



Subjective image quality assessment: experimental aspects and databases

Marius Pedersen

February 2021



Marius Pedersen

- Professor of color imaging at the Norwegian University of Science and Technology (NTNU).
- Deputy head – department of computer science
- Head of the Norwegian Colour and Visual Computing Laboratory
- Research interest: colour imaging, especially image quality



Subjective assessment - introduction

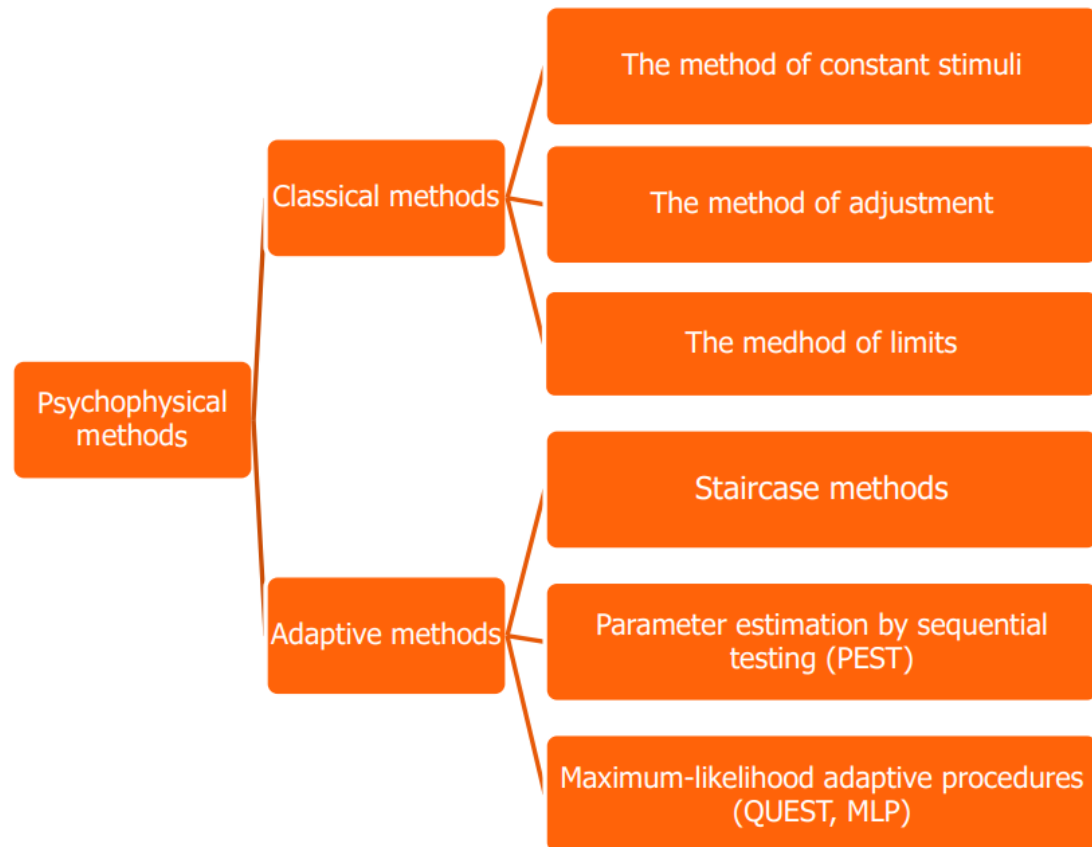
- Subjective assessment involves psychophysical experiments, where human observers are asked to grade a set of images according to a given criterion.
- There are several existing methods for carrying out these types of experiments.





Psychophysical thresholds

- Method of adjustment
- Method of limits
- Method of constant stimuli





Psychometric scaling

- Psychophysical threshold methods are useful to determine color tolerances, compression limits, and more.
- However, the goal is often to determine the scale of perception compared to a single threshold.
- Three common methods for conducting psychophysical scaling experiments:
 - category judgment,
 - pair comparison,
 - and rank order.

