INTERNATIONAL STANDARD

ISO 20462-2

First edition
2005-11-01

Photography — Psychophysical experimental methods for estimating image quality —

Part 2: Triplet comparison method

Photographie — Méthodes psychophysiques expérimentales pour estimer la qualité d'image —

Partie 2: Méthode comparative du triplet

Reference number
ISO 20462-2:2005(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 20462-2 was prepared by Technical Committee ISO/TC 42, *Photography*.

ISO 20462 consists of the following parts, under the general title *Photography — Psychophysical experimental method for estimating image quality*:

— *Part 1: Overview of psychophysical elements*

— *Part 2: Triplet comparison method*

— *Part 3: Quality ruler method*

Introduction

This part of ISO 20462 is necessary to provide a basis for visually assessing photographic image quality in a precise, repeatable and efficient manner. This part of ISO 20462 is needed in order to evaluate various test methods or image processing algorithms that may be used in other international and industry standards. For example, it should be used to perform subjective evaluation of exposure series images from digital cameras as part of the work needed for future revisions of ISO 12232.

The opportunities to create and observe images using different types of hard copy media and soft copy displays have increased significantly with advances in computer-based digital imaging technology. As a result, there is a need to develop requirements for obtaining colour-appearance matches between images produced using various media and display technologies under a variety of viewing conditions. To develop the necessary requirements, organizations, including the CIE and the ICC, are developing methods to compensate for the effect of different viewing conditions, and to map colours optimally across disparate media having different colour gamuts.

Such technical activities are often faced with the need to evaluate proposed methods or algorithms by visual assessment based on psychophysical experiments. K.M. Braun *et al.*[1] examined five viewing techniques for cross-media image comparisons in terms of sensitivity of scaling, and mental and physical stress for the observers. CIE TC1-27 “Specification of Colour Appearance for Reflective Media and Self-Luminous Display Comparisons” proposed guidelines for conducting psychophysical experiments for the evaluation of colorimetric and colour-appearance models[6]. Accordingly, for the design and evaluation of digital imaging systems, it is of great importance to develop a methodology for subjective visual assessment, so that reliable and stable results can be derived with minimum observer stress.

When performing a psychophysical experiment, it is highly desirable to obtain results that are precise and reproducible. In order to derive statistically reliable results, large numbers of observers are required and careful attention should be paid to the experimental setup. Multiple (repeated) assessments are also useful. Observer stress during the visual assessment process can adversely affect the results. The order of image presentation, and the types of questions or questionnaires addressed by the observers, can also affect the results.

Table 1 gives a comparison of three visual assessment techniques commonly used for image quality evaluation. The advantages of the category methods include low stress and high stability, since the observer’s task is to rank each image using typically five or seven categories. However, its scalability within a category is less precise. One of the most common techniques for image quality assessment is the paired comparison method. This method is particularly suited to assessing image quality when precise scalability is required. However, a serious problem with the paired comparison method is that the number of samples to be examined is to be relatively limited. As the number of the samples increases, the number of combinations becomes extensive. This causes excessive observer stress, which can affect the accuracy and repeatability of the results. The third method, commonly known as magnitude scaling, is magnitude estimation. This method is extremely difficult when the psychophysical experiments are conducted using ordinary (non-expert) observers to perform the image quality assessment.

Table 1 — Comparison of typical psychophysical experimental methods

Name of method	Scalability	Stability	Stress
Category	Low	High	Low
Magnitude estimation	Medium	Low	Medium
Paired comparison	High	High	High

G. Johnson *et al.*[3] have proposed “A sharpness rule”, where the magnitude of sharpness was analyzed in terms of resolution, contrast, noise and degree of sharpness-enhancement. Likewise, preferred skin colour may be considered not only from the viewpoint of chromaticity, but also with respect to the lightness, background and white point of the display media[4]. These examples show that image quality is not always evaluated by a single attribute, but may vary in combination with multiple attributes. In cases where a psychophysical experiment is designed for a new application, the experimenter may need to vary many attributes simultaneously during the course of the experiment. In these situations, the number of the samples to be examined becomes excessively large, making it difficult to employ the paired comparison technique.

Photography — Psychophysical experimental methods for estimating image quality —

Part 2: Triplet comparison method

1 Scope

This part of ISO 20462 defines a standard psychophysical experimental method for subjective image quality assessment of soft copy and hard copy still picture images.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
just noticeable difference
JND
stimulus difference that would lead to a 75:25 proportion of responses in a paired comparison task

2.2
psychophysical experimental method
experimental technique for subjective evaluation of image quality or attributes thereof, from which stimulus differences in units of JNDs may be estimated

cf. **categorical sort** (2.5), **paired comparison** (2.3) and **triplet comparison methods** (2.4)

2.3
paired comparison method
psychophysical method involving the choice of which of two simultaneously presented stimuli exhibits greater or lesser image quality or an attribute thereof, in accordance with a set of instructions given to the observer

NOTE Two limitations of the paired comparison method are as follows.

a) If all possible stimulus comparisons are done, as is usually the case, a large number of assessments are required for even modest numbers of experimental stimulus levels [if N levels are to be studied, $N(N-1)/2$ paired comparisons are needed].

b) If a stimulus difference exceeds approximately 1,5 JNDs, the magnitude of the stimulus difference cannot be directly estimated reliably because the response saturates as the proportions approach unanimity.

However, if a series of stimuli having no large gaps are assessed, the differences between more widely separated stimuli may be deduced indirectly by summing smaller, reliably determined (unsaturated) stimulus differences. The standard methods for transformation of paired comparison data to an interval scale (a scale linearly related to JNDs) perform statistically optimized procedures for inferring the stimulus differences, but they may yield unreliable results when too many of the stimulus differences are large enough ($>$ 1,5 JNDs) that they produce saturated responses.

