INTERNATIONAL STANDARD

ISO 20462-1

First edition
2005-11-01

Photography — Psychophysical experimental methods for estimating image quality —

Part 1: Overview of psychophysical elements

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

Partie 1: Aperçu général des éléments psychophysiques

Reference number
ISO 20462-1:2005(E)

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

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 20462-1 was prepared by Technical Committee ISO/TC 42, *Photography*.

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

— *Part 1: Overview of psychophysical elements*

— *Part 2: Triplet comparison method*

— *Part 3: Quality ruler method*

Introduction

There are many circumstances under which it is desirable to quantify image quality in a standardized fashion that facilitates interpretation of results within a given experiment and/or comparison of results between different experiments. Such information can be of value in assessing the performance of different capture or display devices, image processing algorithms, etc. under various conditions. There are a number of psychometric methods described in the literature, such as paired comparison, rank ordering, categorical sort, and magnitude estimation, which might be considered as candidates for experimentally measuring image quality. Several textbooks[1] [3] [4] [5] [9] [12] have reviewed these and other methods and have discussed associated data reduction techniques, which usually are based upon the approach of Thurstone[11] or analogous reasoning. However, the choice of the best method for a particular application may be difficult to make, and interpretation of the rating scales produced by the numerical analyses is frequently ambiguous. Furthermore, none of the commonly used techniques provides an efficient mechanism for calibration of the results against a standardised numerical scale or associated physical references, which is desirable when results of different experiments are to be compared or integrated. The value of new calibrated psychometric methods in developing comprehensive models of imaging system quality has been demonstrated in a recent work[6] that contains more detailed discussions of many of the informative topics superficially considered herein.

The three parts of ISO 20462 address the need for documented means of determining image quality in a calibrated fashion. Part 1 provides an overview of practical psychophysics; specific experimental methods and associated data reduction techniques are described in Part 2 (triplet comparison method[8] [10]) and Part 3 (quality ruler method[6]). Informative Annex A aids in identifying the better choice between the two alternative methods of Parts 2 to 3, which are complementary and together are sufficient to span a wide range of applications. It is the intent of these methods to produce results that are not merely directional in nature, but are expressed in terms of relative or fixed scales that are calibrated in just noticeable differences (JNDs), so that the significance of experimentally measured stimulus differences is readily ascertained.

Photography — Psychophysical experimental methods for estimating image quality —

Part 1: Overview of psychophysical elements

1 Scope

This part of ISO 20462 is part of a multiple-part standard pertaining to the subjective evaluation of pictorial still image quality. This part of ISO 20462

a) defines the units by which image quality is quantified (just noticeable differences, or JNDs);

b) describes the influence of stimulus properties, observer characteristics, and task instructions on results obtained from rating experiments;

c) provides a flow chart for choosing the preferred psychophysical method for determining image quality from among those defined in subsequent parts of ISO 20462.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 3664, *Viewing conditions — Graphic technology and photography*

3 Terms and definitions

For the purpose of this document, the following terms and definitions apply:

3.1
artefactual attribute
attribute of image quality that, when evident in an image, nearly always leads to a loss of overall image quality

EXAMPLE Examples of artefactual attributes include noise and aliasing.

NOTE The commonly used terms *defect* and *impairment* are similar in meaning.

3.2
attribute
aspect, dimension, or component of overall image quality

cf. **artefactual attribute** (3.1) and **preferential attribute** (3.12)

EXAMPLE Examples of image quality attributes include image structure properties such as sharpness and noise; colour and tone reproduction properties such as contrast, colour balance, and relative colourfulness; and digital artefacts such as aliasing, contouring, and compression defects.

3.3
attribute just noticeable difference
attribute JND
measure of the detectability of appearance variations, corresponding to a stimulus difference that leads to a 75:25 proportion of responses in a paired comparison task in which univariate stimuli pairs are assessed in terms of a single attribute identified in the instructions

cf. **quality JND** (3.14)

NOTE 1 As an example, a paired comparison identifying the sharper of two stimuli that differ only in their generating system modulation transfer function (MTF), would yield results in terms of sharpness attribute JNDs. If the MTF curves differed monotonically and did not cross, the outcome of the paired comparison would depend primarily upon the observers' ability to detect changes in the appearance of the stimuli as a function of MTF variations, with little or no value judgement required of the observers. The relationship between paired comparison proportions and stimulus differences is discussed in greater detail in Annex B.

NOTE 2 If observers are instead asked to choose which of a pair of stimuli is higher in overall image quality, and if the stimuli in aggregate are multivariate, such that the observer should make value judgements of the importance of a number of attributes, rather than focussing on one aspect of image appearance, it is observed experimentally that larger objective stimulus differences (for example, MTF changes) are required to obtain a 75:25 proportion of responses, which in this case corresponds to a quality JND.

NOTE 3 A JND is a statistical quantity, derived from a number of observations. An observer assessing a single pair of images differing by one attribute JND is unlikely to be confident that he or she has detected the sample difference. A stimulus difference of approximately three JNDs is usually needed for an observer of average sensitivity to feel reasonably certain of his or her assessment.

3.4
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.