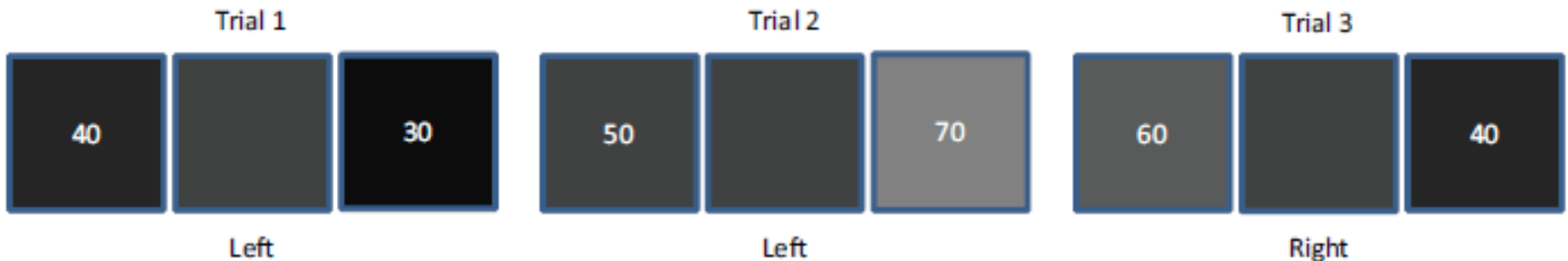
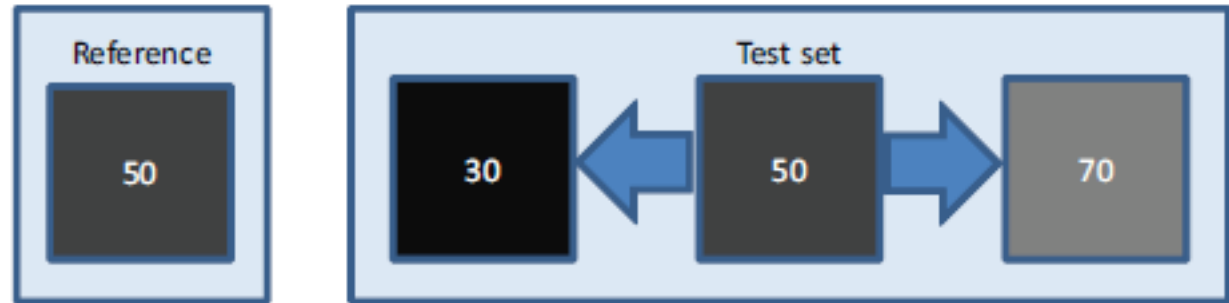


Pair comparison

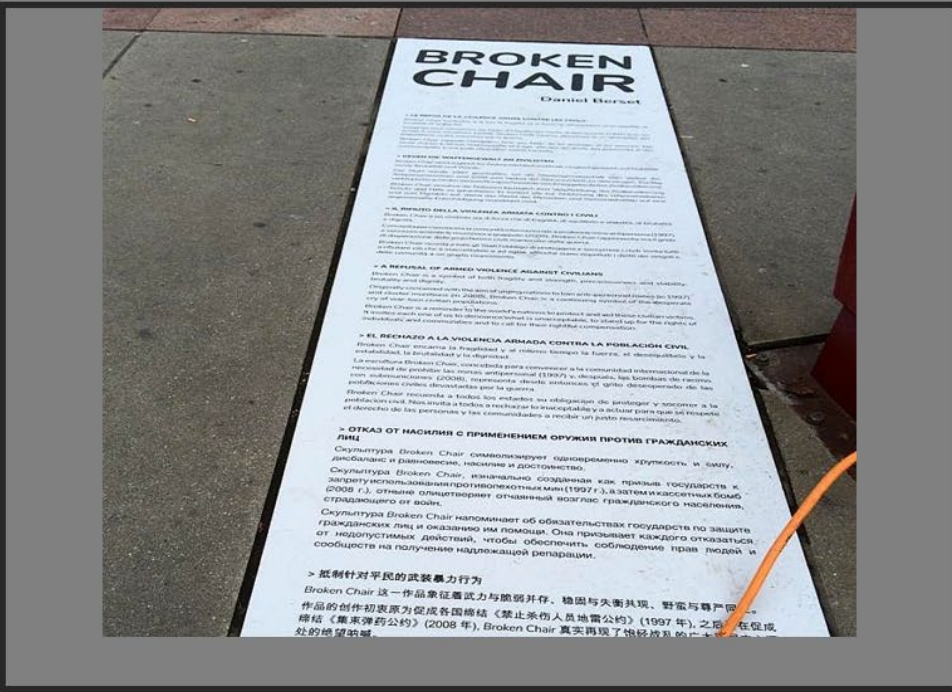
- In pair comparison experiments observers judge quality based on a comparison of image pairs
- The observer is asked which image in the pair is the best according to a given criterion
 - For example which has the highest quality or is the least different from an original.
- These experiments can be either
 - forced-choice, where the observer needs to give an answer, or
 - the observer is not forced to make a decision and may judge the two reproductions as equals (tie).
- In the case of pair comparison experiments no information on the distance between the images is recorded, making it less precise than category judgment, but less complex.
- Pair comparison is the most popular method to evaluate e.g. gamut mapping*, and is often preferred due to its simplicity, requiring little knowledge by the user.