2.4
triplet comparison
psychophysical method that involves the simultaneous scaling of three test stimuli with respect to image quality or an attribute thereof, in accordance with a set of instructions given to the observer

2.5
categorical sort method
psychophysical method involving the classification of a stimulus into one of several ordered categories, at least some of which are identified by adjectives or phrases that describe different levels of image quality or attributes thereof

NOTE The application of adjectival descriptors is strongly affected by the range of stimuli presented, so that it is difficult to compare the results of one categorical sort experiment to another. Range effects and the coarse quantization of categorical sort experiments also hinder conversion of the responses to JND units. Given these limitations, it is not possible to unambiguously map adjectival descriptors to JND units, but it is worth noting that in some experiments where a broad range of stimuli have been presented, the categories *excellent*, *very good*, *good*, *fair*, *poor*, and *not worth keeping* have been found to provide very roughly comparable intervals that average about six quality JNDs in width.

2.6
observer
individual performing the subjective evaluation task in a psychophysical method

3 Two-step psychophysical method

This part of ISO 20462 defines a new psychophysical experimental method, which satisfies the following requirements:

— enables a large number of samples to be examined;

— provides precise scalability;

— provides low observer stress;

— suitable for ordinary (non-expert) observers;

— provides high repeatability of the results.

The method comprises two steps. The first step is a “category step”, and the second step is a “triplet comparison step” which is newly developed for this purpose.

The reason for applying the “category step” is to reduce the number of the samples to an appropriate number which is determined by the purpose of each experiment. Typically this number is less than 27 samples. Category scaling using three categories, such as “favourable”, “acceptable” and “unacceptable” (or “acceptable”, “just acceptable” and “unacceptable”) is used for the first step, and samples are selected according to the number of samples required in the following step. If the number of test samples examined is relatively small, then the first step should be omitted, and the psychophysical experiment should start directly from the second step.

The second step is conducted in order to derive a precise scaling based on an interval scale. The present proposal is to use a newly developed triplet comparison method. In this method three samples are compared at a time, thereby achieving high assessment accuracy while keeping the experimental scale realistic.

NOTE If the normal paired comparison method were used with 21 samples, a total of 210 combinations would need to be examined. This is time-consuming and imposes excessive stress upon the observers. Furthermore, paired comparison methods require a significant number of observers in order that a precise scaling can be derived. This will result in an experiment that is excessively large and unrealizable.

4 Experimental procedure

4.1 Step 1

Proceed as follows.

a) Prepare the test images to be examined.

b) Observe each sample and rank it into 3 categories; “favourable”, “acceptable” and “unacceptable”.

c) Count the number of test images in each category.

d) Select the samples that will be used in Step 2 (4.2) from the upper category. It is recommended that the number of samples, N, be less than 27 in order to avoid observer stress during the experiment. The number of samples should obey the following equations:

$$N = 6K + 1 \text{ or } N = 6K + 3, \quad (1)$$

where

N is the number of samples;

K is an integer number.

NOTE It is possible to use 5 or 7 categories in the case of many samples.

4.2 Step 2

Proceed as follows.

a) Create combinations of samples for use in the triplet comparison step. Each combination shall consist of three samples. If the total number of the samples selected for the triplet comparison step satisfies Equation (1), then it is possible to arrange each combination of samples such that each pair of samples will only ever be viewed together once during the course of the experiment.

b) Observe the samples and rank them into 5 categories;

— 1: favourable,

— 2: acceptable,

— 3: just acceptable,

— 4: unacceptable, and

— 5: poor.

c) Apply Scheffe’s method for statistical analysis to obtain an interval scale.

NOTE See Annex E.

d) Convert interval scale to JNDs.

NOTE See Annex F.

Annex A
(informative)

Comparison between a paired comparison and a triplet comparison technique

The paired comparison method has traditionally been the most popular psychophysical method, capable of providing a high level of reliability and accuracy. However, the reproducibility of Scheffe's method with assessment scales (variations over repeated assessments) and the stress imposed on observers (due to prolonged assessment time caused by the increase in the number of combinations and fluctuation in the assessment scaling for paired comparison, etc.) have not been fully investigated.

The triplet comparison method has the desirable feature of reducing the level of stress on the observer. This is due to shortened assessment times and is expected to improve assessment accuracy and reproducibility. However, no experiments to validate these advantages have been conducted. Furthermore, the triplet comparison method inevitably yields a level of duplication in comparison for certain sample numbers, and the procedure for determining the minimum number of sample combinations has not yet been established. For various reasons, including those cited above, the triplet comparison method has not been commonly used in general subjective assessment experiments.

A series of experiments were conducted in order to assess the two comparison methods from the following aspects:

a) reproducibility (consistency) in terms of order fluctuation over a number of repeated assessments;

b) accuracy evaluated by the correlation between the orders determined by the two methods;

c) degree of difficulty expressed in terms of the degree of fluctuation for each sample, the necessary assessment time and the difficulty reflected in introspective reports;

d) stress on observers reflected in their introspective reports;

e) comparison of expert observers with naïve observers.

A set of experiments to assess favourable skin colour (tones) using the sample set described in References [5] and [6] was conducted for both comparison methods.

The experiments were repeated five times and the results, which are described in detail in Reference [7] of the Bibliography, are summarized as follows.

— In general the overall trends in assessment made by each method are similar.

— The triplet method can accommodate larger scales of assessment and is capable therefore of separating "favourable" samples from "unfavourable" ones more easily than the paired comparison method when assessment deviation is taken into consideration. A method for analysis that is more in agreement with the objectives of the assessment is therefore expected.

— It was generally noted that the assessment result obtained from the first run of the experiment was unreliable. The standard of the assessment scaling and its stability improved with subsequent repetitions of the experiment.

— The time required for assessment with the triplet method was about 1/3 of that required by the paired comparison method.

In conclusion, the two methods are similar with respect to their consistency and accuracy. The level of stress induced by the triplet method on the observer (due to assessment time) was about one third of the stress induced by the paired comparison method. This indicates that the triplet comparison method has the potential of achieving consistent, accurate (reliable) results while simultaneously reducing the level of stress induced on the observer. The primary aim of the triplet comparison technique is therefore fulfilled.

Annex B
(informative)

Number of sample combinations for triplet comparison

The number of sample combinations for paired comparison, N, is expressed by

$$N = {}_nC_2 = n(n-1)/2$$

where n is the number of samples and n = 2, 3, 4, 5, etc.

