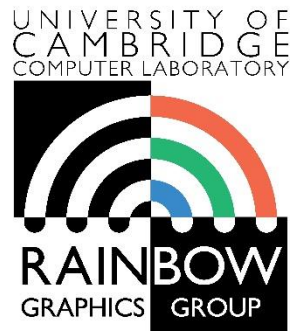


Pairwise comparisons and JND/JOD quality scale 8-year perspective

Rafał Mantiuk
University of Cambridge



Archery competition

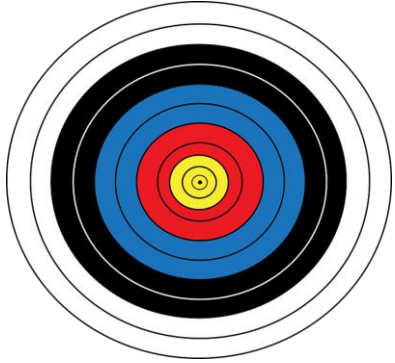
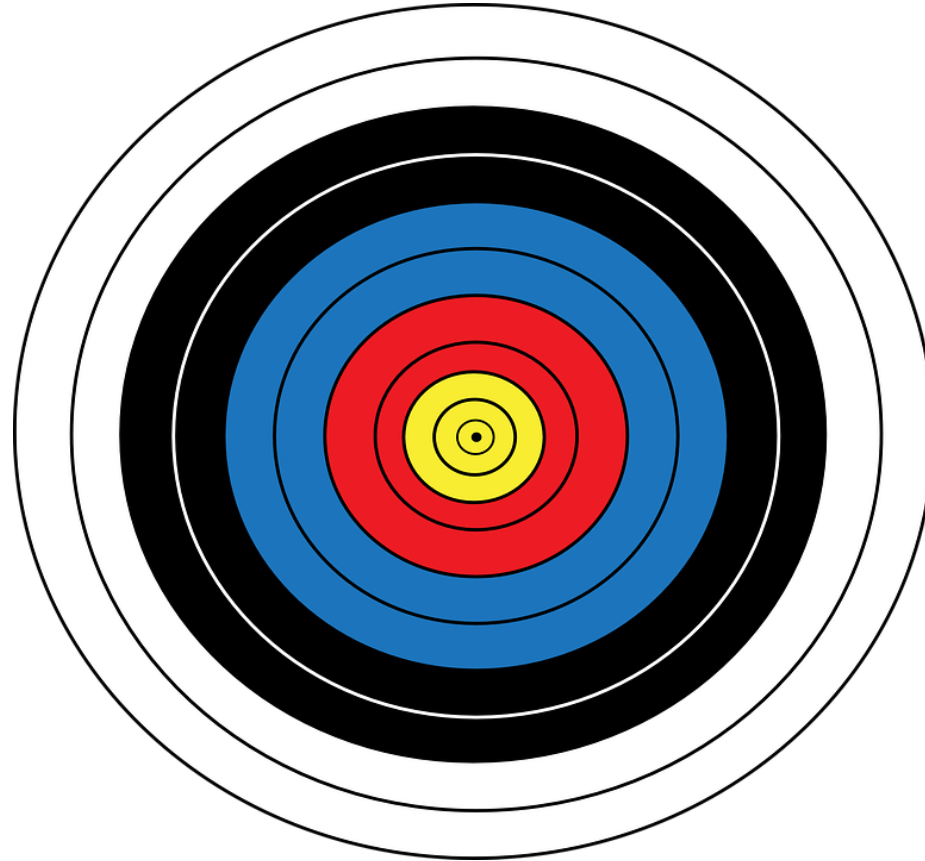