3.5
image quality
impression of the overall merit or excellence of an image, as perceived by an observer neither associated with the act of photography, nor closely involved with the subject matter depicted

NOTE The purpose of defining image quality in terms of third-party (uninvolved) observers is to eliminate sources of variability that arise from more idiosyncratic aspects of image perception and pertain to attributes outside the control of imaging system designers.

3.6
instructions
set of directions given to the observer for performing the psychophysical evaluation task

3.7
just noticeable difference
JND
stimulus difference that leads to a 75:25 proportion of responses in a paired comparison task

cf. **attribute JND** (3.3) and **quality JND** (3.14)

3.8
magnitude estimation method
psychophysical method involving the assignment of a numerical value to each test stimulus that is proportional to image quality; typically, a reference stimulus with an assigned numerical value is present to anchor the rating scale

NOTE The numerical scale resulting from a magnitude estimation experiment is usually assumed to constitute a ratio scale, which, ideally, is a scale in which a constant percentage change in value corresponds with one JND. In practice, modest deviations from this behaviour occur, complicating the transformation of the rating scale into units of JNDs without inclusion of unidentified reference stimuli (having known quality) among the test stimuli.

3.9
multivariate
describing a series of test or reference stimuli that vary in multiple attributes of image quality

3.10
observer
individual performing the subjective evaluation task in a psychophysical method

3.11
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 saturated responses are included in the analysis.

3.12
preferential attribute
attribute of image quality that is invariably evident in an image, and for which the preferred degree is a matter of opinion, depending upon both the observer and the image content

EXAMPLE Examples of preferential image quality attributes include colour and tone reproduction properties such as contrast and relative colourfulness. Because the perceived quality associated with a preferential attribute is dependent upon both the observer and image content, in studies involving variations of preferential attributes, particular care is needed in the selection of representative sets of stimuli and groups of observers.

NOTE The term *noticeable* in just noticeable difference is not linguistically strictly correct when applied to a preferential attribute, but is nonetheless retained in this part of ISO 20462 for convenience. For example, the higher contrast stimulus of a pair differing only in contrast might be readily identified by all observers, whereas there might be a lack of consensus regarding which of the two images was higher in overall image quality. Nonetheless, if the responses from the paired comparison for quality were in the proportion of 75:25, the image chosen more frequently would be said to be one JND higher in quality. The JND is best regarded as a measurement unit tied to the predicted or measured outcome of a paired comparison.

3.13
psychophysical 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** (3.4), **magnitude estimation** (3.8), **paired comparison** (3.11), **quality ruler** (3.15), **rank ordering** (3.16) and **triplet comparison methods** (3.24)

3.14
quality just noticeable difference
quality JND
measure of the significance or importance of quality variations, corresponding to a stimulus difference that leads to a 75:25 proportion of responses in a paired comparison task in which multivariate stimuli pairs are assessed in terms of overall image quality

NOTE 1 See Notes for **attribute JND** (3.3).

NOTE 2 The attribute JND is a measure of detectability of appearance changes, whereas the quality JND is a measure of significance or importance of stimulus differences in terms of their impact on quality. An attribute JND is a useful unit for predicting how observers would react to an advertisement showing images carefully matched in all respects but one, and drawing the attention of the observer to the attribute varying. In contrast, a quality JND is useful for predicting how observers would perceive overall quality as a function of one or more stimulus variations, and so is a more useful quantity in optimizing imaging system design, where different attributes should be balanced against one another. The overall quality of an image may be predicted from a knowledge of the impact of each attribute in isolation, expressed in terms of quality JNDs, whereas the same is not true of attribute JNDs. Therefore, it is often highly desirable to obtain results expressed in quality JNDs, even if the stimuli being assessed are univariate in nature. This can be accomplished if test stimuli are rated against a series of appropriately calibrated reference stimuli, as in the quality ruler method.

3.15
quality ruler method
psychophysical method that involves quality or attribute assessment of a test stimulus against a series of ordered, univariate reference stimuli that differ by known numbers of JNDs

NOTE The quality ruler method is described in more detail in ISO 20462-3.

3.16
rank ordering method
psychophysical method involving the arrangement by an observer of a series of stimuli in order of increasing or decreasing image quality or an attribute thereof, in accordance with the set of instructions provided

3.17
reference stimulus
image provided to the observer for the purpose of anchoring or calibrating the perceptual assessments of test stimuli in such a manner that the given ratings may be converted to JND units

NOTE The plural is reference stimuli.

3.18
scene
content or subject matter of an image, or a starting image from which multiple stimuli may be produced through different experimental treatments

NOTE Typically, stimuli depicting the same scene are compared in a psychophysical experiment, because it is the effect of the treatment that is of interest, and differences in image content could cause spurious effects. In cases where scene content is not matched, a number of scenes should be used so that scene effects may be expected to average out.

3.19
standard quality scale
SQS
fixed numerical scale of quality having the following properties:

a) the numerical scale is anchored against physical standards;

b) a one unit increase in scale value corresponds to an improvement of one JND of quality; and

c) a value of zero corresponds to an image having so little information content that the nature of the subject of the image is difficult to identify

NOTE The standard quality scale is described in more detail in ISO 20462-3.

3.20
stimulus
image presented or provided to the observer either for the purpose of anchoring a perceptual assessment (a reference stimulus) or for the purpose of subjective evaluation (a test stimulus)

NOTE The plural is stimuli.

