A practical guide and software for analysing pairwise comparison experiments

M. Pérez-Ortiz
Computer Laboratory
University of Cambridge
Cambridge, United Kingdom
mp867@cam.ac.uk

R. K. Mantiuk
Computer Laboratory
University of Cambridge
Cambridge, United Kingdom
rkm38@cam.ac.uk

Abstract

Most popular strategies to capture subjective judgments from humans involve the construction of a unidimensional relative measurement scale, representing order preferences or judgments about a set of objects or conditions. This information is generally captured by means of direct scoring, either in the form of a Likert or cardinal scale, or by comparative judgments in pairs or sets. In this sense, the use of pairwise comparisons is becoming increasingly popular because of the simplicity of this experimental procedure. However, this strategy requires non-trivial data analysis to aggregate the comparison ranks into a quality scale and analyse the results, in order to take full advantage of the collected data. This paper explains the process of translating pairwise comparison data into a measurement scale, discusses the benefits and limitations of such scaling methods and introduces a publicly available software in Matlab. We improve on existing scaling methods by introducing outlier analysis, providing methods for computing confidence intervals and statistical testing and introducing a prior, which reduces estimation error when the number of observers is low. Most of our examples focus on image quality assessment.

1   Introduction

One way to measure a perceptual attribute of interest, such as image quality, is to ask experiment participants to rank a set of conditions, for example images. The simplest type of such ranking are pairwise comparisons, where only two conditions are shown at a time and a participant is asked to choose one of them according to some specific criteria. For example, if we want to analyse which of three rendering methods (A, B and C) produces the highest quality results, we could present the images produced by these methods in pairs (AB, BC, AC) and then ask observers which image in each pair has better quality. If enough data is collected, we can then rank the algorithms from the best to the worst, estimate the confidence in such ranking, and scale the ranking scores so they can be easily interpreted in terms of probability of better perceived quality. A representation of this strategy can be seen in Figure 1. Unidimensional scaling methods attempt to represent preference judgments on a line, so as to effectively retain the distance information between the tested objects. This projection may reveal the underlying structure or unique relationships among the objects, allowing to measure and compare them in a meaningful way.

Pairwise comparison experiments are simple to run, but the data analysis step becomes more difficult. Often, data analysis is limited to statistical testing: showing that observed differences are unlikely to be produced by chance. Although this is an important stage of data analysis, it is often insufficient, as statistical significance may not translate into practical significance. The scaling methods presented in
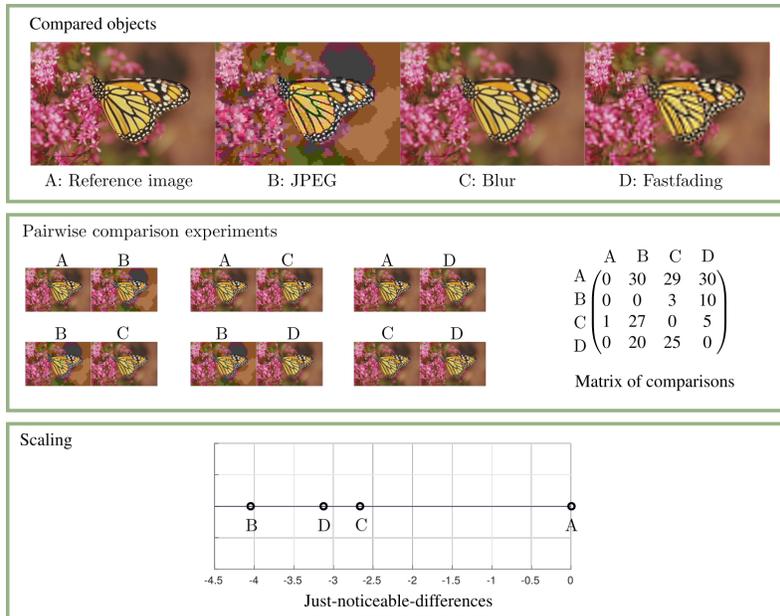
Figure 1: Illustration of scaling pairwise comparison data for evaluating the perceived image quality. 4 conditions (distortion types) are compared in this case, resulting in $\binom{5}{2} = 5$ different comparisons, each comparison repeated 30 times. Scaling algorithms produce a quality scale from the matrix of comparisons, in which distances between conditions can be interpreted as probability of better perceived quality.

this paper can express the results in terms of practical difference: they translate raw comparison data into quality scores that show the magnitude of the difference between tested conditions.

There is a vast amount of literature on scaling or aggregating comparative judgments through pairwise comparison experiments, dating as early as 1927 (Thurstone, 1927; Davidson and Farquhar, 1976). However, studying this literature could be a daunting and time consuming task, which requires a strong background in statistics to understand all the intricacies of these methods. Scaling methods are usually not straightforward to implement. They require a number of precautions to ensure robust results and that errors are not introduced due to insufficient floating point precision and other non-obvious reasons. One of the purposes of this paper is to provide a comprehensive description of how scaling methods work, and accompany this with an open source Matlab toolbox for performing the scaling and statistical analysis.

The scaling method described here has been used in several previous computer graphic projects of our and other groups, including (Karaduzovic-Hadziabdic et al., 2016; Eilertsen et al., 2015; Vangorp et al., 2014; Wanat and Mantiuk, 2014). However, due to space restrictions, we could not explain in those papers all the details and improvements. This paper is meant to serve as a reference for any future work relying on our scaling method.

The contributions of this work are the following: (i) a collection of methods for the analysis of pairwise comparison data, which include outlier analysis, estimation of confidence intervals and statistical testing; (ii) a prior, which improves scaling accuracy when the number of observers is low; (iii) analysis of practical issues concerning the experimental design, such as the use of ties or incomplete designs; and (iv) a Matlab toolbox to perform the analysis.

1.1 Direct rating vs. pairwise comparisons

Direct rating, in which observers assign a score to each condition, may seem to be a simpler and more direct measurement of perceptual attributes (e.g. image quality or taste) than pairwise comparisons. However, direct rating methods have a number of limitations. They require careful training so that participants know what value should be assigned to which condition — to establish a well defined scale for a given experiment. However, even after careful training, such scale can vary substantially

between participants, or even within a single participant when the experiment is repeated on different days. Direct rating experiments are particularly difficult to conduct when compared conditions are substantially different from each other. For example, the popular LIVE image quality dataset (Sheikh et al., 2006) was collected in 7 different experimental sessions, where each session involved only one type of distortion (e.g. JPEG compression, noise, blur, etc.). Isolating each distortion type simplified the experimental task, but it made the quality scale obtained in each session different from one another. To align all scales, the authors had to perform 8 realignment experiments, in which a subset of images from the 7 experimental sessions was assessed again and the collected scores were used to linearly re-scale the previously collected scores. This rather complex procedure demonstrates the challenges of obtaining a unified quality scale in rating experiments.

As opposed to this, the use of pairwise comparison present numerous advantages: i) it leads to a very simple experimental task and is therefore well suited for non-expert participants, ii) it avoids calibration issues frequently encountered in cardinal measurements (Tsukida and Gupta, 2011), iii) it generally provides higher sensitivity and a lower measurement error when compared to direct rating (Shah et al., 2015), and iv) it can be faster to run than direct scaling (particularly since making pairwise comparisons is easier and faster for participants (Stewart et al., 2005) and because the number of comparisons can be reduced using adaptive procedures (Mantiuk et al., 2012; Ye and Doermann, 2014; Xu et al., 2011)).

1.2 Vote counts vs. scaling

The simplest way to report the result of a pairwise comparison experiment is to compute vote counts — the number of times one condition was selected as better than any other condition. Vote counts, however, present the results on an ordinal scale, which would usually produce the correct ranking of the conditions, but it does not correctly capture the magnitude of the differences between conditions. On the other hand, pairwise comparison scaling places those conditions on a continuous interval scale, which captures both the order of conditions and the magnitude of the difference. Zerman et al. compared the results of pairwise comparison scaling and vote counts to the scores obtained in a direct rating experiment (Zerman et al., 2018). They showed that scaled data is more strongly related to rating scores than vote counts, confirming that quality magnitudes are better captured when pairwise comparison data is scaled. Furthermore, vote counting is difficult when not all conditions are compared with each other (incomplete design) or when not all observers compare the same conditions (unbalanced design). Scaling methods can robustly cope with such non-standard experiment designs.