Image quality metric competition

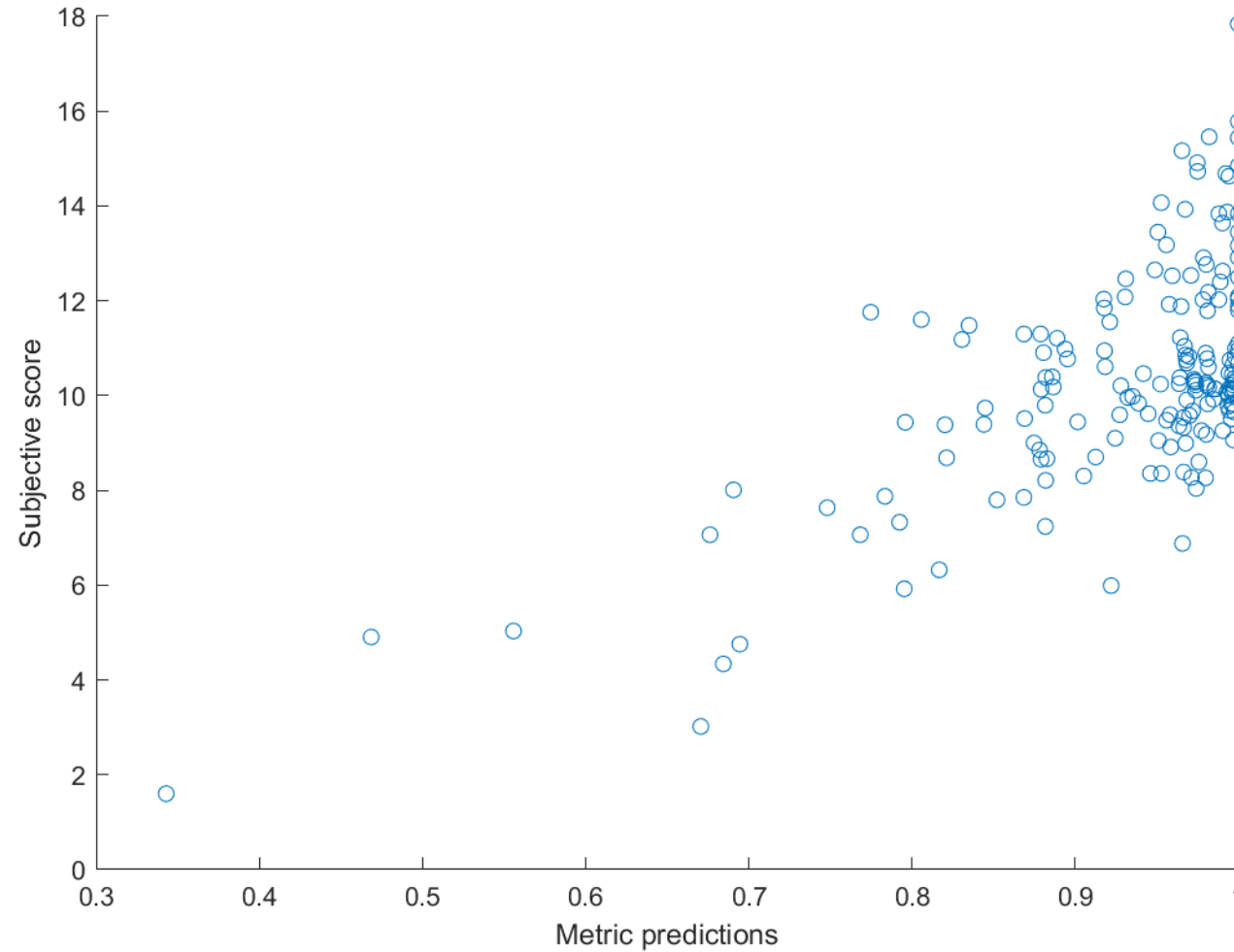
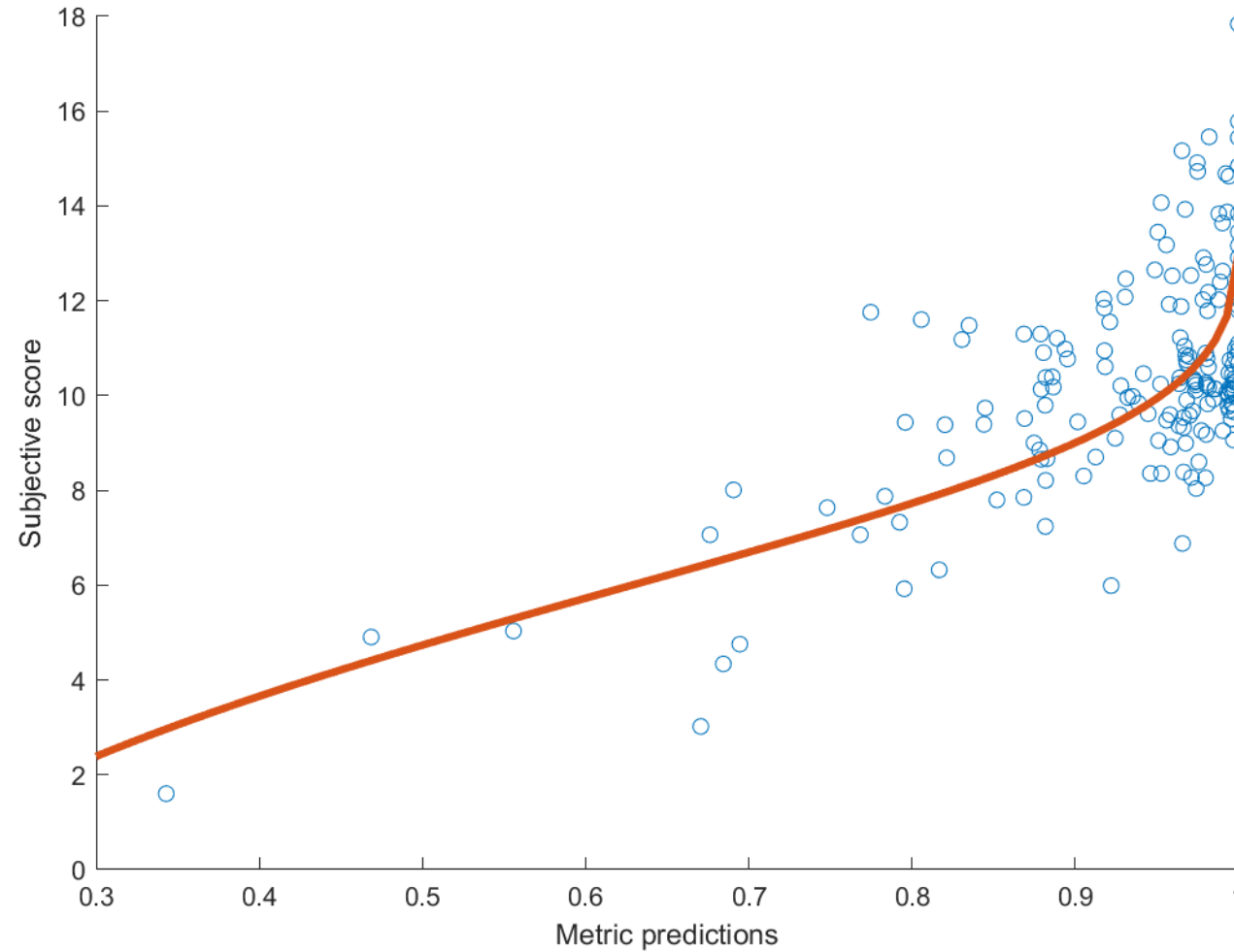
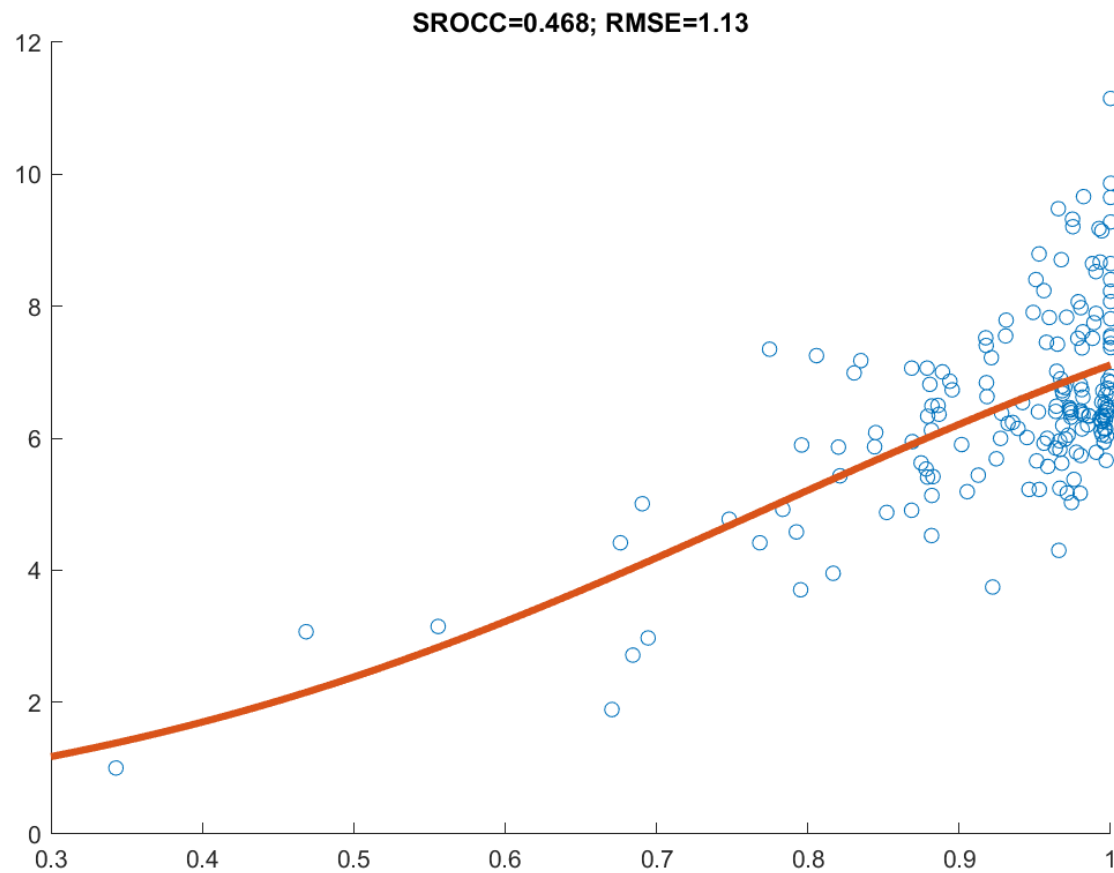
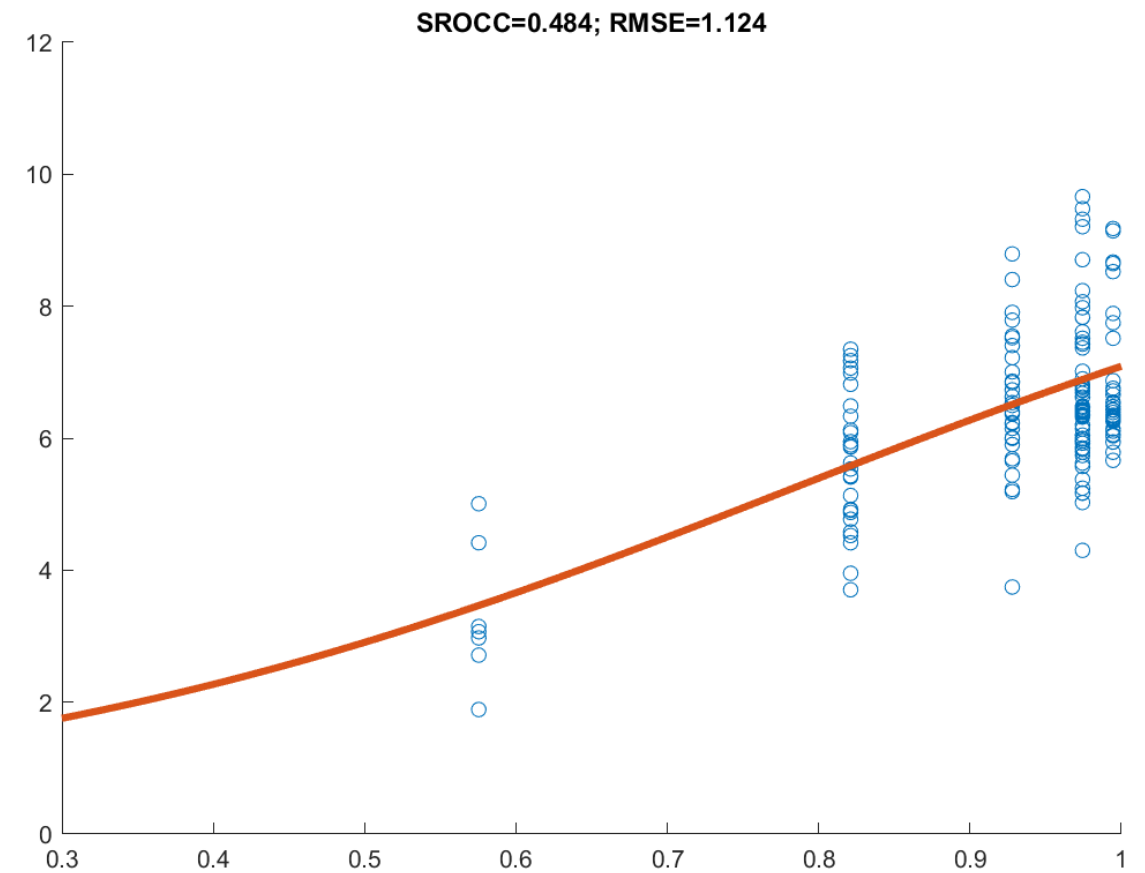


Image quality metric competition



How to improve your metric?