3.21
suppression
perceptual effect in which one attribute is present in a degree that seriously degrades image quality and thereby reduces the impact that other attributes have on overall quality, compared to the impact they would have had in the absence of the dominant attribute

NOTE To generate reference stimuli that are separated by a specified number of JNDs based on variations in one attribute, it will be necessary to ensure that other attributes do not significantly suppress the impact of the attribute varied.

3.22
test stimulus
image presented to the observer for subjective evaluation

NOTE The plural is test stimuli.

3.23
treatment
controlled or characterized source of the variations between test stimuli (excluding scene content) that are to be investigated in a psychophysical experiment

EXAMPLE Examples of treatments include different image processing algorithms, variations in capture or display device properties, changes in image capture conditions (e.g. camera exposure), etc.

NOTE Different treatments may be achieved through hardware or software changes, or may be numerical simulations of such effects. Typically, a series of treatments is applied to multiple scenes, each generating a series of test stimuli. The effect of the treatment may then be determined by averaging the results over scene and observer to improve signal to noise and reduce the likelihood of systematic bias.

3.24
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

NOTE The triplet comparison method is described in more detail in ISO 20462-2.

3.25
univariate
describing a series of test or reference stimuli that vary only in a single attribute of image quality

4 Specification of the experimental conditions and results

4.1 Observer characteristics

Observers shall be free of any personal involvement with the design of the psychophysical experiment or the generation of, or subject matter depicted by, the test stimuli.

Observers shall be checked for normal vision characteristics insofar as they affect their ability to carry out the assessment task. In most cases, observers should be confirmed to have normal colour vision and should be tested for visual acuity at approximately the viewing distance employed in the psychophysical experiment.

The number of observers participating in an experiment shall be reported. If the data of any observers are omitted from the analysis because of indications of difficulties with the task, the number omitted, and the criteria upon which exclusion was based, shall be reported. The percentage of observers excluded should not exceed 15 %. At least 10 observers shall (and preferably 20 should) contribute data to the analysis. Criteria for selection of observers, and notable characteristics of the observer group as a whole, should also be reported.

NOTE Examples of information regarding the observer group that might be reported include demographic data, level of experience in image evaluation, technical training if pertinent to a particular imaging application, etc.

4.2 Stimulus properties

The number of distinct scenes represented in the test stimuli shall be reported and shall equal or exceed 3 scenes (and preferably should equal or exceed 6 scenes). If fewer than 6 scenes are used, each shall be preferably depicted or alternatively briefly described, particularly with regard to properties that might influence the importance or obviousness of the stimulus differences.

The nature of the variation (other than scene content) among the test stimuli shall be described in both subjective terms (image quality attributes) and objective terms (stimulus treatment or generation). Other properties of the stimuli or their generation that might be expected to influence the results obtained even if present at a constant level in the test stimuli should also be reported. If reference stimuli are provided to the observer, as in a quality ruler experiment, their pedigree shall be specified in accordance with ISO 20462-3.

NOTE Examples of stimulus properties or aspects of their generation that might affect the outcome of an experiment even if they were invariant include presence of serious artefacts that might cause suppression, and application of image processing steps that could amplify certain types of signal or noise.

4.3 Instructions to the observer

The instructions shall state what is to be evaluated by the observer and shall describe the mechanics of the experimental procedure. If the test stimuli vary only in the degree of a single artefactual attribute, and there are no calibrated reference stimuli presented to the observer, then the instructions shall direct the observer to evaluate the attribute varied, rather than to evaluate overall quality. A small set of preview images showing the range of stimulus variations should be shown to observers before they begin their evaluations, and the differences between the preview images should be explained.

The task assigned to the observer shall be reported, making clear whether evaluation of overall quality or an attribute thereof was requested, and specifying which psychophysical method was used. The extent to which the nature of the variation of the stimuli was demonstrated and explained to the observers shall also be reported.

NOTE 1 There are various viewpoints regarding the extent to which the instructions should identify the variations in stimuli to be presented to the observer. One danger in not identifying the attributes being varied is that an observer may fail to recognize the nature of the stimulus differences until a particularly obvious example of an attribute is encountered, causing a transition from a state of insensitivity to one of sensitivity. Because the goal of most investigations is to determine responses at a steady state or equilibrium condition, rather than to characterize transient behaviour, dramatic changes in observer perception in the middle of an experiment are normally undesirable.

NOTE 2 If stimuli are univariate or vary only in a small number of attributes, merely asking the observer to assess the overall quality of the stimuli does not guarantee that the observer will make the desired value judgement rather than evaluating the attribute appearances in an artificial and analytical manner, potentially leading to results in units that are intermediate between attribute and quality JNDs. A helpful tactic in such cases is to ask the observer to imagine that the image represents a personally treasured moment, and to compare images (whether test and reference stimuli or multiple test stimuli) based on which they would prefer to own, if they could have only one. This approach may help to place the observer in the proper frame of mind to assess overall quality.

4.4 Viewing conditions

Viewing conditions shall be consistent with ISO 3664 except for the following relaxed criteria.

a) For print viewing, the illuminance level shall be between 375 lx and 2 500 lx.

