This paper was reprinted with the permission of the publisher from: Brian W. Keelan and Hitoshi Urabe, "ISO 20462, A psychophysical image quality measurement standard," in *Image Quality and System Performance*, edited by Yoichi Miyake and D. René Rasmussen, Proceedings of SPIE-IS&T Electronic Imaging, Vol. 5294 (2004), pp. 181-189.

ISO 20462, A psychophysical image quality measurement standard

Brian W. Keelan[a] and Hitoshi Urabe[b]

[a]Eastman Kodak Company, 1700 Dewey Ave., Rochester, NY, USA 14650-1860
[b]Fuji Photo Film Company, Ltd., 26-30, Nishiazabu 2-Chome, Minato-Ku, Tokyo 106-8620, Japan

ABSTRACT

ISO 20462, a three-part standard entitled "Psychophysical experimental methods to estimate image quality," is being developed by WG18 (Electronic Still Picture Imaging) of TC42 (Photography). As of late 2003, all three parts were in the Draft International Standard (DIS) ballot stage, with publication likely during 2004. This standard describes two novel perceptual methods, the triplet comparison technique and the quality ruler, that yield results calibrated in just noticeable differences (JNDs). Part 1, "Overview of psychophysical elements," discusses specifications regarding observers, test stimuli, instructions, viewing conditions, data analysis, and reporting of results. Part 2, "Triplet comparison method," describes a technique involving simultaneous five-point scaling of sets of three stimuli at a time, arranged so that all possible pairs of stimuli are compared exactly once. Part 3, "Quality ruler method," describes a real-time technique optimized for obtaining assessments over a wider range of image quality. A single ruler is a series of ordered reference stimuli depicting a common scene but differing in a single perceptual attribute. Methods for generating quality ruler stimuli of known JND separation through modulation transfer function (MTF) variation are provided. Part 3 also defines a unique absolute Standard Quality Scale (SQS) of quality with one unit equal to one JND. Standard Reference Stimuli (SRS) prints calibrated against this new scale will be made available through the International Imaging Industry Association.

Keywords: psychophysical, quality ruler, triplet comparison, just noticeable difference, JND, image quality, standard quality scale, SQS, standard reference stimuli, SRS

1. INTRODUCTION

It is often desirable to measure image quality in a standardized fashion to facilitate interpretation of results and permit rigorous comparison with other experiments. Calibrated quality measurements are useful in evaluating the performance of capture, output, and display devices, media, image processing algorithms, and entire imaging systems. There are a number of psychometric methods, such as paired comparison, rank ordering, categorical sort, and magnitude estimation that are commonly used in image quality evaluations. Various textbooks[1] have described application of such techniques and subsequent data-reduction methods, the latter of which often follow Thurstone[2] or make use of similar reasoning. Selection of the best psychometric method to use and interpretation of the resulting data is often problematic, and none of these approaches efficiently produce results calibrated against an existing numerical scale or physical standard, rendering comparison between experiments unreliable. Recently, however, the predictive power resulting from integration of multiple, calibrated psychometric experiments has been demonstrated in the development of a comprehensive model of imaging system quality.[3]

ISO 20462, a three-part standard entitled "Psychophysical experimental methods to estimate image quality," defines two perceptual methods, the triplet comparison technique[4] and the quality ruler.[3] These methods yield results calibrated in just noticeable differences (JNDs), so that the significance of experimentally measured stimuli differences is easily understood. This standard is being developed by WG18 (Electronic Still Picture Imaging) of TC42 (Photography). It was

proposed by the Japanese National Body in early 2001 and approved as a new work item late that year. H. Urabe and B. W. Keelan are the project co-leaders, with Urabe authoring Part 2 and Keelan Parts 1 and 3. As of late 2003, all three parts were in the Draft International Standard (DIS) ballot stage, with publication likely during 2004. Part 1, "Overview of psychophysical elements," discusses specifications regarding observers, test stimuli, instructions, viewing conditions, data analysis, and reporting of results. Annex A aids in identifying the better choice between the two alternative methods of Parts 2–3, which are complementary and together are sufficient to span a wide range of applications. Part 2, "Triplet comparison method," describes a technique involving simultaneous, five-point scaling of sets of three stimuli at a time, arranged so that all possible pairs of stimuli are compared exactly once. Part 3, "Quality ruler method," describes a real-time technique optimized for obtaining assessments over a wider range of image quality. A quality ruler is a series of ordered reference stimuli depicting a common scene but differing in a single perceptual attribute. Methods for generating quality ruler stimuli of known JND separation through modulation transfer function (MTF) variation are provided. Part 3 also defines a unique, absolute Standard Quality Scale (SQS) of quality with one unit equal to one JND. Standard Reference Stimuli (SRS) prints calibrated against this new scale will be made available through the International Imaging Industry Association. The contents of Parts 1–3, respectively, are described in the following three sections.