Example pair comparison experiment



- For the first trial the observer judged the left patch to be closer to the reference, the same with the second trial, and in the third trial the right. The observer judges all combinations of pairs.



NEXT



Number of comparisons

- In these experiments the observer evaluates n reproductions for m reference images.
 - Resulting in $(n(n-1) \times m) / 2$ comparisons
 - 10 reference images and 5 reproductions = 100 comparisons
- Each pair of reproductions is usually shown twice, changing the position of the right and left reproductions to avoid bias
 - Resulting in $m \times n(n-1)$ comparisons
 - 10 reference images and 5 reproductions = 200 comparisons
- With an increasing number of reproductions the number of trials increases very rapidly, which makes it unsuitable for experiments involving many reproductions.



Data analysis

- Using Thurstone's Law of Comparative Judgment, data collected from pair comparison experiments can be transformed into interval scale data.
- The results from this transformation represent the distance of a given stimulus from the mean score of the set being evaluated.
- When calculating scaled values several assumptions must be satisfied:
 - Each sample has a single value that can describe its quality.
 - Each observer estimates the quality of this sample with a value from a normal distribution around this actual quality.
 - Each sample has the same perceptual variance.
 - Each comparison is independent.



Confidence Interval

- A **confidence interval** gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. (Valerie J. Easton and John H. McColl's Statistics Glossary v1.1)
- The 95% Confidence Interval (CI) is found by using the number of observations:

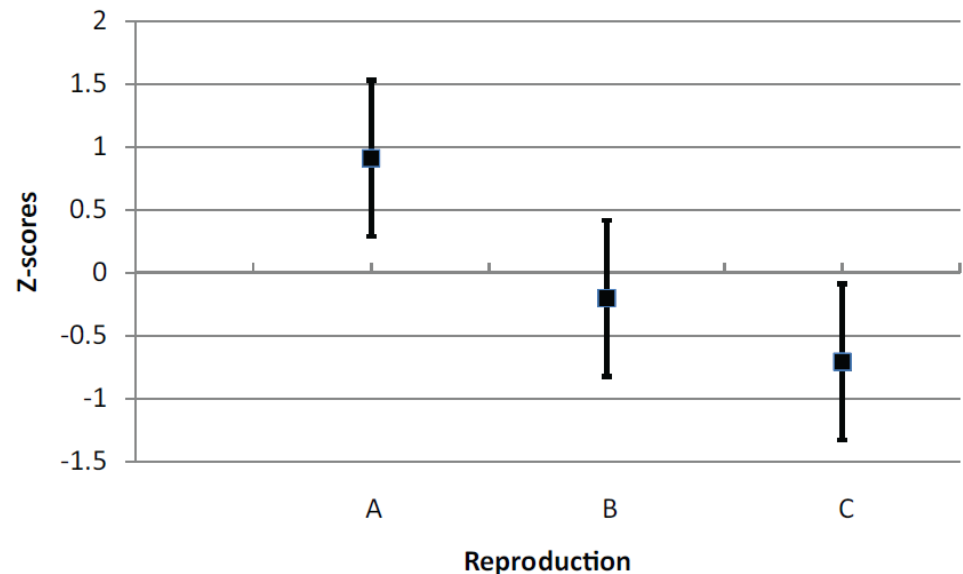
$$CI = 1.96 \times \frac{\sigma}{\sqrt{N}},$$