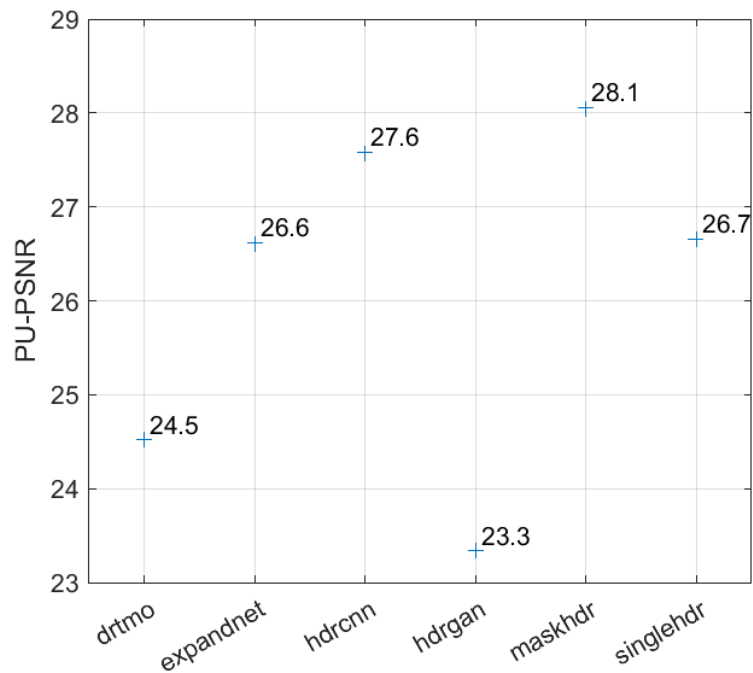
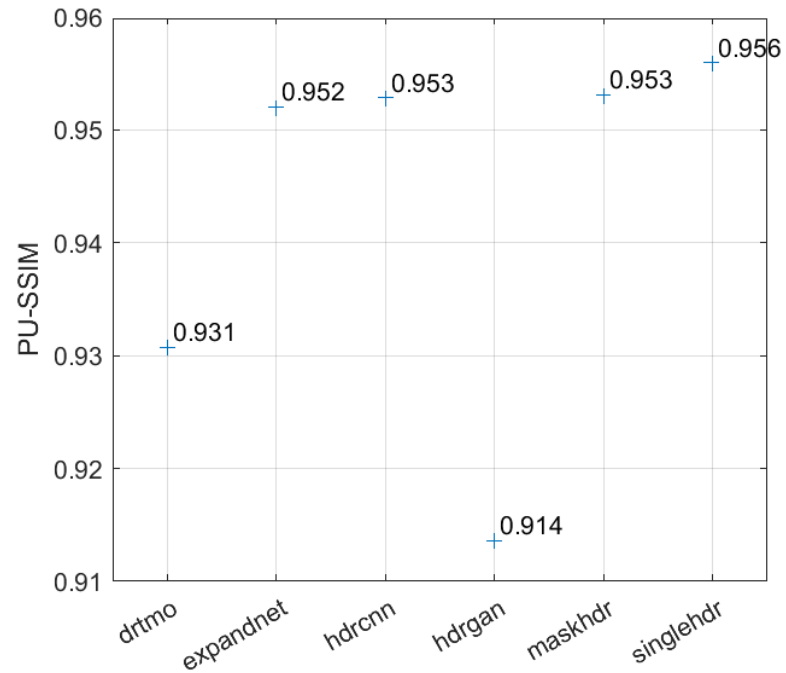
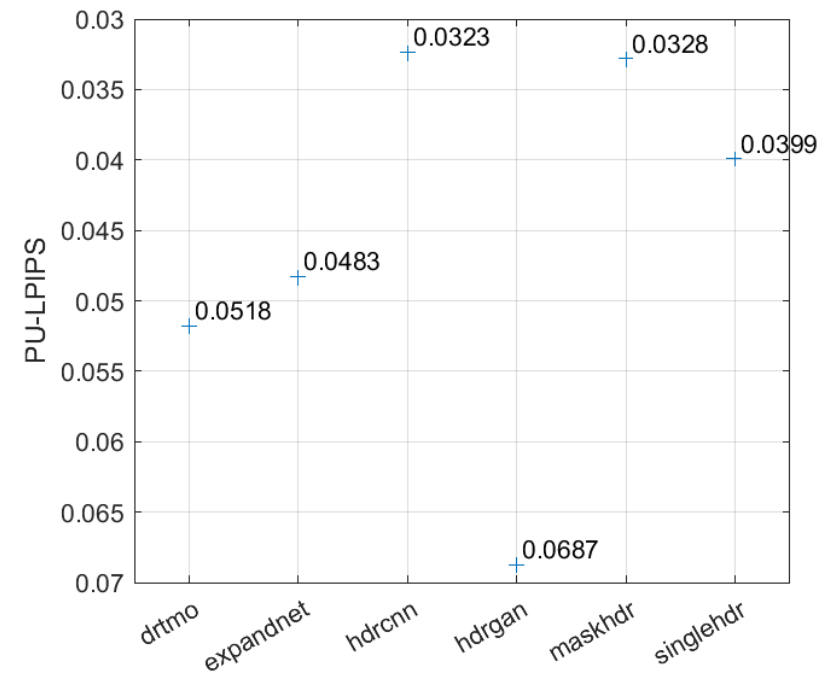
How do we typically report results?

Method	DrTMO	ExpandNet	HDR-CNN	HDR-GAN	MaskHDR	SingleHDR
PU-PSNR $\uparrow$	24.5	26.6	27.6	23.3	28.1	26.7
PU-SSIM $\uparrow$	0.931	0.952	0.953	0.914	0.953	0.956
PU-LPIPS $\downarrow$	0.0518	0.0483	0.0323	0.0687	0.0399	0.0302