NOTE 1 ISO 3664 permits ranges of 375 lx to 625 lx for practical (P2) evaluation and 1 500 lx to 2 500 lx for critical (P1) evaluation. This part of ISO 20462 allows intermediate values as well.

b) For print viewing, the surround surfaces shall exhibit no specular reflections from the observer's position.

NOTE 2 ISO 3664 requires that the surround surfaces be matte, and therefore exhibit no specular reflections at any angle.

c) For print and transparency viewing, the special colour rendering index (CRI) should equal or exceed 80; and the metamerism index should be C or better (visual) and less than 4 (UV). If these criteria are not met, the type of illumination employed shall be specified.

NOTE 3 In this part of ISO 20462 , the unchanged ISO 3664 requirement that the routinely tabulated general CRI value shall equal or exceed 90 is considered sufficient for pictorial image quality evaluation. ISO 3664 further requires that the special CRI and metamerism criteria above be met. These more stringent criteria are appropriate for critical assessment of graphic arts colour appearance matching, but are more restrictive than is necessary for overall image quality evaluation.

d) For monitor viewing, if the white point u′,v′ chromaticities are closer to D50 than D65, the white point luminance shall exceed 60 cd/m^2; otherwise, it shall exceed 75 cd/m^2.

NOTE 4 ISO 3664 requires that the white point luminance exceed 75 cd/m^2 regardless of white point balance, but notes the potential difficulty of meeting this criterion at lower correlated colour temperatures.

The illuminance level (reflection prints) or white point luminance (transparencies, softcopy viewing), and viewing distance (measured from the eye of the observer to the stimulus) shall be reported. Aspects of the viewing conditions not specified in ISO 3664 but thought to influence the perception of the stimulus variations should also be reported.

The viewing conditions at the physical locations assumed by multiple stimuli that are compared simultaneously shall be matched to a degree such that critical observers see no consistent differences in quality between identical stimuli presented simultaneously at each of the physical locations. The observer should be able to view each stimulus merely by changing his or her glance, without having to move their head.

NOTE 5 For example, during a triplet comparison of three test stimuli, or a quality ruler comparison of a test stimulus with two reference stimuli between which the test stimulus falls, three stimuli in three nearby physical locations are considered by the observer at the same time. To avoid bias, at these three physical locations the illumination tristimulus values, viewing distance, viewing angle, flare, etc. should be matched to sufficient tolerances that identical images displayed at each of the three locations would exhibit no consistent differences in attributes of image quality. Trivial differences in perspective, such as whether the left or right side of an image appears minutely foreshortened, are not considered to be of significance in determination of quality.

4.5 Experimental duration

To avoid fatigue, the median duration (over observer) of an experimental session, including review of the instructions, should not exceed 45 min, and shall not exceed 60 min. Individual observers who require more than 60 min to complete the assigned task shall be allowed the option of returning at another time to finish the test.

4.6 Results

Stimulus differences or averages of stimulus differences (over scene) for various treatments shall be reported in JND units, to the nearest 0,1 JND. A value shall not be reportable unless it is computed from at least 30 individual determinations (scenes multiplied by observers and by repetitions). If the instructions direct the observer to assess an attribute rather than overall quality, the JNDs shall be identified as *attribute JNDs*; otherwise, they should be stated to be *JNDs of quality*. If the presence of calibrated reference stimuli permits the determination of SQS values (ISO 20462-3), they shall also be reported to 0,1 units and shall be subject to the same minimum assessment number constraint.

NOTE 1 If 10 observers assess some number of treatments in 3 scenes each, then pooling the data for each treatment over scene and observer will allow the 30-determination requirement to be met exactly. If 20 observers assess treatments in 6 scenes each, separate results could be quoted for subsets of observers and scenes containing at least 25 % of the data.

NOTE 2 The above requirement is intended to adequately reduce potential errors from experimental uncertainty through sufficient averaging of multiple determinations. However, it is the responsibility of the experimental designer to select observer and stimulus groups that are representative, so that biases are not introduced into the analysis. For example, averaging over a number of observers might control the effect of intra-observer variability satisfactorily, so that the mean tendency of the group of observers chosen were accurately quantified, but no amount of averaging could correct for the bias resulting if the range of inter-observer variation of the selected group were unrepresentative of the population as a whole.

4.7 Summary of reported quantities

A list of the information to be reported from a psychophysical experiment is provided in the table below. The order of the tabulation follows that of the discussion in Clause 4 above. An example of a report is provided in Annex C.

Table 1 — Summary of reported quantities

Number of observers participating
Number of excluded observers and reasons for exclusion
Criteria for selection of observers
Vision tests administered to observers
Pertinent characteristics of observer group
Number of scenes
Depiction (preferably) or description of scenes if fewer than six in number
Subjective and objective nature of variation among test stimuli
Other properties of test stimuli that might affect outcome of experiment
Pedigree of reference stimuli if present
Stimulus property observer was instructed to evaluate
Psychophysical method employed
Extent of explanation of the stimuli differences to the observer
Illuminance (prints) or white point luminance (transparencies, monitors)
Stimulus size
Stimulus type (reflection print, transparency, monitor image)
Viewing distance (from the observer's eye to the stimulus)
Unspecified viewing conditions affecting perception of stimulus variations
Treatment differences in attribute JNDs, JNDs of quality, or SQS values

Annex A
(informative)

Selection of an appropriate psychophysical method

There are a number of commonly used psychometric methods such as paired comparison, rank ordering, categorical sort, and magnitude estimation. Each of these methods, as well as those described in subsequent parts of this International Standard, has strengths and weaknesses, which are discussed in Annex D. The triplet comparison method of ISO 20462-2 and the quality ruler method of ISO 20462-3 are complementary in their properties, and one or both should provide a satisfactory approach in most applications. The purpose of this annex is to aid in the choice of the better method, in cases where one or the other is distinctly preferred. This is accomplished by first very briefly describing each method, and then contrasting the methods and providing a simple decision flow chart.

