From pairwise comparisons and rating to a unified quality scale

María Pérez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, *Member, IEEE,* Vedad Hulusic, *Member, IEEE,* Giuseppe Valenzise, *Senior Member, IEEE,* and Rafał K. Mantiuk

***Abstract*—The goal of psychometric scaling is the quantification of perceptual experiences, understanding the relationship between an external stimulus, the internal representation and the response. In this paper, we propose a probabilistic framework to fuse the outcome of different psychophysical experimental protocols, namely rating and pairwise comparisons experiments. Such a method can be used for merging existing datasets of subjective nature and for experiments in which both measurements are collected. We analyze and compare the outcomes of both types of experimental protocols in terms of time and accuracy in a set of simulations and experiments with benchmark and real-world image quality assessment datasets, showing the necessity of scaling and the advantages of each protocol and mixing. Although most of our examples focus on image quality assessment, our findings generalize to any other subjective quality-of-experience task.**

***Index Terms*—Psychometric scaling, pairwise comparisons, rating, image and video quality assessment, dataset fusion**

I. Introduction

AUTOMATIC assessment of image quality is an important problem for many image processing applications, such as image/video compression or reconstruction. Those applications drive the development of computational quality metrics, which predict the level of impairment as perceived by a human observer. Such metrics need to be trained on ground truth data, which are collected in subjective quality assessment experiments. However, it is not always widely recognized that data coming from different quality assessment experiments might be scaled differently, often resulting in very different quality scores. For example, an image rated 4 on a 5-point scale in one experiment could be rated 2 in another experiment because of differences in the training, range and type of considered distortions. Dealing with widely different scales when training quality metrics is problematic, often requires using rank-order correlation as a measure of prediction accuracy, and makes difficult the use of multiple datasets for training [1], [2].

M. Pérez-Ortiz and A. Mikhailiuk contributed equally.

M. Pérez-Ortiz is with the Department of Computer Science at the University College London (UK) (email: maria.perez@ucl.ac.uk)

A. Mikhailiuk and R. Mantiuk are with the Department of Computer Science and Technology at the University of Cambridge (UK) (email: {am2442, rkm38}@cam.ac.uk).

E. Zerman is with the School of Computer Science and Statistics at Trinity College Dublin (Ireland) (email: emin.zerman@scss.tcd.ie)

V. Hulusic is with the Department of Creative Technology, Faculty of Science and Technology at the University of Bournemouth (UK) (email: vhulusic@bournemouth.ac.uk).

G. Valenzise is with the Laboratoire des Signaux et Systemes, CNRS, CentraleSupelec, Université Paris-Sud (France) (email: giuseppe.valenzise@l2s.centralesupelec.fr).

In this paper, we propose a probabilistic model and a scaling procedure that can bring quality scores from different quality assessment experiments into a unified and interpretable quality scale in which a difference of 1 between two conditions corresponds to 75% of observers selecting one condition over another. We denote one unit of difference on this scale as a *just-objectionable-difference* (JOD) and explain how it differs from the more commonly known just-noticeable-difference (JND). The proposed method builds on a well-established field of psychophysics and sensory evaluation and scales together results of two most commonly used experimental protocols: rating and pairwise comparisons. Such scaling can be used for merging existing datasets of subjective nature and for experimental protocols in which both rating and pairwise comparisons are collected. We analyze the requirements necessary for scaling, such as the need for cross-content and with-reference comparisons. Existing quality datasets together with newly collected data are used to justify the assumptions made in the model, such as the linear relation between rating and scaled pairwise comparison data. The utility of the method is demonstrated by re-scaling two existing datasets: TID2013 [3] and the HDR video compression dataset from [4] and mixing TID2013 with LIVE dataset [5] into a unified IQA dataset.

The side-benefit of the joint scaling is that we can compare and analyze sensitivity and time effort for both experimental protocols. Findings from several analyzed real-world datasets show that the standard deviation of the observer model for rating and pairwise comparisons is dependent on the task and dataset, although generally for image/video quality assessment tasks observers confuse measured conditions more often in rating experiments. This emphasizes the need for a pilot study prior to deciding on these two experimental protocols. Finally, we demonstrate using simulations that given the mean times required to rate and compare image quality and the standard deviations found for the observer model, pairwise comparisons on average result in better estimates given the same time effort. We also demonstrate that both protocols can be used together to avoid the need for time-consuming cross-content comparisons and to create larger datasets by means of relatively low experimental effort.

This paper builds on results of our prior work on psychometric scaling [6], cross-content comparisons in pairwise comparison experiments [4] and the practical findings from scaling the TID2013 dataset [7].

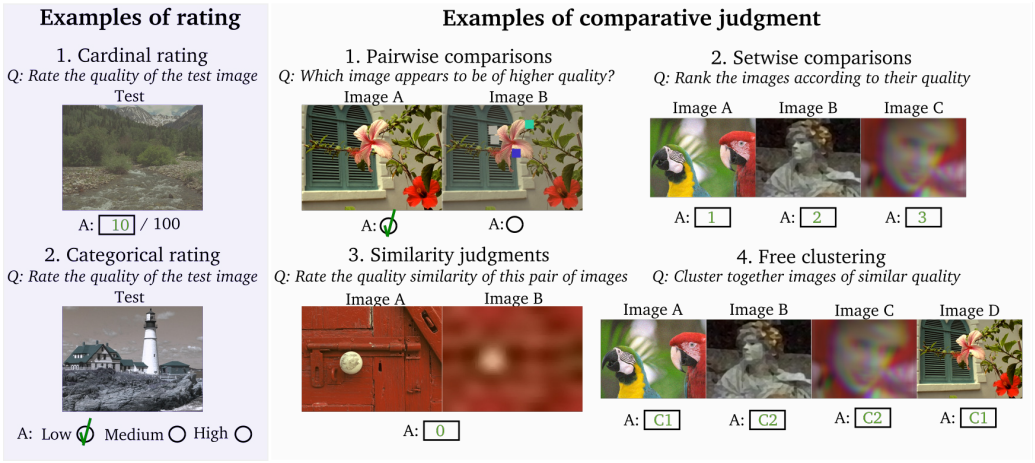

Fig. 1. Examples of different subjective judgment experiments and graphic representation of scaling using pairwise comparisons.

II. Related work

A. Subjective quality assessment methods

Many recommendations were published as guidelines for multimedia or video quality assessment [8]–[10]. These standards thoroughly describe the requirements for subjective experiments, such as set-up, procedure and material selection. Methodologies can be generally classified as rating and ranking (or comparative judgment) methods. Fig. 1 shows some examples of rating and comparative judgment experiments. Rating methods can be single, double, or multi-stimulus, depending on the presentation of the test stimuli. Users are asked to rate the presented stimuli using either a categorical or continuous interval scale. The most commonly used rating methodologies are absolute category rating (ACR) [8] for single-stimulus and double stimulus impairment scale (DSIS) or double stimulus continuous quality scale (DSCQS) [9] for double-stimulus cases. Rating methods generally work better when stimuli are easily distinguishable from one another. In contrast, comparison methods require observers to compare two or more stimuli and rank them [11] and are more suitable for cases in which the visual difference between two stimuli is small. The most commonly used comparative approach is referred to as pairwise comparison (PWC), when only two stimuli are compared at a time. The main advantage of this approach is its simplicity. The weaknesses and strengths of these strategies were compared in several studies [12]–[15].

Essentially, rating has the advantage to provide an interpretable, supra-threshold scale of quality or distortion impairment, but it also requires a careful training of subjects, who might have a different interpretation of the scale adjectives. As a consequence, the rating scale is in general not universal. On the other hand, pairwise comparison experiments have a lower cognitive load, require little training and generally eliminate the bias of the observer. However, the total number of possible comparisons increase quadratically with the number of stimuli, which makes a full comparison approach unfeasible. In practice, not all comparisons are equally useful, e.g., comparing stimuli with too close or too distant impairment levels is generally uninformative [16]. Pairs of stimuli to be compared can be sampled iteratively based on the previously compared stimuli, based on heuristics [3] or, information-theoretic criteria [17]. Recently, Shah et al. [18] compared rating and pairwise comparison experiments by conducting a series of subjective experiments in which ground truth was available – e.g. the correct radius of the presented circle or the word count in a paragraph. Similar to [15], comparison experiments were found to be more accurate in most cases and took less time compared to rating. However, authors also found that performance of rating and pairwise comparison experiments depends on the measurement noise of each experiment.

