Comparison of four subjective methods for image quality assessment

Rafał K. Mantiuk[1] and Anna Tomaszewska[2] and Radosław Mantiuk[2]

[1]Bangor University, United Kingdom

[2]West Pomeranian University of Technology in Szczecin, Poland

Abstract

To provide a convincing proof that a new method is better than the state-of-the-art, computer graphics projects are often accompanied by user studies, in which a group of observers rank or rate results of several algorithms. Such user studies, known as subjective image quality assessment experiments, can be very time consuming and do not guarantee to produce conclusive results. This paper is intended to help design efficient and rigorous quality assessment experiments and emphasise the key aspects of the results analysis. To promote good standards of data analysis, we review the major methods for data analysis, such as establishing confidence intervals, statistical testing and retrospective power analysis. Two methods of visualising ranking results together with the meaningful information about the statistical and practical significance are explored. Finally, we compare four most prominent subjective quality assessment methods: single-stimulus, double-stimulus, forced-choice pairwise comparison, and similarity judgements. We conclude that the forced-choice pairwise comparison method results in the smallest measurement variance and thus produces the most accurate results. This method is also the most time-efficient, assuming a moderate number of compared conditions.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques—

Keywords: image quality, quality metrics, subjective metrics, ranking, user studies, single stimulus, double stimulus, pairwise comparison, similarity judgements

1. Introduction

When developing a new imaging or computer graphics algorithm, there is often a need to compare the results with the state-of-the-art methods. The vast majority of publications in computer graphics rely on rather informal validation, in which several examples included in the paper can be carefully inspected and compared with the results of competitive algorithms. This is an effective method, which often provides a sufficiently convincing proof of superiority of a new algorithm, but only if the visual difference is unquestionably large. If the differences are subtle, such informal comparison is often disputable. There is also a question of how a few very carefully selected and fine-tuned images generalise to the entire population of cases, which the proposed algorithm is claimed to handle. Can the judgement of the authors and several reviewers generalise to the whole population of potential users? There is definitely some lack of rigour in such endeavours. For these reasons there is a strong new trend to support the visual results by user studies, in which a larger group of assessors make judgements about their preference of one method over another.

Such user studies, or subjective quality assessment methods, are the main focus of this paper. They are becoming almost a compulsory part of many research projects. User studies are much more tedious than the informal comparison included in most papers, yet when done improperly, they do not improve generality and strength of the results. There are numerous methods of subjective quality assessment, but it is not clear

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.
The definite version is available at www.interscience.wiley.com

which method is the most effective one and leads to the most accurate results. The experimental results are often noisy and their proper analysis and interpretation is not trivial. Finally, the results for a few selected images may not generalise to another set of images. Therefore, statistical testing is crucial to build a confidence in the data.

In this work we address the problem of designing effective subjective quality assessment experiments and analysing their results. The four most prominent experimental methods selected for this study are described in detail in Section 3. In Section 4 we explain how to compute the relevant score values from raw experimental data. To compare the four experimental methods, a number of experiments have been conducted on two basic image distortions, JPEG2K image compression and unsharp masking. The collected data let us seek to answer the following important questions: Are the measurements reliable (Section 5.2)? Which experimental method is the most efficient and accurate (Section 5.3 and 5.4)? How many observers need to participate in the experiments (Section 5.5)?

The paper offers the following contributions:

- Compares the four most dominant methods of quality assessment by comparing sensitivity and time effort of each method. It helps to make an informed choice when deciding on the most appropriate experimental procedure.
- Introduces the reader to the field of subjective quality assessment and outlines the most important methods for data analysis. Such information is difficult to find in one place and often requires referring to several lengthy standard documents.

2. Related work

The subjective image quality assessment methods originate from a wider group of psychometric scaling methods, which were developed to measure psychological attributes [Tor85]. Image quality is one such attribute that describes preference for a particular image rendering. The interest in image and video quality assessment has been predominantly focused on video compression and transmission applications, resulting in several recommendations for the design of quality assessment experiments [IR02, Kee03, IT08]. The documents recommend experimental procedures (some of them evaluated in this study), viewing conditions, display calibration parameters and the methods for experimental data processing. The goal of these experimental procedures is finding a scalar-valued 'quality correlate' that would express the level of impairment (in case of video compression) or overall quality. In this work we discuss how to interpret such quality correlates in the context of rating rendering methods. We also focus on statistical testing, which is a topic often neglected in the quality assessment literature.

The standards, such as [IR02], describe the detailed procedures, but lack the wider context, statistical background and are limited to the recommended techniques. A more informative explanation of the scaling methods used for image quality assessment can be found in [Eng00] and [CW11]. This paper summarises some practical insights from those formal works in a more concise form. Because of such a concise form, the reader is expected to understand selected concepts of statistical analysis, such as ANOVA or multiple comparisons. References to the statistical textbook [How07] are provided whenever such knowledge is required.

Psychometric methods are not new to computer graphics. Recent SIGGRAPH courses, such as [Fer08] and [SWB09], demonstrate increasing interest in them. Psychometric methods have been used to scale a light reflection model in a perceptually meaningful space [PFG00], or to find the best set of parameters for tone mapping [YMMS06] or colour correction [MMTH09]. But the most prominent application of experimental methods is comparison and validation of the results produced by graphics algorithms. There is a growing share of publications that are accompanied by quality comparison studies, but there is also work devoted explicitly to comparison of existing methods, for example tone mapping operators [LCTS05, vWNA08] or image retargeting methods [RGSS10]. These studies, however, benefit little from the research that has been devoted to image quality assessment. This work is intended to bridge the gap between quality assessment research and practical quality comparison studies in computer graphics.

Relatively little work has been devoted to comparing subjective quality assessment methods. Dijk et al. [vDMW95] compared the direct ranking method (category scaling) with similarity judgements (functional measurement) and found that the results of direct ranking can be biased when the evaluated distortions are of a very different nature. Tominaga et al. [THOT10] compared eight direct rating methods and confirmed very high correlation between their results. However, little work has been done to compare the sensitivity of rating and ranking methods, which is one of the main objectives of this study.

Quality assessment would be a much easier task if it could be performed by a computational algorithm without a need for a subjective experiment. A large number of such algorithms, known as objective quality metrics, have been proposed over the years [WB06, PH11]. Their predictions can correlate well with the subjective experiments if trained for a restricted set of distortions [SSB06], but their accuracy decreases with the growing variety of distortions [PLZ*09]. Given the range of distortions that can be found in computer graphics, variety of content (images, video, geometry, textures) and complex usage scenarios, it is rather unlikely that computational metrics can completely replace the need for subjective experiments in the near future.