2. PART 1: OVERVIEW OF PSYCHOPHYSICAL ELEMENTS

In brief, Part 1 of ISO 20462: (1) defines the JND units by which image quality is quantified; (2) describes the influence of stimuli properties, observer characteristics, and task instructions on results obtained from rating experiments; and (3) provides guidance in choosing which of the psychophysical methods of Parts 2–3 to use.

Part 1 defines 25 terms, most of which are new to ISO standards. Of particular importance are the definitions of different types of JNDs. The standard defines a JND to be a stimulus difference that yields a 75%:25% proportion in a forced-choice paired comparison. This is an arbitrary but common choice, falling halfway between random guessing or equality (50%:50%) and certainty or unanimity (100%:0%). The probability observed in a forced-choice paired comparison is usually converted to an interval scale (such as JNDs) based on logic like that of Thurstone,[2] which involves assumption of a normal perceptual response distribution. However, the extended tails of a Gaussian function do not always accurately match that of real perceptual response distributions, and the low slopes of the tails lead to divergent uncertainties at larger stimuli differences. (For example, if 39 of 40 observers agreed, the deduced stimuli difference would be about two standard deviations or three JNDs, but if the lone observer changed his or her mind, the deduced stimulus difference would become infinite.) Given both the accuracy and precision problems in normal distribution tails, which we will refer to as "saturation effects," it is convenient to adopt a different function with bounded behavior. Annex B of Part 1 gives the following equation for converting forced-choice paired comparison probability p to JNDs based on an angular (arcsine) function, which closely resembles an integrated normal distribution but has finite tails.[3]

$$\text{JNDs} = \frac{12}{p} \cdot \sin^{-1}\left(\sqrt{p}\right) - 3 \qquad (1)$$

The possible JND values from Eq. 1 range from –3 to +3, and have fairly uniform uncertainty, so that deduced JND values may conveniently be analyzed without variance weighting.[5] Values outside the range of approximately –1.5 to +1.5, though constrained to be reasonably precise, may be inaccurate because of the arbitrary tail shape and should be considered less reliable. This observation is the basis for the bounds placed on the range of usefulness of triplet comparison and binary techniques such as paired comparison and rank ordering.

Two types of JND units, attribute and quality JNDs, are distinguished. An attribute JND is a measure of detectability of appearance differences. For example, if observers were asked to identify which of two images, differing only in tone scale, were higher in contrast, the task would be one of detection of lightness gradient differences, and observer efficiency might be limited by properties of the human visual system. If Image A were chosen 75% of the time and Image B 25%, Image A would be one attribute JND higher in contrast than Image B. The second type of JND is a measure of the significance or importance of perceived differences on overall image quality. Assessment of quality involves not only detection of appearance differences but also value judgment and personal preference. Consider again two stimuli differing only in contrast, but this time by an amount equivalent to many attribute JNDs of contrast. Essentially all observers could detect the difference and correctly identify which sample was higher in contrast but, if the task were to choose the preferred or higher quality image, it might be that some observers preferred one rendition and a

comparable number had the opposite opinion. If the two images were selected exactly equally frequently, they would differ by zero JNDs of quality, despite being many attribute JNDs different. The standard requires that the task instructions, stimuli differences, and reported JND type be consistent so that the results are unambiguously interpretable.

Other than definitions of terms, most of the normative portion of Part 1 is devoted to specifications regarding experimental conditions and reporting of results. Individual sections address properties of the observers, stimuli, and instructions; viewing conditions; session duration; and results requirements. Highlights of the specifications include the following. Relative quality values in JNDs must be based upon data from a minimum of ten observers and three scenes. Absolute quality values on the SQS scale must derive from at least twenty observers and six scenes. Observers will be tested for normal color vision and/or acuity depending upon the task requirements. The use of preview or practice images with an explanation of their differences is encouraged to avoid unstable observer criteria. To prevent undue fatigue, the median duration of an experimental session may not exceed one hour and observers requiring more time shall be allowed to return subsequently to complete a session. Viewing conditions generally follow ISO 3664 but are somewhat relaxed with regard to illuminance or white-point luminance, special color rendering, and metamerism index.