B. Fusing rating and pairwise comparisons data

It is useful in practice to aggregate quality scores obtained from different quality evaluation experiments, e.g., to create larger annotated datasets. While this aggregation of subjective quality scores is usually done for rating (i.e. mean opinion scores) [1], [2], [19] or pairwise comparisons [20], [21] individually, little has been done to study the fusion of scores obtained by both these two methodologies. In this regard, Ye and Doermann [17] proposed a unified probabilistic model, aggregating rating and pairwise comparisons together. However, they used a categorical MOS test and cutoff values for these categories. This makes the optimization procedure more difficult, which needs to be extended to experiments using a continuous interval scale rather than categories. Moreover, they did not consider the relationship between both scales, meaning that the final mixed scale could not be interpreted in terms of probabilities.

Watson [16] studied the correlation between rating scales and results of pairwise comparisons, in the context of psychometric scaling of pairwise preference probabilities. He found that the degree of agreement between two scales, for the case of video compression, is relatively high, indicating that quality scores obtained from comparisons experiment are at least as valid as double-stimulus rating scores. Differently from that work, which reports a quadratic relationship between MOS and scaled PWC (although with a very small quadratic coefficient), we assume in this work that this relationship is

linear. Nevertheless, our results might easily generalized to more complex functional forms, provided that this relation is known.

III. Towards a unified quality scale

A. Observer model

In order to map data collected in experiments into a unified quality scale, we need to make certain assumptions about how observers respond. Such assumptions are encapsulated in the observer model. It is often assumed in quality assessment experiments that quality is a one-dimensional variable, i.e., observers assign a scalar quality value to each condition. However, observers might vary in their notions of quality among them (inter-observer variance), and their opinions are also likely to change when they repeat the same experiment (intra-observer variance). Thus, quality is not a deterministic value, but a random variable, which accounts for the subjective nature of these experiments.

In *rating* experiments the random variable associated with the quality can be expressed using the following model of observer rating behavior [22]:

$$\pi_{ik} = m_i + \delta_k + \xi_{ik}, \tag{1}$$

meaning that the rating π_{ik} for observer k and condition i depends on: m_i, the ground truth quality score; δ_k, the subject bias; and ξ_{ik} the subject inaccuracy and stimulus scoring difficulty. All components in the model are assumed to be independent random variables that are Normally distributed and ξ_{ik} is assumed to have a zero mean. This makes rating π_{ik} also Normally distributed.

As for *pairwise comparisons*, the two most widely used observer models are Thurstone [23] and Bradley-Terry [24]. In practice, both lead to similar solutions. Within the Thurstone model the perceived quality of condition i is modeled as a random variable:

$$\omega_i \sim N(q_i, \sigma_i) \tag{2}$$

where the mean of the distribution is assumed to be the true quality score q_i and the standard deviation σ_i accounts for combined inter- and intra-observer variance. Individual quality scores of compared conditions can be inferred from the relative distances, calculated as:

$$\omega_j - \omega_i \sim N(q_{ij}, \sigma_{ij}) \tag{3}$$

where σ_{ij} is the standard deviation of a new distribution obtained from the difference between two quality distributions and $q_{ij} = q_i - q_j$.

Five cases of the original Thurstone model are distinguished, based on simplifying assumptions imposed on σ_{ij}:

1) The original Thurstone model, referred to as Case I, assumes that only one participant is performing the experiment and the standard deviation of the difference between random variables $\omega_i - \omega_j$ is $\sigma_{ij} = \sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho\sigma_i\sigma_j}$, where ρ is the correlation between individual scores. Despite being general, Thurstone Case I is insolvable, as every new observation will introduce a new unknown, making the number of unknowns always greater than the number of equations [23].
2) Thurstone Case II assumes that the law of comparative judgment can be applied to a group of participants, i.e. the results of individual participants can be mixed together.
3) Thurstone Case III assumes that $\sigma_{ij} = \sqrt{\sigma_i^2 + \sigma_j^2}$, that is $\rho = 0$.
4) Thurstone case IV further assumes that σ_i and σ_j are approximately equal, resulting in further simplification $\sigma_{ij} = \frac{\sigma_i + \sigma_j}{\sqrt{2}}$.
5) Thurstone Case V assumes σ_{ij} to be constant across all conditions.

If we compare Case V Thurstone, where $\omega_i \sim N(q_i, \sigma)$, to the rating model in Equation 1 we can see that it eliminates the observer bias δ_i (since pairwise comparisons are relative) and that it assumes the same standard deviation σ for different comparisons. It is important to note that the standard deviation σ describes the inherent inter- and intra-observer variations, and it is not an estimate of the measurement noise due to a limited sample size (standard error of the mean). As both are often confused in the context of pairwise comparison experiments, we will discuss these differences in detail in Section V-D.

The main difference between Thurstone Case V and Bradley-Terry models is that in the latter the difference between quality scores is expressed using a logistic distribution instead of a normal distribution. This leads to a more efficient numerical solution when optimizing quality scores. When a logistic distribution describes the difference, individual quality measurement can be described by the Gumbel distribution [25], shown in Fig. 2. It can be seen in that figure that the Bradley-Terry observer model is not symmetric. However, it leads to a very similar description of the difference in quality scores, as shown in Fig. 3. In this paper we focus on Thurstone Case V, however our findings also generalize to Bradley-Terry model.

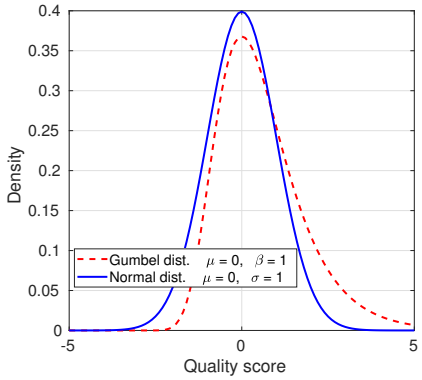

Fig. 2. Different observer models for quality assessment.

B. Pairwise comparisons and psychometric scaling

The results of a pairwise comparison experiment are usually arranged in a matrix $\mathbf{C}$, in which element c_{ij} counts the number of times stimulus i was chosen as better than j. This

subsection describes a way of converting such a matrix to an interpretable quality scale.

Probabilities p_{ij} of $\omega_i > \omega_j$ can be empirically estimated:

$$\hat{p}_{ij} = \frac{c_{ij}}{c_{ij} + c_{ji}}, \quad i \neq j. \tag{4}$$

In practice, when scaling pairwise comparison data, we can only recover the distance $q_i - q_j$ between underlying quality scores q_i and q_j, since scores are relative. The difference of two Gaussians ω_i and ω_j is also a Gaussian random variable (for Gumbel distributions a logistic), as shown in Eq. 3.

The probability of choosing condition i over j can be computed using the cumulative distribution over the difference $\omega_i - \omega_j$:

$$P(\omega_i > \omega_j) = F(q_{ij}, s_{ij}) \approx \hat{p}_{ij}, \tag{5}$$

where F is the cumulative distribution function associated to the chosen observer model and s_{ij} the parameter associated to the distribution (σ_{ij} for the Normal distribution in Thurstone model and s_{ij} for the logistic function in Bradley-Terry model). $P(\omega_i > \omega_j)$ is approximated using $\hat{p}_{ij}$. The inverse of F is shown in Fig. 3. Note that the choice of s_{ij} determines the relationship between distances in the quality scale and probabilities of better perceived quality.

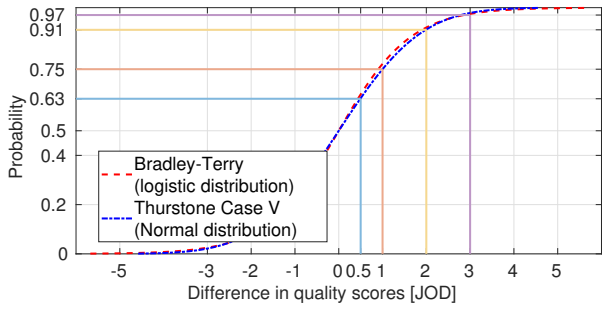

Fig. 3. Different cumulative distributions mapping probabilities into distances in the scale. Parameters for Thurstone and Bradley-Terry models were chosen such that the difference in 1 unit correspond to 75% probability of one condition being better than another.