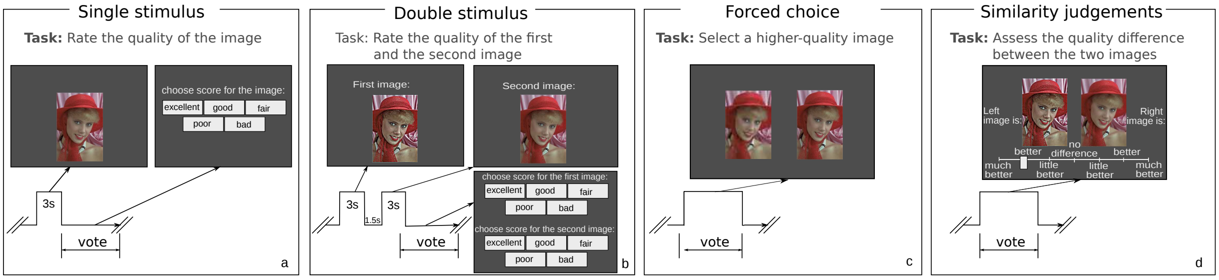


Figure 1: Overview of the four subjective quality assessment methods we investigate in this work. The diagram shows the timeline of each method and the corresponding screens.

Figure 2: The reference images (scenes) from the public domain Kodak Photo CD used in the experiments.

3. Subjective quality assessment

To avoid ambiguities, let us introduce a consistent naming convention for all subsequent sections. Our goal is to compare the quality between the results of several algorithms, each using several sets of parameters. Following [IR02], the combination of the algorithm and its parameters is called a **condition**. Quality is assessed for several **scenes**, each depicting different content, and each scene is rendered at several conditions. The experiment is run by many **observers**, and each observer can make several **repetitions** of the experiment. The image produced for a particular condition and for a particular scene is a **test image** while the original undistorted image is a **reference**.

3.1. Experimental methods

We investigate four experimental methods of quality assessment, illustrated in Figure 1. The methods were selected to represent the broad spectrum of experimental procedures, as well as reflect the most common practice in quality assessment. *Single* and *double stimulus* methods represent categorical rating, in which observers judge the quality of a single and a pair of images on a fixed 5-point scale. Both methods are dominant in video quality assessment [IR02, IT08]. *Forced-choice* pairwise comparison is an ordering method, in which observers decide which of the two displayed images has higher quality. The method is popular in computer graphics [LCTS05], but is very tedious if large number of conditions needs to be compared. In the pairwise *similarity judgement* method observers not only choose which image has higher quality, but also estimate the difference in quality on a continuous scale. Such method is used in the functional measurement approach [dRM90], which relies on relative judgements. In the following paragraphs each method is discussed in more detail.

Single stimulus categorical rating involves displaying an image for a short and fixed duration and then asking an observer to rate it using one of the five categories: excellent, good, fair, poor or bad (see Figure 1a). Such adjectives are commonly used in quality assessment as they give intuitive meaning to the numbers on an abstract quality scale. The five-point Likert-type scale is a widely used approach for scaling responses. But it must be also noted that some methods favour continuous rather than categorical scales to avoid quantisation artifacts [IR02, sec. 5.4]. The experimental method is also known as *Absolute category rating with hidden reference* [IT08]. Although 5–10 s presentation time is recommended for video, we found in the pilot study that 3 s presentation is sufficient to assess image quality, yet it does not slow-down the experiment too much. Fixing presentation time ensures that a comparable amount of attention is devoted to each image. However, presentation time is a variable that also affects the overall length of the experiment and thus the efficiency of the experimental method. All images are shown in random order and include reference images. There is no time limit in the voting stage but no image is shown during that time. The method is efficient as it requires only $n+1$ trials to assess n conditions (one additional trial for the reference image).

Double stimulus categorical rating is analogous to the single-stimulus method, but a reference image and a test image are presented in random order one after another for 3 seconds each (see Figure 1b). Following that, a voting screen is displayed on which both images are assessed separately using the same scale as for the single stimulus method. The method requires *n* trials to assess *n* conditions.

Ordering by force-choice pairwise comparison. The observers are shown a pair of images (of the same scene) corresponding to different conditions and asked to indicate an image of higher quality (see Figure 1c). Observers are always forced to choose one image, even if they see no difference between them (thus a forced-choice design). There is no time limit or minimum time to make the choice. The method is straightforward and thus expected to be more accurate than rating methods. But it also requires more trials to compare each possible pair of conditions: $0.5\,(n\cdot(n-1))$ for *n* conditions. The number of trials can be limited using balanced incomplete block designs [GT61] in which all possible paired comparisons are indirectly inferred. But even more effective reduction of trials can be achieved if a sorting algorithm is used to choose pairs to compare [SF01].

Efficient sorting algorithms, such as *quicksort*, can reduce the number of comparisons necessary to order a set of conditions to approximately $n \log n$, which could be significantly less than the full comparison, especially if the number of conditions *n* is large. When incorporated into an experiment, the sorting algorithm decides in an on-line fashion which pairs of images to compare based on the previous comparisons made in the same experimental session. Each comparison necessary to sort a set of conditions requires one trial with a two-alternative-forced-choice decision. Because such decisions are noisy and non-deterministic, such sorting rarely reflects the ranking of the true means. However, Monte-Carlo simulations have shown that gains in performance outweigh the loss of accuracy due to the incomplete design [SF01]. This is because sorting tends to concentrate comparisons around very similar images, which are the most sensitive to subjective variations. For our experiments we used the sorting algorithm based on the self balancing binary trees, as it results in low and stable number of comparisons.

Pairwise similarity judgements. While the forced-choice method orders images according to quality, it does not tell how different the images are. In pairwise similarity judgements observers are not only asked to mark their preference, but also to indicate on a continuous scale how large the difference in quality is between the two images (see Figure 1d). Observers can choose to leave the marker in the '0' position if they see no difference between the pair. The sorting algorithm used for the pairwise comparisons can also be used for the similarity judgements. The position of the marker (on the left or right side of '0') decides on the ranking of the image pair. If '0' is selected, the images are ranked randomly.

3.2. Experiment design

To compare the effectiveness of subjective quality assessment methods we conduct experiments based on the four selected methods.

Observers The images were assessed by naïve observers who were confirmed to have normal or corrected to normal vision. The age varied between 22 and 43. 17 observers completed two ranking experiments and 11 observers completed the two pairwise comparison experiments. Different groups of observers completed the experiment for each experimental method. For additional reliability, all observers repeated each experiment three times, but no two repetitions took place on the same day in order to reduce the learning effect.

Display conditions The experiments were run in two separate laboratories on two different displays: 26" NEC SpectraView 2690 and 24" HP LP2480zx. Both are high quality, 1920 × 1200 pixel resolution, LCD displays offering very good color reproduction. The display responses were measured with the Minolta CS-200 colorimeter and Specbos 1201 spectroradiometer. The measurements were used to calibrate the displays and ensure that all images were reproduced in the sRGB colour space. We increased the peak luminance of the sRGB colour space from the suggested $80\,cd/m^2$ to $180\,cd/m^2$ to reflect current capabilities of LCD displays rather than an average peak brightness of a CRT. The observers were free to adjust the viewing distance to their preference. The illumination in the room was subdued by blackout curtains to minimise the effect of display glare. Images were shown on 50%-grey background. The same background was used for the intervals between images and the voting screen. Note that quality experiments are rarely performed in controlled conditions, where viewing distance is restricted by a chin-rest and the display angular resolution (in pixels per degree) is kept constant. This is because in real-world applications images are seen from varying distance on screens of different resolutions. Therefore, the data is more representative for real-world conditions if the variability due to uncontrolled viewing conditions is included in the measurements.