The triplet comparison method, described in ISO 20462-2, involves scaling each of a minimum number of sets of three stimuli each, which encompass all possible pairs of treatments of a given scene. The computation of JNDs after data collection is complete may be compromised if gaps between adjacent stimuli exceed $\approx$1,5 JNDs, or if too many pairs of stimuli differ by $>$ 1,5 JNDs, causing saturated responses. The method is most efficient for numbers of stimuli of the form $N = 6K + 1$ or $N = 6K - 3$ (where K is a positive integer), because there is no duplication of pairs within the triplets. This method is particularly suitable for precisely determining small quality or attribute differences.

In the quality ruler method, described in ISO 20462-3, test stimuli are rated against a series of reference stimuli having known separations in JNDs, so the observers' responses can be converted to JNDs in real time. The quality ruler method entails an initial effort to generate or to obtain reference stimuli and to set up the supporting hardware or software; however, once an experimental set-up is in place, it may be reused for subsequent studies. This method is particularly suitable for characterizing stimuli that vary significantly in quality and/or for producing results calibrated to the standard quality scale (SQS), facilitating comparison of results between different experiments.

In many applications, either the triplet comparison or the quality ruler approaches may be satisfactory. The triplet comparison method is better for precisely measuring smaller quality differences (on the order of one JND), whereas the quality ruler method is preferred when larger stimulus differences (multiple JNDs) are to be quantified. The triplet comparison method is somewhat simpler, although most of the effort in setting up a first quality ruler experiment need not be repeated in subsequent studies. The quality ruler method is sometimes advantageous in allowing observer responses to be converted to JNDs instantly with a lookup table, rather than waiting until all data are collected and can be analysed as a set. With suitable reference stimuli, the quality ruler method permits the results to be reported using the SQS. With effort, SQS can also be determined using the triplet comparison method, by including reference stimuli of known SQS that have quality nearly identical to that of the test stimuli, and constraining the viewing distance and other parameters as required in the quality ruler method.

The flow chart given in Figure A.1 provides some guidance regarding which method is more suitable for a given application; however, the use of either method is permitted as long as all other requirements are met.

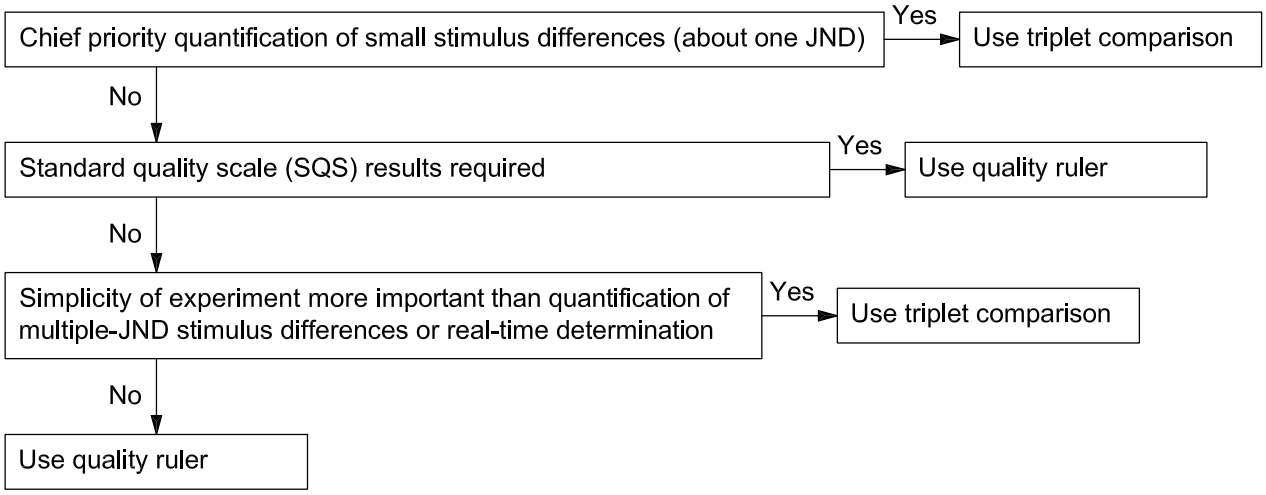

Figure A.1 — Selection of an appropriate psychophysical method

Annex B
(informative)

Stimulus differences, paired comparison proportions, and JNDs

Following Thurstone[11], perception is generally modelled as a variable or noisy process, rather than a deterministic phenomenon. Image quality or an attribute thereof may be regarded as a dimension of stimulus variation that modulates a corresponding perceptual continuum. If the perception from an individual assessment is quantified by the stimulus value that on average would have produced that same sensation, then the outcomes of a number of assessments of a single stimulus can be represented as a distribution of stimulus values. This perceptual distribution will be peaked near the actual stimulus value and will have some breadth reflecting the uncertainty in the perceptual process. A paired comparison is modelled as a perception of one stimulus value, then a perception of the other stimulus value, followed by a comparison of the resulting two values, and a decision based upon the sign of their difference.