Fig. 4 shows a graphic representation of different steps in psychometric scaling via pairwise comparisons. Psychometric scaling aims to find estimated scores $\hat{\mathbf{q}}$ such that distances between scores closely resemble distances $\hat{q}_i - \hat{q}_j$.

The probability of observing pairwise comparisons c_{ij} given latent quality scores q_i is explained by the Binomial distribution:

$$P(\mathbf{C}|\mathbf{q}, \sigma) = \prod_{i,j} \binom{n_{ij}}{c_{ij}} F\left(q_{ij}, s_{ij}\right)^{c_{ij}} \left(1 - F\left(q_{ij}, s_{ij}\right)\right)^{n_{ij} - c_{ij}}, \tag{6}$$

where $n_{ij} = c_{ij} + c_{ji}$ and F is the cumulative distribution from Eq. 5. Under Thurstone Case V assumptions, F is the cumulative normal distribution and $s_{ij} = \sqrt{2}\sigma$, where σ is the standard deviation of the observer model. σ is often selected so that when conditions are 1 unit apart in the quality scale, 75% of observers select one condition over another. This corresponds to $\sigma = 1.0484$ and $s_{ij} = 1.4826$ for normal distribution. Given the posterior probability in Eq. 6, the latent quality scores $\mathbf{q}$ can be found using the maximum likelihood estimation. More information on this formulation can be found in [6].

It should be noted that in some works the scaling of quality scores is avoided and the quality estimates are computed directly by summing up columns (or rows) of the comparison matrix. For example, the quality scores for the TID2013 dataset were computed as the average number of votes (wins in pairwise comparisons) that each condition received [3]. For that reason, we will refer to this approach as vote counts (VC). Such an approach works only if each condition was compared the same number of times and it is unsuitable for imbalanced experiment designs. We discuss shortcomings of vote counts in [7] and in Section V-D.

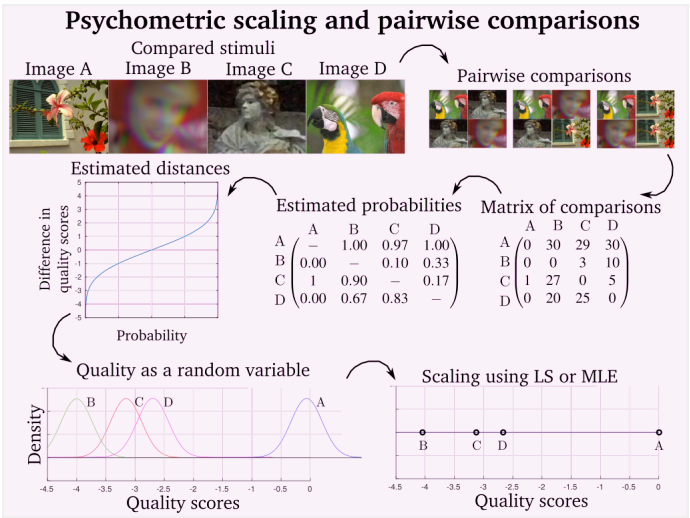

Fig. 4. Examples of different subjective judgment experiments and graphic representation of scaling using pairwise comparisons.

IV. Proposed Unified Quality Scale

A. Requirements for a unified quality scale

The vast majority of image quality assessment studies employing pairwise comparisons compare only images depicting the same content, e.g. comparing different distortion levels applied to the same original image. This "apple-to-apple" comparison simplifies the observers' task, but it comes with some limitations. Firstly, assessing and scaling each content independently makes it impossible to obtain scores that correctly capture quality differences between conditions across different contents on a common quality scale. Secondly, pairwise comparisons capture only relative quality relations. Therefore, in order to assign an absolute value to such relative measurements, the experimenter needs to assume a fixed quality for a certain condition which is then used as a reference for the scaling. As a result, the scaling error accumulates as conditions get farther from the reference on the quality scale.

Furthermore, pairwise comparison experiments can be viewed as a graph, in which conditions represent nodes and comparisons edges. To scale the quality scores for such a

graph in a consistent manner all conditions must be connected, i.e. there should be no disconnected components in the graph of comparisons. However, when each content is assessed individually, this forms a set of disconnected graphs, each with its own relative quality scale. We could potentially anchor each content by assuming that reference image for each content has a fixed quality score, for example, 0. However, we then suffer from the second mentioned problem, where conditions far away in quality from the reference accumulate large measurement error. Thus, connecting these disconnected parts is an essential step for unifying quality scale.

To address these problems, cross-content pairs can be used to connect the disconnected 'nodes' together and to eliminate the error accumulation. Additionally, assuming that all the undistorted reference stimuli are equivalent to each other (i.e. having pristine quality with "0" quality score) this graph can be connected at the reference 'node'. All the distorted images would then have negative quality values after scaling, corresponding to the distortions compared to the undistorted reference stimuli (unless enhancement is considered).

As a concept, this 'distance' to the undistorted reference stimulus is very similar to the differential mean opinion scores (DMOS) found after some rating experiments [8], [9]. Essentially, DMOS also represents the amount of impairment from the reference stimulus, similar to the scaling results. Therefore, we use DMOS in this study to compare rating scores and pairwise comparisons scaling results together.

B. JNDs and JODs

Results of pairwise comparisons are typically scaled in Just-Noticeable-Difference (JND) units [26]. Usually, the scale is constructed such that two stimuli are 1 JND apart when 75% of observers can see the difference between them. However, we believe that considering measured differences as "noticeable" leads to an incorrect interpretation of the experimental results. Let us take as an example two distorted images shown in Fig. 5: one image is distorted by noise, another by blur. They are definitely noticeably different and intuitively they should be more than 1 JND apart. However, the question we ask in an image quality experiment is not whether they are different, but rather which one is closer to the perfect quality reference. Note that a reference image does not need to be shown to answer this question as we usually have a mental notion of how a high quality image should look like. Therefore, the data we collect is not related to noticeable differences between images, but rather to image quality difference in relation to a perfect quality reference. For that reason, we describe this quality measure as Just-Objectionable-Differences (JODs) rather than JNDs. Note that JOD is the measure of impairment and not overall image aesthetics and, therefore, is related to DMOS rather than to mean opinion score (MOS). Note also that JOD does not replace JND, and the term JND is still more appropriate for all the tasks that involve direct discrimination between a pair of conditions.

The relation between JOD values and the probability of selecting condition A over condition B is illustrated in Fig. 3. When equal number of observers vote for both conditions, the probability is 0.5 and JOD difference between the conditions is 0. The differences of 1 JOD, 2 JOD and 3 JOD correspond to the probabilities $P(A > B)$ of 0.75, 0.91, and 0.97. The negative JOD values indicate that more observers preferred B over A. In all our examples we assume that the reference condition is at 0 JOD. Because of that most JOD scores we report are negative (worse than the reference).

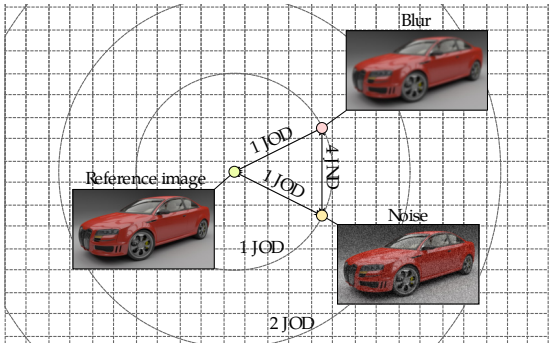

Fig. 5. Illustration of the difference between just-objectionable-differences (JODs) and just-noticeable-differences (JNDs). The images affected by blur and noise may appear to be similarly degraded in comparison to the reference image (the same JOD), but they are noticeably different and therefore several JNDs apart. The mapping between JODs and JNDs can be very complex and the relation shown in this plot is just for illustrative purposes.

C. Combination of rating and pairwise comparisons

When results of both ranking and rating experiments are available for the same set of contents, it may be desirable to use all information when constructing the quality scale. In this section we propose a simple way of combining both types of measurements. As we will show, this is also another alternative for constructing a unified quality scale.