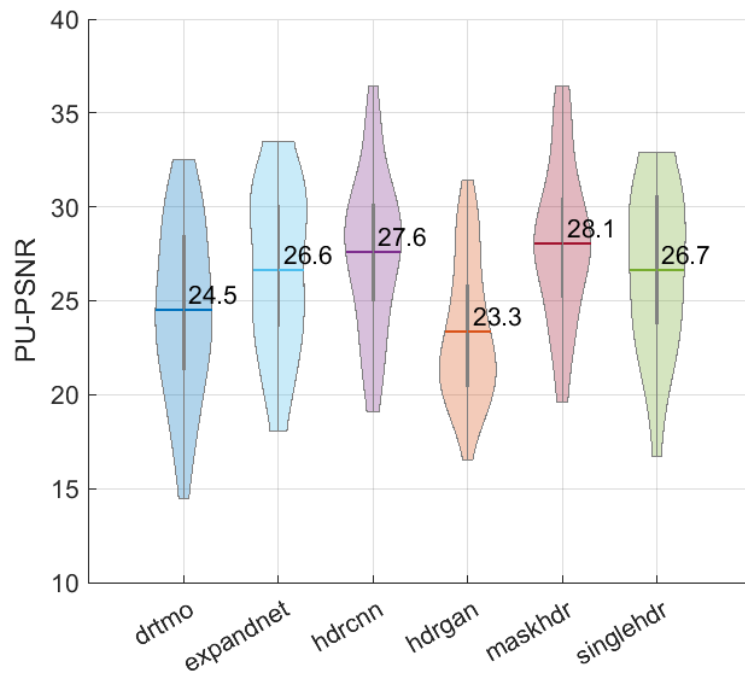
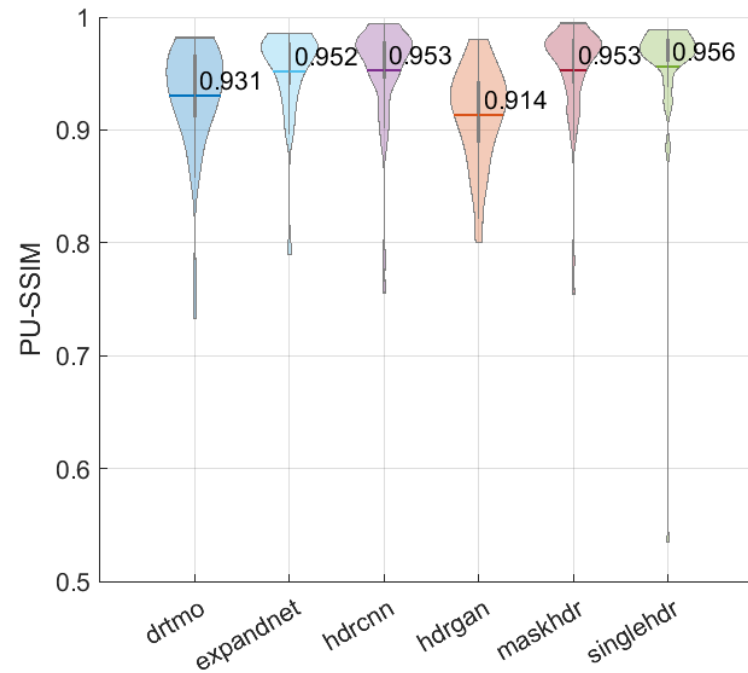
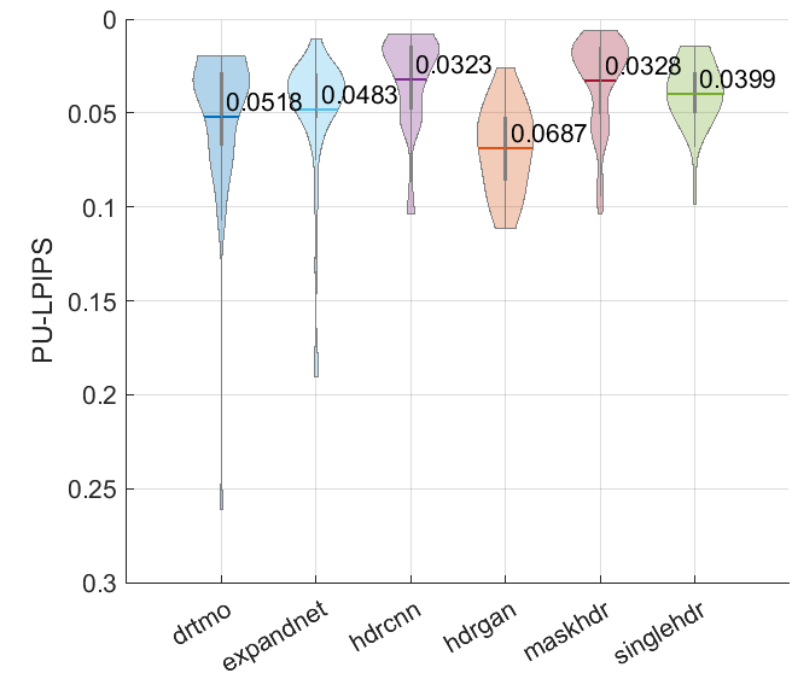
How do we typically report results?

Method	DrTMO	ExpandNet	HDR-CNN	HDR-GAN	MaskHDR	SingleHDR
PU-PSNR $\uparrow$	24.5	26.6	<u>27.6</u>	23.3	28.1	26.7
PU-SSIM $\uparrow$	0.931	0.952	<u>0.953</u>	0.914	<u>0.953</u>	0.956
PU-LPIPS $\downarrow$	0.0518	0.0483	<u>0.0323</u>	0.0687	0.0399	0.0302

The same results, but better presented

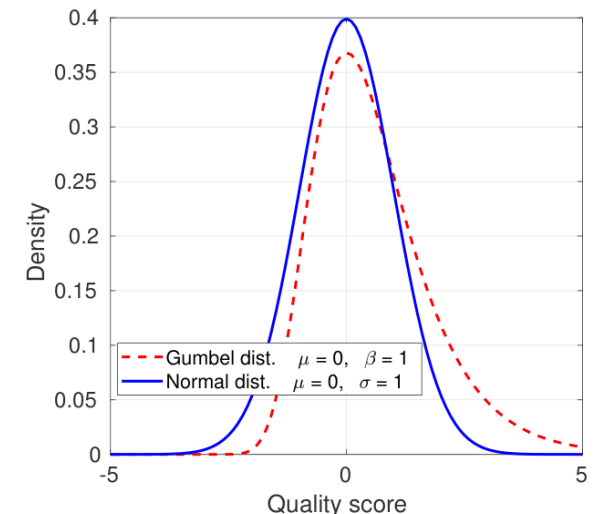


The same results, but even better presented



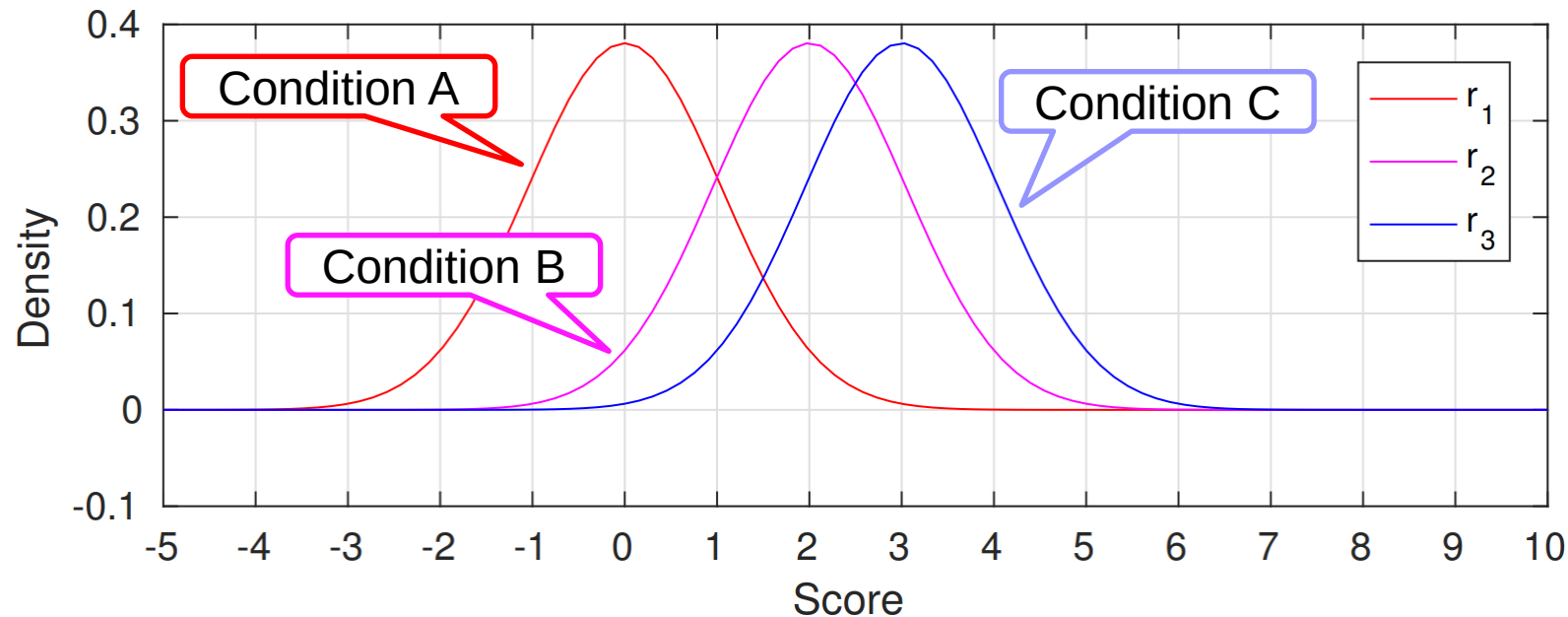
Absolute quality scale

- MOS – different scale in every experiment
 - This is why we need to fit a logistic function
 - The “Excellent – Poor” scale meaningless without context
 - The context sets a different scale for every experiment
- Can we have an absolute scale that is (approx.) consistent across the experiments?
 - Thurstone model – Case V (Normal distribution)
 - Bradley-Terry model (Gumbel distribution)