Table 1 summarizes the quantities that must be reported to facilitate understanding and interpretation of experimental results by other researchers. Annex C of Part 1 shows an example of a results report.

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 (lux) or white-point luminance (cd/m^2)
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 stimuli variations
treatment differences in attribute JNDs, JNDs of quality, or SQS values

Table 1. Summary of reported quantities. Used with the permission of ISO.

Annex A of Part 1 describes the strengths and weaknesses of the triplet comparison and quality ruler methods and provides a flow chart for deciding which method might be preferred for a given application. In brief, the triplet comparison method is an experimentally simple technique that is well suited for very precisely determining small quality differences.. As discussed above, stimuli differences larger than about 1.5 JNDs are not reliably measured by this technique (or common binary methods, such as paired comparison and rank ordering). The quality ruler technique is optimized for quantification of larger quality differences, thereby complementing the triplet comparison method. It is more complex experimentally but yields JND values in real time (rather than after all data is collected and analyzed) and permits determination of absolute SQS values.

3. PART 2: TRIPLET COMPARISON METHOD

Small stimuli differences are often measured using forced-choice comparison of all pairs of stimuli or a single simultaneous rank ordering of all stimuli. These two methods provide similar information; the results of a rank order can

be used to predict what would happen in paired comparisons by assuming that the higher ranked sample would always be chosen. In actual paired comparison experiments, the responses are not always consistent in this manner; for example, Sample A may be chosen over B, B over C, and C over A, creating a contradictory cycle. Furthermore, the methods employed by many observers in ranking sets of stimuli are often haphazard rather than systematic, particularly as the size of the set increases. Consequently, rank order data are often noisier than paired comparison data. However, the dramatically reduced sample handling and assessment times of the rank ordering task may more than compensate for the lower quality of the data, so that better data are obtained per unit observer time. Rank ordering of multiple small sets of stimuli probably represents the optimum performance balance between paired comparison (multiple rankings of sets of two stimuli) and rank order (a single ranking of all stimuli simultaneously). The small number of stimuli simultaneously considered allows the observer to assess the samples systematically, so that the quality of the data is good, while still obtaining the benefit of faster evaluations. In our experience, good results are likely when the set size is between three and eight.

The triplet comparison method is based upon this type of reasoning, adopting a set size of three. Performing all triplet comparisons requires only about one-third as many assessments as do paired comparisons. Published work indicates that the time required per assessment is comparable between the two methods (see comments below), so that total triplet assessment time is only about one-third as large, a result achieved at minimal cost in terms of the quality of the data obtained.[4] Two modifications to a strict ranking procedure are made to address the limited dynamic range (ca. 1.5 JNDs) of binary methods and the escalating length of experimental sessions as the total number of stimuli increase. First, optionally, the stimuli may be presorted into three or more categories, and only those stimuli falling within certain ranges are subsequently rated against one another. Secondly, instead of rank ordering the images in each triplet, the samples are simultaneously rated against a five-category scale. In principle, this change would be expected: (1) to reduce performance for very similar stimuli because ties are allowed (i.e., two or three stimuli of the triplet can be given the same categorical ratings); and (2) to improve performance for less similar stimuli because the rough magnitude of differences (and not just the sign) could now be distinguished. In practice, these differences have not been demonstrated, and the published data suggests that these effects are small. Consequently, the standard adopts the conservative position of assuming that triplet comparison is subject to similar dynamic range limitations as are rank ordering and paired comparison. It seems likely that the allowance of ties reduces the time assessment because observers need not belabor distinction of very similar stimuli. This may explain why triplet assessments have been observed to take no longer than paired assessments, despite the greater number of stimuli involved.

It is well known that the number of paired comparisons N required to span a total of n stimuli is given by

$$N = {}_nC_2 = \frac{n \cdot (n-1)}{2} \tag{2}$$

There are certain numbers of total stimuli n′ for which triplet comparisons can be formulated that include each possible paired comparison exactly once; for other total number of stimuli, redundancy is unavoidable. When there is no redundancy, the number of triplet comparisons N′ that are needed is given by

$$N' = \frac{n' \cdot (n'-1)}{6} \tag{3}$$

which is one-third as many as for paired comparisons. The numbers of stimuli n′ involving no duplication are given by

$$n' = 6 \cdot k - 3 \text{ or } n' = 6 \cdot k + 1 \qquad (k = 1, 2, 3, \ldots) \tag{4}$$