For the method of triplet comparison, if the number of samples selected, n' = 7, 9, 13, 15, 19, 21, 25 and 27, then it is possible to select sample combinations that eliminate the duplication of samples across combinations. More generally, the number of samples, n', can be expressed as:

$$n' = 6k + 1,\ 6k + 3\ (k = 1, 2, 3, 4, 5, 6, \text{etc.})$$

For any value n', the number of sample combinations, N', is calculated as:

$$N' = n'\ (n' - 1)/6$$

Let us place n' points on the circumference of a circle to form an n'-sided regular polygon. Each apex of the polygon, is assigned an integer value 1, 2, 3, . . . , n'. We define the notation whereby (p, q, r) represents a triangle comprising the apices p, q and r, and where the triangle apices represent a combination of samples for the triplet comparison method.

Examples of combinations without duplication are shown in Table B.1. In this table function f is defined as follows;

$$f(i) = 1 + \text{modulo}(i - 1, n')$$

where modulo(i – 1, n') represents the remainder for the division of (i – 1) by n'.

For the case of n' = 7, congruent triangles represented by (1, 2, 4), (2, 3, 5), (3, 4, 6), (4, 5, 7), (5, 6, 1), (6, 7, 2) and (7, 1, 3) give combinations without duplication.

For the case where n' = 6k + 1 and k = 1, 2, 3, 4, etc., combinations without duplication are represented by combining k differently shaped triangles chosen from the n' congruent acute angle triangles and the n' congruent obtuse angle ones.

For n' = 9, combinations without duplication can be achieved by triangles (1, 2, 4), (4, 5, 7), (7, 8, 1), (2, 3, 5), (5, 6, 8), (8, 9, 2), (1, 3, 6), (4, 6, 9), (7, 9, 3), (1, 5, 9), (4, 8, 3) and (7, 2, 6).

For n' = 15, the first thirty combinations are specified as follows: two combinations correspond to the apices of triangles (1, 3, 9) and (1, 2, 5). Twenty-eight combinations are formed when the apices of triangles (1, 3, 9) and (1, 2, 5) are moved to their adjacent apices, respectively, to form a further 14 triangles each. The remaining five combinations are specified by the apices of the five regular triangles (with a side length of 5) that are formed by shifting apex (1) of the regular triangle (1, 6, 11) to 2, 3, 4 and 5 respectively. The sample combinations are obtained without duplication.

For the case where n' = 6k + 3 and k = 2, 3, 4, etc., combinations without duplication are represented by combining k differently shaped triangles chosen from the n' congruent acute angle triangles, the n' congruent obtuse angle triangles and using the n'/3 regular triangles.

Table B.1 — Examples of combinations without duplication

n'	N'	Possible combinations	Alternative combinations
n' = 7	7 × 1 = 7	[i, $f(i+1)$, $f(i+3)$] for i = 1 to 7	
n' = 13	13 × 2 = 26	[i, $f(i+2)$, $f(i+7)$] for i = 1 to 13	
		[i, $f(i+1)$, $f(i+4)$] for i = 1 to 13	
n' = 19	19 × 3 = 57	[i, $f(i+2)$, $f(i+10)$] for i = 1 to 19	[i, $f(i+3)$, $f(i+10)$] for i = 1 to 19
		[i, $f(i+3)$, $f(i+7)$] for i = 1 to 19	[i, $f(i+2)$, $f(i+8)$] for i = 1 to 19
		[i, $f(i+1)$, $f(i+6)$] for i = 1 to 19	[i, $f(i+1)$, $f(i+5)$] for i = 1 to 19
n' = 25	25 × 4 = 100	[i, $f(i+2)$, $f(i+12)$] for i = 1 to 25	[i, $f(i+2)$, $f(i+12)$] for i = 1 to 25
		[i, $f(i+3)$, $f(i+11)$] for i = 1 to 25	[i, $f(i+5)$, $f(i+11)$] for i = 1 to 25
		[i, $f(i+4)$, $f(i+9)$] for i = 1 to 25	[i, $f(i+1)$, $f(i+9)$] for i = 1 to 25
		[i, $f(i+1)$, $f(i+7)$] for i = 1 to 25	[i, $f(i+3)$, $f(i+7)$] for i = 1 to 25
n' = 9	9 + 3 = 12	[i, $f(i+1)$, $f(i+3)$] for i = 1, 4, 7	
		[i, $f(i+1)$, $f(i+3)$] for i = 2, 5, 8	
		[i, $f(i+2)$, $f(i+5)$] for i = 1, 4, 7	
		[i, $f(i+4)$, $f(i+8)$] for i = 1, 4, 7	
n' = 15	15 × 2 + 5 = 35	[i, $f(i+2)$, $f(i+8)$] for i = 1 to 15	
		[i, $f(i+1)$, $f(i+4)$] for i = 1 to 15	
		[i, $f(i+5)$, $f(i+10)$] for i = 1 to 5	
n' = 21	21 × 3 + 7 = 70	[i, $f(i+1)$, $f(i+10)$] for i = 1 to 21	[i, $f(i+2)$, $f(i+10)$] for i = 1 to 21
		[i, $f(i+3)$, $f(i+8)$] for i = 1 to 21	[i, $f(i+3)$, $f(i+9)$] for i = 1 to 21
		[i, $f(i+2)$, $f(i+6)$] for i = 1 to 21	[i, $f(i+1)$, $f(i+5)$] for i = 1 to 21
		[i, $f(i+7)$, $f(i+14)$] for i = 1 to 7	[i, $f(i+7)$, $f(i+14)$] for i = 1 to 7
n' = 27	27 × 4 + 9 = 117	[i, $f(i+1)$, $f(i+13)$] for i = 1 to 27	[i, $f(i+2)$, $f(i+13)$] for i = 1 to 27
		[i, $f(i+3)$, $f(i+11)$] for i = 1 to 27	[i, $f(i+4)$, $f(i+12)$] for i = 1 to 27
		[i, $f(i+4)$, $f(i+10)$] for i = 1 to 27	[i, $f(i+3)$, $f(i+10)$] for i = 1 to 27
		[i, $f(i+2)$, $f(i+7)$] for i = 1 to 27	[i, $f(i+1)$, $f(i+6)$] for i = 1 to 27
		[i, $f(i+9)$, $f(i+18)$] for i = 1 to 9	[i, $f(i+9)$, $f(i+18)$] for i = 1 to 9