For example, consider two stimuli differing only in the degree to which a common initial image has been blurred. A numerical scale of blur might be defined so that increasing blur (and therefore decreasing sharpness) corresponds to increasing value. Suppose Stimulus No. 1 has a blur value of 7,0 and Stimulus No. 2 has a blur value of 9,0. If perception were a deterministic phenomenon, then Stimulus No. 1 would always be identified as being the sharper stimulus in a paired comparison of the two stimuli. In the probabilistic model, however, the perceived degree of blur of each stimulus is represented by a peaked distribution centred upon their actual blur values. If the perceptual distributions both were normal (Gaussian) distributions with standard deviations of 2,0 blur value units, the two distributions would overlap considerably, and in some pairs of samples drawn from the distributions, the perceived blur value of Stimulus No. 1 would exceed that of Stimulus No. 2, causing the observer to identify the objectively more blurred Stimulus No. 2 as being sharper, an error.

The distribution of differences between samples drawn from two distributions is given by the convolution of the two distributions. The convolution of two normal distributions is again a normal distribution having a mean equal to the difference of the starting distribution means and a variance (square of the standard deviation) equal to the sum of the variances of the two starting distributions. In our present example, the distribution of differences of blur values between those drawn from the Stimulus No. 1 and the Stimulus No. 2 perceptual distributions would have a mean value of $7 - 9 = -2$ and a dispersion of $(2^2 + 2^2)^{1/2} \approx 2{,}8$ blur value units (making the stimulus difference $2{,}0/2{,}8 \approx 0{,}70$ standard deviations). As long as the difference in perceived blur values has a negative value, the observer would correctly identify Stimulus No. 1 as being sharper. The probability of a correct answer is equal to the fractional area of the difference distribution lying to the left of the origin ($x = 0$) and hence having a negative sign.

To generalize this result, it is convenient to refer to a normal distribution having zero mean and unit standard deviation. In terms of this distribution, the probability of a correct response is the fractional area of the distribution to the left of the absolute stimulus difference, expressed in multiples of the standard deviation of the difference distribution, which are usually called z-values. The fractional area under a distribution that lies to the left of a particular value is given by a cumulative distribution function (CDF). The CDF of a normal distribution does not have an analytical expression but is tabulated in statistical and other works, and is related to what is called the error function. In the present example, the normal CDF at $z = 0{,}70$ is 0,76, so the stimulus difference just slightly exceeds one JND, which would yield a 75:25 proportion.

In practice, the outcomes of paired comparison experiments are not always accurately described by assuming a normal difference distribution. Systematic discrepancies are observed at larger stimulus differences, which imply that the tails of the difference distribution are relatively more extensive than those of a normal distribution, so that larger stimulus differences are needed to drive paired comparison proportions towards unanimity than would be anticipated. In addition to this inaccuracy, there is also imprecision in the determination of values in the tails of a normal distribution. In a typical application, a paired comparison proportion would be used to determine the unknown stimulus difference (e.g. the quality difference). Because the slope is low in the tails of the distribution, a small change in the observed proportion can cause large changes in the deduced stimulus difference. For example, if 39 of 40 observers agree which stimulus is higher

in quality, the deduced stimulus difference would be about two z-units or about three JNDs. However, if the lone observer changed his or her single assessment so that the result was unanimous, the deduced stimulus difference would become infinite. The combination of inaccuracy and imprecision in the tails of a difference distribution begins to compromise the reliability of deduced stimulus differences from direct paired comparison when the two stimuli differ by more than approximately 1,5 JNDs.

Given the inaccuracy and imprecision in the tails of difference distributions, and the lack of an analytical expression for the CDF of a normal distribution, it is convenient to adopt a different function to model the difference distribution. The angular distribution closely resembles the normal distribution at small stimulus differences, but has truncated tails at larger stimulus differences, which bound deduced stimulus differences and their uncertainties, allowing regression analyses of the data to be performed rigorously without requiring signal-dependent weighting[2]. The relationship between the angular deviate z_{a} and the probability of a given stimulus being selected in a paired comparison p is given by Reference [7]:

$$z_{\mathrm{a}}(p) = \sqrt{2\pi}\left[\sin^{-1}\left(\sqrt{p}\right) - \frac{\pi}{4}\right] \tag{B.1}$$

where the angle from the arcsine function is in radians. Angular (or other) deviate units may be converted to JNDs by dividing the observed stimulus difference by the difference corresponding to one JND, which is obtained by setting p equal to 0,75. Using Equation (B.1), this procedure yields:

$$\mathrm{JNDs} = \frac{z(p)}{z(0{,}75)} = \frac{12}{\pi}\sin^{-1}\left(\sqrt{p}\right) - 3 \tag{B.2}$$

where the deduced JND values range from $+3$ to -3. For modest stimulus differences, the number of computed JNDs will depend only slightly upon which distribution is assumed, but for larger stimulus differences, significantly different answers result from different distributions, reflecting the variations in their tail shapes.