For larger numbers of stimuli, it can be complicated to derive triplet sets of stimuli indices that exactly span the samples with no redundancy. Informative Annex B of Part 2 provides formulas for generating such stimuli indices, as summarized in Table 2. As an example, consider the case of n′ = 13 stimuli. This meets the second criterion of Eq. 4 because 6·2+1 = 13. From Eq. 3 we expect that 26 triplet comparisons will be needed. From Table 2, the stimuli indices are given by

$$(i, f(i+1), f(i+4)) \quad \text{and} \quad (i, f(i+2), f(i+7)) \qquad (i = 1, 2, 3, ..., 13) \tag{5}$$

where the function f(j) is defined by:

$$f(j) = 1 + \text{modulo}(j-1, n') \quad (6)$$

For i = 1, the two triplets from Eq. 5 are (1, 2, 5) and (1, 3, 8); i = 13 yields (13, 1, 4) and (13, 2, 7). Indexing *i* from 1 to 13 yields 13 pairs of triplets, for a total of 26 comparisons, as expected.

n′	Triplet Combinations	n′	Triplet Combinations
7	(i, f(i+1), f(i+3)) for i = 1 to 7	21	(i, f(i+1), f(i+10)) for i = 1 to 21
9	(i, f(i+1), f(i+3)) for i = 1, 4, 7		(i, f(i+3), f(i+8)) for i = 1 to 21
	(i, f(i+1), f(i+3)) for i = 2, 5, 8		(i, f(i+2), f(i+6)) for i = 1 to 21
	(i, f(i+2), f(i+5)) for i = 1, 4, 7		(i, f(i+7), f(i+14)) for i = 1 to 7
	(i, f(i+4), f(i+8)) for i = 1, 4, 7	25	(i, f(i+2), f(i+12)) for i = 1 to 25
13	(i, f(i+2), f(i+7)) for i = 1 to 13		(i, f(i+3), f(i+11)) for i = 1 to 25
	(i, f(i+1), f(i+4)) for i = 1 to 13		(i, f(i+4), f(i+9)) for i = 1 to 25
15	(i, f(i+2), f(i+8)) for i = 1 to 15		(i, f(i+1), f(i+7)) for i = 1 to 25
	(i, f(i+1), f(i+4)) for i = 1 to 15	27	(i, f(i+1), f(i+13)) for i = 1 to 27
	(i, f(i+5), f(i+10)) for i = 1 to 5		(i, f(i+3), f(i+11)) for i = 1 to 27
19	(i, f(i+2), f(i+10)) for i = 1 to 19		(i, f(i+4), f(i+10)) for i = 1 to 27
	(i, f(i+3), f(i+7)) for i = 1 to 19		(i, f(i+2), f(i+7)) for i = 1 to 27
	(i, f(i+1), f(i+6)) for i = 1 to 19		(i, f(i+9), f(i+18)) for i = 1 to 9

Table 2. Stimuli indices in triplet comparisons without duplication. Used with the permission of ISO.

Two complementary techniques for analyzing the rating data obtained in the triplet comparisons are described in Part 2. The first technique is a type of analysis of variance based on the method of Scheffé.[6] This technique, described in detail in informative Annex E, provides information regarding variability arising from observers, stimuli, interactions, error, etc., and is useful in understanding the reliability of the measured data and the source of the measured effects. The second technique is analogous to the commonly employed Thurstone Case V analysis[7] and provides a JND calibration. Although this analysis is described in informative Annex F, a slightly different but equivalent formulation is presented here. A matrix is prepared in which the element in the i^{th} row and j^{th} column, P_{ij}, is equal to the fraction of times that Stimulus j is chosen over Stimulus i by all observers. If two stimuli are given the same rating against the five-point categorical scale, it is considered a tie, and half a "vote" is credited to each stimulus. Although a stimulus is never compared to itself, it is assumed that a tie would result, so $P_{i=j}$ is assigned a fraction of one-half. The sum $P_{ij} + P_{ji} = 1$ for all i and j. Each element of the P matrix is transformed to JNDs using Eq. 1 for improved robustness compared to use of an integrated normal distribution. The resulting Q matrix has $Q_{i=j} = 0$ and $Q_{ij} + Q_{ji} = 0$ for all i and j. The overall estimate for the JNDs of the j^{th} sample is obtained by averaging the elements of the j^{th} column. The sum of all sample JNDs will be zero when this computation is done correctly. Except for the treatment of ties and the use of the angular distribution, the analysis described above is the standard method for converting paired comparison or rank order data to an interval scale.[7]

The JND values from the Thurstone Case V analysis can be regressed against the scale values obtained from the Scheffé analysis to map the latter into JND units. An example of such a regression is shown in Fig. 1. If curvature is detected, the linear portion closest to the origin should be used to determine the constant of proportionality between the scales. In this way, saturation in the Thurstone Case V analysis will not influence the Scheffé to JND mapping.