Annex C
(informative)

Standard portrait images

C.1 Preparation of standard images

It is important to decide on a suitable set of standard images for use in visual assessment during the course of a psychophysical experiment. Image and subject composition, and colour casts are examples of factors that may significantly affect the results of visual assessment. In order to minimize unwanted bias in psychophysical experiments, factors that should be taken into account when establishing a set of standard images, include the following.

a) Choose a model whose skin tone is close to that of a typical Japanese person, described as Type A skin in C.2, based on its spectral reflectance.

b) Choose neutral grey for clothes and background in order to avoid unwanted casts on skin tone and to remove any unwanted biases which may be introduced into the visual assessment.

c) Shoot a head and shoulders portrait with the model facing you and compose the scene such that the face of the model is reproduced at an appropriate size for visual assessment.

d) Select camera equipment that is widely used in professional portrait studio photography.

e) Adjust the lighting conditions to those that are typically used in studio portrait photography. Relatively soft lighting with an illumination ratio of 1:2 can be used to avoid dark shadows.

f) Select professional use 4 in × 5 in photographic films, that are known to provide excellent image quality from the point of view of sharpness and graininess.

A set of pictures was taken under the conditions described above and 2L size photographic prints were made using a typical optical printer. Skilful operators optimized the density and colour balance during printing. The equipment and materials used during the experiment are listed in Table C.1.

Table C.1 — Equipment and photographic materials used for the preparation of standard portrait images

Equipment	Description of materials used
Camera and lens	Sinar[a] P 4×5, Fujinon[a] 250 mm F:6.3
Strobe	Photona[a] PH 2501× 3 with umbrellas
Shooting film	Fujicolor[a] NS 160 (a colour negative film)
Photographic colour paper	Fujicolor[a] paper FA-P

[a] These are examples of suitable products available commercially that have been used for this example. This information is given for the convenience of users of this part of ISO 20462 and does not constitute an endorsement by ISO of these products.

A standard digital file was required in order to enable high quality reproductions of the standard image to be made for use in future experiments. The requirement for preparing a standard digital file was met using the workflow shown in Figure C.1. In order to minimize image degradation the reflection print image was scanned using a drum scanner capable of scanning at a bit depth of 12 bits per colour. The integral spectral densities

for each pixel were calculated using a prepared conversion table. CIELAB values under D65 were derived from integral densities due to the fact that the standard illuminant for the colour space defined by ITU-R BT.709-3 is D65. Finally the CIELAB values were converted to sRGB and a digital file that conformed to the TIFF 6.0 format was created.

Figure C.1 — Workflow of preparing standard digital files

C.2 Data on CD-ROM

The image data for the 3 portrait images were encoded as 8-bit sRGB and were contained on one CD-ROM. The portrait images are identified as P1 to P3, respectively, and each image has been assigned a descriptive name that is associated with content of the picture. For this example, the descriptive names Type A, Type B and Type C skin were used. Figure C.2 shows a reduced size monochrome reproduction of the images. The portrait images have the following characteristics:

Picture size: 1400 × 1 900 pixels

NOTE The images (1 400 × 1 900 pixels) produce a physical image size of 116,67 mm by 158,33 mm when rendered at 12 pixels/mm.

Interleaving: Pixel interleaving

Colour sequence: R, G, B

Colour values: RGB data consists of three 8-bit values.

Image data orientation: Horizontal scanning starting from top left of the image and ending at bottom right.

a) Type A skin **b) Type B skin** **c) Type C skin**

Figure C.2 — Standard images

C.3 Check-sum data

The check-sums given in Table C.2 may be used to check the integrity of the data. These values were calculated by summing each image plane (R, G, B) with a one-byte accumulator and ignoring the overflow bit of the accumulator. The total accumulation, T, for all three planes is also shown. The data are shown in both hex and decimal notation. The check-sums apply only to the image data and exclude any headers.

Table C.2 — Check-sum

Image	Decimal				Hex			
	R	G	B	T	B	G	B	T
Type A skin	185	88	223	240	B9	58	DF	F0
Type B skin	182	62	46	34	B6	3E	2E	22
					B6			
Type C skin	90	255	217	50	5A	FF	D9	32

C.4 CD-ROM operating system compatibility

The format used for the format layer on the CD-ROM is as follows:

— Physical format layer — ISO/IEC 10149

— Volume and file formats layer — ISO 9660, interchange level 1 and implementation level 1

— Application format layer — TIFF, Revision 6.0 for RGB image data

Special TIFF structured file format based on TIFF 6.0

Figure C.3 in C.5 shows the TIFF 6.0 file headers of images: P1RGB.TIF

The RGB image files are compatible with TIFF Revision 6.0, Section 6 and Section 20.

C.5 Example of TIFF file header of the RGB image

Figure C.3 shows the TIFF file header for portrait images P1RGB, “TYPE A SKIN” of the image set recorded on the CD-ROM.

The TIFF file header encoding of the colour picture file named “P1RGB.TIF” is shown in Figure C.3. This encoding uses tags defined as TIFF 6.0.

The following fields are not included and take their default values.

NewSubfileType = 0

Orientation = 1 (load from top left, horizontally)

RowsPerStrip = $2^{32} - 1$ (only one strip)

PlanarConfiguration = 1 (pixel interleaving)

The symbol “n” represents a null byte, and “x” represents a “don't care” hexadecimal digit for padding data.