We assume a linear relationship between random variables ω_i representing quality scores obtained from a pairwise comparison experiment (Eq. 2), and the random variables obtained from a rating experiment π_i:

$$\omega_i = a \cdot \pi_i + b. \tag{7}$$

We could instead assume a more complex relationship between the quality scores, for example quadratic [16]. However, we found that a linear assumption is sufficient for large-scale quality datasets (more details in Section V). We further assume that the standard deviation of the observer model may differ between both experimental protocols: people can confuse two conditions more often in one protocol than the other. Given that, the relationship is expanded to:

$$N(q_i, \sigma) = a \cdot N(m_{ik}, c \cdot \sigma) + b = N(a \cdot m_{ik} + b, a \cdot c \cdot \sigma), \tag{8}$$

where m_{ik} is the collected opinion score for the condition i and observer k. q_i is the latent quality score, which we want to recover. a, b and c are the unknown parameters that control the relationship between the rating and pairwise comparison data. Our goal is to find the values of the latent variables given the observed opinion scores m_{ik} and pairwise comparisons c_{ij}.

Since opinion scores are generally continuous, we express the probability of observing m_{ik} using the density function of the Normal distribution:

$$f(m_{ik}|q_i,a,b,c) = \frac{1}{\sqrt{2\pi a^2c^2\sigma^2}} e^{-\frac{((a\cdot m_{ik}+b)-q_i)^2}{2a^2c^2\sigma^2}}. \quad (9)$$

Assuming independence between observers, the likelihood of observing the whole set of opinion scores $\mathbf{M}$ is:

$$P(\mathbf{M}|\mathbf{q},\sigma,a,b,c) = \prod_{i=1}^{N}\prod_{k=1}^{J} f(m_{ik}|q_i,\sigma,a,b,c). \quad (10)$$

Similarly, the likelihood of observing pairwise comparisons $P(\mathbf{C}|\mathbf{q},\sigma)$ is given in Eq. 6. One advantage of this probabilistic formulation is that missing data, for example when observers rate only a portion of all conditions, can be simply omitted from the above product.

To recover latent quality scores $\mathbf{q}$ from both measurements, we use the maximum likelihood estimator with the posterior probability:

$$\arg\max_{\mathbf{q},a,b,c} P(\mathbf{q},a,b,c|\mathbf{C},\mathbf{M},\sigma), \quad (11)$$

where $P(\mathbf{q},a,b,c|\mathbf{C},\mathbf{M},\sigma) \propto P(\mathbf{C}|\mathbf{q},\sigma)\cdot P(\mathbf{M}|\mathbf{q},\sigma,a,b,c)\cdot P(\mathbf{q})$ and $P(\mathbf{q})$ is a Gaussian prior included to enforce convexity:

$$P(\mathbf{q}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi N\sigma^2}} e^{-\frac{(\mu_{\mathbf{q}}-q_i)^2}{N\sigma^2}}, \quad (12)$$

$\mu_{\mathbf{q}}$ being the mean of quality scores $\mathbf{q}$.

Likelihood functions are scale-invariant, i.e. $P(\mathbf{M}|q,\sigma) = P(\mathbf{M}|tq,t\sigma)$ for a constant $t \neq 0$. Thus, without loss of generality, we can fix σ to an arbitrary value. As before, since scales are relative, we need to set an anchor, e.g. $q_1 = 0$.

Note that if we wish to mix different datasets, e.g. several datasets for which rating measurements have been collected, we can do so by collecting pairwise comparisons that link the data and running the optimization procedure previously presented. In this case, different standard deviation of the observer model and scaling parameters (a, b and c) should be assumed for different datasets.

V. Experiments: Scaling existing datasets

In this section, we validate our assumptions using two real-world image quality assessment datasets. We first test the linear relationship between subjective quality scores coming from pairwise comparisons and rating and estimate the time effort and the standard deviation of the observer model in both measurements. We also validate the use of Thurstone Case V and summarize findings on an example.

To scale data, we use psychometric scaling with maximum likelihood estimation using the Thurstone Case V model, described in Section III, using the Matlab code[1] given in [6]. The code for mixing both types of measurements is available online[2].

[1]https://github.com/mantiuk/pwcmp

[2]https://github.com/mantiuk/pwcmp

A. HDR video compression dataset

As the first real-world example, we use a high dynamic range (HDR) video compression dataset[3] collected in one of our previous works [4]. This dataset contains 60 compressed HDR videos. As it was created to analyse the relationship between rating and PWC scaling, this dataset includes rating (DSIS) and PWC experiments with and without cross-content pairs.

In order to both have comparable quality PWC scaling values across different contents and to improve the effectiveness of the PWC scaling, we proposed to use additional cross-content comparisons for PWC experiments and reported the effects of having additional cross-content pairs [4]. For this purpose, four different subjective quality assessment experiments were conducted using compressed HDR video sequences and the same experimental conditions. Three of these subjective experiments were pairwise comparisons experiments with incomplete design of pair selection, with or without cross-content pairs. The results show that there is a strong linear relationship between MOS and PWC scaling results, and adding cross-content comparisons is beneficial on three different aspects: i) It reduces the content dependency, ii) increases the linear relationship between MOS values and PWC scaling results, and iii) reduces error accumulation as it reduces the confidence intervals.

Fig. 6 shows the relationship and correlation coefficients between both scales: JOD PWC scale (using psychometric scaling with pairwise comparisons) and DMOS (difference mean opinion scores from rating), where it can be seen that a linear relationship between both scales fits the data well. We performed mixed scaling and estimated the value of the parameter c from Equation (8), which we found to be 1.5 for this HDR video dataset. This means that the standard deviation of the observer model in rating experiments is 50% higher for this problem than with pairwise comparisons. The relationship between the JOD mixed scale, incorporating both rating and ranking, and JOD PWC with only ranking is shown in Fig. 6. The relation shows that rating data has little influence on the final mixed scale, which could be explained by the higher standard deviation of the observer model in the rating data.

The decision times were recorded for each participant during the subjective experiment. For HDR video compression dataset, the subjects were not able to skip the presentation of the stimuli, therefore the viewing time is the same for all subjects (10 seconds). Average decision time for the rating experiment is 6.1 seconds per rated conditions and 1.2 seconds per pair for the pairwise comparison experiment.

B. TID image quality dataset

TID2013 is one of the largest subjective image quality assessment dataset. The dataset contains over 3000 measured conditions [3]. Although there are larger datasets, such as Live Challenge [27] and KonIQ-10k [28] with over 10,000 images and natural distortions, they either do not contain pristine reference images or lack a variety of distortion types

[3]https://scss.tcd.ie/~zermane/docs/hdrVideoCompressionDB.zip

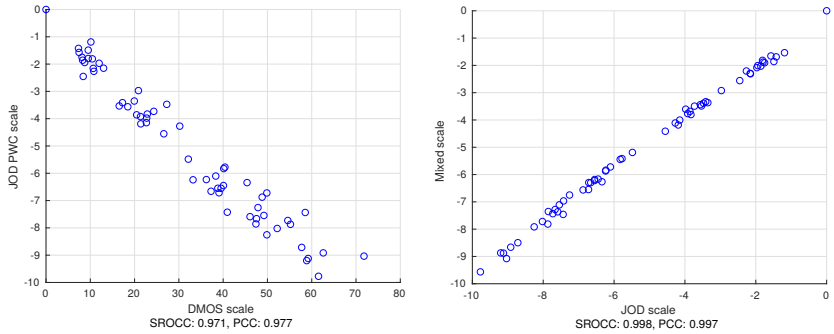

Fig. 6. Constructed scales for the HDR video dataset and correlation coefficients: Spearman (SROCC) and Pearson (PCC). From left to right: JOD PWC versus DMOS scale and JOD mixed versus JOD PWC scale.

and levels, which are both useful when establishing a ground truth image quality scale. Because of that, we focused on TID2013 in this paper. TID2013 dataset has also proven to be a challenging test for objective quality metrics. Quality scores in this dataset were obtained by collecting pairwise comparison judgements using the Swiss tournament system. In this method, all conditions are compared the same predefined number of times. The first comparisons are chosen at random. In later stages, conditions are sorted based on the number of times they were previously selected by an observer, and conditions having similar quality compete in pairs. The quality scale can then be obtained by averaging votes of observers (vote counts). However, this approach differs from the usual analysis of multiple pairwise comparisons, which involves psychometric scaling of the comparison data using either Thurstone or Bradley-Terry models. Because the initial matrix of comparisons had disconnected components and the data could not be scaled, we extended TID2013 with additional 15.000 cross-content and with-reference comparisons [7]. Furthermore, for this work we conducted an additional subjective experiment to complement the original pairwise comparison data with rating in order to analyze the relation between rating and rating protocols. Details of this experiment can be found in the Appendix.