Annex C
(informative)

Example of a report of a psychophysical experiment

In this example, a psychophysical experiment has been performed using the triplet comparison method. The stimuli depict 5 scenes, the 13 treatments of which differ in the exposure provided by a digital still camera. Such an experiment could be useful in determining recommended exposure indices for the camera.

Table C.1 — Example report of a psychophysical experiment

a) ***Observers:*** Seventeen observers participated in this study. Data from one observer were excluded from the analysis because the ratings provided did not vary systematically with treatment nor did they correlate well with those of the remaining 16 observers. All observers were tested for normal colour vision and were confirmed to have high visual acuity with corrective lenses, if any, at the viewing distance employed in this study (406 mm). All observers were chosen based upon usage of digital images in their profession (10 of 16) or otherwise.

b) ***Scenes*****:** Five scenes were represented in the test stimuli; these depicted:

1) a house and yard;

2) a dinner party inside a restaurant (flash illumination);

3) a couple standing at a scenic overlook;

4) an indoor high school basketball game; and

5) a formal portrait.

Scenes 1), 3), and 5) contained slowly varying areas of blue sky [1) and 3)] or skin tones 5) in which variations of noise level were particularly evident. Scenes 2) and 5) contained important highlight detail the appearance of which was sensitive to overexposure.

c) ***Treatments:*** The treatments of each scene consisted of a series of 13 camera exposures separated by one-third-stop increments. The resulting images were rendered such that a selected scene midtone mapped to the same visual density. The final prints varied subjectively in noisiness and in the degree of detail in the highlights and shadows at the more extreme exposures.

d) ***Instructions*****:** The observers were directed to rate the stimuli based upon their overall quality. The range of variations in noise level and shadow and highlight detail were demonstrated to the observers using preview samples, and the differences in appearance that they might see were described verbally while the preview samples were being studied. The study was carried out using the triplet comparison method.

e) ***Viewing*****:** The glossy reflection prints (101 mm × 152 mm in size) were placed on a rack mounted on a drafting table, which was angled to produce a ≈ 45° illumination angle and a perpendicular viewing angle. Viewing distance of the stimuli was restricted to 406 mm through the use of a padded headrest. The illuminance at the plane of the stimuli was 700 lx.

f) ***Results*****:** The treatment ratings in JNDs of quality, relative to the peak quality observed, are tabulated as a function of camera exposure index. The data for professional and amateur observers were computed separately but were not found to be statistically different, and so the results presented are averaged over all 5 scenes and 16 observers.

Results

Exposure index	50	64	80	100	125	160	200	250	320	400	500	640	800
JNDs of quality	−2,4	−1,5	−0,8	−0,3	0,0	0,0	−0,1	−0,2	−0,4	−0,7	−1,1	−1,6	−2,2

The results shown in Table C.1 do not correspond to any actual digital still camera and are provided for illustrative purposes only. The quality JND values at exposure indices of 50 and 800 may be less reliable because of saturation effects (see Annex B).

Had the results differed for professional and amateur observers, they could have been reported separately, because the minimum requirement for number of determinations (4.6) would have been met (6 amateur observers times 5 scenes equals 30 determinations per treatment).

Annex D
(informative)

Comparison of selected psychometric methods

The triplet comparison method, described in detail in ISO 20462-2, incorporates elements of paired comparison, rank ordering, and categorical sort methods. In a classic paired comparison test, all $N(N-1)/2$ distinct pairs of N stimuli (usually corresponding to each of N treatments applied to a given scene) are compared. This yields results that are convertible to JNDs, but as N increases, the number of comparisons grows rapidly, and the task is inefficient because of the amount of sample handling required. In a rank ordering, all N stimuli are simultaneously ranked by the observer, which may be impractical with softcopy or projected image display, and can place stringent requirements on the size of an observation area, which should provide uniform and equivalent viewing conditions. Furthermore, in practice, the task becomes difficult as N increases, and the quality of data obtained suffers; however, the observer time required per stimulus is much less than with paired comparison. In the triplet comparison method, sets of three stimuli at a time, encompassing all possible paired comparisons, are rated on a five-point scale. Compared to paired comparison, the amount of sample handling is dramatically reduced, but without the loss of data integrity in larger rank orderings, and only a slight increase in display complexity.

As discussed in Annex B, paired comparisons and rank orderings share the limitation of saturation when stimulus differences exceed approximately 1,5 JNDs. However, if there are intermediate stimuli such that no gaps exceed this limit, larger differences can be estimated with care by avoiding potentially saturated responses and adding together smaller stimulus intervals. This requires either pilot experiments to establish estimated quality differences or discarding of sometimes sizeable fractions of data collected; in either case, the approach is inefficient if the total range of stimuli exceeds a few JNDs. Triplet comparison is subject to similar limitations, although the use of a five-point scale rather than a binary response may delay the onset of saturation at larger stimulus differences and provide slightly greater dynamic range. This method, which represents an excellent compromise between the data quality of paired comparison and the efficiency of rank ordering [8] [10] has been selected for inclusion in this part of ISO 20462 as the recommended method for quantifying stimuli that are very similar in quality.