Offsets	Value					Description
	TIFF File Header					
00000000	4D4D					Byte order “MM”(big-endian)
00000002	002A					Version number: 42
00000004	00000008					Pointer to the 1st: the 1st IFD begins in 8th byte in a file
	the 1st IFD					
00000008	0011					Number of in this IFD: 17 entries in this IFD
	Tag#	Type	Count	Value-offset		
0000000A	0100	0003	00000001	076Cxxxx	256	ImageWidth: 1900 pixels/line
00000016	0101	0003	00000001	0578xxxx	257	ImageLength: 1400 lines/image
00000022	0102	0003	00000003	00000200	258	BitsPerSample: pointer to the area of 00000200h
0000002E	0103	0003	00000001	0001xxxx	259	Compression: 1(no compression)
0000003A	0106	0003	00000001	0002xxxx	262	PhotometricInterpretation: 2 (for RGB image)
00000046	010E	0002	00000009	00000206	270	ImageDescription: pointer to the area of 00000206h
00000052	010F	0002	0000000E	00000220	271	Make(Vendor name): pointer to the area of 00000220h
0000005E	0111	0004	00000001	00000400	273	StripOffsets: 00000400h (pointer to the image data)
0000006A	0115	0003	00000001	0003xxxx	277	SamplesPerPixel: 3
00000076	0117	0004	00000001	0079C3E0	279	StripByteCounts: 7,980,000 bytes in the strip
00000082	011A	0005	00000001	00000230	282	XResolution: pointer to the area of 00000230h
0000008E	011B	0005	00000001	00000238	283	YResolution: pointer to the area of 00000238h
0000009A	0128	0003	00000001	0003xxxx	296	ResolutionUnit: cm
000000A6	0132	0002	00000014	00000240	306	DateTime: pointer to the area of 00000240h
000000B2	013E	0005	00000002	00000258	318	WhitePoint: pointer to the area of 00000258h
000000BE	013F	0005	00000006	00000268	319	PrimaryChromaticities: pointer to the area of 00000268h
000000CA	8298	0002	00000029	00000298	33432	Copyright: pointer to the area of 00000298h
000000D6	00000000					Pointer to next IFD: None
	Value area					
00000200	0008	0008	0008			BitsPerSample: 8,8,8 (8-bits/sample for each separation)
00000206	41 53 49 41 4E 52 47 42 00 xx xx xx xx xx xx xx xx xx xx xx xx xx xx xx					ImageDescription: “TYPE A SKINRGBn”
00000220	49 53 4F 20 54 43 34 32 2F 57 47 31 38 00 xx xx					Make(Vendor name): “ISO TC42/WG18n”
00000230	00002EE0	00000064				XResolution: 12000/100 (120 pixels/cm)
00000238	00002EE0	00000064				YResolution: 12000/100 (120 pixels/cm)
00000240	32 30 30 32 3A 30 34 3A 30 31 20 31 30 3A 30 30 3A 30 30 00 xx xx xx xx					DateTime: “2002:04:01 10:00:00n” (April 1, 2002 at 10:00:00)
00000258	00000C37	00002710	00000CDA	00002710		D_{65} white point: x=3127/10000 y=3290/10000
00000268	00000280	000003E8	0000014A	000003E8		sRGB primary colours: R(640/1000,330/1000) G(300/1000,600/1000) B(150/1000,60/1000)
	0000012C	000003E8	00000258	000003E8		
	00000096	000003E8	0000003C	000003E8		
00000298	43 6F 70 79 72 69 67 68 74 20 32 30 30 30 20 49 53 4F 2C 20 41 6C 6C 20 72 69 67 68 74 73 20 72 65 73 65 72 76 65 64 2E 00 xx					Copyright: “Copyright 2000 ISO, All rights reserved.n”
000002C2 – 000003FF						not used
	Image data					
00000400 – 0079C7DF						Image data area is from 00000400h to 0079C7DFh

Figure C.3 — TIFF file header encoding of the colour picture file named "P1RGB.TIF for Type A skin

Annex D
(informative)

Performance of the triplet comparison method

D.1 General

In 4.1, it states that the proposed method comprises two steps as shown in Figure D.1. The first step is a “category step”, and the second step is a “triplet comparison step”. The reason for the first step is to reduce the number of the samples to the appropriate number determined by the purpose of each experiment.

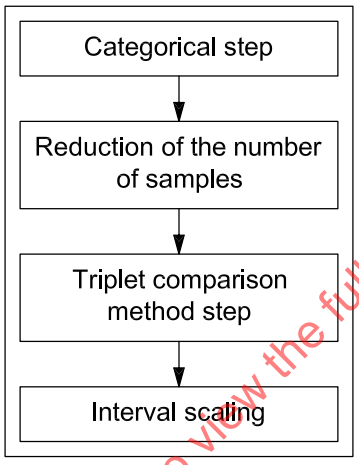


Figure D.1 — Flow of the proposed method

Category scaling using three categories, such as “favourable”, “acceptable” and “unacceptable” (or “acceptable”, “just acceptable”, “unacceptable”) is used for the first step, and samples are selected according to the number of samples required for the next step. If the number of test samples to be examined is relatively small, then this first step should be omitted and the psychophysical experiment should be started directly from the second step.

The second step is conducted in order to derive a precise scaling based on an interval scale. Three samples are compared at a time, achieving high assessment accuracy while keeping the experimental scale realistic.

D.2 Experimental

D.2.1 General

To examine the visual technique employed for psychophysical experiments in more detail, a case study was conducted in order to derive the preferred skin colour reproduced on photographic paper. The standard portrait image, Type A skin, was designed and details of procedure taken to prepare the image are described in Reference [6]. The reliability of the proposed method was investigated by conducting psychophysical experiments using both the “categorical step” and “triplet comparison step” processes respectively.

D.2.2 Step 1: Categorical step

A psychophysical experiment was conducted during which a total of 102 reflection print samples were prepared by changing the hue and chroma (17 combinations) as well as lightness (6 steps) of the facial area of the portrait image in CIELAB space. A total of 18 observers participated in the experiment. Each observer was asked to apply category scaling using three categories. For example, “favourable”, “acceptable” and “unacceptable”. Wherever possible, the viewing conditions applied were based on those specified in ISO 3664. However, fluorescent lamps for colour evaluation purposes were used. Illumination level was set to 1 000 lx. The rank order, with respect to skin colour preference, was obtained by assigning a score of +1, 0, and −1 to each of the categories.