Images and distortions. Selected 10 images from the Kodak Photo CD photo sampler collection, shown in Figure 2. This is the subset of images used to collect data for the LIVE quality database [SSB06]. They contain a broad range of content type, including faces, animals, man-made objects and nature.

We selected JPEG 2000 (JP2K) compression distortions and unsharp masking based on the bilateral filter ($\sigma_s = 8$, $\sigma_r = 50$) as the two evaluated algorithms, both at three levels of either distortion (JP2K) or enhancement (unsharp masking). The JP2K test images are the same as in the LIVE quality database [SSB06] while unsharp masking is a new algorithm that we decided to include in our study. Unlike

JP2K, unsharp masking can potentially improve image quality, though the quality will degrade if the filter is applied in excessive amounts. Because of this, unsharp masking is a very difficult case for computational (objective) quality metrics, which rely on the difference between test and reference images. Unsharp masking is also a common component of many computer graphics algorithms, such as tone-mapping. We intentionally selected two well known and understood algorithms, which are not comparable, so that we can focus on the subjective assessment methods rather than on the problem of finding a better algorithm.

Experimental procedure Observers were asked to read a written instruction before every experiment. Following the ITU-R500 recommendation [IR02], the experiment started with a training session in which observers could familiarise themselves with the task, interface, and typical images. The training session included 5 trials with images from the original data set, which were selected to span a wide range of distortions.. After that session, they could ask questions or start the main experiment. To ensure that observers fully attend the experiment, three random trials were shown at the beginning of the main session without recording the results. The images were displayed in a random order and with a different randomisation for each session. Two consecutive trials showing the same scene were avoided if possible. No session took longer than 30 minutes to avoid fatigue.

4. Computing scores

Once we have collected experimental data, our goal is to find a scalar measure for each test image that would rate its quality on a continuous interval scale. The following sections describe how this can be done for each experimental method.

4.1. Rating methods

Differential scores. We may be tempted to directly use the rating results: excellent, good, fair, etc. However, it was found in many studies that such estimates are very unreliable. One reason for this is that observers tend to assign a separate quality scale for each particular scene and even distortion type [vDMW95]. Instead of directly using rating results, modern quality assessment methods focus on assessing differences in quality between pairs of images. Following this approach, we compute the difference mean opinion score (DMOS) as the difference between reference and test images

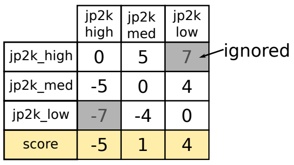
$$d_{i,j,k,r} = r_{i,\mathrm{ref}(k),k,r} - r_{i,j,k,r}. \tag{1}$$

The indices correspond to i-th observer, j-th condition, k-th scene and r-th repetition. $\mathrm{ref}(k)$ is the reference condition for scene k.

Z-scores. Different people are likely to use different adjectives when rating images, resulting in different scale associated with each observer. The easiest way to unify the scales across observers, and thus make their data comparable, is to apply a linear transform that makes the mean and the standard deviation equal for all observers. The result of such a transform is called z-score and is computed as

$$z_{i,j,k,r} = \frac{d_{i,j,k,r} - \overline{d_i}}{\sigma_i}. \tag{2}$$

The mean DMOS, $\overline{d_i}$, and standard deviation σ_i are computed across all images rated by an observer i. More sophisticated scaling procedures, such as Thurstone scaling, can account for non-linear scales that differ between observers [Tor85]. These methods, however, require a large number of measurements which are usually not available for smaller scale quality assessment experiments.



	jp2k high	jp2k med	jp2k low
jp2k_high	0	5	7
jp2k_med	-5	0	4
jp2k_low	-7	-4	0
score	-5	1	4

Figure 3: Example of projecting pairwise similarity scores into a 1-dimensional scale. The table contains the dissimilarity judgements (differences), where positive values mean that the condition in the column was selected as better than the condition in the row. The conditions are ordered from the lowest to the highest quality, so that the values just above the diagonal line are all positive. The final score is the sum of columns but computed only for the nearest (in terms of ranking) condition pairs, which lie just above and below the diagonal. All other difference values are ignored.

4.2. Pairwise methods

Transitive relation. Since a reduced pairwise comparison design was used for both the force-choice and the similarity judgement methods, several assumptions must be made to infer data for missing comparisons. The most obvious assumption is that the quality estimates are in the transitive relation: if image A is better than image B and B is better than C, then A is better than C. It must be noted that this does not need to be true for actual data collected in the full pairwise comparison experiment. Cyclic relations, in which the assumption is violated, are quite common in the full design, especially when images are similar.

Forced choice. Assuming the transitive relation, it is not difficult to compute the number of votes for each condition — the number of times one condition is preferred to other assuming that all pairs of conditions are compared. The vote count is also equivalent to the position in the ranking.

Similarity judgements data contains signed quality differences, where the sign indicates which image was judged as better. Because each observer could use a different range of values, the quality differences are divided by the overall standard deviation for a particular observer. Such scaling is similar to the z-score transformation (Equation 2), with the difference that there is no correction for the mean value $\bar{d_i}$.

We need to make one more assumption to find unambiguous quality scores from the quality differences between pairs of images. The quality differences given by the observers rarely correspond to distances in one-dimensional space. That is, if $||\cdot||$ is the magnitude estimate between a pair of conditions, $||AB|| + ||BC||$ is rarely equal to $||AC||$. We could use Multidimensional Scaling (MDS) to project the difference data to one dimensional space under the least-square criterion, but this would introduce further complications in the analysis of variance. Instead, to find the quality scores in the 1D space, we take into account only the quality difference values between the closest (in terms of ranking) pairs of images while ignoring all other magnitude estimates in the data. This is motivated by the fact that the magnitude estimates of the most similar images should be the most reliable. This simplification gives a unique (up to a constant) projection to a 1D space for each set of conditions. An example of computing such a projection is shown in Figure 3.

It is possible to use more advanced scaling methods to compute quality correlates from the distance data, for example using the Bayesian approach proposed in [SF01]. However, this would complicate the analysis of variance and estimation of confidence intervals, which are the key points of our analysis.

5. Results and analysis

The following sections are meant to discuss the features of the analysis using the collected data as an example, rather than to compare a case of J2PK and unsharp masking distortions. The detailed results and some discussion of them is included in the supplementary materials, where we also compare our results with the LIVE image quality database [SSB06] to demonstrate that the quality assessment studies can be reproduced with high consistency.

Figure 4 shows the result of the single stimulus experiment for all images, averaged over all observers and shown individually for two selected observers. Such visualisation is useful to understand the variations in the data due to scene content and observers. In our example, we can notice that the two observers gave different opinions about the unsharp masking operator. The observer with the ID 13 showed preference for moderate unsharp masking, with the z-scores in some cases exceeding that of the reference image, while the observer 16 indicated dislike for unsharp masking with most z-scores significantly lower than for the reference image. However, the *mean* observer data holds the opinion of the second observer with lower quality for images treated with unsharp masking.