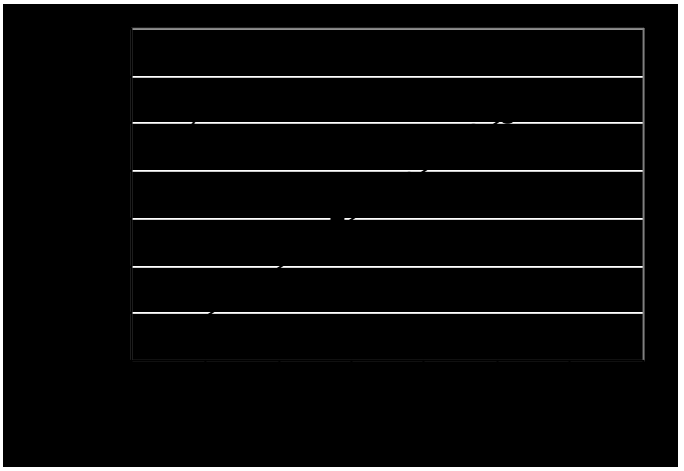

Figure 1. Mapping Scheffé scale values to JNDs. Used with the permission of ISO.

Although Part 2 of the standard does not directly address the issue of saturation, the matrix approach described above allows investigation of potential problems. After the JNDs for each sample are computed, the stimuli can be assigned new indices equal to their overall quality ranks, the highest quality sample being assigned one and the lowest, n'. If a new Q matrix is now generated, the more positive values will cluster in the lower-left corner and the more negative values in the upper right. Entries with values outside the range –1.5 to 1.5 are less reliable and should be highlighted. If significant portions of the matrix are highlighted, better estimates may result when overlapping blocks along the diagonal are analyzed independently. For example, if $n' = 9$ and there are a number of potentially saturated values in the upper-right and lower-left corners of the matrix, breaking up the 9×9 matrix into two 5×5 overlapping matrices may eliminate most of the problematical values. The first matrix would have i and $j \leq 5$ and the second i and $j \geq 5$, so they would overlap on the diagonal at $Q_{5,5}$. These two smaller matrices would be used to compute the relative JNDs of samples 1–5 and 5–9, respectively. Because Stimulus 5 is included in each scale, the two values can be equated, and the scales stitched together. An offset should be added to all values in the combined scale so that they sum to zero, as if they were analyzed from a single matrix. In this way, a single scale could be built up without using potentially saturated values. This procedure can involve omission of significant fractions of the data from the analysis if the stimuli span a wide range of quality.

Finally, Annex C of Part 2 provides a set of three standard digital portrait images that may be used to generate stimuli for psychophysical studies.

4. PART 3: QUALITY RULER METHOD

The quality ruler method described in Part 3 is particularly suitable for measuring larger quality differences. The observer ratings can be converted to JNDs in real time, rather than requiring collection and analysis of the entire experimental data set. Furthermore, using calibrated reflection print sets of Standard Reference Stimuli (SRS), the quality ruler method yields results on the standard quality scale (SQS), a fixed numerical scale that: (1) is anchored against physical standards; (2) has one unit corresponding to one JND; and (3) has a zero point corresponding to an image having little identifiable information content. Part 3 describes how users can generate their own quality ruler images with known relative JND values and, if desired, calibrate them absolutely against the SRS.

A quality ruler is a series of reference images depicting the same scene, but varying widely in a single attribute, with known quality differences (in JNDs) between the samples. The ruler is arranged in a fashion facilitating: (1) identification of the ruler stimuli closest in quality to the test stimulus; and (2) comparison between the test stimulus and the closest reference stimuli under very closely matched viewing conditions. Because the JND values of the reference stimuli are known, identification of the point of equality on the ruler allows a numerical value to be associated with the test stimulus immediately upon assessment by the observer. Both hardcopy and softcopy implementations of the quality ruler have been tested extensively and are discussed in more detail below. The two implementations have similar performance characteristics, as reported in Ch. 8 of Ref. 3. The RMS uncertainty in a single assessment (one observer rating one test stimulus, one time) is approximately 2.5 JNDs. Therefore, for example, the mean of 6 determinations (by different observers and/or against different rulers and/or of different scenes) has a standard error of $2.5/\sqrt{6} \approx 1.0$ JND. The median assessment time is approximately 30 seconds. While some other methods such as categorical sort and magnitude estimation are faster, they have larger RMS uncertainties and are still disadvantaged when considered on an equal-time basis, particularly when the necessity of included reference stimuli for calibration is considered. If a researcher generates his or her own quality rulers, the scene depicted in the ruler and the test stimulus can be matched, which makes the assessment task easier for the observer. Surprisingly, however, use of unmatched rulers and test stimuli does not cause a bias nor does it adversely affect experimental noise.