2 Related work

The bibliography in papers that review how to aggregate pairwise comparison data testifies the widespread interest of the scientific community on this type of methods: more than 350 papers in (Davidson and Farquhar, 1976) and more than 100 in (Cattelan, 2012). There is a wide range of applications in which this approach has shown to be successful, e.g. to study consumer preference, in sport rankings, econometrics or perceived image/video quality. For a detailed discussion on the topic the reader should refer to one of the aforementioned review papers or to the monograph of David (David, 1963). An accessible introduction to the topic of scaling can be found in (Tsukida and Gupta, 2011) and (Dunn-Rankin et al., 2004).

The scaling procedure depends on the selection of the model relating observers' answers to the abstract quality scale. Two of the most common models are that of Thurstone (Thurstone, 1927) (considered in this work) and of Bradley and Terry (Bradley and Terry, 1952). The differences between the two models are minor (Tsukida and Gupta, 2011) and the choice is a matter of preference. Many extensions of the two models can be found in the literature. For example, some models give the observers an additional option of choosing tie (no preference) (Davidson, 1970) or let them express strong, mild, or no preference judgment for a pair conditions (Agresti, 1992). While the answers are typically considered to be independent, some literature focuses on learning from dependent data (Cattelan, 2012), where either condition and observer covariates are accounted for. Condition dependencies stem from the fact that generally the same condition is involved in multiple paired comparisons. Modeling observer covariates assumes that the comparisons made by the same person are dependent. Allowing ties or accounting for covariates is not free from shortcomings. Those more complex models usually require more data as more parameters need to be estimated. Other type of

models introduce a temporal component (Herbrich et al., 2006), e.g. for ranking tournament data, where players take part in different matches during a prolonged period of time and most recently played matches need to have more influence on the ranking to account for changes in the skills of different players.

It is also worth noting that pairwise comparison experiments are not only used for measuring stimuli on interval scales, as presented in this paper, but can also be used to discover unknown perceptual attributes. This is, they can be used to discover explanatory variables that affect the results of the comparisons (Springall, 1973). This can be performed by standard statistical approaches (e.g. regression analysis), multi-dimensional scaling (Pellacini et al., 2000) or using more advanced ranking machine learning methods (Wauthier et al., 2013). Paired comparisons are usually expected to be consistent, which may not hold in practice. In some cases, preferences can be naturally intransitive (i.e. A>B, B>C but C>A), which usually originates from the fact that the conditions have more than one aspect of interest, and different aspects prevail in different comparisons. As said, some approaches account for this (Causeur and Husson, 2005; Usami, 2010), and project the data to more than one dimension by multi-dimensional scaling. This approach might simplify the task of scaling but makes more difficult the interpretation of the final solution.

The pairwise comparison scaling method discussed in this paper is only suitable when the quality differences between compared conditions are small so that the observers vary in their answers. When a perceptual attribute must be scaled over a larger range, the difference scaling method (Maloney and Yang, 2003) could be more appropriate. In this method observers are asked to judge the magnitude of a difference for two pairs of stimuli and select the pair of higher difference.

Most of the software facilitated to work with paired comparison data is implemented in R, where apart from most traditional models, one can also find more specific techniques. The *eba* package (Wickelmaier and Schmid, 2004) adapts one of the most popular models (the Bradley-Terry model) to consider that different conditions might present various aspects that account for their worth (referred to as elimination-by-aspects models) and includes different functions to check the consistency of the answers (i.e. violations of transitivity). The *prefmod* package (Hatzinger and Dittrich, 2012) implements also different versions of the Bradley-Terry model. No preference options (ties) can be included and specifically modelled, as well as incomplete designs. The *Bradley-Terry2* package (Turner and Firth, 2012) includes different probability functions and models data covariates. This package also allows the use of tournament data. The package *choix* in python presents inference algorithms based on an extension of Bradley-Terry model (Placket, 1975), which allow to explain and model comparisons between items, not only in a pairwise manner, but also setwise (Maystre and Grossglauser, 2015). Finally, the tutorial in (Tsukida and Gupta, 2011) also includes some basic scaling code in Matlab. Although the mentioned software serves a similar purpose as our proposed method, none of the packages offers a complete set of methods for analysis, including outlier analysis, the computation of confidence intervals and statistical testing.

This work is inspired by the previously mentioned papers, however, our focus is on more practical issues of scaling, such as experimental design, statistical analysis and low sample scenarios, providing guidance for the end-user of this type of methods and software for performing the scaling and analysis of results.

3 Example of pairwise comparison data analysis

We start by presening an example[1] of a typical pairwise comparison data analysis session using our software[2], in which we analyse the data from the video tone mapping evaluation project presented in (Eilertsen et al., 2013).

We recommend to keep the data in a tabulated format, such as comma-separated-files (CSV), in which each condition is described by meaningful labels. Such files are easy to read with any software and can be easily interpreted even long after the data have been collected. Table 1 shows a few rows from the analysed dataset.

---

[1]The code for the example can be found in the *examples* folder, under the name of *video_TMO_analysis_example*.

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

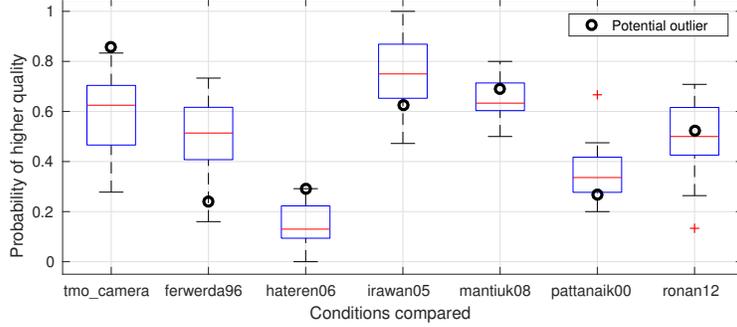| Observer	Session	Scene	Condition_1	Condition_2	Selection
1	1	Window	TMO_Camera	Ferwerda96	1
1	1	Exhibition	Ronan12	Irawan05	2
1	1	Corridor	Irawan05	Ferwerda96	1
2	2	Corridor	Ronan12	TMO_Camera	2

Table 1: Formatting example recommended to store pairwise comparison data.

Figure 2: Distribution of the general perceived quality for each condition. This plot should be interpreted with care and never be used as a replacement of the scaling. In this case, the probabilities plotted here and the scaling in Figure 3 show similar behaviour because the dataset uses full design, but they would be very different for incomplete designs.

The first step is to convert the answers from the table into a set of comparison matrices M, one matrix per each observer. In such a matrix, columns and rows correspond to compared conditions and matrix value $c_{ij} = n$ means that condition O_i was n times selected as better than condition O_j. If there is a reference condition, such as a non-distorted image, it should be put in the matrix as the first condition in the first row and column. The first condition will be assigned a fixed quality value of 0.

The second step is to perform outlier analysis to detect potential observers who performed very differently from the rest. The function to perform this analysis is *[L,L_dist]=pw_outlier_analysis(M)*, which receives a matrix M with the responses per observer and returns the likelihood L of observing the data of each observer and a inter-quartile-normalised score L_dist, which indicates the observers that should be further investigated. Since there is no objective threshold that could distinguish outliers with high confidence, we advise to investigate all observers whose L_dist score is close or above the customary threshold of 1.5. The results for the 18 observers in the analysed dataset indicate that there is one observer with a score of 2.72, which requires further attention. To compare the answers of the indicated observer (observer number n_obs) to the rest of observers, we use the function *compare_probs_observer(M,n_obs)*, which plots the probabilities of selecting one condition over all others, shown in Figure 2. Note that this presentation of the data does not involve scaling, which could obscure the patterns that are specific to an outlier. The black circles in the plot represent the answers of the potential outlier. The plot indicates that the potential outlier had a different opinion about operators *Ferwerda96*, *Hateren06* and *TMO_Camera*, but the patterns of his answers were not much different from the rest of observers. Therefore, although the observer was not fully consistent with the rest of observers, we could not justify removing her/his answers from the dataset. We recommend performing such detailed per-observer analysis, rather than using arbitrary measure to exclude observers. The details of the outlier analysis can be found in Section 9.