5.1. Screening observers

The visualisation in Figure 4 is also useful to screen the observers whose results are not coherent with the rest of the data. The observers may report implausible quality scores because they misunderstood the experiment instruction or they did not engage in the task and gave random answers. If the number of participants is low, it is easy to spot unreliable observers by inspecting the plots. However, when the number of observers is very high or it is difficult to scrutinise the plots, the ITU-R-BT.500-11 standard [IR02], Annex 2.3.1 provides a numerical screening procedure. The procedure involves counting the number of trials in which the result of the observer lies outside ±2× standard deviation range and rejecting those observers for which a) more than 5% of the trials are outside that range; and b) the trials outside that range are evenly distributed so that the absolute difference between the counts of trials exceeding the lower and the upper bound of that range is not more than 30%. We executed this procedure on our data and we did not find any participants whose data needed to be removed.

5.2. Confidence intervals and significance

Horizontal bars, such as the ones shown in Figure 5, are a common way to visualise rating experiment results. Most studies are expected to report in addition to the mean scores also the 95% confidence interval for the mean, i.e. the range of values in which the true mean score resides with the 95% probability [IR02]. However, such confidence intervals do not explain whether the difference in scores is statistically significant or not. Even if such confidence bars overlap by small amount, the probability that the true mean lies within the overlapping region is very small, usually lower than the assumed 0.05 threshold. Thus, contrary to the convention, the thin horizontal lines in Figure 5 denote the confidence interval for the *differences in means*. If such a bar overlaps with the mean score of another condition, we have no sufficient evidence to say which condition produced higher quality image at $\alpha = 0.05$ level. However, this does not necessarily mean that both conditions produce equally good results. It can be only said that there is no statistical evidence that they differ in quality.

When comparing several pairs of conditions, it is important to adjust the confidence intervals for multiple-comparisons. This is because as more comparisons are made, the chance of the Type I error (falsely rejecting H_0) increases and is no longer $\alpha = 0.05$, but instead is closer to the sum of probabilities $\alpha = 1-(1-0.05)^c$, where c is the number of comparisons. The multiple-comparisons adjustment ensures that Type I error rate is below the desired confidence criterion α. For our analysis we used Tukey's honestly significant difference criterion [How07, sec. 12.6] available in *multcompare* Matlab Statistical Toolbox function.

To visualise significant differences using bars in Figure 5, it

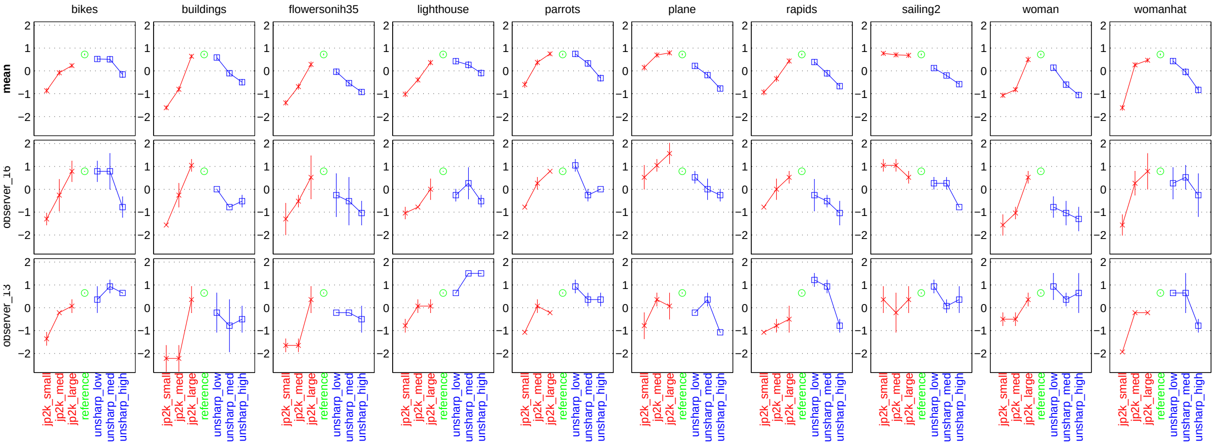


Figure 4: The results of the single stimulus experiment for all images. The top row shows the data averaged across all observers and the remaining rows shows the data for two selected observers. The error bars denote the standard error of the means.

is necessary to assume that the variance is the same for all conditions. This assumption greatly simplifies the analysis, the visualisation and gives better estimate of the true variance.

Although the bars in Figure 5 are a convenient way to show statistical significance, they do not convey the practical importance of the difference in quality. Two conditions can be statistically different, but the ostensibly better condition may be in fact selected only marginally more often than its competitor. To better visualize this aspect, Figure 6 contains additional information. It is a ranking and rating graph for the same data as shown in Figure 5. The x-axis corresponds to score values and the conditions are plotted alternatively on higher and lower position on the y-axis. Such layout lets us better depict the relations between a condition and up to four of its neighbors on the ranking scale. The relations are marked as lines between two conditions: solid blue if there is a statistically significant difference between them, and dashed red if there is no evidence to choose a better method (though one condition may still appear as better in the ranking).

The percent numbers shown on the relation lines in Figure 6 are the key feature of the graph. They represent the estimate of the probability that the condition on the right is selected as better than the condition on the left. When two methods are indistinguishable, such probability is 50%, if one methods is always selected, the probability is 100%. If the mean scores for both conditions are u_i and u_j and they have a common variance σ^2 and equal sample sizes, such probability is equal to

$$P = 1 - \frac{1}{\sigma\sqrt{4\pi}} \int_{-\infty}^{0} e^{\frac{-(u_i-u_j)^2}{4\sigma^2}} dt. \qquad (3)$$

The P value is computed from the normal cumulative distribution function assuming that the variance of the score difference is $2\sigma^2$. Such probability is very useful as it estimates in what percentage of cases an average observer will select one method (condition) over another. For example, the average quality for the *unsharp_low* is statistically different and better than for the *unsharp_med* (both bottom plots in Figure 6), but the method will be selected as better only about 6 times out of 10 (64% probability) in the two alternative forced choice scenario. Such information is very useful, as it not only tells whether the difference is statistically significant, but also whether it is significant from a practical point of view.

The scaling methods that transform scores to the scale of just-noticeable-differences (JND), such as case IV and V or Thustone's law of comparative judgements [Tor85, Eng00], or Bayesian methods, [SF01] serve similar purpose as reporting the proposed probabilities. The difference of 1 on such a JND-scale usually corresponds to the 'probability of winning' from Equation 3 equal to 75%. The scaling, however, requires much stronger assumptions about the distribution of quality scores, is ineffective when the comparisons are unanimous, it reduces the effect size (see Section 5.3), and adds an additional level of complexity to the analysis. Thus reporting the "probability of winning" is often a more convenient alternative to the scaling methods.