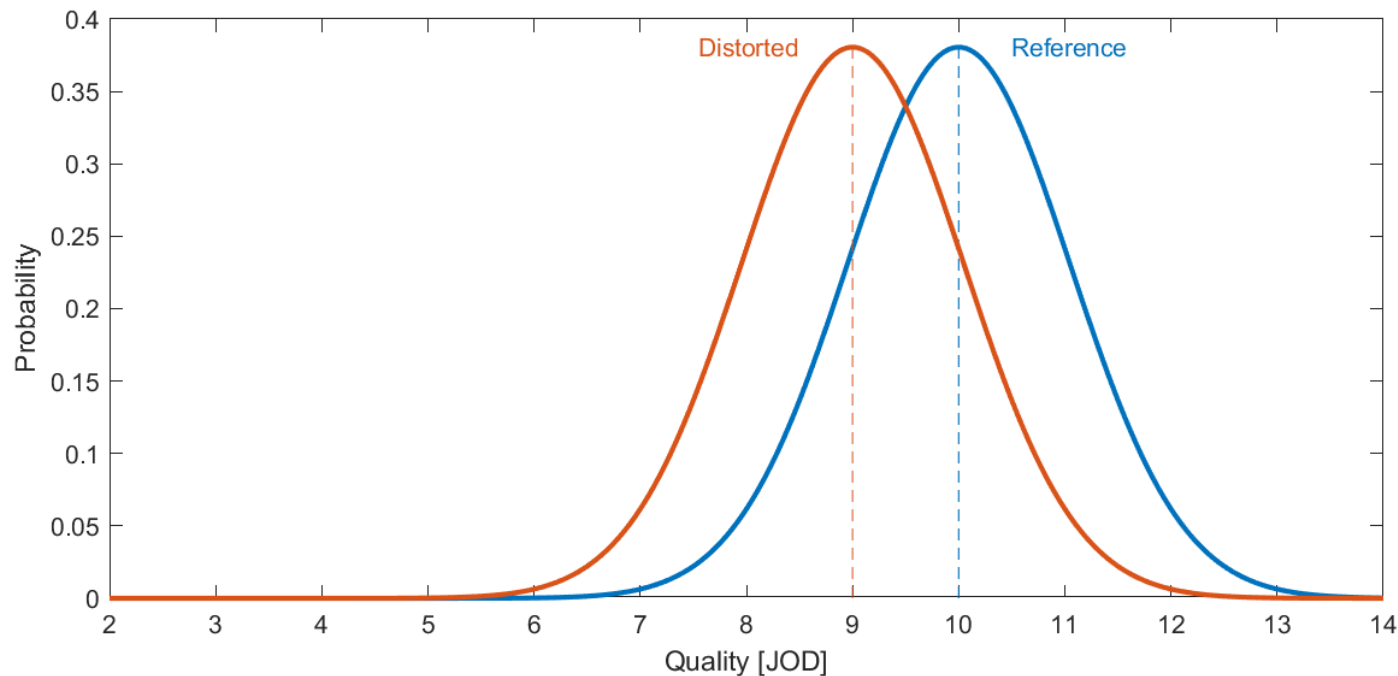
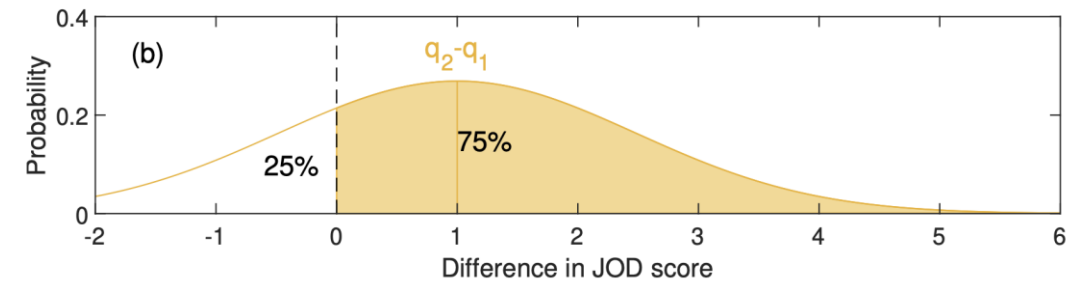
Thurstone (observer) model - Case V

- Two assumptions:
 - Quality scores for a given condition are normally distributed across the population
 - The variance of that distribution is the same for each condition and the judgements are independent



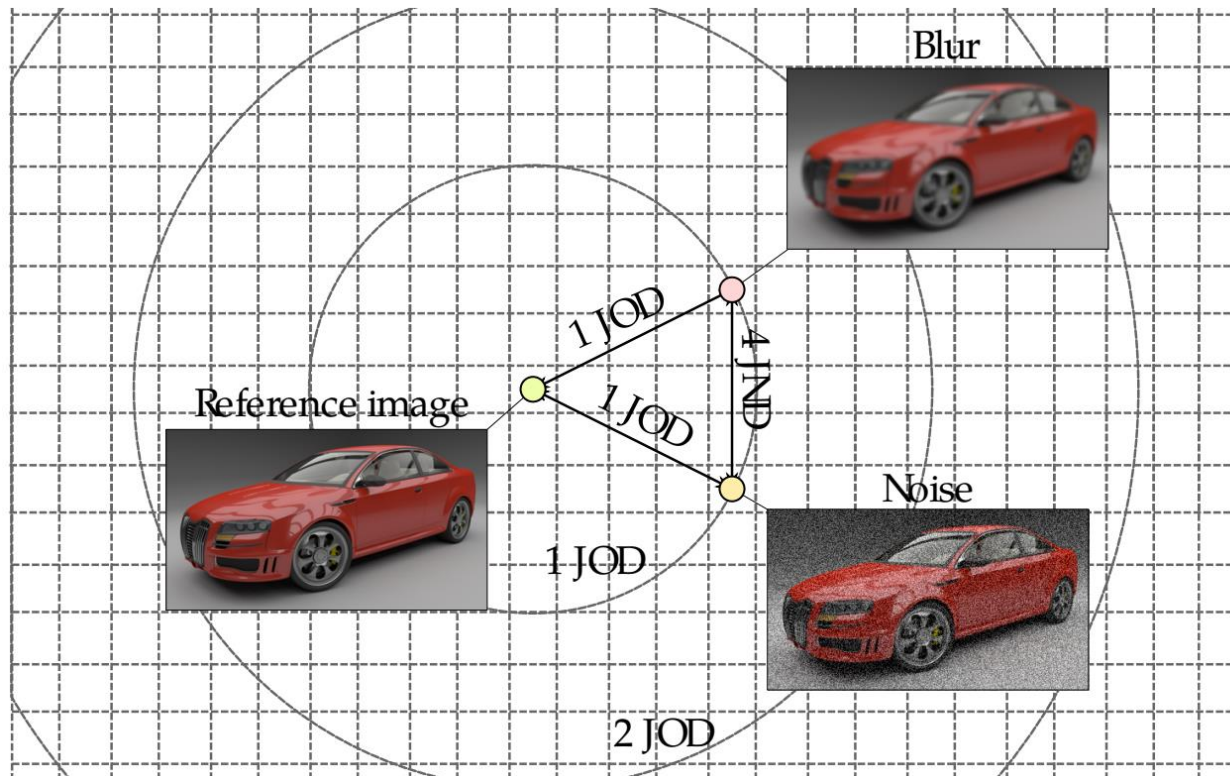
Just Objectionable Difference scale

- Thurstone case V + two more assumptions:
 - Reference is at 10 [JOD]
 - The difference of 1 JOD means that 75% of the population prefers one condition over another ($\sigma = 1.0484$)



JND vs JOD

- Just Noticeable Differences
- Just Objectionable Differences



- JND – is one visually different from another
- JOD – is the **quality** of one different from the quality of another (relative to the reference)

Pair-wise comparison method

- Example: video quality
- Task: Select a video sequence that has a higher quality



Comparison matrix

- Results can be stored in a comparison matrix

$$C = \begin{matrix} & \begin{matrix} C1 & C2 & C3 \end{matrix} \\ \begin{matrix} C1 \\ C2 \\ C3 \end{matrix} & \begin{bmatrix} 0 & 3 & 1 \\ 3 & 0 & 2 \\ 5 & 4 & 0 \end{bmatrix} \end{matrix}$$