- where σ is the standard deviation and N the number of observations.
- Since the z-scores have a scale with units equal to $\sigma/\sqrt{2}$ the standard deviation can be set to 1, giving the confidence interval of the mean to be $1.96 \times (1/\sqrt{N})$
- The 95% confidence interval is then the mean z-scores $\pm CI$



Visualization of z-scores

- The most common way to visualize z-scores is by an error bar plot.
 - The mean z-score value is indicated by the center square, and the whiskers on each line show the 95% CIs
- If two CIs overlap the two reproductions are not considered to be significantly different with a 95% confidence (as seen between reproduction B and C)
- If they do not overlap the difference between the two CIs they are statistically significant with 95% confidence (as seen between reproduction A and C)



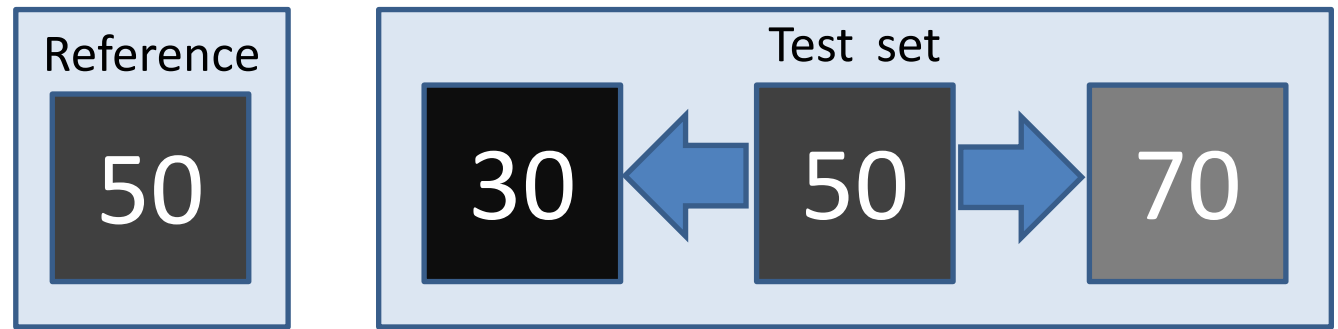


Category judgement

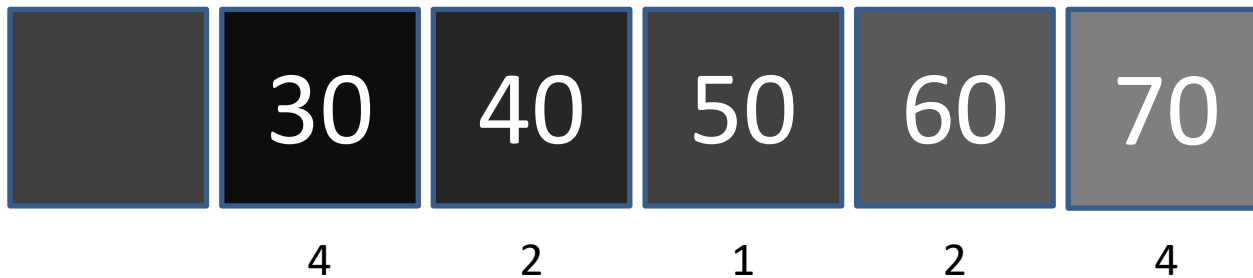
- In category judgment the observer is instructed to judge an image according to a criterion, and the image is assigned to a category.
- Five or seven categories are commonly used, with or without a description of the categories.
- One advantage of category judgment is that information on the distance between images is recorded, but the task is more complex than pair comparison for the observers.
- Category judgment experiments are often faster than pair comparison, with fewer comparisons necessary.