Once we are confident there are no outliers in the dataset, we can scale the results and compute confidence intervals using *[jod, stats]=pw_scale_bootstrp(M)* function. The function expects the same matrix of comparison per observer M as the outlier analysis and returns the scaling solution and a set of statistics. The scaling and the confidence intervals have been plotted for our dataset in Figure 3. Confidence intervals represent the range in which the estimated quality values lie with 95% confidence. The confidence intervals, however, should not be used to infer statistical significance of the difference. The statistical tests are performed by the function *pw_plot_ranking_triangles(jod,stats)*, which produces a plot shown in Figure 4. The continuous lines in that plot indicate statistically significant difference between the pair of conditions and the dashed lines indicate the lack of evidence for
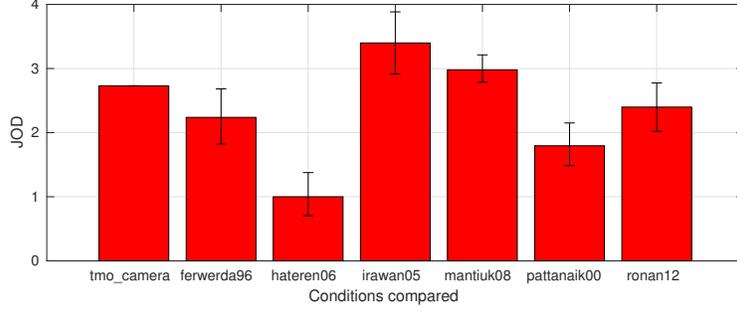
Figure 3: Visualization of the scaling results and confidence intervals for the chosen dataset. Note that there is no confidence interval for the first condition, as this is always set up to a fixed value (since scores are relative). The difference of 1 JOD indicates that 75% of observers selected one condition as better than the other.
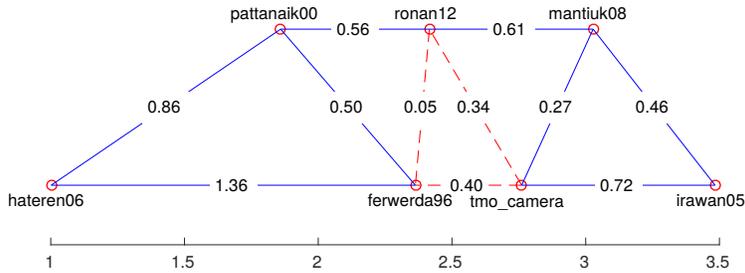

Figure 4: Graphical representation of the scaling. Red points represent conditions, and they are only connected to their neighbours, as these are usually the comparisons in which we are most interested. Blue solid lines represent statistically significant differences, as opposed to red dashed lines. The x-axis shows the scaling.

statistically significant difference. More information on the scaling and statistical analysis can be found in Sections 5 and 7, respectively.

4   Designing pairwise comparison experiments

Planning for pairwise comparison experiments requires taking into account several considerations to ensure that sufficient data is collected with possibly small experimental effort. The number of required comparisons depends on the number of compared conditions n (e.g. different algorithms or distortion levels), the number of different pieces of content k (e.g. images or video clips) and the number of repetitions t of the experiment. If each observer is asked to compare each condition with the rest, they would need to perform $\frac{1}{2} \cdot n \cdot (n-1) \cdot k \cdot t$ comparisons. This number grows quickly, especially for large n.

An important issue is the choice of compared conditions, since not all comparisons are equally useful. The comparisons that produce obvious results, e.g. comparing the highest and lowest distortion levels, do not contribute much to the outcome of the experiment and can be obviated. The experiments, in which only selected pairs are compared, are referred to as incomplete design, as opposed to a complete design, in which every pair is compared. If not all observers compare the same set of conditions, but instead, every observer has a different experimental design, the experiment is said to have an imbalanced design. Note that this imbalanced design is generally non-advisable (Cattelan, 2012), at least when not accounting for observer covariates. If we do not have any apriori information about the potential ordering of our conditions, we could use an efficient sorting algorithm (e.g. quicksort) (Maystre and Grossglauser, 2017) or other specifically designed techniques, such as active sampling (Ye and Doermann, 2014; Jamieson and Nowak, 2011). This results in less variance given the same number of trials (Silverstein and Farrell, 2001; Shah et al., 2015). In many cases, however, we know in advance the most likely order of the conditions, e.g. in image compression we know that lower bit-rate images will have worse quality than those of higher bit-rate. In such cases, we can restrict
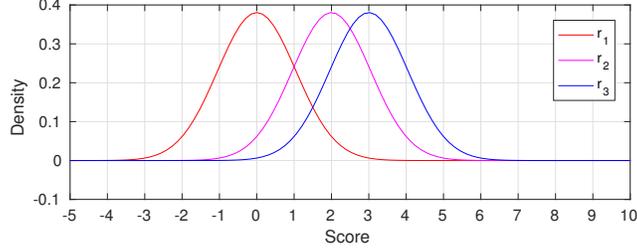
Figure 5: The probability of assigning a score to three conditions. Quality scores are assumed to be random variables.

comparisons to neighbours in the scale of distortion level. It is important to ensure that the quality levels of compared images are relatively similar, so that they are confused in certain number of cases. If all observers give the same response, we will not be able to reliably estimate the scaled difference between them. This is further discussed in Sections 8 and 10.1.

Finally, it is possible to offer a third answer in the experiment (i.e. ties). This, however, usually makes modelling more difficult. We discuss this problem in more detail in Section 10.3. Our general recommendation is to run two-alternative-force-choice experiments without ties.

Discussion of other important factors related to experimental design, such as control of the viewing conditions, reducing learning effects, training, experimental fatigue, are out of the scope of this report. Readers can refer to (Engeldrum, 2000) or psychophysics textbooks, such as (Kingdom and Prins, 2016) or (Lu and Dosher, 2013).

5 Problem formulation

Suppose we aim to compare n conditions $O_1, \ldots, O_n$ (e.g. n images, generally with the same content, each processed with a different algorithm) with unknown underlying true quality scores $q = (q_1, \ldots, q_n), q_i \in \mathbb{R}$. The aim of this analysis is to estimate scores $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_n)$ that approximate the true quality scores q. This can be obtained from the pairwise comparisons collected from m observers in t trials (and possibly k pieces of content, each processed separately). Because the pairwise comparisons are relative, we also assume that $q_1 = \hat{q}_1 = 0$.

5.1 Comparison matrix

A pairwise comparison experiment is usually represented in a count matrix $\mathbf{C}$, where each element c_{ij} measures the number of cases in which condition O_i has been selected as better than condition O_j (considering m observers and t trials). For example, in an experiment with three conditions, the resulting matrix could be as follows:

$$\mathbf{C} = \begin{bmatrix} 0 & 3 & 0 \\ 27 & 0 & 7 \\ 30 & 23 & 0 \end{bmatrix} \tag{1}$$

This is, $c_{12} = 3$ tells us that condition O_1 has been selected three times as being better than condition O_2, and $c_{21} = 27$ tells us that condition O_2 has been selected 27 times as better than O_1. The probability that one condition is selected as better than another (denoted as p_{ij} for O_i and O_j) can be estimated using the empirical information in matrix $\mathbf{C}$ (Tsukida and Gupta, 2011):

$$\hat{p}_{ij} = \frac{c_{ij}}{c_{ij} + c_{ji}}, \quad i \neq j \tag{2}$$

e.g. the probability that O_2 is selected as better than O_1 can be estimated as $\frac{c_{21}}{c_{21}+c_{12}} = \frac{27}{27+3} = 0.9$.

5.2 Observer model

In this paper we use the model proposed by L. L. Thurstone (Thurstone, 1927; Engeldrum, 2000). This model assumes that observers make quality judgements by assigning a single quality value to
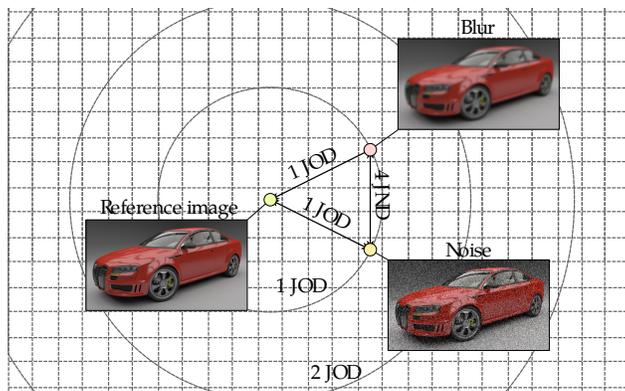
Figure 6: Illustration of the difference between just-objectionable-differences (JODs) and just-noticeable-differences (JNDs). The images affected by blur and noise may appear to be similarly degraded in comparison to the reference image (the same JOD), but they are noticeably different and therefore 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.

each condition and that the condition's quality is a random variable, so as to account for the subjective nature of these experiments. This is, the perceived quality of a condition O_i is modeled as a random variable: $r_i \sim N(q_i, \sigma)$ (i.e. the mean of the distribution is assumed to be the true quality score q_i). This is illustrated on an example of three conditions in Figure 5. Observers 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). Thurstone Case V model assumes that both inter- and intra-observer variance can be explained by a Normal distribution, and that the variance of that distribution is the same for each condition (the noise parameter σ is the same for all items and accounts for the uncertainty in the comparisons). The goal of the pairwise comparison experiment is to find the expected values $\hat{q}$ of the distribution of the scores for each condition. In practice, since scores are relative, we are interested in recovering the distances among them.