We first compare in Fig. 7a the quality values obtained from scaling using only pairwise comparisons (JOD PWC scale) and those obtained from the rating experiment (DMOS scale). To obtain DMOS scores, the MOS scores given to the distorted images were subtracted from the scores given to the corresponding reference images. The plot shows that the relation between DMOS and JOD values can be well explained by a linear function with the exception of a few values at the extreme end of the quality scale. For those extreme points, JOD scale predicts stronger quality degradation than the DMOS scale. However, we do not have sufficient evidence to justify a non-linear relationship even though TID2013 is one of the largest quality datasets.

We performed mixed scaling and estimated the value of the parameter c from Equation 8, which we found to be 1.24. This suggest that in a typical image quality assessment experiment, the pairwise comparison protocol results in less confusion between observers. Fig. 7b shows that adding rating data (JOD mixed scale) has little impact on the final scale, maybe because the rating experiment contains much less measurements than the original set of pairwise comparisons.

Fig. 7c shows the relationship between JOD and vote-count (VC) scale. It demonstrates that psychometric scaling and additional cross-content and with-reference comparisons result in substantially different scores than those reported in the original TID2013 paper [3]. In our previous work [7] we demonstrated that the JOD scale indeed produces more consistent quality estimates and made the re-scaled TID2013 available[4].

The additional experiment let us also estimate the time effort needed for each protocol. We measured an average response time for the rating experiment to be 7.7 ± 0.9 seconds per rated condition and 3.4 ± 1.8 seconds per pair for the pairwise comparison experiment (combined viewing and decision time).

C. Validation of Thurstone Case III vs. V

In Section III-A we stated that the most commonly used assumption for the observer model, Thurstone Case V, stipulates that the standard deviation for each pair of measured conditions is the same. This would imply that the difficulty of assessing each pair of conditions and the level of confusion is the same. However, cross-content comparisons are clearly more difficult for observers to perform than within-content comparisons. It is thus reasonable to expect that more difficult types of comparisons will have a higher variability in human judgments and Case V model assumption is no longer valid.

In order to determine whether Thurstone Case V assumption is valid for cross-content and within content comparisons, we run an additional experiment on ten groups of six conditions each coming from two contents in the TID2013 dataset. Each group, shown in Fig. 8, consisted of all possible comparisons: with-reference, within-content, cross-content, within-distortions and cross-distortions. Distortions and distortion levels were the same across two contents. In the experiment, each of ten participants performed ten comparisons: six within-content comparisons and four cross-content comparisons, on every group of six conditions as illustrated in Fig. 8.

To validate whether the type of comparisons has an effect on the level of confusion (s_{ij} in Eq. 6), we performed MLE-based scaling in which s_{hard} for all "hard" comparisons (shown

[4]TID2013 scaled in JOD units: https://doi.org/10.17863/CAM.21517

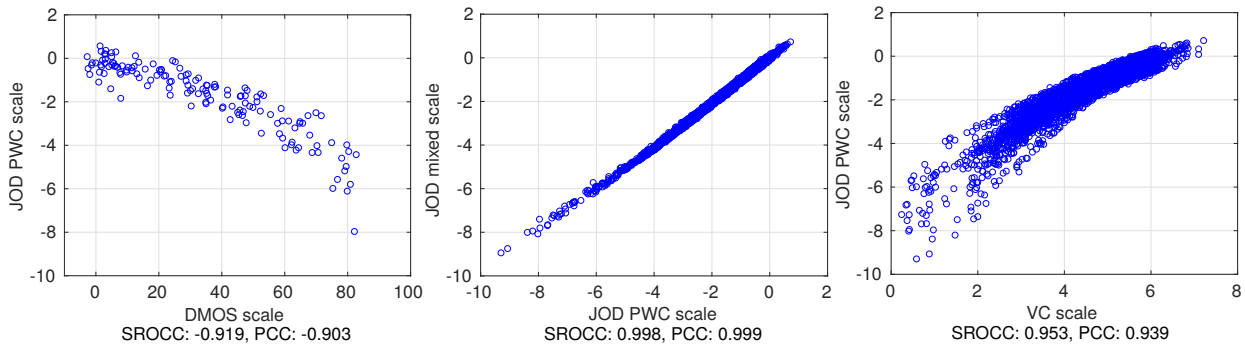

Fig. 7. Constructed scales for TID2013 and correlation coefficients: Spearman (SROCC) and Pearson (PCC). From left to right: JOD versus vote count scale, mixed versus JOD scale and mixed scale versus DMOS.

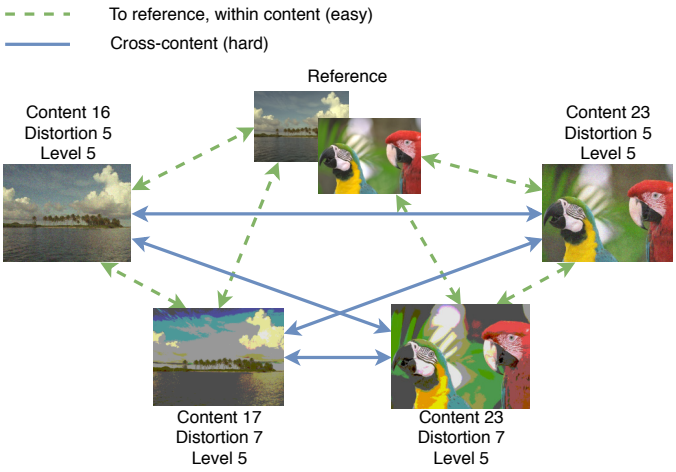

Fig. 8. Example of different comparisons types for images selected from the TID2013 dataset.

as solid lines in Fig. 8) was a free parameter. The standard deviation for all "easy" comparisons was fixed to the usual value of $s_{easy} = 1.4826$. The estimated value of s_{hard} for all ten groups is shown in Fig. 9a. The result of t-test (t(1)=-1.0, $p_{0.05} = 0.5$) indicates that we do not have evidence to suggest that the comparisons of different difficulty result in a different standard deviation s_{ij}. Therefore, contrary to our expectations, we cannot reject the assumptions of the Thurstone Case V model.

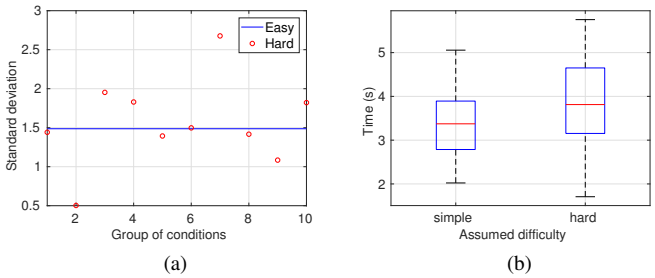

Fig. 9. (a) The estimated standard deviation of the "hard" comparisons (s_{hard}) for ten groups of conditions. The blue line represents the fixed standard deviation of the "easy" comparisons. (b) Time to complete each comparison for both difficulty levels.

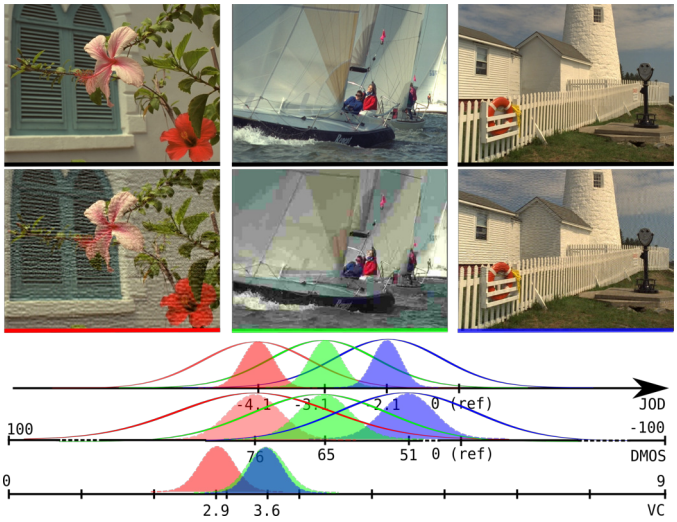

Fig. 10. The comparison of three quality scales (JOD, DMOS, VC), underlying observer model distributions (lines) and estimate distributions (filled shapes). Colors used in scales correspond to the underlines below each image. The top row shows reference images, which correspond to (ref) condition on the scale.