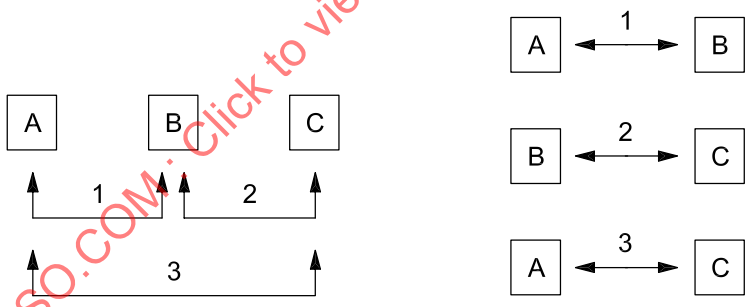
D.2.3 Step 2: Triplet comparison step

In order to improve the assessment accuracy and repeatability of the judging without imposing excessive stress on the observer during the visual assessment, a triplet comparison method[2], shown in Figure D.2, was developed. Psychophysical experiments conducted using the triplet comparison method can be designed using a higher number of samples than with the paired comparison method. This is due to the fact that triplet comparison invariably always reduces the number of comparisons relative to paired comparison. To determine the reliability and usefulness of the proposed triplet comparison method, psychophysical experiments were conducted and the results compared against those obtained by the paired comparison method. The following points were considered:

a) repeatability of the psychophysical scale;

b) similarity of the results between the methods;

c) observer stress (evaluated in terms of the validity of the rank for each sample, and the assessment time required).



a) Triplet comparison (1 set) **b) Paired comparison (3 sets)**

A, B and C are samples.

Figure D.2 — A new triplet comparison and conventional paired comparison

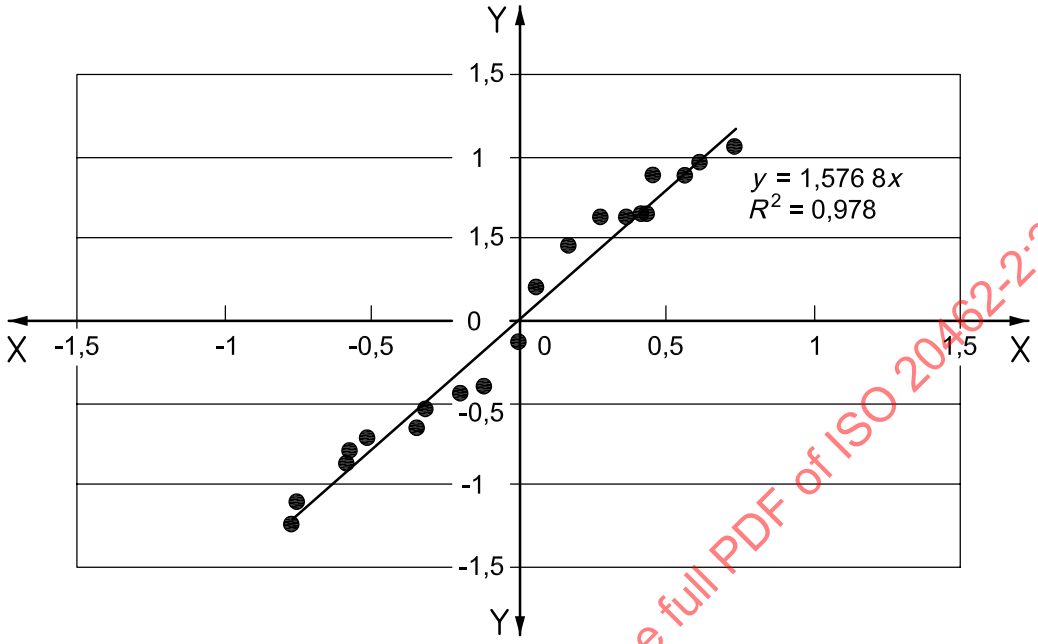
D.2.4 Procedure

The experiments were conducted as follows. The total number of samples, N, used in the triplet comparison experiment was selected such that the equations, $N = 6K + 1$ or $N = 6K + 3$, where K is a positive integer, were satisfied. This is recommended as it ensures that combinations of samples can be selected without unnecessary duplication of sample combinations. A total of 21 samples were selected from 102 print samples and 15 observers took part in the experiment.

All the observers were requested to perform both the paired comparison and the triplet comparison experiments. They were also encouraged to repeat the same experiment 5 times. The viewing conditions did not vary between experiments and were held constant throughout each experiment.

D.2.5 Results

Scheffe's method was used to derive an interval scale using statistical analysis. The results are shown in Figure D.3. The correlation between scale values derived by the paired comparison and the triplet comparison is examined and is shown in Figure D.4.