5.3 JNDs and JODs

The results of paired comparisons are typically scaled in Just-Noticeable-Difference (JND) units (Engeldrum, 2000; Silverstein and Farrell, 2001). Two stimuli are 1 JND apart if 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 the two distorted images shown in Figure 6: one image is distorted by noise, the other 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 visual 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 the measure of JOD is more similar to visual equivalence (Ramanarayanan et al., 2007) or to the quality expressed as a difference-mean-opinion-score rather than to JNDs.

6 Scaling methods

Pairwise comparisons can be viewed as noisy samples of the underlying quality difference between two conditions. The goal of scaling is to estimate these latent differences based on noisy data in the form of pairwise comparisons. Given the observer model, we can use one of the following methods to transform collected probabilities $\hat{p}_{ij}$ into scaled quality scores $\hat{q}$.
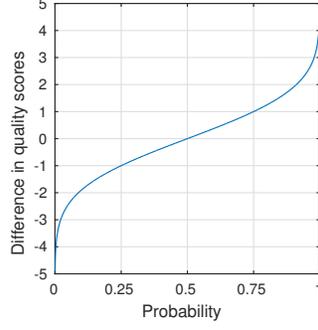
Figure 7: Inverse cummulative Normal distribution mapping probabilities into distances on the JOD scale. When the observers make a random guess ($p_{ij} = 0.5$), the JOD distance between conditions i and j is 0. When the observers select the first condition as better 75% of the times ($p_{ij} = 0.75$), the JOD distance is 1. The distance is negative when the observers select the second condition more often as being better.

6.1 From probabilities to distances

When scaling data, we are mostly interested in recovering the distance $q_i - q_j$ between underlying quality scores q_i and q_j (since scores are relative). This distance is linked to the probability of condition O_i having a higher quality than condition O_j. Note that the difference of two Gaussians r_i and r_j is also a Gaussian random variable:

$$r_i - r_j \sim N(q_{ij}, \sigma_{ij}), \tag{3}$$

where $q_{ij} = q_i - q_j$ and $\sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2 = 2\sigma^2$.

The probability of choosing O_i over O_j can be computed using the cumulative Normal distribution Φ over the difference $r_i - r_j$:

$$P(r_i > r_j) \quad = P(r_i - r_j > 0) = \Phi\left(\frac{q_i - q_j}{\sigma_{ij}}\right) \tag{4}$$

$$= \frac{1}{\sigma_{ij}\sqrt{2\pi}} \int_{-\infty}^{q_i - q_j} e^{\left(\frac{-x^2}{2\sigma_{ij}^2}\right)} dx.$$

The mapping from probabilities into score differences is given by the inverse of Φ (know as the probit and shown in Figure 7):

$$q_i - q_j = \sigma_{ij}\Phi^{-1}(P(r_i > r_j)). \tag{5}$$

Thurstone's model assumes that the noise parameter σ is known and constant for all conditions, so that $\sigma_{ij} = \sigma\sqrt{2}$. However, we do not know its value. A common approach is to select σ_{ij} so that a probability of 0.75, in the midway between a random guess and being completely certain, is mapped to a score distance of 1 JOD unit. The difference of 2 JODs corresponds to the probability of 0.91 and so on. The inverse cumulative distribution crosses the value of 1 for $p_{ij} = 0.75$ when the standard deviation σ_{ij} is 1.4826.

6.2 Least-square distance solution

Once that we have established the relation between probabilities and score differences, we can substitute $P(r_i > r_j)$ by the estimate $\hat{p}_{ij}$ in Eq. (2) to obtain an estimate of the distance:

$$d_{ij} = \sigma_{ij}\Phi^{-1}(\hat{p}_{ij}). \tag{6}$$

When these probabilities are transformed into score differences, we obtain the following distance matrix:

$$\mathbf{D} = \begin{bmatrix} 0 & -2.7190 & -Inf \\ 2.7190 & 0 & -1.0792 \\ Inf & 1.0792 & 0 \end{bmatrix} \tag{7}$$
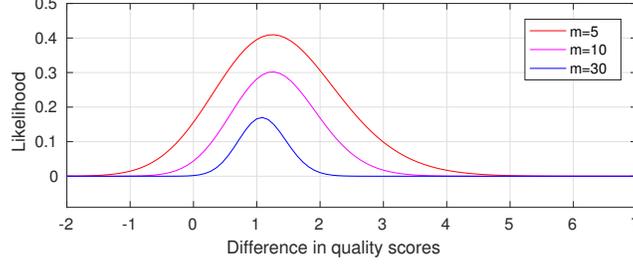
9

Figure 8: The likelihood function of JOD distance for three sample sizes and when $p_{ij} = 0.75$. Note that the range of likely distance values gets smaller with the number of samples.

Our aim is to find an estimation $\hat{q}$ such that the distances between the different scores closely resemble the distances in matrix $\mathbf{D}$. Such quality scores are often found by solving an optimisation problem of the form (Engeldrum, 2000):

$$\underset{\hat{q}_2, \ldots, \hat{q}_n}{\arg\min} \sum_{i=1}^{n-1} \sum_{j=k}^{n} ((\hat{q}_i - \hat{q}_j) - d_{ij})^2. \tag{8}$$

This formulation is similar to the problem of multi-dimensional-scaling when we scale to a single dimension, except that our distances are signed. Since it is not possible to optimise the absolute score values given only distances between them, one of the scores is usually fixed (most commonly $\hat{q}_1 = 0$).

Unfortunately, the solution of Eq. (8) is unfeasible in our example because of the infinite values in $\mathbf{D}$. The two infinite values correspond to the cases when all observers gave the same (unanimous) answer and the probability is equal to 0 or 1. As the inverse cumulative Normal distribution reaches one of its asymptotes at 0 and 1, the corresponding distances in scores are infinite. The distance of plus or minus infinity is definitely an incorrect estimate, but it is also impossible to tell exactly what the true distance should be, given the data. Having unanimous answers is common in experiments, so it is highly important to devise a method to deal with those cases. Sometimes unanimous answers are ignored, but this removes valid observations from the data. In other cases the range of distances is restricted, for example to be between -3 and 3, but this introduces a bias in the estimate. In the next section we present an optimisation method more suitable for these cases.

6.3 Maximum likelihood estimation

A more elegant and robust solution for scaling is offered by Maximum Likelihood Estimation (MLE). Instead of minimising stress in distances in Eq. (8), MLE looks for the difference in quality scores that maximise the probability of observing our data $\mathbf{C}$. To do so, we need to connect the quality differences with the data collected in the comparison matrix $\mathbf{C}$. If we know the true probability of selecting O_i as better than O_j ($P(r_i > r_j)$), the probability that O_i was selected over O_j in exactly c_{ij} trials from the total number of $n_{ij} = n_{ji} = c_{ij} + c_{ji}$ trials is given by the binomial distribution:

$$\begin{aligned}
L(\hat{q}_i - \hat{q}_j | c_{ij}, n_{ij}) &= \binom{n_{ij}}{c_{ij}} P(r_i > r_j)^{c_{ij}} (1 - P(r_i > r_j))^{n_{ij} - c_{ij}} \tag{9} \\
&= \binom{n_{ij}}{c_{ij}} \Phi\left(\frac{\hat{q}_i - \hat{q}_j}{\sigma_{ij}}\right)^{c_{ij}} \left(1 - \Phi\left(\frac{\hat{q}_i - \hat{q}_j}{\sigma_{ij}}\right)\right)^{n_{ij} - c_{ij}}
\end{aligned}$$

Note that, as shown earlier in Eq. (4), the probability $P(r_i > r_j)$ depends on the difference in quality scores and is given by the cumulative Normal distribution Φ.

To scale all compared conditions, we maximise the product of the likelihood for all pairs of conditions:

$$\underset{\hat{q}_2, \ldots, \hat{q}_n}{\arg\max} \prod_{i,j \in \Omega} L(\hat{q}_i - \hat{q}_j | c_{ij}, n_{ij}) \tag{10}$$

where Ω is the set of all pairs for which at least one comparison has been made: $c_{ij} + c_{ji} > 0$. Note that, in practice, it is more convenient to maximise the log of the likelihood function.