Different image attributes may be varied in quality rulers, but attributes of color and tone reproduction that are matters of preference (such as contrast) are unsuitable because different observers might rank order the ruler stimuli for quality differently. Image structure attributes are generally artifactual in nature and so are good candidates for use in quality rulers; essentially all observers agree that higher levels of artifacts are less desirable. Part 3 describes how rulers varying in sharpness may be generated. Sharpness is a good reference attribute because it: (1) is readily varied by image processing; (2) is correlated with MTF, which can be quantified by measurements from standard targets; (3) exhibits relatively low variability between different observers and scenes; and (4) has a strong affect on image quality in many practical imaging systems.

A hardcopy implementation of the quality ruler is shown in Fig. 2. The ruler proper (1) easily slides sideways in a Teflon® track, allowing the observer to bring any ruler image into direct comparison with the test stimulus, which is held in a fixture (2) next to the ruler. This fixture can be translated up and down to accommodate rulers of different heights (e.g., for horizontal and vertical images). A headrest bar (5) constrains the viewing distance, which is a requirement when the ruler varies in an image structure attribute. The illumination system (5) and base (3) together produce approximate 45°/0° viewing geometry. The observer slides the ruler so that the two ruler stimuli closest in quality to the test stimulus (usually one higher in quality and one lower) are located below the test stimulus in the triangular configuration shown in Fig. 2. In this configuration, viewing distance, viewing angle, and illumination are essentially matched. The observer rates the test image using the numerical values of the ruler images, interpolating or rarely extrapolating to the nearest JND. It is recommended that the ruler images be spaced by roughly three JNDs for the best compromise of the observer readily seeing the difference between adjacent samples, without losing interpolation accuracy.

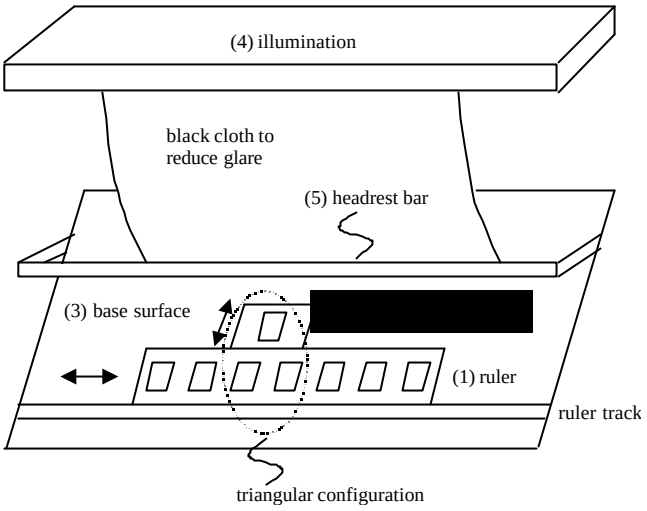

Figure 2. Hardcopy quality ruler. Used with the permission of ISO.

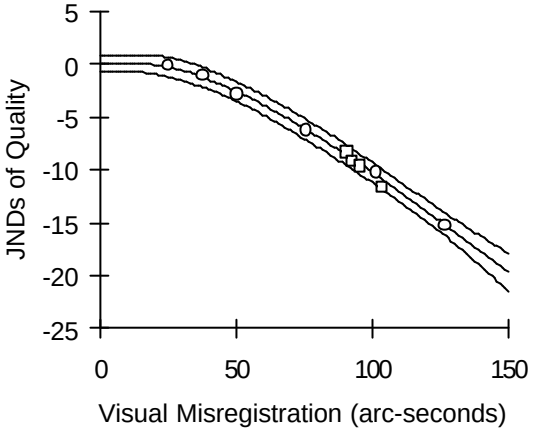

Figure 3. Typical quality ruler results. Adapted from Ref. 3, p. 202, by courtesy of Marcel Dekker, Inc.