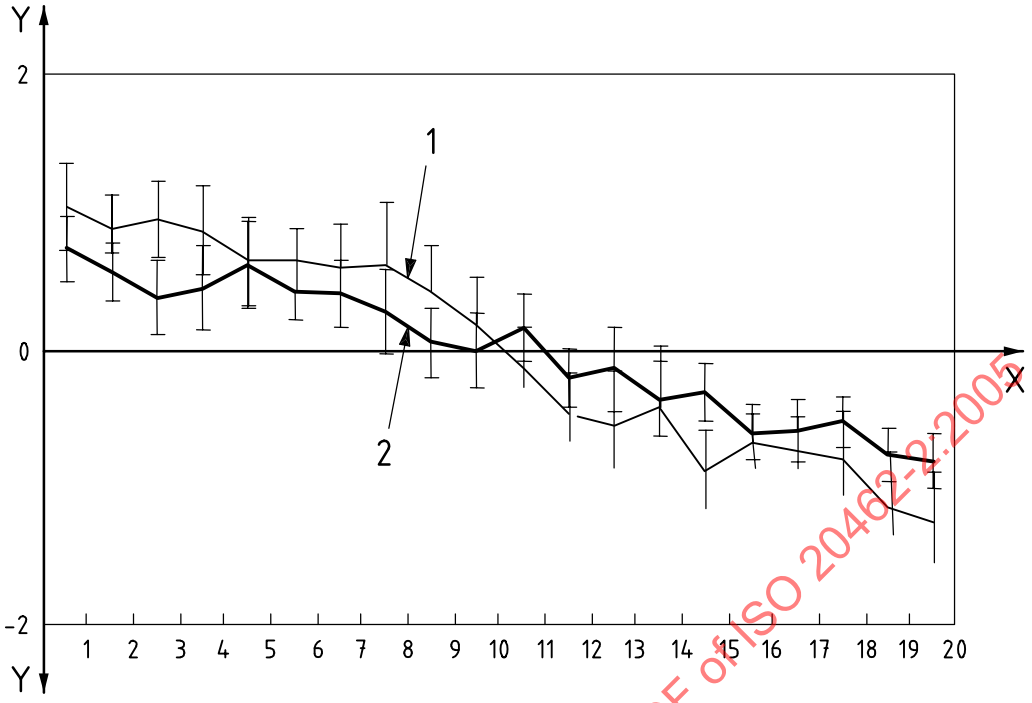


Key

X paired comparison

Y triple comparison

Figure D.3 — Comparison of the experimental results

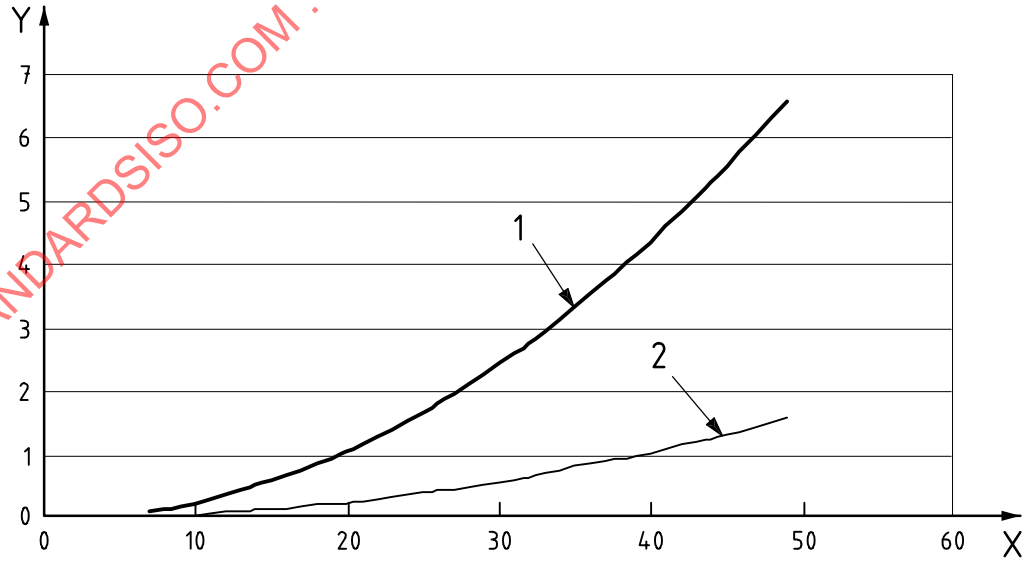


Key

X	sample number	1	triplet
Y	interval scale	2	paired

Figure D.4 — Correlations between two psychophysical methods

It was found that a good level of correlation exists between the two methods. It was also found that in the triplet comparison method, the scale values were distributed across a relatively wide range, suggesting that the observers were able to clearly distinguish the difference between samples. Figure D.5 compares the assessment time required for observations.



Key

X	number of samples	1	paired comparison
Y	assessment time, hours	2	triplet comparison

Figure D.5 — Estimated time required for visual assessment

D.2.6 Conclusions

A series of psychophysical experiments were performed to examine the reliability of the proposed method. The following conclusions were drawn.

— The scale values obtained by the triplet comparison method correlate highly with those obtained by the pair comparison method and suggest that the reliability of the triplet comparison method is sufficiently high.

— The repeatability of the results derived by the triplet comparison is the same as that derived by the paired comparison.

— The triplet comparison reduced the assessment time by almost 50 % compared to the pair comparison.

— The new psychophysical experimental method can, therefore, be considered as a useful technique for estimating image quality.

It can be concluded from a review of the subjective assessment results of the paired comparison and triplet comparison methods that triplet comparison has the following features.

a) Due to the reduction in the number of combinations required for assessment, the stress on the observer can be reduced.

By establishing a set of sample numbers from which combinations (of three samples) can be created that eliminate duplication, the total number of combinations in the experiment is reduced to about one-third of that for paired comparison. Correspondingly, the time required for the assessment is reduced by almost the same factor. All the observers stated that the reduction in total assessment time was the most effective contributor to stress reduction.

b) The degree of difficulty experienced by observers who assessed and judged samples varied among the individual observers.

The assessment time taken for each combination of samples was longer for the triplet comparison method than for paired comparison. However, all indications were that the assessment was generally performed with little difficulty. A slight concern was the level of difficulty in judging experienced by an observer varied depending on the individual observer. There was no clear correlation between the psychological difficulty and the instability in the assessment result, and even observers who exhibited stable assessments stated that paired comparison was superior with regards to the ease of judgment.

c) The stability in repeated assessments is almost equivalent for the two comparisons.

The standard deviation of repeated assessments was slightly larger for triplet comparisons than for paired comparisons. However, the number of samples for which the assessed order varied with repeated assessments was almost the same in both triplet and paired comparisons. For those observers for whom the use of the different comparison methods resulted in large differences in the assessed ordering of samples, the assessed ordering tended to be stabilized by arbitrarily defining one of the simultaneously presented samples in the triplet comparison to be a reference sample.

d) The distribution on the assessment scale tends to expand.

In the triplet comparison method, the assessment scaling was seen to occupy a range of values that included both ends of the assessment scale (+2, - - -, 0, - - -, –2). This indicated that the judging of sample quality was easier with triple comparison than with the paired comparison method. This can be explained by the fact that an observer taking part in a paired comparison experiment is more likely to assume that one of the samples in the sample pair is a reference. Secondly, the observer may be more reluctant to assign samples with values at the extreme end of the assessment scale.