Solving MLE in Eq. (10) has a number of advantages over the least square distance solution in Eq.(8):
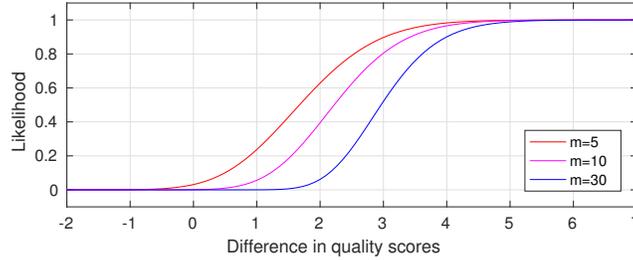
10

Figure 9: The likelihood function of JOD distance for three sample sizes and when one condition is always selected as better ($p_{ij} = 1$).

- MLE accounts for the number of comparisons and thus the measure of confidence we have in our data. Figure 8 shows the likelihood for three sample sizes for $p_{ij} = 0.75$. The larger the sample, the narrower is the range of likely differences between the scores. This property of the MLE solution is in particular useful when the experimental design is not balanced.

- MLE solution (almost) gracefully handles the cases with unanimous answers. Figure 9 plots the likelihood L as the function of difference in quality scores, for the case when p_{ij} is equal to 1 and the number of observers m is 5, 10 and 30. In each case, the most likely distance is greater than 5, but there is also a likelihood of a smaller distance, especially when m is small.

- MLE allows us to work with incomplete experimental designs, when only a subset of pairs is compared.

7 Statistical analysis

Since any experiment gives only estimates of the true quality values, it is important to analyse and report the level of uncertainty in the data. In this section we show how to compute confidence intervals and test for statistical differences.

7.1 Confidence intervals

Computing confidence intervals for scaled quality scores using analytical methods is difficult because multiple conditions influence each other. The original formulation of Thurstone Case V does not allow the computation of confidence intervals. Different authors have change the base model to account for this (Montag, 2003), but this is at the cost of the simplicity of the model. However, confidence intervals can be computed using numerical methods, e.g. resampling (see for example ch. 18.1 in (Howell, 2009)). Resampling is generally used as a statistical method for estimating the sampling distribution. It represents a robust alternative to inference based on parametric assumptions when those assumptions are in doubt. A common example is the use of the bootstrapping technique. This method always resamples from the sample, therefore relying on the generation of pseudo-samples from the sample collected. Given a measured sample (result of a pairwise comparison experiment), we generate a new sample of the same size by randomly replicating data for some participants and removing data for others. The procedure is know as random sampling with replacement. To compute confidence intervals, a large number of pseudo-samples in generated (usually more than 500), then each sample is scaled using the MLE method from Section 6.3, and finally the 2.5-th and 97.5-th percentiles of JOD values are computed for each condition across all samples. This gives the 95% confidence intervals for the mean JOD scores.

Figure 10 shows three examples of confidence intervals computed for simulated experiments. Assuming a set of fixed true scores, we can simulate the randomness of observers' judgments by drawing simulated answers from distributions, such as those shown in Figure 5. In our examples, ten virtual observers (n=10) performed three repetitions (t=3) of the experiment, in which all pairs were compared. There are a few conclusions that can be drawn from the plot:
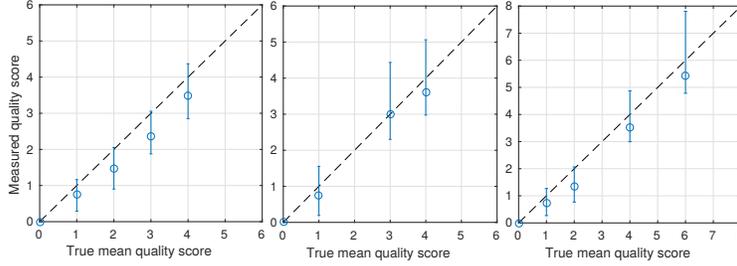
Figure 10: Results of the simulated experiments, in which the pairwise comparison results have been generated by randomising scores using probability distributions similar to those shown in Figure 5. The known means of those distributions are shown on the x-axis and the results of JOD-scaling on the y-axis. The error bars denote 95% confidence intervals computed by bootstrapping.
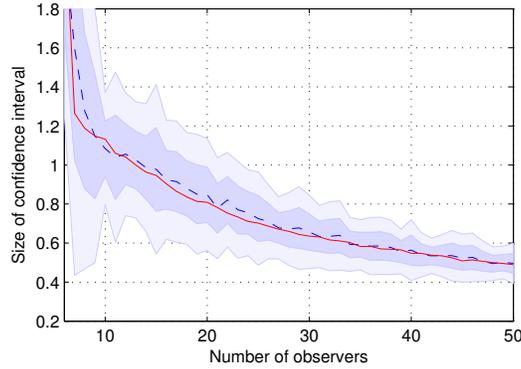

Figure 11: The confidence interval computed from experiment data via bootstrapping compared with the actual confidence interval found in a Monte-Carlo simulation (10,000 simulated experiments). The bootstrapping on average (dashed-blue line) follows very closely the true confidence interval (red continous line). The darker and brighter blue-shaded area represent the standard deviation and 95% interval for 50 experiment runs for which bootstrapping was computed.

- Confidence intervals are larger for quality scores that are farther from the reference point 0. Since the absolute scores are estimated from distances between the pairs, the estimation error between the first and second condition is propagated to the third condition, and so on.

- Confidence intervals become larger as the distance between conditions increases. Intuitively, Figure 7 shows that larger distances are projected onto smaller differences in probability. Thus, when a JOD distance is large, a small error in the estimation of probabilities can cause a large error in estimated distance.

To analyse how accurate bootstrapping is for estimating confidence intervals in our problem we analyse its performance in a simulation, where true confidence intervals can be estimated with high precision. We assume we know that the true quality scores are $q = (0, 1, 2, 3, 4)$. Then, we simulate 10,000 runs of an experiment by randomising answers of a certain number of observers (adding random Gaussian noise $N(0, 1.4826)$ to q), generating corresponding comparison matrices and running our scaling method. We compute the mean size of the confidence interval (mean of the distance between 97.5th percentile and the mean; and the distance between the mean and 2.5th percentile) and plot it for experiments with different number of observers in Figure 11 (red continuous line). Then, we use the same procedure to simulate 50 experiments (for each number of observers) for which we run bootstrapping and compute the mean size of the confidence interval in the same way. The distribution of bootstrapping results is shown as the blue-shaded areas and the blue dashed line in Figure 11. It can be seen that on average bootstrapping gives us a correct estimate. However, we need to keep in mind that bootstrapping is just an estimate, and the computed interval can be easily both under- and over-estimated, especially when the number of observers is small. Therefore, we need to have limited confidence even in the confidence intervals.
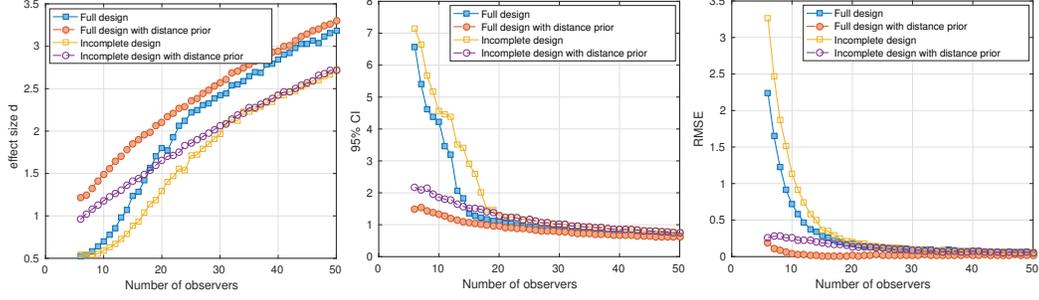
12

Figure 12: The difference in precision with and without the distance prior. Each plot represents a different measure (see text for details).
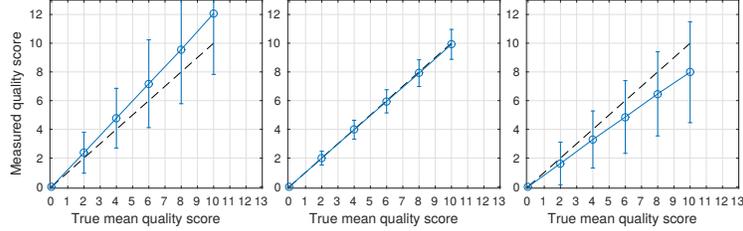

Figure 13: Averaged JOD values from running 1000 simulated (independent) experiments. Left: when the original MLE formulation from Eq. (10) is used, the mean shows bias towards higher JOD differences. Middle: when the prior is added in Eq. (14), bias is reduced. Right: when all comparison with unanimous answers ($p_{ij} = 0$ or $p_{ij} = 1$) are removed, the bias is tipped to the other side, i.e. JOD distances are under-estimated.