- $C_{ij} = n$ means that condition C_i was preferred over C_j n times

Full and reduced designs

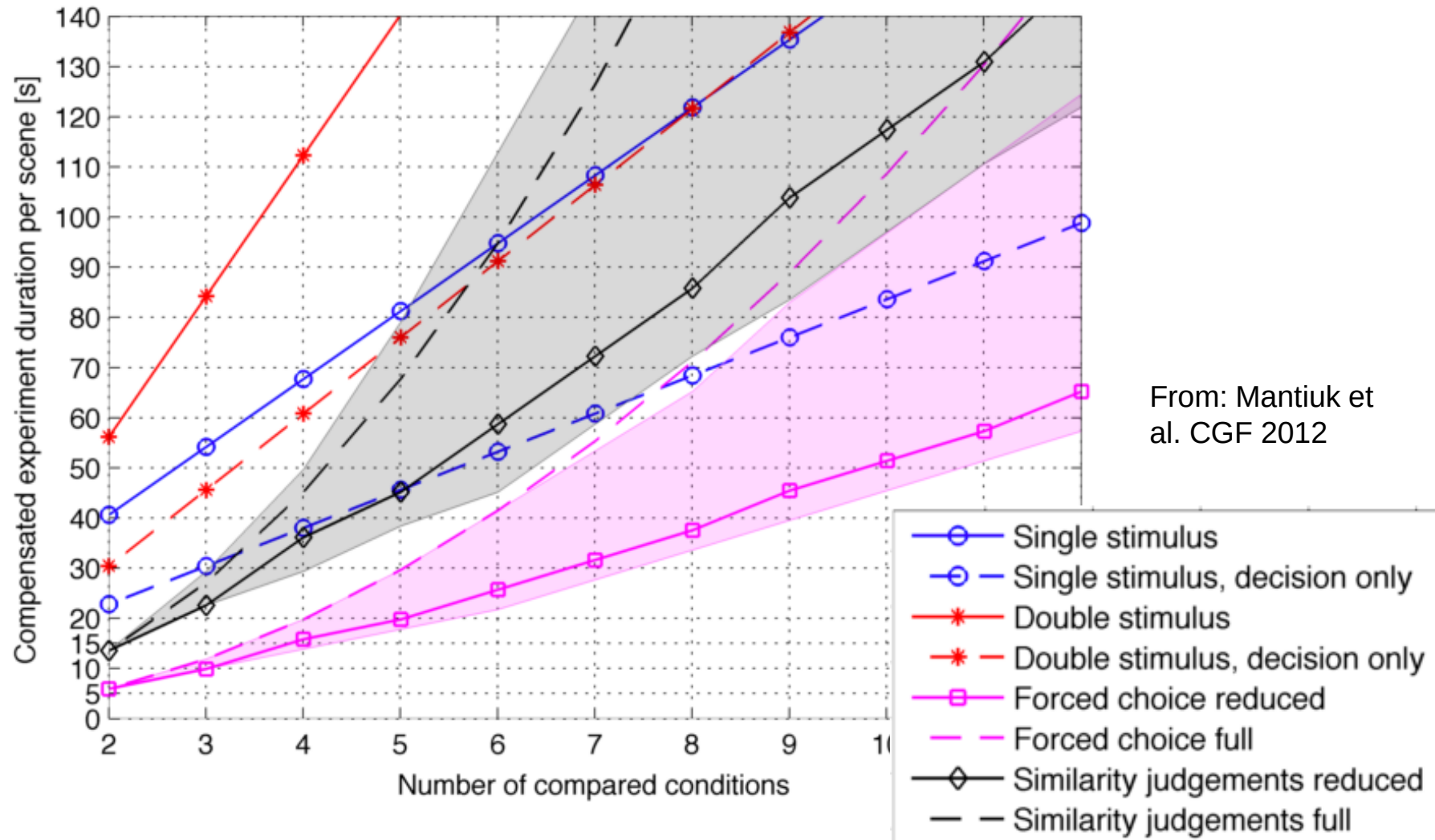
- Full design
 - Compare all pairs of conditions
 - This requires $\binom{n}{2} = \frac{n(n-1)}{2}$ comparisons for n conditions
 - Tedious if n is large
- Reduced design
 - We assume transitivity
 - If $C1 > C2$ and $C2 > C3$ then $C1 > C3$
 - no need to do all comparisons
 - There are numerous “block designs” (before computers)
 - But the task is also a sorting problem
 - The number comparison can be reduced to $n \cdot \log(n)$ for a “human quick-sort”
 - And many others: Swiss chess system, active sampling ...

$$C = \begin{matrix} & \begin{matrix} C1 & C2 & C3 \end{matrix} \\ \begin{matrix} C1 \\ C2 \\ C3 \end{matrix} & \begin{bmatrix} 0 & 3 & 1 \\ 3 & 0 & 2 \\ 5 & 4 & 0 \end{bmatrix} \end{matrix}$$

Pairwise comparisons vs. rating

- The method of pairwise comparisons is **fast**
 - More comparisons, but
 - It takes less time to achieve the same sensitivity as for direct rating methods
- Has a higher sensitivity
 - Less “external” variance between and within observers
- Provides a unified quality scale
 - The scale is transferrable between experiments
- Simple procedure
 - Training participants is much easier
 - Less affected by learning effects
- Especially suitable for non-expert participants
 - e.g., crowdsourcing experiments

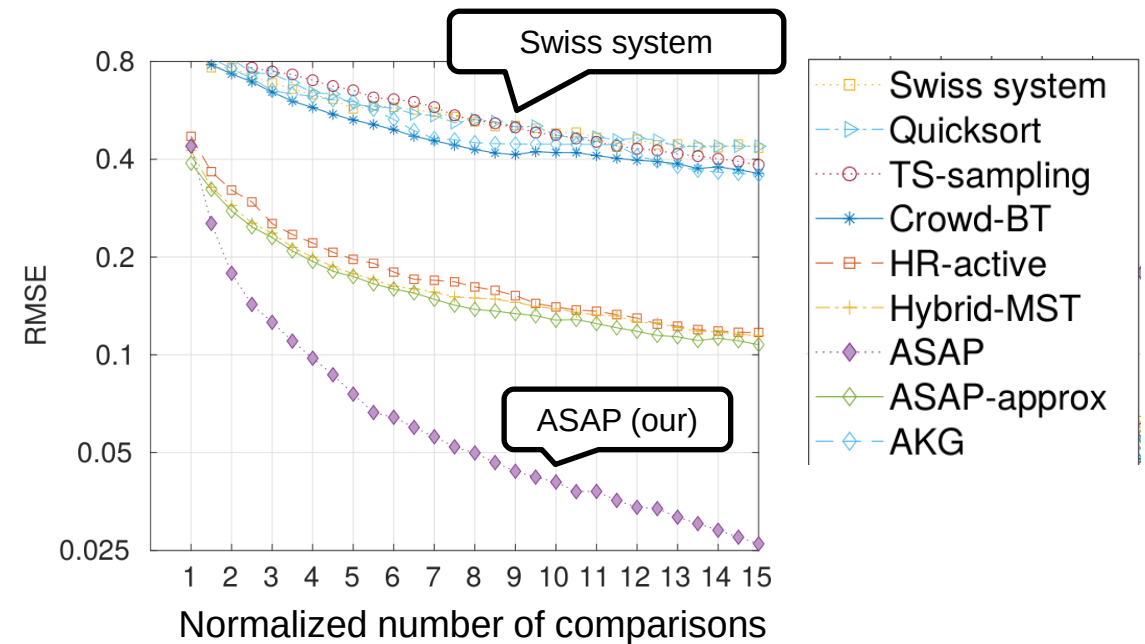
Time-efficiency



From: Mantiuk et al. CGF 2012

Active sampling can make the experiments even faster

- Active sampling
 - For each trial, select a pair of conditions that maximizes the information gain
 - Information gain - DK-divergence between the prior and posterior distributions
- Mikhailiuk, A., et al. “Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization.” In *International Conference on Pattern Recognition*, 2020. [Best Student Paper Award]