Figure 6 lets us also compare the results of four experimental methods with each other. All methods produced almost identical ranking, confirming that all of them can reliably measure quality. If there are any inconsistencies in ranking, such as the order of JP2K-affected images for the image 'sailing2', the multiple-comparison test correctly identifies no statistical difference. The dashed-lines serve as a warning that the ranking order should not be trusted, and that there is

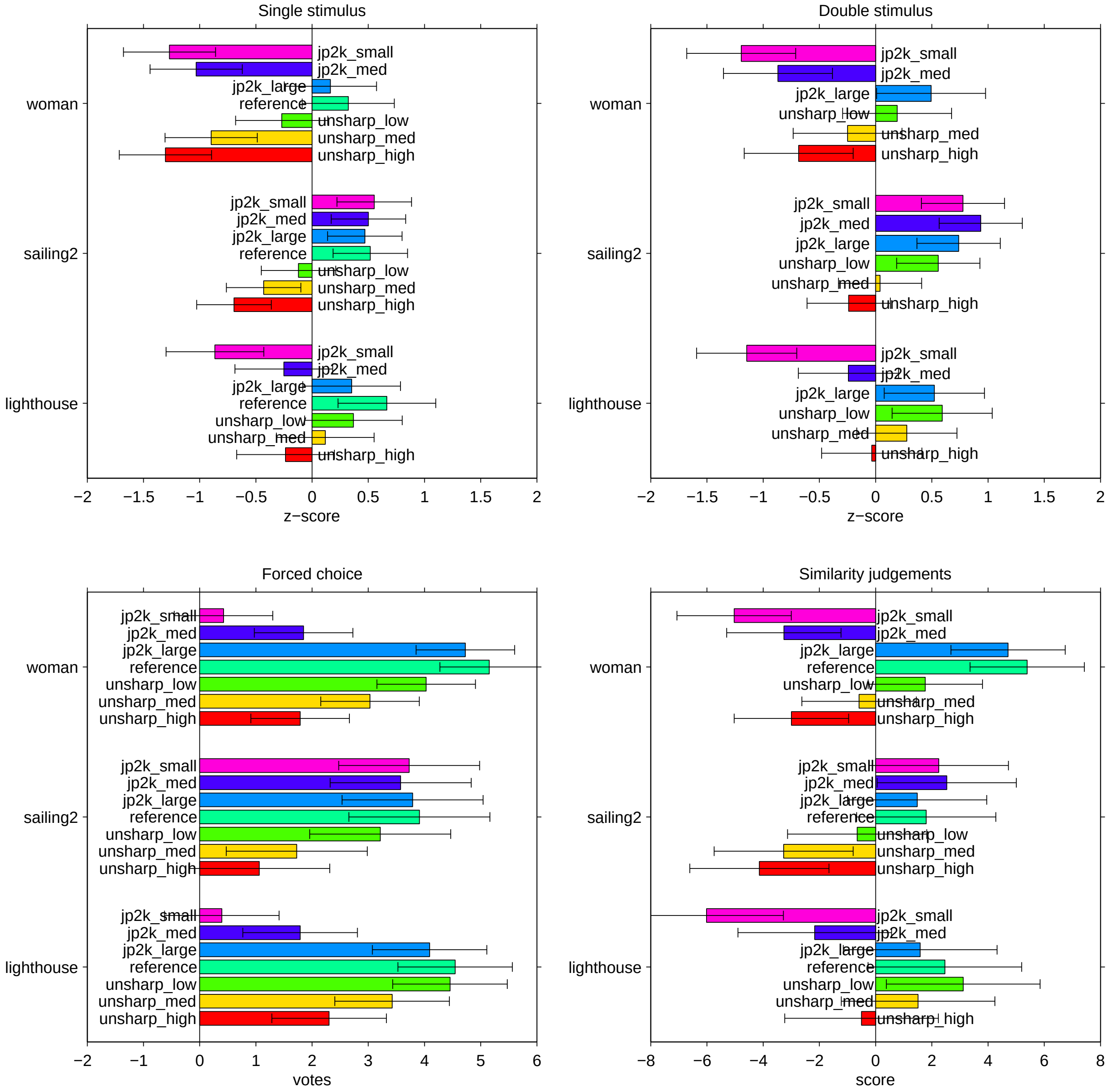


Figure 5: Comparison of quality scores, for four different experimental procedures and the three representative scenes. The thin black error bars visualise pair-wise statistical testing. If the two thin bars from two different conditions overlap at any point, the difference between them is too small to be statistically significant.

either no visible difference, or the difference is not measurable given the collected data.

5.3. Sensitivity and reliability

One of the main goals of this study was to evaluate sensitivity and reliability of each experimental method. A more accurate method should reduce randomness in answers, making the pair of compared conditions more distinctive. A more accurate method should result in more pairs of images whose quality can be said to be different under a statistical test. Previous studies used the width of confidence intervals [RLA*10] or the standard deviation [PLZ*09] to compare experimental methods. Such measures, however, are not the most suitable as the scale of quality values can vary significantly between experimental methods. Even if the data is