7.2 Statistical difference between two conditions

The analysis of confidence intervals for pairwise comparison data is more complicated than for a typical direct rating experiment because the computed JOD values are not independent. Since all conditions are "linked" to each other by pairwise comparisons, changing the value of one condition will "push" the values of all directly or indirectly linked conditions. This correlation between conditions can be captured in a covariance matrix Σ, such as one shown below:

$$\Sigma = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.431 & 0.486 \\ 0 & 0.486 & 0.683 \end{bmatrix} \tag{11}$$

The first row and column have 0s because O_1 is always fixed at 0 and cannot vary. Values $\Sigma_{22} = 0.431$ and $\Sigma_{33} = 0.683$ represent variance for O_2 and O_3. The value $\Sigma_{23} = \Sigma_{32} = 0.486$ represents the variance between a pair of conditions. If we want to reject H_0 that the difference in JOD scores between two conditions is 0, we need to compute the variance for that difference as:

$$v_{ij} = \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij} . \tag{12}$$

Using the variance and the difference in JOD scores, a two-tailed test can be used to test H_0 (David, 1963) for a given level of confidence.

8 Finite distance prior

Unanimous answers are problematic for scaling methods as they put no upper bound on the distance between two conditions and thus introduce a bias in the estimation. The problem is most noticeable when the sample size (number of observers) is small. This is because i) the probability of having unanimous answers increases with few observers, and ii) the smaller the sample, the wider is the range of likely differences between the scores (see Figure 8). However, the scaling can be made more robust by adding a simple distance prior to the likelihood function.
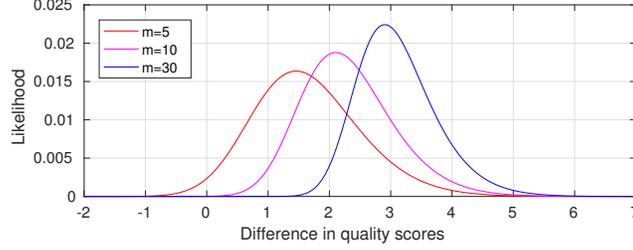
Figure 14: The likelihood function modulated by a distance prior (posterior) for the case when all answers are unanimous.
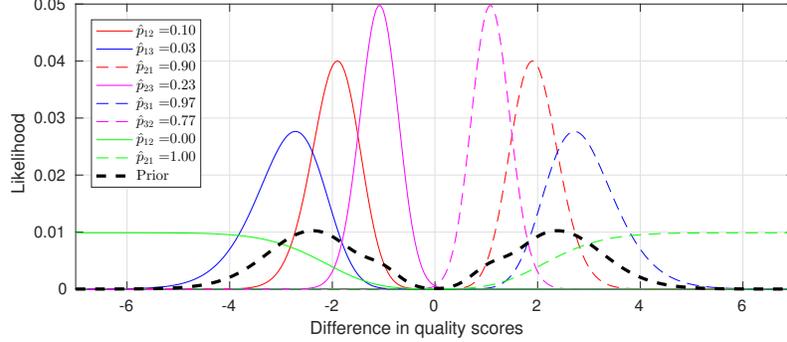

Figure 15: The proposed prior (dashed-line) as the normalised sum of probabilities of observing a difference for all compared pairs of conditions. The distribution are computed for our toy-example from Eq. (1).

This problem is easier to see on the example shown in Figure 13, where 1000 runs of an experiment are simulated for the true scores of $q = (0, 2, 4, 6, 8, 10)$. If there was no bias in the method, the experiments should, on average, give the correct answer, exactly on the black-dashed line in the left side of Figure 13. However, due to the bias, the measured scores are larger than the true scores. The reason for that are the cases of unanimous answers, which put no upper bound on the distances between conditions. The likelihood for those cases, as shown in Figure 9, ensures that distances are larger than a certain value, but they do not restrict the maximum distance values. Such cases are pushing conditions on the quality scale away from each other. It may seem that it would be much easier to ignore the cases of unanimous answers from the comparison matrix. However, as we show in the right plot in Figure 13, this leads to under-estimated JOD values.

Although the likelihood functions in Figure 9 allow distances between conditions to be infinity, we know that in practice all distances are finite and usually moderate numbers. Such knowledge of finite distances will be our prior. We can define as our prior the likelihood of observing a particular distance in quality scores for any randomly selected pair of conditions. Such likelihood for a given pair of conditions is expressed in Eq. (9). Given our toy-example comparison matrix from Eq. (1), we plot the likelihood for all pairs of conditions in Figure 14. The probability of observing any difference is a normalised sum of all plotted probabilities. The problem is, however, that the likelihood for unanimous answers ($\hat{p}_{12}$ and $\hat{p}_{21}$, green lines in the plot) has infinite support and thus cannot be normalised. To avoid this issue, we transform this answer to the closest non-unanimous response. After this step, we can compute the probability of observing a distance z between any two random conditions as:

$$l(z) = \frac{1}{|\Omega|} \sum_{i,j \in \Omega} L(z|c_{ij}^*, n_{ij}), \tag{13}$$

where c_{ij}^* is the non-anonymous closest version of c_{ij}. The main term in the sum is given by Eq. (9) and our prior depends on the estimated distances in the current iteration of the optimisation method and changes in an iterative manner. This probability is shown as a dashed-black line in Figure 14. It shows that the most probable difference between two randomly chosen conditions is about 2.5 JODS, and the support of this probability function is finite. We can add our distance prior to the likelihood
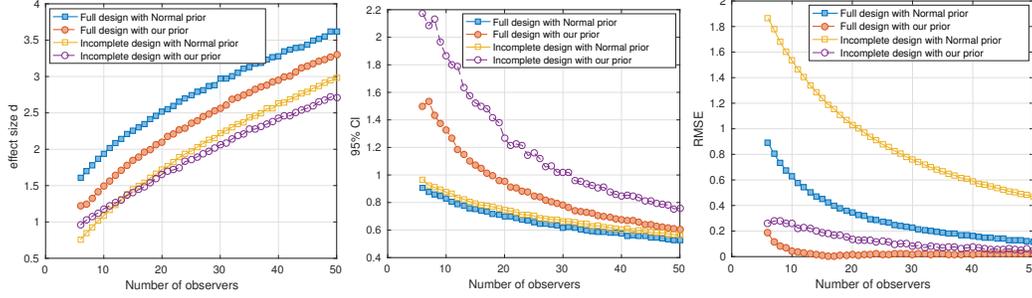
14

Figure 16: The difference in precision with the prior in (Tsukida and Gupta, 2011) for the previously presented metrics.

function from Eq. (10):

$$\underset{\hat{q}_2,\ldots,\hat{q}_n}{\arg\max} \prod_{i,j\in\Omega} L(\hat{q}_i - \hat{q}_j | c_{ij}, n_{ij}) \cdot (l(\hat{q}_i - \hat{q}_j) + \gamma) \,. \tag{14}$$

Note that this is just a prior modulating distances, not a constraint. To allow the selection of other distances, we add a small offset of $\gamma = 0.1$ to our prior. The centre plot in Figure 13 demonstrates how the bias is reduced when the prior is included in the likelihood function.

Figure 14 shows the likelihood function from Figure 9 when it is multiplied by the prior. The likelihood has no longer plateau and has a single maximum, which also improves the stability of the optimisation.

To evaluate the improvement in estimates brought by the prior, we analyse how the precision of the estimation varies with the number of observers. We perform a Monte Carlo simulation of $10,000$ runs for the true quality scores $q = (0, 1, 2, 3, 4)$ and with the same assumption as for estimation of confidence intervals in the Section 7.1. We run the simulation for both complete design (in which we compare all conditions) and incomplete design (in which only nearest neighbours are compared). For each simulation we obtain a set of $10,000$ estimated quality scores $\hat{q}$, which we aim to compare to the true quality scores in q. We define the mean for our estimation of q_i as $\bar{q}_i$. The results are shown in Figure 12 for three different measures:

- Effect size d: ratio of the difference between estimated quality scores divided by the standard deviation of the estimation error:

$$d = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{(\bar{q}_{i+1} - \bar{q}_i)}{\sigma_{\hat{q}_i}} \,, \tag{15}$$

  where $\sigma_{\hat{q}_i}$ is the standard deviation of each individual estimated result from the mean of the distribution $\bar{q}$. The effect size (David, 1963) is a useful measure of the sensitivity of an experimental method, computing whether it can detect a difference between a pair of conditions and prove its statistical significance.

- The average size of 95% confidence interval, computed as in Section 7.1.

- Root Mean Squared Error (RMSE), which measures the deviation from the ground truth, as follows: RMSE$= \sqrt{\frac{1}{n-1} \sum_{i=2}^{n} (q_i - \hat{q}_i)^2}$.