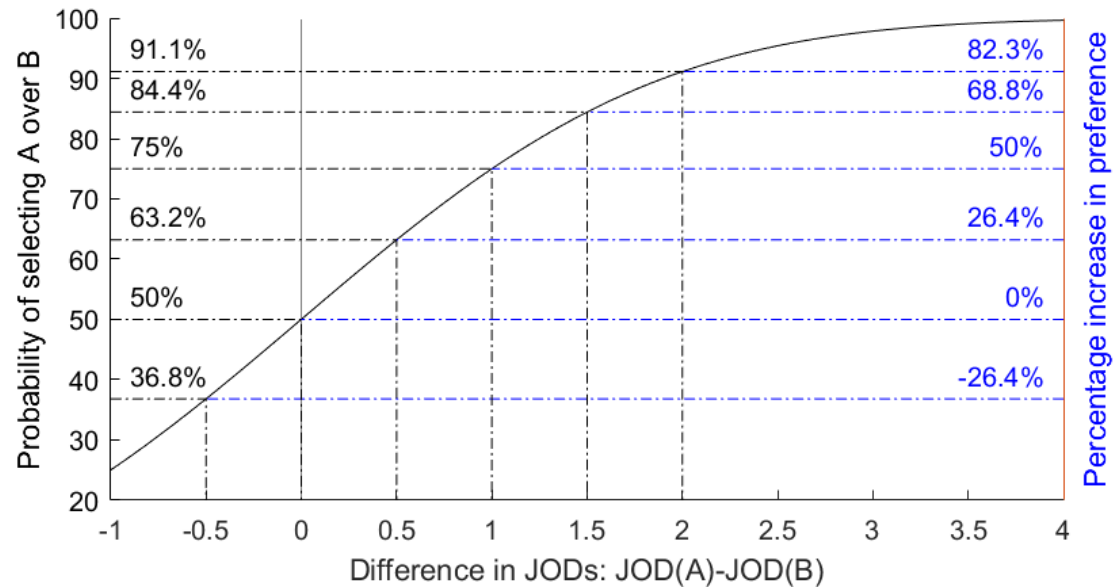
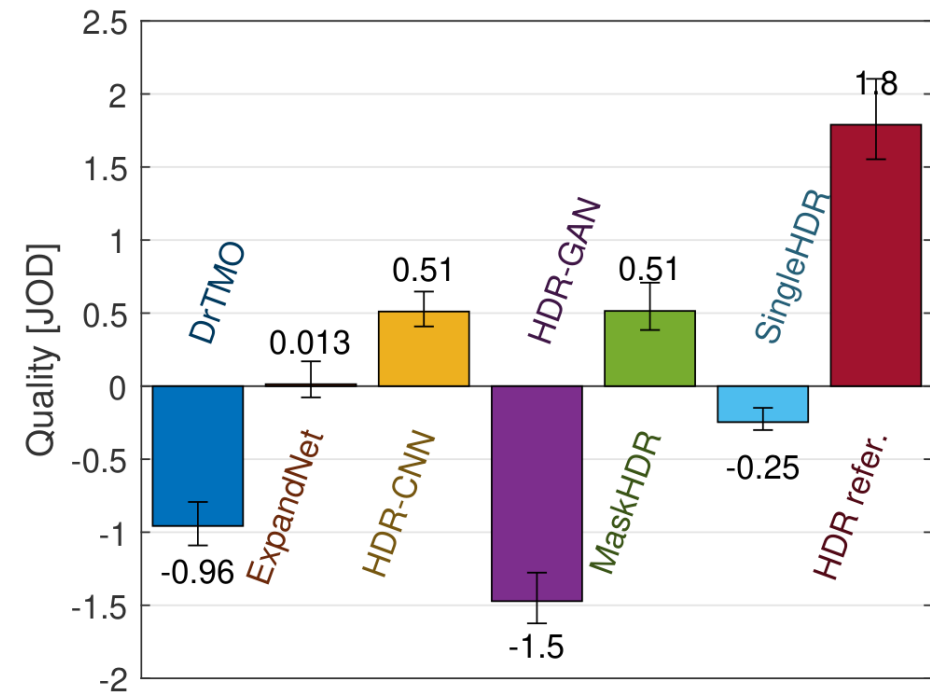


Software for pairwise comparisons active sampling

- ASAP active sampling [Matlab + Python]
 - State-of-the-art method
 - <https://github.com/gfxdisp/asap>
- The best starting point:
 - https://github.com/gfxdisp/asap/blob/master/matlab/ASAP/PwcmpASAPScheduler_example.m
- Recommended to work on batches of comparisons
 - Minimum-spanning-tree (Hybrid-MST)
 - If there are N conditions to compare, a single batch requires $N-1$ comparisons

Practical significance - scaling

- Scaling: to map user judgments into meaningful interval scale
- Useful to show “practical” significance



Scaling pairwise comparison data

- Given a matrix of comparisons, for example

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

- Infer the quality scores for all compared conditions
 - Using Maximum Likelihood Estimation (MLE)
- We start from an observer model, then link it to the observations

From the observer model to probabilities

- Given the observer model for two conditions:

$$r_i = N(q_i, \sigma^2) \quad r_j = N(q_j, \sigma^2)$$

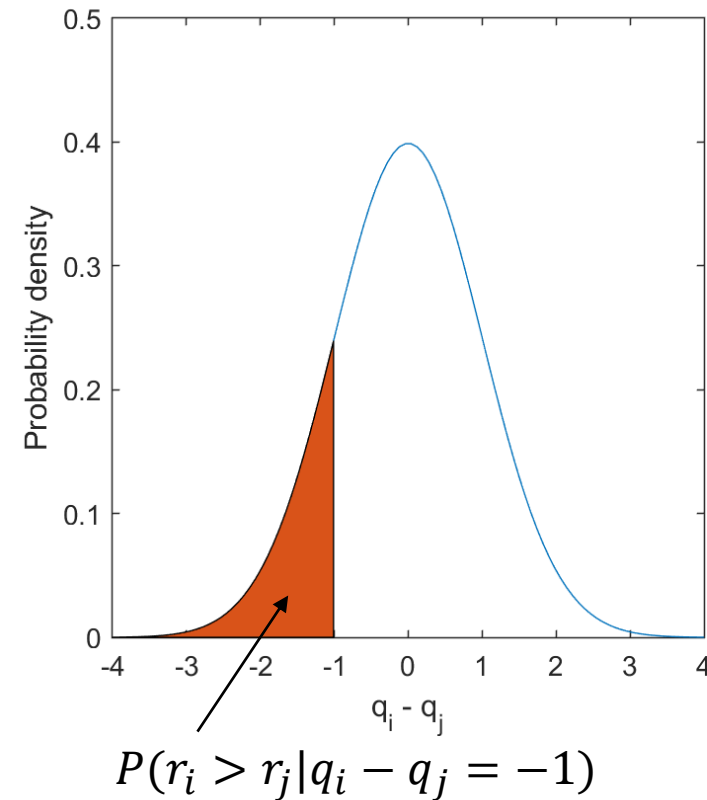
- The difference between two quality scores is:

$$r_i - r_j = N(q_i - q_j, 2\sigma^2)$$

- Then, the probability of the judgment is explained by the cumulative normal distribution

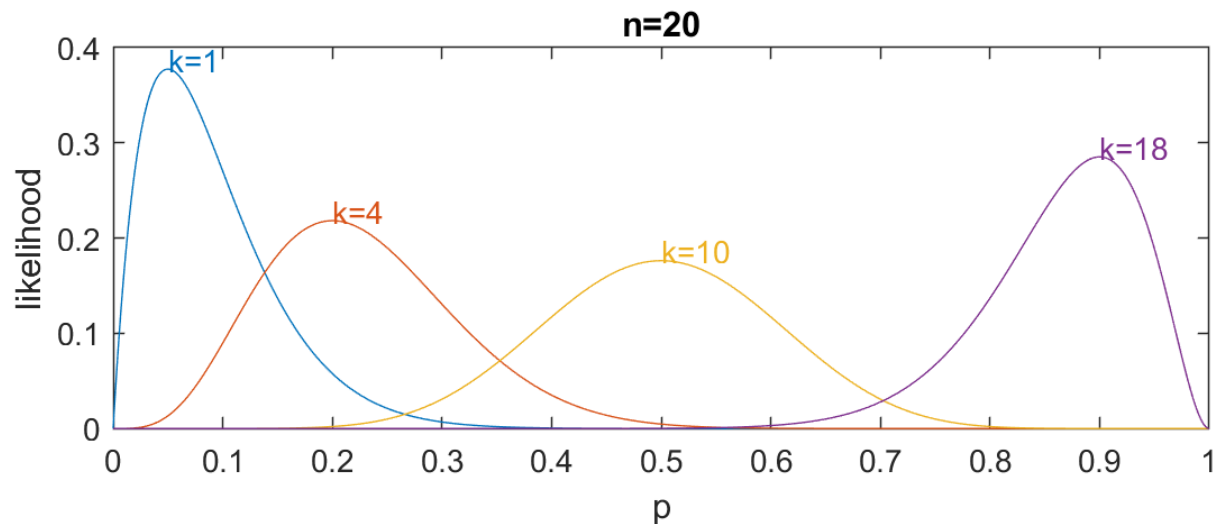
$$P(r_i > r_j) = P(r_i - r_j > 0) = \Phi\left(\frac{q_i - q_j}{\sigma_{ij}}\right)$$

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



Binomial distribution

- Given that the true probability of selecting A over B is p , what is the likelihood that A was selected k times in n trials.



$$P(r_i > r_j | n, k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Maximum Likelihood Estimation

- Given our observations (comparison matrix) what is the likelihood of the quality values q_i :

