).
33. T. J. W. M. Janssen, and F. J. J. Blommaert, "Visual metrics: discriminative power through flexibility," submitted to *Spatial Vision*.
34. R. M. Evans, "Eye, film, camera in color photography," Wiley, New York, 1959.
35. D. B. Judd, "A flattery index for artificial illuminants," *Illum. Engineering,* **42**, 593, (1967).
36. G. Wyszecki, and W. S. Stiles, "Color Science: Concepts and Methods, Quantitative Data and Formulae", 2nd edition, John Wiley and Sons, New York, 1982.
37. S. N. Yendrikhovskij, F.J.J. Blommaert, and H. de Ridder, "Color reproduction and the naturalness constraint," *Color Research and Application*, **24,** 3, (1999).
38. R. M. Boynton, and C. X. Oslon, "Locating basic colors in the OSA space," *Color Research and Application*, 12, 94, (1987).