In the softcopy implementation of the quality ruler, the test image and one ruler image at a time are displayed on matched monitors arranged side by side or on a single monitor if the display is sufficiently large. Controlling software first randomly selects the test image to be evaluated and the side on which it will be displayed. The first ruler image to be displayed is also selected randomly. The observer indicates on a small keypad whether the left or right image is of higher quality. Based on his or her response, a new ruler image is selected and displayed to the same position. The new ruler image is chosen by the software using a binary search algorithm. The new image is selected to be approximately halfway between the lowest quality ruler image that has "won" a paired comparison and the highest quality ruler image that has "lost." Before the first win or loss, the highest or lowest quality images of the rulers, respectively, are used as initial defaults. The process is terminated when the test image has been rated differently against two adjacent ruler stimuli, and the quality is computed as the arithmetic mean of the two sample JND values. So that adequate resolution is obtained, softcopy ruler images are spaced by about one JND. The controlling software randomly selects and displays new test and ruler images and the cycle begins again. As in the hardcopy ruler, viewing distance is constrained by a headrest, and careful matching of the image properties and viewing conditions between the two monitors or image positions is required. Sample instructions are provided in Annex A (hardcopy) and Annex B (softcopy). An example of a binary search routine for the softcopy ruler is given in Annex C.

Some typical results from a quality ruler experiment are shown in Fig. 3. The x-axis is an objective measure of the degree of misregistration (misalignment) between the color records of an image. The y-axis shows the quality loss in JNDs, with zero corresponding to an absence of (or subthreshold response to) the misregistration artifact. The circles

show results for different degrees of green misregistration, and the squares represent differing amounts and relative directions of simultaneous blue and red misregistration. The solid lines show a regression fit and the 95% confidence interval of the regression. The data shown are an average of results for 20 observers rating 12 scenes. A number of other examples of quality ruler results are shown in Annex E of Part 3 and in Ref. 3.

In many applications, researchers will want to make their own quality rulers so that most attributes can be matched between the test and ruler samples, simplifying the assessment task. Rulers having known separations between stimuli (when averaged over several scenes) can be generated by system MTF modification, as described in Section 7 of Part 3. Using spatial filtering, the system MTF is shaped to approximately mimic the response of a monochromatic, on-axis, diffraction-limited lens (Eq. 7), which is chosen for mathematical convenience and because many systems naturally have MTFs of approximately this shape.

$$\begin{aligned} m(n) &= \frac{2}{p}\cdot\left(\cos^{-1}(kn) - kn\cdot\sqrt{1-(kn)^2}\right) \qquad kn \le 1 \\ m(n) &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad kn > 1 \end{aligned} \tag{7}$$

In an actual lens, the constant k would correspond to the product of the lens aperture and the wavelength of light, but in this application, k is simply regarded as a reciprocal measure of system bandwidth. With the spatial frequency ν in cycles per degree, k has units of degrees. Part 3 provides details regarding how orientation (vertical and horizontal) and field position (on-axis and off-axis) MTFs are weighted to compute the system MTF, and specifies how closely the system MTF must match that of Eq. 7 on a frequency-by-frequency basis. If the shape conforms sufficiently to that of Eq. 7, the value of k fitting the system MTF is used to compute the relative JNDs of a scene, the quality of which is influenced by sharpness to an average degree. Because the scenes used may not be exactly average in their properties, it is recommended that the results from several rulers of different scenes be averaged to reduce potential bias. Extreme scenes, dominated by or virtually lacking fine detail, should be avoided, as their quality versus MTF relationship is likely to be atypical. If desired, rulers generated by researchers may be rigorously calibrated one time by rating them as test stimuli in a ruler experiment using SRS.

The SRS will be made available on the International Imaging Industry Association (I3A) website: www.i3a.org. The SRS will consist of three ruler sets depicting different scenes, each set consisting of seven 4 × 6-inch prints of varying sharpness, separated by approximately three JNDs of quality. The images will be marked with nominal SQS values and an accompanying letter will provide the calibrated SQS values for each image at a viewing distance of 16 inches. The SQS scale is defined as follows: (1) one unit is one JND; and (2) the zero point corresponds to an image of such low quality that the nature of the subject matter is not readily apparent. Images classified as being excellent would typically have SQS values on the order of thirty. Insofar as we are aware, the SRS and SQS represent the first calibrated physical standards and numerical scale of absolute pictorial image quality.