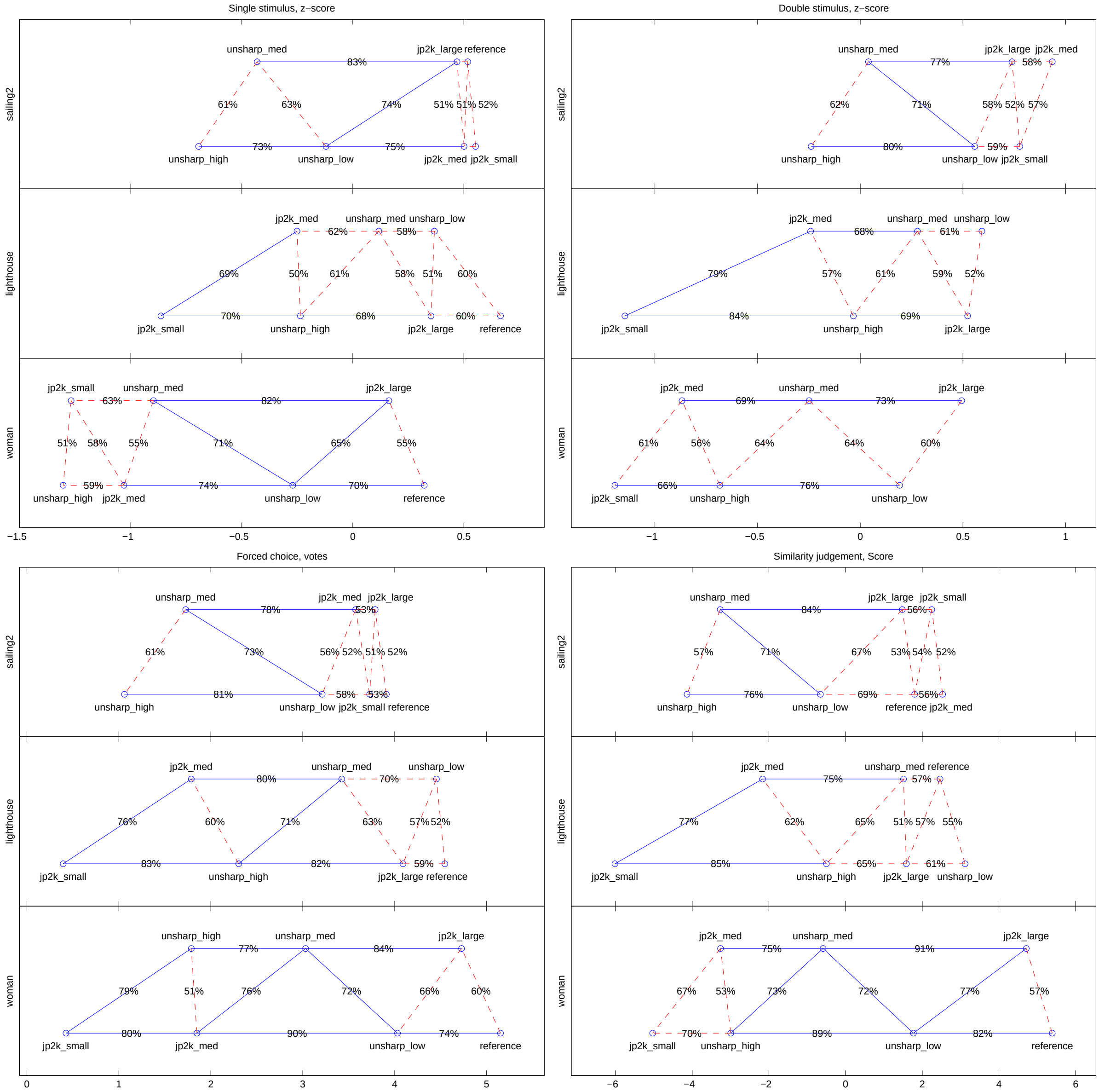


Figure 6: Ranking and rating graph illustrating the results of four tested experimental methods. Each blue circle represents tested condition and they are ordered according to their ranking, with the least preferred condition on the left. X-axis represents rating of each condition, expressed as z-score or the mean number of votes. The percentages indicate the probability that an average observer will choose the condition on the right as better than the condition on the left. If the line connecting two conditions is red and dashed, it indicates that there is no statistical difference between this pair of conditions (H_o could not be rejected for $\alpha = 0.05$ and adjusted for multiple comparisons). The probabilities close to 50% usually result in the lack of statistical significance. However, with the increasing sample size, the confidence intervals will shrink (not shown) and dashed-lines will start to disappear from the plot, while the percentage values will get more accurate (but not necessarily lower).

linearly scaled to match the same range of values, there is no guarantee that the score distribution is the same for each method. A more robust, yet still very simple of measure of performance is the *effect size*, d, which is the difference between quality scores normalised by a common standard deviation:

$$d = \frac{|u_i - u_j|}{\sigma}. \tag{4}$$

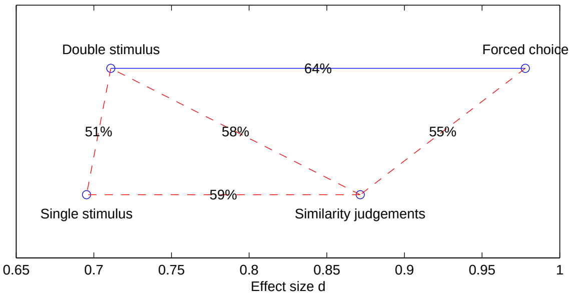


Figure 7: The comparison of effect size for each experimental method. The larger the effect size, the more accurate is the method. The notation is identical as in Figure 6.

The larger the effect size is, the higher the statistical power is, and thus we are more likely to find the statistically significant difference between a pair of conditions.

To compare the methods, we computed the effect size between the pairs of conditions jp2k_small ↔ jp2k_med, jp2k_med ↔ jp2k_large, unsharp_small ↔ unsharp_med, and unsharp_med ↔ unsharp_large. These are the pairs of small but noticeable differences in quality. Since the effect size is very sensitive to the estimate of the standard deviation σ, that estimate was computed for each pair of conditions using bootstrapping of 1000 samples from the original measurements. Then, we ran 3-way analysis of variance [How07, ch.13] (ANOVA: condition pair × experimental method × scene) on the computed d-values. The results of paired comparison test are shown in Figure 7. If one method has a higher effect size and the test confirms statistical difference (blue line), there is a high likelihood that the same method will produce higher effect size also for other set of images or pair of conditions. The percentage numbers on the lines indicate the probability that the method on the right will produce a larger effect size given a random image and a random pair of conditions.

The highest sensitivity was achieved by the forced-choice pairwise comparison method, whose effect size was statistically different from both ranking methods, which scored the worst. However, the performance improvement is not as high as the four-fold reduction of standard deviation between single stimulus and forced-choice methods reported in [PLZ*09]. We suspect that this discrepancy comes from using different measure of performance: standard deviation versus effect size. There was almost no difference between the effect size of both ranking methods. The similarity judgement method placed in-between ranking methods and the forced-choice method, but the difference with respect to all other methods was not significant.

The good performance of the forced-choice method confirms that the method is a good choice when the sensitivity is the major concern. The forced-choice method was also reported to be the easiest for the observers, as it only requires directly comparing two simultaneously shown images and a quick decision. The performance of the similarity judgement method was rather disappointing given that each trial collects more information than in the case of the forced-choice method, the task is more difficult and the experiment takes more time in overall. Surprisingly, there is no difference between the double and single stimulus methods. Although the double stimulus method should result in higher sensitivity because it provides a reference image, it also makes the task more difficult for the observers, requiring a rating of two instead of a single image. Overall, the methods that require a simpler task from observers tend to give more coherent results.

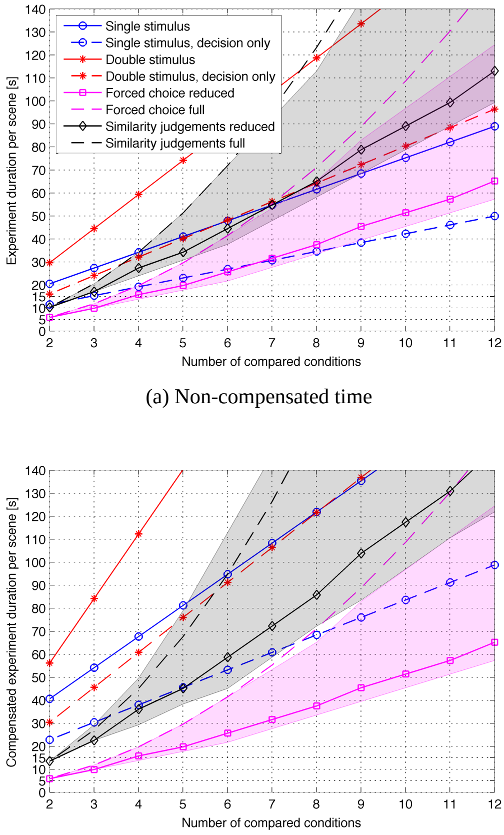
Note that the improved effect size is not necessarily reflected in smaller confidence intervals in Figures 5 and 6. This is because fewer measurements were collected for both pairwise methods as compared with the ranking methods (33 vs. 51).

5.4. Time effort

Conditions	2	3	4	5	6	7	8	9	10	11
Single stimulus	65	52	43	37	32	29	26	23	21	20
Double stimulus	40	30	24	20	17	15	13	12	11	10
Forced choice	170	123	97	80	63	52	44	39	34	31
Similarity judgements	130	87	65	52	40	32	27	23	21	18

Table 1: The maximum number of measurements (e.g. scenes × repetitions) that can be assessed within 30 minute session to compare a given number of conditions. The numbers correspond to the 80th-percentile of our experimental data, i.e. 80% of observers are expected to finish the experiment with so many measurements in 30 or less minutes.