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

- where $n_{ij} = c_{ij} + c_{ji}$
- To estimate the values of q_i , we maximize:

$$\arg \max_{\hat{q}_2, \dots, \hat{q}_n} \prod_{i, j \in \Omega} L(\hat{q}_i - \hat{q}_j | c_{ij}, n_{ij})$$

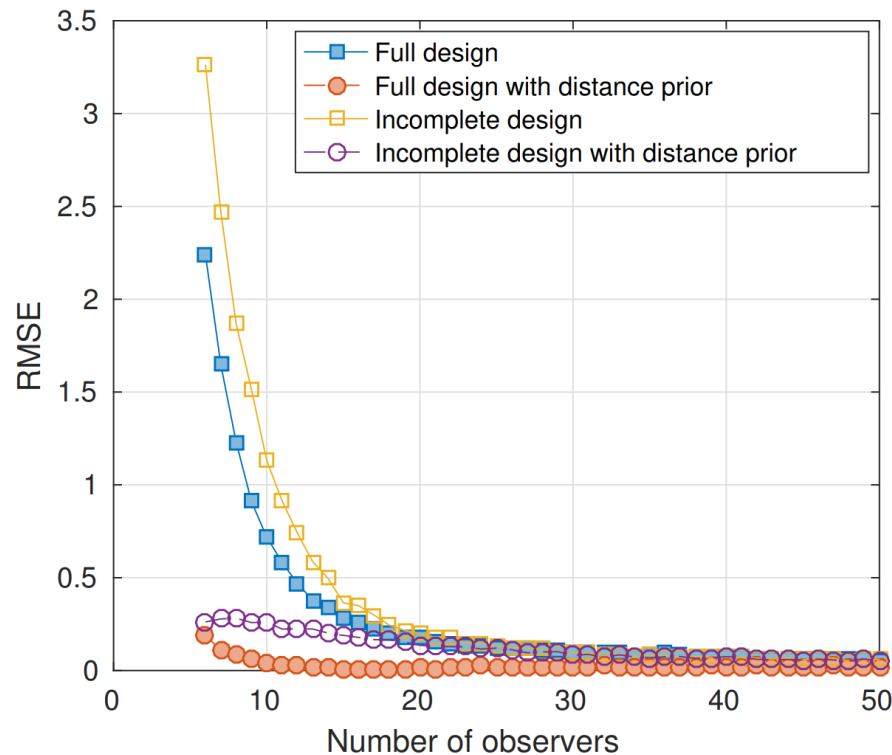
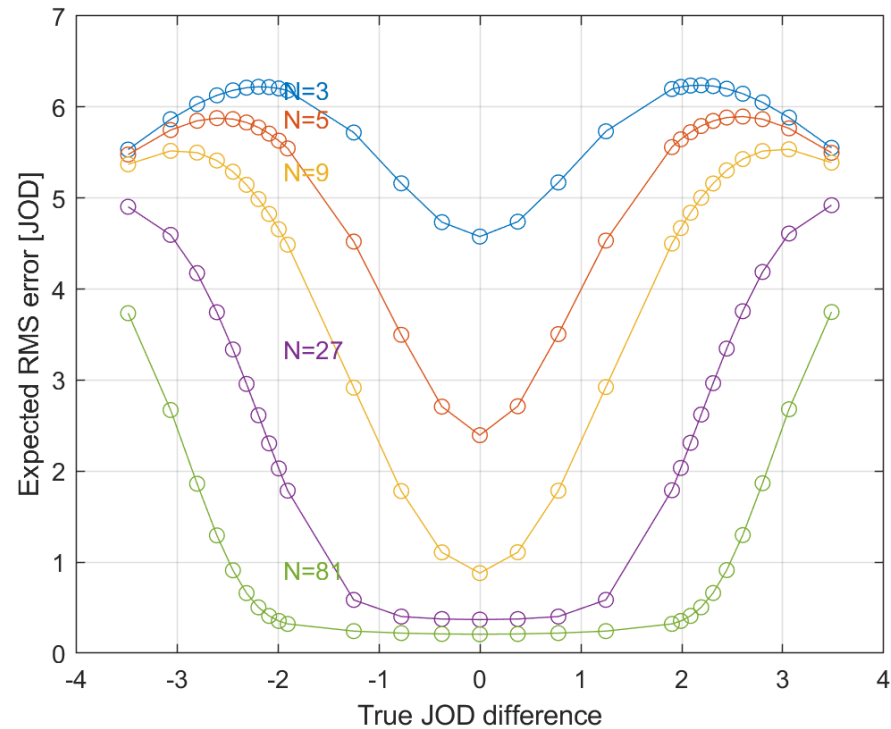
- The entire derivation is gently explained in: <https://arxiv.org/abs/1712.03686>

Software for scaling pairwise comparisons

- Scaling of pairwise comparison data [Matlab]
 - With prior
 - Can bootstrap confidence intervals
 - Fast
 - Used in 80+ projects
 - <https://github.com/mantiuk/pwcmp>

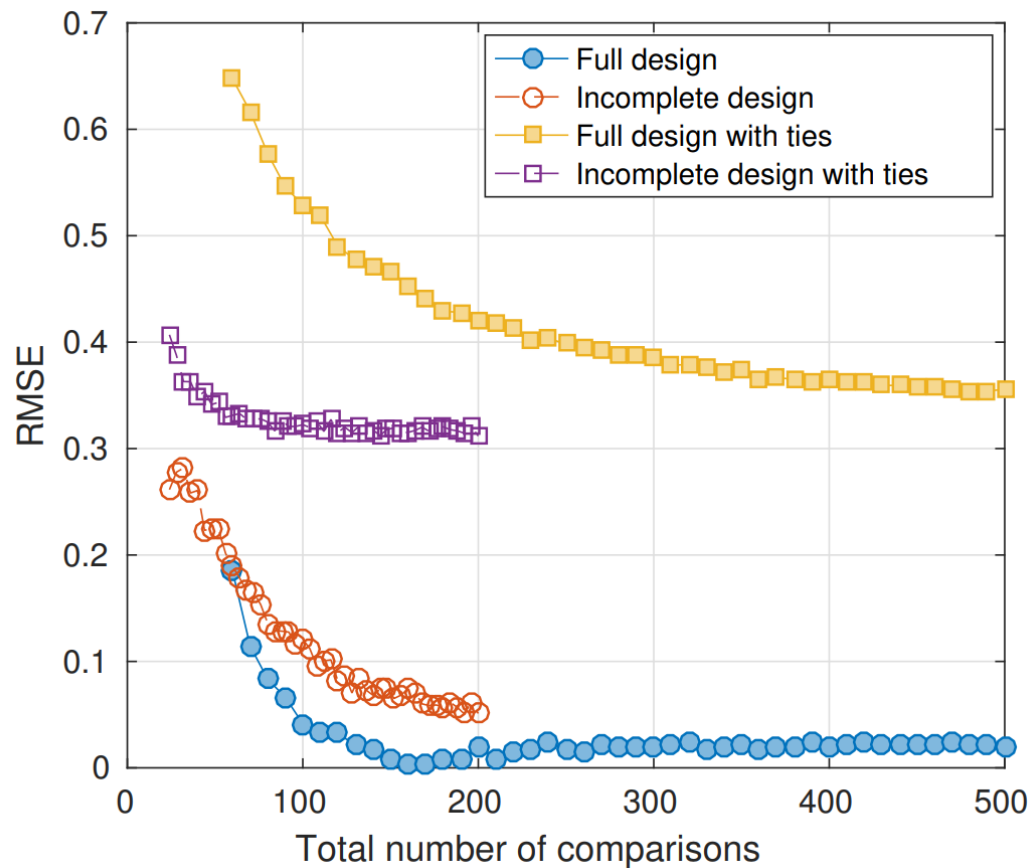
Practicalities of MLE scaling

- At least 15-20 comparisons per each pair are needed to obtain stable results (prior helps)



2-alternative forced choice vs. comparison with ties

- Giving a “tie” option is usually a bad idea



- Scaling the results with ties requires a more complex observer model with more parameters to estimate

Mean Opinion Scores and JODs

- MOS and JODs can be scaled together
 - We need a part of data to be collected using both protocols
 - MOS and JODs are (linearly) related
 - Scaling method
 - [\[Pérez-Ortiz et al. From pairwise comparisons and rating to a unified quality scale, TIP 2019\]](#)
 - Example: UPIQ dataset
 - Four dataset brought to the same quality scale
 - [\[Mikhailiuk et al. UPIQ: Consolidated dataset and metrics for high-dynamic-range image quality, IEEE TM 2021\]](#)