Figure 12 shows how the measures improve as we increase the number of observers. It also shows that both the RMSE and the confidence intervals can be very large if the number of observers is less than 20. The proposed distance prior significantly improves accuracy and robustness of estimation, specially for small samples.

To demonstrate the challenge of selecting the right prior, we compare scaling using our distance prior to the prior proposed in (Tsukida and Gupta, 2011). The authors introduced a prior that assumed the quality scores to be drawn from a normal distribution. Figure 16 shows that even though their prior strongly reduces confidence intervals (as most priors do), it also introduces a large error in the estimates (large RMSE).

9 Outlier detection

In practice, some observers may not fully understand or follow the instructions of the experiment, in particular in less controlled crowdsourcing experiments. It is important to detect these observers that fall outside of the overall pattern because their answers can push the scaling towards an incorrect solution. This section presents a new method to detect those outlier observers. Note that this approach is only intended to support the experimenter, who makes the final decision on whether the observer should be considered an outlier and removed from the dataset.

To indicate if a specific observer can be considered as an outlier, we compare her/his answers to the rest of the sample. First, we exclude a given observer (one by one) from the dataset and use MLE method to find the scaled distances and thus probabilities $P(r_i > r_j)$. Given these probabilities for the rest of the sample, we use the product of likelihoods (Eq. (9)) to calculate the probability of observing the answers of the considered observer. If the considered observer is consistent with the rest of the population, the corresponding probability will be high. In practice, we use the sum of logarithmic likelihoods as it not only simplifies the subsequent analysis, but it also helps numerically, since the product of a large number of small probabilities can easily underflow the numerical precision of floating point numbers.

Different rules can be used to detect outliers, most of them taking into account the distance to a central measure of the distribution and the range of the data. In our case we consider J. Tukey's rules on quartiles. We transform log-likelihoods into the scores, which express the distance to the central range of the distribution in the multiples of the interquartile range. The interquartile range is the distance between 75th and 25th percentiles. We only consider outliers on the left side of the distribution, i.e. cases which show a significantly low likelihood of belonging to the sample, computing the distance to the first quartile.

We ran a series of Monte-Carlo simulations to determine how the presence of an outlier affects the results of scaling and whether our method could be used to automatically determine outliers. As expected, an outlier can introduce the highest error when the number of valid observers (non-outliers) is small. But more interestingly, we observed that the outliers are more difficult to detect when the number of repetitions t is small. Thus, we recommend that each observer repeats the same comparisons at least 3 times. When investigating the actual (non-simulated) datasets from previous papers, we found that given the subjectivity of the experiments this automatic criteria for detecting outliers might not be always accurate. Therefore, we recommend leaving this decision to an experimenter, who should investigate answers of flagged observers whose outlier scores are high (as discussed in Section 3).

10 Practical issues

In this section we explore three relevant issues concerning pairwise comparison experiments: the comparison between complete and incomplete designs, the distance between quality scores and the allowance of ties in the experiment.

10.1 Complete or incomplete design

When designing an experiment we have a choice of comparing all possible pairs of conditions (full or complete design), or only selected pairs, usually those that are the most similar to each other (incomplete design). We are interested in knowing which approach is more efficient and leads to more accurate results.

We have already shown some results for full and incomplete results in Figure 12 when discussing the importance of a prior for small sample sizes. In the plots in that figure incomplete design results in similar accuracy but lower stability in general. However, the plots do not account for the fact that in the full design each participant needs to run many more comparisons. Given $n = 5$ compared conditions in our simulation, the full design requires comparing $(5 \cdot 4)/2 = 10$ pairs, but in the case of incomplete design we compare just 4 pairs: $q_1 \Leftrightarrow q_2$, $q_2 \Leftrightarrow q_3$, $q_3 \Leftrightarrow q_4$ and $q_4 \Leftrightarrow q_5$.

We replotted the data as the function of the number of comparisons instead of observers in Figure 17. Please ignore for now "with ties" curves and focus on the blue and red lines of full and incomplete designs. The plots show that incomplete design results in more stable and similarly accurate estimates
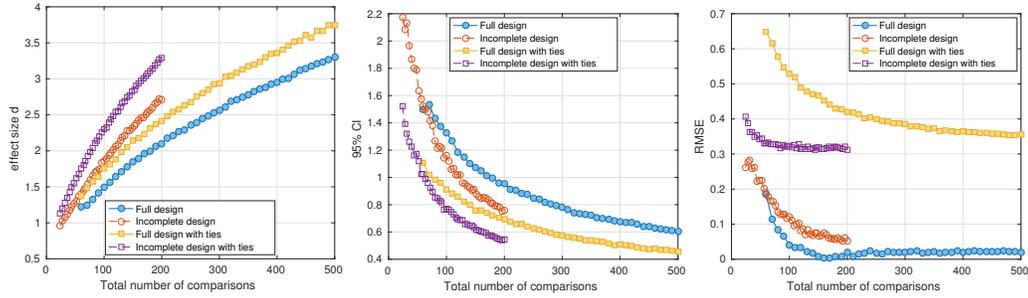
Figure 17: The effect of the total number of comparisons on the effect size (left), confidence interval (middle) and RMSE (right). The values computed by simulating 10,000 experiment runs assuming true quality values to be separated by 1 JOD. Note that these results are valid only for those conditions.
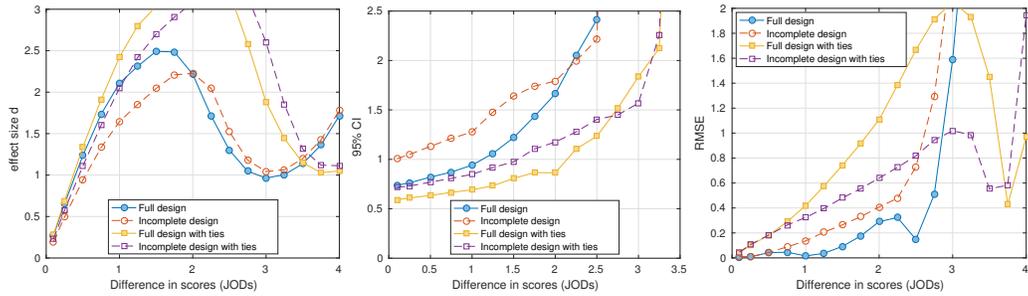

Figure 18: The effect of difference between quality scores on the effect size (left), confidence interval (middle) and bias (right). The more measured conditions are separated on the quality scale, the larger is the error and bias. But this also depend whether ties were allowed and whether the all (full design) or only neighbouring conditions (incomplete design) were compared. The values computed by simulating 10,000 experiment runs for 20 observers for each point on the plot.

given the same experimental effort. The gain will depend on the number of conditions to compare. For example, if we had 10 conditions, full design would require comparing 45 pairs, but only 9 pairs would need to be compared in incomplete design, resulting in much larger gain. Similar conclusions have been drawn when a sorting algorithm was used (Silverstein and Farrell, 2001; Maystre and Grossglauser, 2017).

10.2 Distance between quality scores

The accuracy of scaling methods depends on the distances between quality scores. The scaling becomes especially unreliable if the distance between quality scores is larger than 2 JODs (i.e. $p_{ij} > 0.91$). When we suspect that perceptual attributes will be scaled over a larger range than 2 JODS, the difference scaling method (Maloney and Yang, 2003) could be more appropriate. To test this effect, we run a Monte-Carlo simulation for different assumed distances (all equal) between true JOD scores, and summarised them in Figure 18. Let us focus on the full design (blue-continuous lines) and ignore all other curves for now. As shown in Figure 18, both the RMSE and confidence intervals increase rapidly, even although the distances between conditions are changed smoothly. It is obvious that measures such as RMSE will increase, since the range of true values increases along x-axis, however, the increase is more abrupt than the linear increase expected.

10.3 Experiments with ties

Allowing observers to select a third "no preference" option when they can not see a difference, is a controversial issue in pairwise comparison experiments, still disputed and researched (Ennis and Ennis, 2012b,a; Chapman and Lawless, 2005; Davidson, 1970).

There are different ways in which ties can be introduced in the statistical analysis (Ennis and Ennis, 2012b). For our next experiment we choose the equal-split method: if an observer chooses "no-

preference", we split the vote in two and add a half-vote to each condition. This may result in a non-integer number of votes, which we round up to the nearest smaller or larger integer (randomly selected and taking into account the number of comparisons needs to be consistent). We simulate observers who make the "no-preference" choice when the difference between the two conditions is less than a certain threshold. As different observers are unlikely to have the same and consistent opinion when the two conditions are the same, our "no-preference" threshold is a random variable $N(0.7, 0.3)$ in the space of JOD units. The result of simulating 10,000 experiment runs with ties are compared with the same experiments but without the tie option in Figures 17 and 18.