The calibration of the SRS and SQS are described in Annex D of Part 3. In brief, a large number of reflection prints representative of consumer and professional pictorial photography were assembled. These images spanned the full range of quality encountered in practice, and myriad multivariate combinations of different levels and types of attributes were represented. To provide a preliminary numerical scale, these samples were rated by experts, using magnitude estimation. Small clusters of very similarly rated stimuli were abstracted from the full range of the numerical scale and were rank-ordered by both trained observers and representative consumers. The rank order data were converted to JNDs as described in Sect. 3, from which the change in scale value corresponding to one JND was computed. These JND increments were plotted against the scale value and were fit well by a linear function, providing a conversion from scale value to JNDs over the full range of the numerical scale. Had the numerical scale been a perfect ratio scale as often assumed to result from magnitude estimation, the JND increment would have been directly proportional to scale value but, instead, a distinct intercept was present, so the relationship was linear but not perfectly proportional. The numerical scale was then converted to an absolute JND scale using the equation[3]:

$$Q(s) = Q_r + \int_{s_r}^{s} \frac{ds'}{\Delta s_J(s')} \tag{8}$$

where s is the numerical scale value, $Q(s)$ is the quality in JNDs at s, s_r is a reference value that maps to Q_r, and $\Delta s_J(s)$ is the JND increment, which in this instance was linearly related to s. The integrand is an infinitesimal numerical scale value change divided by the scale value units per JND and therefore corresponds to infinitesimal JNDs. These are accumulated by the integral starting at the reference position. The reference values were chosen so that the position on the rating scale, where it became difficult to identify the nature of the principal subject matter in the image (essentially zero on the magnitude estimation scale), was mapped to an SQS value of zero JNDs, effectively making the SQS scale an absolute scale of quality.

5. CONCLUSION

ISO 20462 should be of utility to researchers wishing to measure image quality in a relatively or absolutely calibrated sense. The standard defines two psychometric methods that are most effective over different ranges of quality and thereby complement one another. Insofar as we are aware, the SRS and SQS represent the first calibrated physical standards and numerical scale of absolute pictorial image quality. If published results are reported using the SQS scale, much more rigorous comparison and integration of experimental results from different studies and laboratories should be possible. The SRS, which will be available from the International Imaging Industry Association website, www.i3a.org, will serve to anchor the SQS and should prove useful in quality ruler and other psychophysical experiments.

ACKNOWLEDGMENTS

We thank Kenneth Parulski, Andrew Juenger, Motokazu Ohkawa, and James Peyton for helpful discussions regarding the ISO 20462 standard.

REFERENCES

1. For example, see: C. J. Bartleson. and F. Grum, *Optical Radiation Measurements*, Vol. 5, Academic Press, New York, 1984; P. G. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Press, Winchester, MA2000; G. A. Gescheider, *Psychophysics, Method, Theory, and Application*, 2nd ed., Lawrence Erlbaum Associates, Inc., New Jersey, 1985; J. P. Guilford, *Psychometric Methods*, McGraw-Hill, New York, 1954; J. C. Nunnaly and I. H. Bernstein, *Psychometric Theory*, 3rd ed., McGraw Hill, New York, 1994; and W. S. Torgerson, *Theory and Methods of Scaling*, Wiley, New York, 1958.
2. L. L. Thurstone, "A Law of Comparative Judgment," *Psych. Rev.* **34**, pp. 273–286, 1927.
3. B. W. Keelan, *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, 2002.
4. K. Kanafusa, K. Miyazaki, H. Umemoto, K. Takemura, and H. Urabe, "A Standard Portrait Image and Image Quality Assessment," *Proc. IS&T PICS 2000 Conf.*, Society for Imaging Science and Technology, Springfield, VA, pp. 317–320, 2000; K. Miyazaki, K. Kanafusa, H. Umemoto, K. Takemura, H. Urabe, K. Hirai, K. Ishikawa, and T. Hatada, "A standard portrait image and image quality assessment (II) – Triplet comparison," *Proc. SPIE* **4300**, International Society for Optical Engineering, Bellingham, WA, pp. 309–313, 2001; K. Takemura, K. Miyazaki, H. Urabe, N. Toyoda, K. Ishakawa, and T. Hatada, "Developing a new psychophysical experimental method to estimate image quality," *Proc. SPIE* **4421**, International Society for Optical Engineering, Bellingham, WA, pp. 906–909, 2001.
5. R. D. Bock and L. V. Jones, *The Measurement and Prediction of Judgment and Choice*, Holden-Day, San Francisco, pp. 71–75 and 134–136, 1968.
6. D. C. Montgomery, *Design and Analysis of Experiments*, 2nd ed., John Wiley and Sons, New York, pp. 62–64, 1984.
7. P. G. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Press, Winchester, MA, Sect. 8.2.1 and 8.3.3, 2000.