When choosing an experimental method it is important to consider not only the sensitivity of a statistical test, but also the time that observers need to complete the experiment. After all, even a less accurate method may result in smaller confidence intervals if more measurements are collected. For each run of the main sessions of the experiment we recorded the total time as well as the number of trials. The data was averaged over all observers to compute the mean time required for a single trial. We used this data as well as the expected number of trials for each method to plot in Figure 8 the time required to compare a single scene at a given number of conditions. It is important to note that the experiment time for the reduced pairwise design can vary depending on the number of comparisons that is required to sort conditions. Therefore the timing for these methods is shown as the upper and lower bound of the sorting algorithm complexity (shaded region), together with the average complexity (the line with markers). The timing for the full design (all pairs



(a) Non-compensated time

(b) Compensated time

Figure 8: (a) — the time required to compare the given number of conditions (x-axis) using each experimental method. (b) — the same time that is compensated to result in the same relative width of the confidence intervals. The plots are based on the average time recorded in our experiments. Because the number of trials in reduced pairwise methods depends on the complexity of a sorting algorithm, the shaded regions represent the bounds between the best- and worst-case scenario. The continuous lines indicate the times based on the average complexity. The times include the assessment of a reference image for all methods, i.e. 2 conditions point corresponds to the assessment of two test images and one reference image. Non-smooth shape is due to rounding to an integer number of comparisons.

compared with each other) is shown as the dashed lines. Because the total amount time for the single and double stimulus methods varies with the fixed presentation time, we report the times both including (dashed lines) and excluding (continuous lines) the 3-second presentation of images.

Plot 8a shows that the forced choice is the fastest method for a moderate number of conditions. When the reduced design with sorting is used, the method is even faster than the single stimulus. The simplicity of the task of choosing one of the two images definitely contributes to the relatively short times for this method. But the time for this method raises rapidly with the number of conditions if the full design is used. Thus the reduced pairwise design brings significant savings in time compared to the full design starting with 6 or more conditions. Both the double stimulus and the similarity judgements methods are relatively slow, arguably because the observers' task for those methods is more complex. The time required for these methods will also depend on the choice of the presentation time.

Table 1 gives a more practical view of the same data. It shows how many measurements can be run within a recommended 30-minute session for each experimental method. To give a conservative estimates, the data corresponds to the 80-th percentile rather than average time. The table can be used as a guideline on how many scenes to include in the experiment so that each session is not excessively long. Note that the times are valid for the experimental procedures described in Section 3.1.

Since the methods differ in their sensitivity, some methods may require more measurements to result in the same confidence intervals as the other methods. To account for this difference, Figure 8b shows the times compensated for the difference in the effect size. To produce this plot, the relative standard deviation for each method was estimated from the average effect size (Section 5.3). The confidence intervals for the experimental methods A and B are equal when

$$\sigma_A\sqrt{\frac{1}{N_A}} = \sigma_B\sqrt{\frac{1}{N_B}}, \tag{5}$$

where σ_A and σ_B are the relative standard deviations for the corresponding methods, and N_A is the number of measurements per condition. Then, the increase in the number of measurements can be estimated as

$$\frac{N_A}{N_B} = \frac{\sigma_A^2}{\sigma_B^2}. \tag{6}$$

The time for worse performing methods was increased relative to the most accurate method — the forced choice pairwise comparison. After compensating the times, it is clear that both rating methods are significantly less effective than the pairwise comparison methods, especially the forced-choice method. If the reduced design is used, the single stimulus method does not seem to be more effective even if a large number of conditions is considered.

5.5. Retrospective power analysis

The key question about the quality experiment design is how many observers and/or how many repetitions are necessary to collect reliable data. Fewer samples will result in wider confidence intervals so that small quality differences will be indistinguishable in statistical terms. In an attempt to find the minimum number of observers, Winkler [Win09] collected data from 5 different quality assessment experiments and ran simulations to find that at least 10 observers are needed to measure quality score variability (standard deviation) with sufficient accuracy. But this number does not necessarily guarantee that the quality difference between a pair of conditions is statistically significant. Statistical power analysis is the method for estimating the sample sizes that give desirable sensitivity levels. The problem is that the power analysis requires prior knowledge of the differences in quality scores and their variance, which are usually unknown in advance. Therefore, in this section we employ a retrospective power analysis on our results in order to estimate typical sample sizes that are required to distinguish small and larger differences in image quality. We hope that these values will be a useful guidance when designing experiments using one of the described subjective methods.

Statistical power is the probability of correctly rejecting false H_0 (image quality is the same) when the alternative is true (image quality is different). If there is actual difference in quality, the high value of statistical power (0.8 or more) will ensure us that this difference will be detected by our test. The values in Table 2 were computed using the *sampsizepwr* Matlab Statistical Toolbox function and Equation 4 for the effect size. The desirable sample sizes $N_{0.8}$ in the table tell us how many measurements are necessary, so that the probability of finding a statistically significant difference is 0.8. However, this value is meaningful only if such a difference actually exists.

The values in the table include the cases for which any statistical difference is unlikely to be found, such as the difference between *small* and *med* JP2K conditions for the scene *sailing2*. Because of the tiny effect size for this pair of conditions, over 2,700 measurements are needed for the pairwise-comparison method to prove a potential difference. For such a case, it is usually safe to assume that the difference cannot be found. The number of measurements needed for both pairwise comparison methods is on average lower than for the ranking methods, which confirms the result of the effect size comparison from Section 5.3. The median condition data suggests that the experiment must be repeated at least 26–38 times on a single scene and a set of conditions to collect sufficient evidence for ranking images in case of a larger quality difference. But even 29–66 measurements (observers × repetitions) are needed in case of smaller quality differences. These numbers are accurate only for a single statistical t-test, and the values can be expected to be higher if the confidence levels are adjusted to account for multiple-comparisons.

Condition pair	Method	d	N	power	$N_{0.8}$
sailing2 small - med	Single stimulus	0.061	51	0.071	2115
	Double stimulus	0.14	51	0.17	381
	Forced choice	0.054	33	0.06	2736
	Similarity judgements	0.079	33	0.072	1271
sailing2 med - large	Single stimulus	0.019	51	0.052	21791
	Double stimulus	0.18	51	0.24	248
	Forced choice	0.075	33	0.07	1397
	Similarity judgements	0.18	33	0.17	253
woman small - med	Single stimulus	0.17	51	0.22	276
	Double stimulus	0.22	51	0.34	161
	Forced choice	0.53	33	0.84	30
	Similarity judgements	0.27	33	0.32	111
woman med - large	Single stimulus	0.91	51	1	12
	Double stimulus	0.93	51	1	12
	Forced choice	1.1	33	1	9
	Similarity judgements	1.3	33	1	7
Median small - med	Single stimulus	0.47	51	0.91	38
	Double stimulus	0.57	51	0.98	27
	Forced choice	0.58	33	0.89	26
	Similarity judgements	0.58	33	0.9	26
Median med - large	Single stimulus	0.41	51	0.73	66
	Double stimulus	0.49	51	0.92	36
	Forced choice	0.58	33	0.85	29
	Similarity judgements	0.44	33	0.67	48

Table 2: Power analysis for quality differences between images compressed at different JP2K settings. In addition to the values for selected scenes, median values are computed across all scenes. d represents standardised effect size, N is the sample size in our experiments and $N_{0.8}$ is the sample size required to achieve statistical power greater than or equal to 0.8.