Fig. 9b shows the average time spent on easy and hard comparisons. Although the results for 10 groups do not indicate a statistically significant difference (t(18) = -0.92, $p_{0.05} = 0.36$), we noted that the observers spend on average 3.9s on hard and 3.3s on easy comparisons.

We do not have sufficient evidence that harder difficulty of comparisons results in higher level of confusion. It may be impractical to collect sufficient data to estimate sigmas individually for each difficulty level. Therefore, even though the sigmas could potentially be different, Case V is a good simplifying assumption and a pragmatic choice.

D. Comparison of quality scales

To summarize our findings, we show the differences between the JOD, DMOS and vote count (VC) quality scales in an example in Fig. 10. The figure shows three images from the TID2013 dataset and their corresponding quality scores in each scale. We plot above each scale the distribution associated with the observer model as a solid line and one associated with the distribution of the estimate of the mean as a filled area.

The observer distribution explains how the quality estimates vary across the population and it combines inter- and intra-observer variations. The standard deviation of this distribution is fixed for the JOD scale so that the difference of 1 unit corresponds to 75% of the population selecting one condition over another. Since DMOS scale is approximately linearly related to the JOD scale (as we show in Fig. 6 and 7), the observer model distribution for DMOS has also approximately constant standard deviation across all conditions, but its value is larger than for the JOD scale ($c = 1.24$ found for TID2013 in Section V-B). This means that the observer model and its distribution differs between experimental procedures and that observers are more likely to confuse image quality in a rating experiments than in a pairwise comparison experiment. The main difference between JOD and DMOS scales is that the distances in the JOD scale are well defined and directly related to the standard deviation of the observer model. In contrast, such distances are arbitrary for DMOS scale and vary between experiments. This is because there is no strict definition of quality ratings such as "poor" or "excellent" used in those experiments and their interpretation depends on the type of distortions that are considered, training of the participants and other factors.

The filled-shape distributions in Fig. 10 tell us how confident we are in the estimate of the mean quality score associated with our observer model. If we were to run the experiment multiple times with the same number of observers, the mean quality values across all repetitions would be distributed according to the filled shapes. Such estimate distribution can be readily calculated for DMOS scale as the standard error of the mean. Finding such distribution for JOD scale is more complex and can be obtained, for example, by bootstrapping [6]. As we collect more data, the standard deviation of that estimate distribution decreases, while the standard deviation of the observer model converges to the same constant value of σ. The estimation distribution is typically used to determine whether we have enough data to say that the quality means are different from each other (statistical significance). The observer model distribution can explain a practical significance: tell what portion of the population will make a particular judgment.

Fig. 10 also shows limitations of vote counts used as a quality measure. Firstly, there is no associated observer model that could explain quality values on a continuous quality scale. Secondly, the scale does not have the absolute 0 point assigned to reference images. Finally, the lack of cross-content comparisons makes the absolute quality estimate inaccurate when more than one content is considered.

E. On the choice of the protocol

Authors in [17] have developed a probabilistic framework for choosing either a pairwise comparison or a rating protocol, based on the information gain. The method relies on heavy computations and is not feasible for large scale datasets. Our model can be used for large scale experiments, however does not allow for a dynamic choice of the protocol. We believe, however, that our proposed approach can be very useful when combined with pilot studies. More specifically, the value of c given the results of the pilot study can guide the choice. This is, both pairwise comparisons and rating scores can be obtained for a subset of conditions and the estimation of the value of c can suggest which experiment to use, i.e. for $c < 1$ it is recommended to use rating and for $c > 1$ pairwise comparisons, and provide an estimation of the ranking of conditions so that more informed experimental designs in the case of pairwise comparisons could be used.

VI. Experiments: Validation

In this section we analyze the effect of combining rating and pairwise comparison through a set of experiments on benchmark datasets and simulations, for which ground truth is available. We use two measures for evaluating the errors: 1) Spearman Rank Order Correlation Coefficient (SROCC), which accounts for the ranking and 2) Root Mean Squared Error (RMSE), which takes the distance between conditions into account. For some experiments we also report Pearson's Linear Correlation Coefficient (PLCC).

A. Berkeley datasets

In order to find the relationship between rating scores and estimations from PWC, Shah et al. [18] conducted seven different experiments for various tasks. The tasks were estimating areas of ***circles***, ***age*** of people from photos, ***distances*** between cities, number of ***spelling*** mistakes in text, finding the frequency of ***piano*** sounds, rating ***tag-lines*** for a product and rating the ***relevance*** of image search results. Some of these datasets include ground truth, we use those for our analysis.

The measurements from each dataset were used to estimate scores for a) rating data alone, b) pairwise comparison data alone using the scaling procedure from Section III and c) mixed measurements, combining both rating and pairwise comparison data using the scaling method from Section IV-C. When both protocols were combined, we could also estimate factor c, explaining how much observer variance differs between rating and pairwise comparisons (Equation 8). We also include the total time effort spent collecting each type of experimental measurement. Note that since time effort differs, we can not directly compare both protocols in terms of accuracy. However, note that variance decreases as sample size increases, which means that estimated parameter c not only takes into account observer variance but also number of measurements.

It should be noted that we could not scale pairwise comparison results for the *Age* dataset as it contained disconnected components. However, we could use pairwise comparisons when the data from both protocols was combined. This illustrates one of the benefits of mixing both types of data: It allows to have disconnected components in the graph of comparisons, as long as conditions from both components are rated.

Results of scaling all four datasets are shown in TABLE I, together with the total time needed to collect the data. Several conclusions can be drawn from these results. Firstly, we can see that SROCC and PLCC are similar for both rating and pairwise comparisons. This indicates that both protocols are

capable of estimating the ranking between conditions correctly. However, with pairwise comparisons these ranking results are achieved with less time effort. Secondly, when RMSE is considered, the performance of both protocols depends on the standard deviation of the observer model associated with each protocol, as suggested in [18]. Note that if the c parameter is larger than 1, the rating protocol results in a larger standard deviation of the observer model than the pairwise comparison protocol. For example, since c is larger than 1 in the *Piano* dataset, pairwise comparisons result in the smaller RMSE. In the rest of the cases, c was lower than 1, which meant that rating had better results. Finally, concerning the mixing of both protocols we can see that in most cases this approach has better performance or achieves a good trade-off between both measures. This is expected, as the total amount of measurements is significantly increased when mixing both sources. However, it can also be seen that the result of the mixing highly depends on the accuracy of both types of measurement, achieving the mixing worse results in cases in which one of the protocols achieved significantly worse result than the other (e.g. case of *Spelling* for RMSE).

B. Combining LIVE and TID datasets

We further validate our method by merging two of the largest IQA datasets, i.e. TID2013 [3] and LIVE [5] datasets. LIVE contains 779 distorted images, with the scores obtained using rating. For such rating-based datasets, we need to collect two types of pairwise comparison measurements: within and cross-dataset. Within-dataset comparisons help to set the relationship between JODs and rating. Cross-dataset comparisons are necessary to put all datasets in a common unified scale. Thus, to supplement the rating data obtained for the original LIVE dataset we collected additional pairwise comparisons and re-used the data collected from the study in [17], where authors collected a total of 35700 pairwise comparisons for 7140 pairs of conditions. In our additional experiment we collected a set of 1158 pairwise comparisons for 193 pairs of images of similar quality within LIVE. We also collected cross-dataset comparisons in a similar way, where images similar in quality from the TID2013 and LIVE datasets were compared together. We collected a total of 946 comparisons for 158 pairs of conditions.

The new scale is plotted in Figure 11 versus the original scores. The plot shows substantial changes in the quality scores resulting from jointly scaling both datasets. A value of $c = 0.8$ for the LIVE dataset indicates that the rating was a more accurate protocol than pairwise comparisons. However, the opposite could be observed for TID2013 ($c = 1.24$), where pairwise comparisons resulted in more accurate measurements. Distances in the new scale have a 0.764 correlation (in terms of SROCC) with the measured cross-dataset pairwise probabilities, meaning that the mixing is able of representing the collected information properly. Fig. 12 shows a visual example to appreciate how the final mixed scale group together images of similar quality (images at the top row are 0.004 apart in the scale, images at the bottom row are 0.041 apart.