Our simulation shows that offering a "no-preference" option reduces the size of confidence intervals and improves the effect size. But this happens at the cost of a larger error (see the third plot in Figure 17). Taking a closer look at the results, we observe that the solution is always under-estimated. There is an intuitive interpretation of this result: offering "no-preference" option results in more "no difference" responses while the difference is actually there, giving smaller JOD distances and negative bias (under-prediction). The bias is large enough to offset any gains in the reduced confidence intervals. The bias can be potentially eliminated, but it requires modeling the "no-preference" selection (Davidson, 1970) and finding the parameters of that model: how likely will observers select "no-preference" where there is actually no difference (Ennis and Ennis, 2012a). This in turn requires collecting extra data: observer responses for two identical conditions. The current version of the pwcmp software does not support modeling ties when scaling, therefore we cannot recommend offering a "no-preference" option when this software is used for scaling.

11   Conclusions and limitations

The choice of pairwise comparison data and scaling methods over a more simplistic analysis presents several advantages: (i) it can be used to compare and rank items that present similar quality (as opposed to direct ordinal rating), (ii) it allows the potential use of incomplete designs to decrease the data to collect (while presenting accurate predictions), (iii) the scaling can be interpreted (especially since the difference measure units can provide information about the probabilities) and (iv) measurement noise in the comparisons can be addressed in a principled way.

Concerning general guidelines for the experimental design, our results show that incomplete designs can achieve competitive performance if comparisons are appropriately chosen (e.g. neighbours in the quality scale), that the use of ties generally results in an under-estimation of the scaling solution and that it is crucial to ensure that differences between compared conditions are relatively small. Our experiments have also shown the importance of the finite distance prior and screening outlier observers.

The limitation of our work is the assumption of a simple Thurstone Case V observer model, which does not account for dependence between repetitions, observers and conditions, and assumes quality to be explained by a single scalar value.

As future work, we would like to extend our analysis and software to include adaptive sampling procedures to reduce the number of required comparisons, and more advanced machine learning techniques that link explanatory variables to the scaling, to facilitate the process of knowledge extraction.

References

Agresti, A. (1992). Analysis of ordinal paired comparison data. *Applied Statistics*, 41:287–297.

Bradley, R. A. and Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27(3):412–433.

Causeur, D. and Husson, F. (2005). A 2-dimensional extension of the bradley-terry model for paired comparisons. *Journal of Statistical Planning and Inference*, 135(2):245–259.

Chapman, K. W. and Lawless, H. T. (2005). Sources of error and the no-preference option in dairy product testing. *Journal of Sensory Studies*, 20(5):454–468.

David, H. (1963). *The method of paired comparisons*, volume 298. Charles Griffin and Co.

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *J. Amer. Statist. Assoc.*, 65:317–328.

Davidson, R. R. and Farquhar, P. H. (1976). A bibliography on the method of paired comparisons. *Biometrics*, 32(2):241–252.

Dunn-Rankin, P., Knezek, G., Wallace, S., and Zhang, S. (2004). *Scaling Methods*. Taylor & Francis.

Eilertsen, G., Mantiuk, R. K., and Unger, J. (2015). Real-time noise-aware tone mapping. *ACM Transactions on Graphics*, 34(6):1–15.

Eilertsen, G., Wanat, R., Mantiuk, R. K., and Unger, J. (2013). Evaluation of Tone Mapping Operators for HDR-Video. *Computer Graphics Forum*.

Engeldrum, P. G. (2000). *Psychometric scaling: a toolkit for imaging systems development*. Imcotek Press.

Ennis, D. M. and Ennis, J. M. (2012a). Accounting for no difference/preference responses or ties in choice experiments. *Food Quality and Preference*, 23(1):13–17.

Ennis, J. M. and Ennis, D. M. (2012b). A comparison of three commonly used methods for treating no preference votes. *Journal of Sensory Studies*, 27(2):123–129.

Hatzinger, R. and Dittrich, R. (2012). prefmod: An R Package for Modeling Preferences Based on Paired Comparisons, Rankings, or Ratings. *Journal of Statistical Software*, 48(10):1–31.

Herbrich, R., Minka, T., and Graepel, T. (2006). TrueSkill: A Bayesian Skill Rating System. *Advances in Neural Information Processing Systems*, 19:569–576.

Howell, D. C. (2009). *Statistical Methods for Psychology*. Cengage Learning.

Jamieson, K. G. and Nowak, R. D. (2011). Active ranking using pairwise comparisons. In *Neural Information Processing Systems (NIPS)*, pages 2240–2248.

Karaduzovic-Hadziabdic, K., Telalovic, J. H., and Mantiuk, R. (2016). Subjective and Objective Evaluation of Multi-exposure High Dynamic Range Image Deghosting Methods. In *Eurographics 2016 - Short Papers*.

Kingdom, F. A. and Prins, N. (2016). *Psychophysics: A Practical Introduction*. Academic Press, 2nd editio edition.

Lu, Z.-l. and Dosher, B. (2013). *Visual Psychophysics: From Laboratory to TheoryNo Title*. MIT Press.

Maloney, L. T. and Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, 3(8):5–5.

Mantiuk, R. K., Tomaszewska, A., and Mantiuk, R. (2012). Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum*, 31(8):2478–2491.

Maystre, L. and Grossglauser, M. (2015). Fast and accurate inference of plackett-luce models. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 172–180, Cambridge, MA, USA. MIT Press.

Maystre, L. and Grossglauser, M. (2017). Just sort it! A simple and effective approach to active preference learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2344–2353. PMLR.

Montag, E. D. (2003). Louis leon thurstone in monte carlo: creating error bars for the method of paired comparison. volume 5294, pages 222–230. SPIE.

Pellacini, F., Ferwerda, J. A., and Greenberg, D. P. (2000). Toward a psychophysically-based light reflection model for image synthesis. In *Proc. SIGGRAPH '00*, pages 55–64. ACM Press.

Placket, R. L. (1975). The analysis of permutations. *Applied Statistics*, 24:193–202.

Ramanarayanan, G., Ferwerda, J., and Walter, B. (2007). Visual equivalence: towards a new standard for image fidelity. *ACM Transactions on Graphics (TOG)*, 26(3):76.

Shah, N., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. (2015). Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 856–865. PMLR.

Sheikh, H., Sabir, M., and Bovik, A. (2006). A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451.

Silverstein, D. and Farrell, J. (2001). Efficient method for paired comparison. *Journal of Electronic Imaging*, 10:394.

Springall, A. (1973). Response surface fitting using a generalization of the bradley-terry paired comparison model. *Applied Statistics*, pages 59–68.

Stewart, N., Brown, G. D., and Chater, N. (2005). Absolute identification by relative judgement. *Psychological Review*, 112(4):881–911.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Tsukida, K. and Gupta, M. R. (2011). How to Analyze Paired Comparison Data. Technical Report UWEETR-2011-0004, Department of Electrical Engineering University of Washington.

Turner, H. and Firth, D. (2012). Bradley-terry models in r: The bradleyterry2 package. *Journal of Statistical Software, Articles*, 48(9):1–21.

Usami, S. (2010). Individual differences multidimensional bradley-terry model using reversible jump markov chain monte carlo algorithm. *Behaviormetrika*, 37(2):135–155.

Vangorp, P., Mantiuk, R. K., Bazyluk, B., Myszkowski, K., Mantiuk, R., Watt, S. J., and Seidel, H.-P. (2014). Depth from HDR: depth induction or increased realism? In *ACM Symposium on Applied Perception - SAP '14*, pages 71–78. ACM Press.

Wanat, R. and Mantiuk, R. K. (2014). Simulating and compensating changes in appearance between day and night vision. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 33(4):147.

Wauthier, F. L., Jordan, M. I., and Jojic, N. (2013). Efficient ranking from pairwise comparisons. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, volume 28, pages 109–117. JMLR.org.

Wickelmaier, F. and Schmid, C. (2004). A matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, & Computers*, 36(1):29–40.

Xu, Q., Jiang, T., Yao, Y., Huang, Q., Yan, B., and Lin, W. (2011). Random partial paired comparison for subjective video quality assessment via hodgerank. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 393–402. ACM.

Ye, P. and Doermann, D. (2014). Active sampling for subjective image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4249–4256.

Zerman, E., Hulusic, V., Valenzise, G., Mantiuk, R. K., and Dufaux, F. (2018). The relation between MOS and pairwise comparisons and the importance of cross-content comparisons. In *Proc. of Human Vision and Electronic Imaging*.