Annex E
(informative)

Scheffe's method

E.1 The data processing for the paired comparison according to Scheffe's method is described below:

a) First the cumulative sum of the assessment value for each sample by each observer is calculated. The sum is then divided by the product of the sample number with the number of the observers to give an average assessment value.

b) Next, a dispersion analysis table is made by deriving the following values; the effect on the samples (this effect, which is represented by S_a, is obtained by dividing the sum of the square of the assessment values by the product of the sample number with the number of the observers), the effect on both of the samples and observers (this value is obtained by first dividing the sum of the square of the assessment values by each observer and then subtracting S_a from the resulting sum), the effect of combination (this value is obtained by first dividing the square sum of the assessment value by the number of the observers and then subtracting S_a from the resulting sum), and the overall square sum for the total effect. Then, the error is determined by subtracting the square sum for individual effects from the overall square sum.

c) From these results, the square sum for each effect is divided by its degree of freedom to give an unbiased estimate of variation.

d) The degree of freedom for each effect is determined as follows.

e) The degree of freedom for the samples is equal to (sample number − 1); for the samples and the observers it is equal to (sample number − 1) × (observer number − 1); and for the overall square sum it is equal to [observer number × sample number × (sample number − 1)]/2. Accordingly, for the case where the value obtained by dividing the unbiased estimate of variation for each effect by the error is larger than the F value of the F-distribution, a significant difference is assumed to exist.

f) From the above premises (under these conditions), it is necessary to establish a yardstick in order to investigate the differences between the samples. The yardstick can be calculated from the square root of the value that is derived from the range of the standardized observer (refer to the numerical tables defined by the degrees of freedom for the sample number, the number of observers, etc.) and the unbiased estimate of variation divided by the product of the sample number with the observer number. Then, the confidence interval between the samples is obtained.

For the case of triplet comparison, where a single comparison is equivalent to three paired comparisons, steps a) to d) described above are conducted.

E.2 The data processing for paired comparison according to Scheffe's method is described below.

a) Let the sample number be expressed by t, the number of panellists by N, the number of repetitions by R, and the value of the sensory test for any sample pair (i, j) by each panellist by X_{ijkr}.

Here, k, i, j and r, respectively, are positive integers in the range.

k = 1 to N

i = 1 to t

j = 1 to t

and

$r = 1$ *to* R.

Then, the values represented by $X_{i.k.}$, $(X_{i.k.})^2$, $X_{i...}$, $(X_{i...})^2$, $X_{ij..}$ and $(X_{ij..})^2$ are calculated as follows:

$$X_{i.k.} = \sum_{i=1,t} \sum_{r=1,R} X_{ijkr}$$

$$(X_{i.k.})^2 = X_{i.k.} \times X_{i.k.}$$

$$X_{i...} = \sum_{i=1,N} X_{i.k.}$$

$$(X_{i...})^2 = X_{i...} \times X_{i...}$$

$$X_{ij..} = \sum_{k=1,N} \sum_{r=1,R} X_{ijkr}$$

and

$$(X_{ij..})^2 = X_{ij..} \times X_{ij..}$$

The average value of the sensory test a_i for Sample i, (i.e. cumulatively summing up the sensory test value for each sample by each panellist and dividing the cumulative sum by the product of the sample number, the number of panellists and the number of repetitions), is obtained from the following calculation;

$$a_i = X_{i...}/(t \times N \times R)$$

NOTE In the present analysis, the average of all the a_i's is set to zero. Since samples are presented simultaneously in the present subjective image quality evaluation, X_{ijkr} is always equal to X_{jikr}.

b) Next, calculations to make a variance analysis table are carried out.

The effect between samples, S_a, (the value obtained by dividing the sum of the squares of the sensory test values by the product of the sample number, the number of panellists and the number of repetitions) is calculated as follows:

$$S_a = \sum_{i=1,N} (X_{i...})^2/(t \times N \times R)$$

The interaction between samples and panellists $S_{a(B)}$ is calculated by first dividing the sum of the squares of the sensory test values for each sample by the product of the number of panellists with the number of sample number repetitions and then subtracting S_a from the resulting value.

$$S_{a(B)} = \left(\sum_{i=1,t} \sum_{k=1,N} (X_{i.k.})^2 / t \times R \right) - S_a$$

The combination effect S_c is calculated by first dividing the sum of the squares of the sensory test values for each sample by the number of panellists and then subtracting S_a from the resulting value.

$$S_c = \left(\sum_{i=1,t} \sum_{i=1,t} (X_{ij..})^2 /(N \times R) \right) - S_a$$

where $j > i$.

The overall sum of squares S_T for the total effect is calculated by

$$S_T = \sum_{i=1,t} \sum_{j=1,t} \sum_{k=1,N} \sum_{r=1,R} X_{ijkr} \times X_{ijkr}$$

where $j > i$.

The error S_e is obtained by subtracting the sum of the squares for the individual effects from the overall sum of the squares.

$$S_e = S_T - S_a - S_{a(B)} - S_c$$

c) The variance analysis table is formulated (arranged) as shown in Table E.1.

Table E.1 — Variance analysis

Effect	Sum of squares	Degrees of freedom	Unbiased variance	F_0 value
Main effect	S_a	$\phi_a=(t-1)$	$V_a=S_a/\phi_a$	V_a/V_e
Interaction between samples and panellists	$S_{a(B)}$	$\phi_{a(B)}=(t-1)(N-1)$	$V_{a(B)}=S_{a(B)}/\phi_{a(B)}$	
Combination effect	S_c	$\phi_c=(t-1)(t-2)/2$	$V_c = S_c/\phi_c$	
Error	S_e	$\phi_e=\phi_T-\phi_a-\phi_{a(B)}-\phi_c$	$V_e = S_e/\phi_e$	
Overall sum of squares	S_T	$\phi_T = NR(t-1)/2$		

In cases where the value obtained by dividing the unbiased variance for each effect by the unbiased variance for error is larger than the F value for the F-distribution (the degree of freedom for each effect, and the degree of freedom for error: significance probability), a significant difference is assumed to exist.

The variance analysis table for a subjective image quality skin colour assessment experiment that is similar to the one described in Annex D is shown below. In this experiment, the sample number t is 21, the panellist number N is 9 and the repetition number R is 3.

Table E.2 — Variance analysis table for a subjective image quality skin colour assessment

Effect	Sum of squares	Degrees of freedom	Unbiased variance	F_0 value
Main effect	5 840,38	20	292,02	231,30[a]
Interaction between samples and panellists	6 143,27	160	38,40	
Combination effect	615,03	190	3,24	
Error	6 691,32	5 300	1,26	
Overall sum of squares	19 290	5 670		

[a] Denotes that the F_0 value is within a 1 % significance probability.

The F value is confirmed to have the following numerical values:

F (20, 5 300 : 0,01) = 1,882

F (20, 5 300 : 0,05) = 1,573

The main effect is significant with a 1 % significance probability.