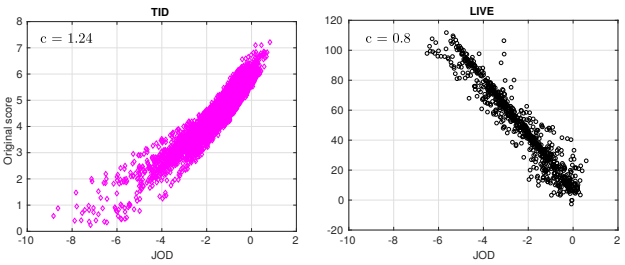

Fig. 11. Original scores of TID and LIVE datasets versus JOD values after scaling the datasets together.

Fig. 12. Two examples of images from distinct datasets that are close in the new mixed scale. The top row shows two images with associated quality close to the reference image (-0.013 and -0.009 respectively). The bottom row shows two images with quality of -5.954 and -5.913.

C. Simulations

Our goal now is to analyze which measurement is more appropriate given the same time budget. In this section we rely on Monte Carlo simulations, which assume ground truth quality scores and can be used to easily test a range of experimental strategies. For every method the simulation was set to run 100 times. We found this number of Monte Carlo iterations sufficient due to the stability of the results. The first 30 conditions of TID2013 (i.e. associated to content 1) were used as underlining true quality scores for the simulation. We use Thurstone case V observer model, described in Section III-A, to generate simulated pairwise comparison data. Swiss system was used to guide the search for the pairs to compare using 9 rounds, as done in TID2013 [3]. This means that each observer of pairwise comparison experiments measured $9 \cdot (N/2)$ comparisons in total. To generate simulated ratings we add Gaussian-distributed noise to ground truth data, i.e. assuming that the same observer model is used for both pairwise comparisons and rating. Each observer measured N conditions for rating. The same experimental procedure is used for all simulations in this paper. In our simulation we test how the standard deviation of the observer model for each protocol (related to c in our model) affects the results.

We simulated pairwise comparison, rating and mixed exper-

TABLE I
RESULTS OBTAINED BY RATING, PAIRWISE COMPARISONS AND MIXED EXPERIMENTS IN FOUR PUBLICLY AVAILABLE DATASETS. THE TABLE SHOWS PLCC, SROCC AND RMSE MEASURES AND THE FITTED C PARAMETER EXPLOITING THE RELATION BETWEEN THE STANDARD DEVIATION OF THE OBSERVER MODEL FOR BOTH PROTOCOLS. TOTAL TIME FOR DATA COLLECTION FOR EACH TYPE OF EXPERIMENTS IS ALSO SHOWN.

	PLCC			SROCC			RMSE			c	Total time (secs.)	
Dataset	Rating	Pairw. comp.	Mix	Rating	Pairw. comp.	Mix	Rating	Pairw. comp.	Mix	Mix	Rating	Pairw. comp.
Distances	0.982	0.951	0.981	0.982	0.977	0.979	0.258	0.304	0.189	0.911	15176	12844
Age	0.886	-	0.913	0.805	-	0.875	0.442	-	0.388	0.762	6462	2790
Piano	0.889	0.944	0.938	0.830	0.927	0.939	0.602	0.316	0.334	1.737	7431	5218
Spelling	0.568	0.481	0.546	0.667	0.667	0.667	0.785	0.953	0.892	0.810	9706	17505

iments with varying number of measurements. In the case of the mixed scale, half of the observers performed a pairwise comparison experiment and the other half performed rating. In our simulations, we tested i) $c = 0.5$ (rating results in less confusion than PWC), ii) $c = 1$ (both measurements result in the same confusion), iii) $c = 1.24$ (the ratio found in TID2013) and iv) $c = 2$ (rating has double the standard deviation of PWC). The error measures are plotted according to the total time effort needed in Fig. 13, where time effort corresponds to the number of measurements multiplied by the average time required per measurement found with TID2013.

From the figures, we can conclude that the measurement with the lowest standard deviation of the observer model obviously achieves better performance and is preferred in all scenarios, although most measurements converge with enough time effort. When measurement noise is unknown, mixing represents a suitable approach, achieving reasonable performance and a trade-off between both experimental protocols. Mixing also behaves well when data coming from rating is much more noisy, achieving performance close to PWC. We can also see that for the case of $c = 1.24$ (found with TID2013) pairwise comparisons are more efficient, supporting the use of such pairwise comparisons for image quality assessment.

Next, we study the case of disconnected components in the graph of comparisons and missing rating data when mixing both scales. Here we do not assume the same budget of comparisons, but rather use fixed number of observers. The same configuration for the simulation, explained at the beginning of this subsection, is used. TABLE II shows the case of three approaches: Rating, rating with data missing at random (20% of the rating data is missing), pairwise comparisons with connected components (PWC) and mixing with data missing at random (again, same 20%) and disconnected components (here we break the graph of comparisons so that there is always two disconnected components). We perform 100 runs for each method and test it with 10, 20 and 30 observers. We report RMSE, SROCC and total time effort. The same standard deviation of the observer model as in TID2013 (c=1.24) is assumed. Analyzing these results, we can conclude that mixing is possible even when dealing with disconnected components and missing rating data, showing similar performance to the sole use of pairwise comparisons at similar time cost. Being able to handle such experimental designs is a highly desirable feature, given that this can simplify the pairwise comparison experimental procedure for large-scale datasets or when mixing different quality assessment datasets, for which missing rating data is common.

VII. CONCLUSIONS

In this work we propose a probabilistic model that can bring the results of pairwise comparison and rating experiments into a unified quality scale. The model is based on the Thurstone Model V assumptions and our observation that the relation between DMOS ratings and scaled pairwise comparison quality scores is approximately linear. The units in that scale, which we denote as just-objectionable-differences (JODs), are scaled accordingly to the combined inter- and intra-observer variations so that 1 unit corresponds to 75% of observers selecting one condition over another. Our model can be used to estimate observer variation for each experimental protocol and bring measurements to the scale determined by the variation in a side-by-side pairwise comparison experiment. We use the pairwise comparison protocol as a base-line, as we found it to result in a lower standard deviation of the observer model and also lower RMSE given the same time effort to collect data.

We test our model on several real datasets and in a number of simulations. Tests have confirmed our assumption and further revealed interesting observations about the two experimental protocols. Given the same time effort there is no clear conclusion what experimental protocol to use. The decision should rely on the noise of both scales, measured by parameter c in our model. We also found that mixing both protocols can be beneficial i) to mix datasets that use either rating or pairwise comparisons, ii) to avoid disconnected components in pairwise comparison experiments, iii) if cross-content comparisons must be avoided and iv) if both types of measurements were previously collected.

APPENDIX
COLLECTED DATA FOR TID2013

A. MOS experiment

In order to obtain mean opinion scores, an experiment was conducted using the absolute category rating with hidden reference (ACR-HR) methodology [8]. In this experiment, a subset of color images from TID2013 color image dataset [3] were presented with a mid-grey background on a standard display in a dark room, following the ITU recommentations [9]. The participants were seated at the distance equal of 3 display heights (∼1m). The stimuli were shown for 5 seconds and the observers were allowed to confirm their answer either during or after displaying the stimulus. The participants were then asked to rate the quality of the color image presented on the display using a continuous scale ([0,100], 100 corresponding to the best quality). ACR-HR was selected to take also the reference images and some quality enhancements (e.g.