6. Supplementary materials

The complete set of measurements is included in the supplementary materials, which we hope to serve as a reference for validating objective and subjective image quality metrics. Although there are several publicly available image quality data sets, they usually provide only aggregated data, which are difficult to use for proper statistical analysis.

7. Conclusions

This study is meant to provide a better understanding of subjective quality assessment methods and their potential in ranking computer graphics algorithms. The central theme of this study is the comparison of four most common quality assessment methods. The forced-choice pairwise comparison method was found to be the most accurate from the tested methods. This method was also found to be the most time-efficient if used in combination with a sorting algorithm that reduces the number of comparisons. Surprisingly, we found no benefit of using a more complex similarity judgement method as compared with a straightforward forced-choice. We also found no evidence that the double stimulus method is more accurate than the single stimulus method. It is important to note that these results are valid only for the experimental procedures, images and distortions used in our study, and they do not necessarily generalise to differently designed experiments. For example, if the visible differences between conditions are larger, the single stimulus rating method may become more effective.

This work emphasises the need for analysis of variance and statistical testing. If such analysis is missing, the mean rating or ranking results alone do not provide evidence that there is actually a difference between the tested algorithms. Collecting such evidence, though, requires relatively large number of measurements, often exceeding 30–60 repetitions per condition.

Quality experiment results are not easy to visualise because of the inherent uncertainty associated with the subjective measurements. This work intends to promote unambiguous reporting of such results, which does not hide uncertainty in averaged scores, but clearly indicates both statistical and practical significance of quality differences.

Acknowledgements

We would like to thank Karol Myszkowski and anonymous reviewers for their helpful comments. This research was supported by EPSRC grant no. EP/I006575/1 (UK) and grants no. N N516 193537 and N N516 508539 founded from MNiSW (Poland), years 2010-2012.

References

[CW11] Cunningham D., Wallraven C.: *Experimental Design: From User Studies to Psychophysics*, vol. 2011. Taylor & Francis, 2011. 2

[dRM90] de Ridder H., Majoor G.: Numerical category scaling: an efficient method for assessing digital image coding impairments. In *Proceedings of SPIE* (1990), vol. 1249, p. 65. 3

[Eng00] Engeldrum P.: *Psychometric scaling: a toolkit for imaging systems development*. Imcotek Press, 2000. 2, 7

[Fer08] Ferwerda J. A.: Psychophysics 101: how to run perception experiments in computer graphics. In *ACM SIGGRAPH 2008 classes* (2008), ACM, pp. 1–60. 2

[GT61] Gulliksen H., Tucker L.: A general procedure for obtaining paired comparisons from multiple rank orders. *Psychometrika 26* (1961), 173–184. 4

[How07] Howell D. C.: *Statistical Methods for Psychology*, 6th edition ed. Thomas Wadsworth, 2007. 2, 6, 10

[IR02] ITU-R.Rec.BT.500-11: Methodology for the subjective assessment of the quality for television pictures, 2002. 2, 3, 5, 6

[IT08] ITU-T.Rec.P.910: Subjective audiovisual quality assessment methods for multimedia applications, 2008. 2, 3

[Kee03] Keelan B. W.: ISO 20462: a psychophysical image quality measurement standard. In *Proceedings of SPIE* (2003), vol. 5294, SPIE, pp. 181–189. 2

[LCTS05] Ledda P., Chalmers A., Troscianko T., Seetzen H.: Evaluation of tone mapping operators using a high dynamic range display. *ACM Trans. Graph. 24*, 3 (2005), 640–648. 2, 3

[MMTH09] Mantiuk R., Mantiuk R., Tomaszewska A., Heidrich W.: Color correction for tone mapping. *Computer Graphics Forum 28*, 2 (2009), 193–202. 2

[PFG00] Pellacini F., Ferwerda J. A., Greenberg D. P.: Toward a psychophysically-based light reflection model for image synthesis. *Proceedings of the 27th annual conference on Computer graphics and interactive techniques - SIGGRAPH '00* (2000), 55–64. 2

[PH11] Pedersen M., Hardeberg Jon Y.: Full-Reference Image Quality Metrics: Classification and Evaluation. *Foundations and Trends in Computer Graphics and Vision 7*, 1 (2011), 1–80. 2

[PLZ*09] Ponomarenko N., Lukin V., Zelensky A., Egiazarian K., Carli M., Battisti F.: TID2008 - A database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics 10* (2009), 30–45. 2, 8, 10

[RGSS10] Rubinstein M., Gutierrez D., Sorkine O., Shamir A.: A comparative study of image retargeting. *ACM Trans. on Graph. 29*, 6 (2010), 160. 2

[RLA*10] Redi J., Liu H., Alers H., Zunino R., Heynderickx I.: Comparing subjective image quality measurement methods for the creation of public databases. In *SPIE 7529* (2010), vol. 7529, pp. 752903–752903–11. 8

[SF01] Silverstein D., Farrell J.: Efficient method for paired comparison. *Journal of Electronic Imaging 10* (2001), 394. 4, 6, 7

[SSB06] Sheikh H., Sabir M., Bovik A.: A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing 15*, 11 (2006), 3441–3452. 2, 4, 6

[SWB09] Sundstedt V., Whitton M., Bloj M.: The whys, how tos and pitfals of user studies. In *ACM SIGGRAPH 2009 Courses* (2009), no. 25, ACM. 2

[THOT10] Tominaga T., Hayashi T., Okamoto J., Takahashi A.: Performance comparisons of subjective quality assessment methods for mobile video. In *2nd Int. Workshop on Quality of Multimedia Experience (QoMEX)* (2010), pp. 82–87. 2

[Tor85] Torgerson W. S.: *Theory and methods of scaling*. Wiley, 1985. 2, 5, 7

[vDMW95] van Dijk A. M., Martens J. B., Watson A. B.: Quality assessment of coded images using numerical category scaling. *Proc. SPIE 2451* (1995), 90–101. 2, 5

[vWNA08] Čadík M., Wimmer M., Neumann L., Artusi A.: Evaluation of HDR tone mapping methods using essential perceptual attributes. *Computers & Graphics 32*, 3 (2008), 330–349. 2

[WB06] Wang Z., Bovik A. C.: *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 2006. 2

[Win09] Winkler S.: On the properties of subjective ratings in video quality experiments. In *1st Int. Workshop on Quality of Multimedia Experience (QoMEX)* (2009), pp. 139–144. 11

[YMMS06] Yoshida A., Mantiuk R., Myszkowski K., Seidel H.-P.: Analysis of Reproducing Real-World Appearance on Displays of Varying Dynamic Range. *Computer Graphics Forum 25*, 3 (2006), 415–426. 2