It is often desirable to be able to characterise stimuli varying over a wider range of quality, as in the determination of the full quality distribution produced by an imaging system during practical use by untrained consumers. Two psychometric methods that are commonly used in the study of more widely varying stimuli are categorical sort and magnitude estimation procedures. In a categorical sort experiment, stimuli are considered one at a time and are classified into one of several categories, at least some of which are characterized by adjectival descriptors such as excellent, good, poor, etc. In a magnitude estimation experiment, stimuli are usually compared to a single reference stimulus, which is assigned an arbitrary numerical value. The observer is instructed to give a value twice as high if the stimulus is twice as good in quality, half as high if half as good, etc. Because neither type of experiment involves direct comparison of multiple test stimuli, conversion of the resulting rating scales to JNDs is problematic, requiring assumptions that may be arbitrary and difficult to test. In addition, the categorical sort method is subject to severe range effects, in which the range of stimuli presented influences how the classification categories are applied. Typically, there is a tendency for the observer to use each of the categories (except sometimes the two end categories, which may be held in reserve), regardless of the adjectival descriptors associated with them. Essentially, the observers adapt to the range of stimuli presented, and adjust their sensitivity so that most or all of the categories are employed, without excessive use of end categories.

To reliably convert the results of categorical sort or magnitude estimation experiments to JNDs usually requires either that selected stimuli be subsequently assessed by paired comparison to calibrate the resulting scales, or that stimuli with known differences (from previous paired comparison experiments) be included among the test stimuli. Both methods are inefficient, significantly increasing the amount of observer effort required. The quality ruler method was developed by building upon the idea of the reference stimulus in magnitude estimation [6]. Instead of single reference stimulus, the observer is provided with a series of closely spaced stimuli of known separation (usually one to three JNDs), which vary in a single attribute of image quality, and depict a single scene (often the same scene as that of the test stimulus). A mechanism is

provided for readily bringing any of the reference stimuli into direct comparison with the test stimulus, under matched viewing conditions. Comparison of the results from representative experiments yielded root-mean-square (r.m.s.) uncertainties in a single assessment (one observer rating one stimulus) of 7,8 JNDs for categorical sort, 4,3 JNDs for magnitude estimation, and 2,5 JNDs for the quality ruler. The time per assessment is higher with the quality ruler (approximately 30 s compared to about 15 s for the other two techniques), but to produce results calibrated in terms of JNDs requires that many extra stimuli be assessed in the categorical sort and magnitude estimation experiments, so their speed advantage is largely nullified. In a quality ruler experiment in which data are averaged over 3 scenes and 10 observers, an r.m.s. uncertainty of $2{,}5/(3 \times 10)^{1/2} \approx 0{,}5$ JNDs may be expected.

One potential advantage of the quality ruler method is that if the reference stimuli are calibrated against a fixed, standard numerical scale of quality, observer ratings may be converted to that scale (in real time), permitting rigorous comparison between, or integration of, separate experiments. Although it is true that such reference stimuli could be included among the test stimuli in other methods, and a regression analysis done after the completion of the experiment to permit transformation of the results to the standard scale, such a procedure is inefficient because of the extra observer effort and data analysis required. The quality ruler method has been selected for inclusion in this part of ISO 20462 as the recommended method when larger stimulus variations are to be quantified, or when results calibrated to a standard quality scale (SQS) are desired.

Bibliography

[1] BARTLESON, C.J. and GRUM, F. *Optical Radiation Measurements*, Vol. 5, Academic Press, New York, 1984.

[2] BOCK, R.D. and JONES, L. V. *The Measurement and Prediction of Judgment and Choice*, Holden-Day, San Francisco, 1968, pp. 71–75 and pp. 134–136

[3] ENGELDRUM, P.G. *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Press, Winchester, MA, USA, 2000

[4] GESCHEIDER, G.A. *Psychophysics, Method, Theory, and Application*, 2nd ed., Lawrence Erlbaum Associates, Publishers, New Jersey, 1985

[5] GUILFORD, J.P. *Psychometric Methods*, McGraw-Hill, New York, 1954

[6] KEELAN, B.W. *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, 2002, ISBN 0-8247-0770-2

[7] KEELAN, B.W. *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, 2002, pp. 29-32

[8] MIYAZAKI, K., KANAFUSA, K., UMEMOTO, H., TAKEMURA, K., URABE, H., HIRAI, K., ISHIKAWA, K. and HATADA, T. A standard portrait image and image quality assessment (II) — Triplet comparison. *Proc. SPIE,* **4300***,* 2001, pp. 309–313

[9] NUNNALY, J.C. and BERNSTEIN, I.H. *Psychometric Theory*, 3rd ed., McGraw Hill, New York, 1994

[10] TAKEMURA, K. MIYAZAKI, K., URABE, H., TOYODA, N., ISHIKAWA, K. and HATADA, T. Developing a new psychophysical experimental method to estimate image quality. *Proc. SPIE,* **4421**, 2001*,* p. 906

[11] THURSTONE, L.L.A Law of Comparative Judgment, *Psych. Rev.* **34**, 1927, pp. 273–286

[12] TORGERSON, W.S. *Theory and Methods of Scaling*, Wiley, New York, 1958

ICS 37.040.01

Price based on 17 pages