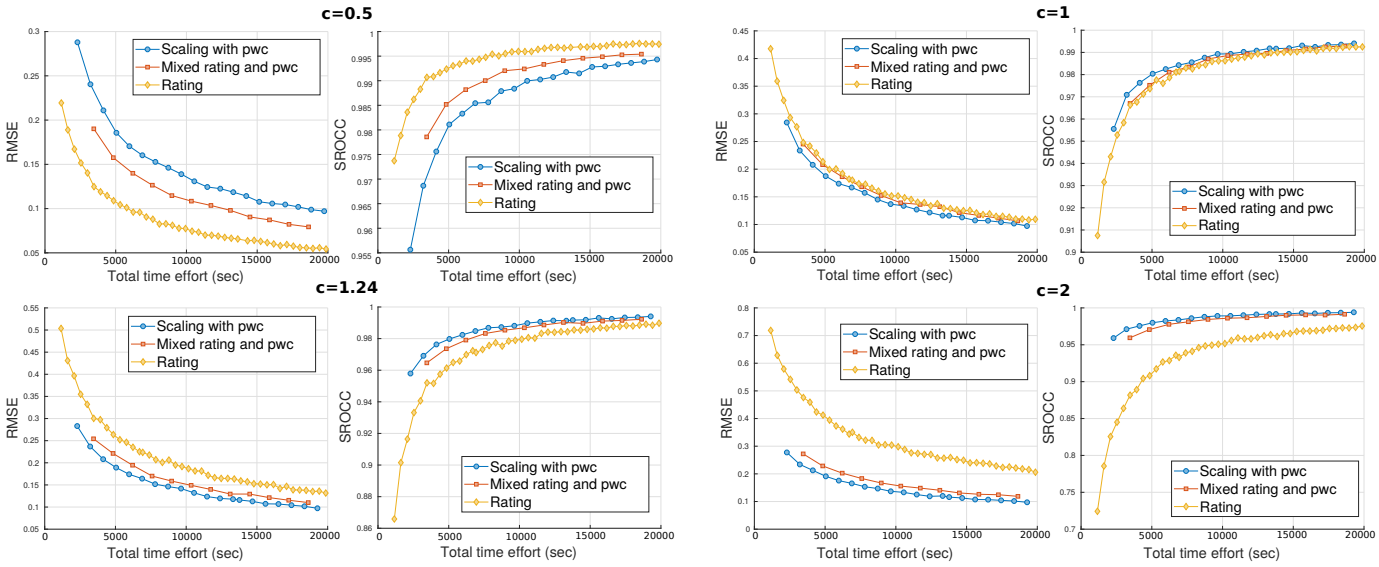

Fig. 13. Simulation of mixed scale for different values of standard deviation of the observer model (parameter c).

TABLE II
RESULTS FOR THE EXPERIMENT WITH DATA MISSING (DM) AND DISCONNECTED COMPONENTS (DC) FOR RMSE, SROCC AND TOTAL TIME EFFORT (IN SECS).

	Obs = 10			Obs = 20			Obs=30		
Type of measurement	RMSE	SROCC	Time effort	RMSE	SROCC	Time effort	RMSE	SROCC	Time effort
Rating	0.367	0.926	2310	0.277	0.958	4620	0.220	0.973	6930
Rating with DM	0.415	0.908	1848	0.311	0.947	3696	0.249	0.966	5544
PWC	0.200	0.978	4590	0.143	0.988	9180	0.116	0.991	13770
Mix with DM and DC	0.207	0.976	4677	0.151	0.987	9333	0.126	0.990	13956

increase in the contrast for 'contrast change' distortion type). The participants spent on average 3.9 ± 1.5 seconds on viewing an image and 3.8 ± 2.3 seconds on assigning a score.

In order to avoid fatigue and to keep the experiment under 30 minutes, a subset of images was used. Two distortion types were selected for each content through random permutation of the 24 different distortion types. A total of 175 images (25 contents $\times$ 2 distortion types $\times$ 3 distortion levels + 25 original images) were voted during the experiment. Looking at the quality values provided with TID2013 color dataset, we notice that some of the distortion types (e.g. non-eccentricity pattern noise and contrast change) have different behavior compared to the other compression types. In order to capture the uncommon behavior of these distortion methods, distortion levels of $\{2, 4, 5\}$ were used for non-eccentricity pattern noise and contrast change distortion type, as well as JPEG compression to have a more varying quality values. For the rest of the distortion types, distortion levels of $\{1, 3, 5\}$ were selected. To minimize context effects, the images were ordered randomly for each subject, and consequent images were selected from different contents.

Before the experiment, participants were screened for visual acuity and correct color vision using Snellen and Ishihara charts, respectively. A training session was conducted prior to the experiment to familiarize the subjects with the test procedure and distortion levels. Images used for training were not used in the experiment. Subjects were asked to rate "the overall quality of the presented image". In total, 22 people (4 female and 18 male) with the average age of 30.6 participated in the experiment. After outlier detection [29], 1 of the 22 subjects was removed. MOS, standard deviation, and confidence intervals are calculated for each stimulus as described in ITU-T Rec. P.1401 [29].

ACKNOWLEDGMENT

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725253–EyeCode), from EPSRC research grant EP/P007902/1 and from a Science Foundation Ireland (SFI) research grant under the Grant Number 15/RP/2776. María Pérez-Ortiz did part of this work while at the University of Cambridge and University College London (under MURI grant EPSRC 542892).

REFERENCES

[1] M. Pinson and S. Wolf, "Techniques for evaluating objective video quality models using overlapping subjective data sets," Tech. Rep., US Department of Commerce, National Telecommunications and Information Administration, 2008, NTIA Technical Report TR-09-457.

[2] E. Zerman, G. Valenzise, and F. Dufaux, "An extensive performance evaluation of full-reference HDR image quality metrics," *Quality and User Experience*, vol. 2, no. 5, 2017.

[3] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[4] E. Zerman, V. Hulusic, G. Valenzise, R. Mantiuk, and F. Dufaux, "The relation between MOS and pairwise comparisons and the importance of cross-content comparisons," in *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XXII*. International Society for Optics and Photonics, 2018.

[5] H. Sheikh, M. Sabir, and A. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[6] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," 2017.

[7] A. Mikhailiuk, M. Pérez-Ortiz, and R. K. Mantiuk, "Psychometric scaling of TID2013 dataset," in *10th International Conference on Quiality of Multimedia Experience (QoMEX)*, 2018.

[8] ITU-T, "Subjective video quality assessment methods for multimedia applications," ITU-T Recommendation P.910, Apr 2008.

[9] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R Recommendation BT.500-13, Jan 2012.

[10] EBU, "SAMVIQ - subjective assessment methodology for video quality," Tech. Rep., European Broadcasting Union, 2003, BPN 056.

[11] P. Dunn-Rankin, *Scaling methods*, L. Erlbaum, 1983.

[12] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Visual Communications and Image Processing (VCIP)*. International Society for Optics and Photonics, 2003, pp. 573–582.

[13] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *2nd International Workshop on Quality of Multimedia Experience (QoMEX)*. Jun 2010, IEEE.

[14] D. M. Rouse, R. Pépion, P. Le Callet, and S. S. Hemami, "Tradeoffs in subjective testing methods for image and video quality assessment," in *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XV*. International Society for Optics and Photonics, 2010, pp. 75270F–1–75270F–11.

[15] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Computer Graphics Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[16] A. B. Watson and L. Kreslake, "Measurement of visual impairment scales for digital video," *SPIE Electronic Imaging, Human Vision and Electronic Imaging VI*, vol. 4299, pp. 79–89, 2001.

[17] P. Ye and D. Doermann, "Active sampling for subjective image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4249–4256.

[18] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright, "Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence," *Journal of Machine Learning Research*, vol. 17, no. 58, pp. 1–47, 2016.

[19] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, and P. Le Callet, "Aligning subjective tests using a low cost common set," in *Euro ITV*, 2011.

[20] T. Pfeiffer, X. A. Gao, Y. Chen, A. Mao, and D. G. Rand, "Adaptive polling for information aggregation.," in *The 26th Conference on Artificial Intelligence (AAAI'12)*, 2012.

[21] L. Skorin-Kapov, M. Varela, T. Hoßfeld, and K.-T. Chen, "A survey of emerging concepts and challenges for QoE management of multimedia services," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2s, pp. 29:1–29:29, 2018.

[22] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, Dec 2015.

[23] L. L. Thurstone, "A law of comparative judgement," *Psychological Review*, vol. 34, no. 4, pp. 273–286, 1927.

[24] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons.," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[25] K. Tsukida and M. R. Gupta, "How to analyze paired comparison data," Tech. Rep. UWEETR-2011-0004, Department of Electrical Engineering University of Washington, 2011.

[26] P. G. Engeldrum, *Psychometric scaling: A toolkit for imaging systems development*, Imcotek Press, 2000.

[27] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, Jan 2016.

[28] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10K: Towards an ecologically valid and large-scale IQA database," 2018.

[29] ITU-T, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," ITU-T Recommendation P.1401, Jul 2012.