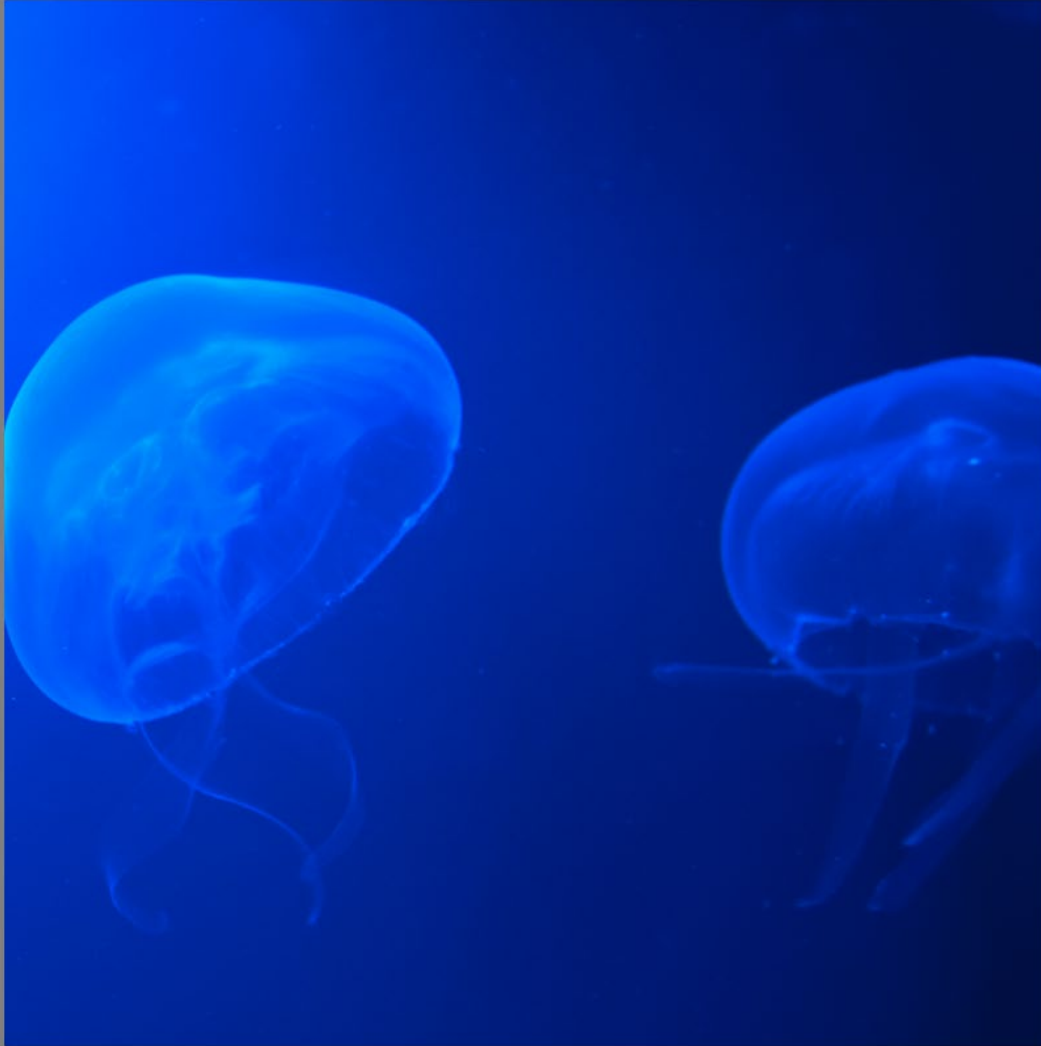
Category judgment experiment



Trial 1



Categories: 1-7



5 - Excellent

4 - Good

3 - Fair

2 - Poor

1 - Bad

NEXT (1/110) >



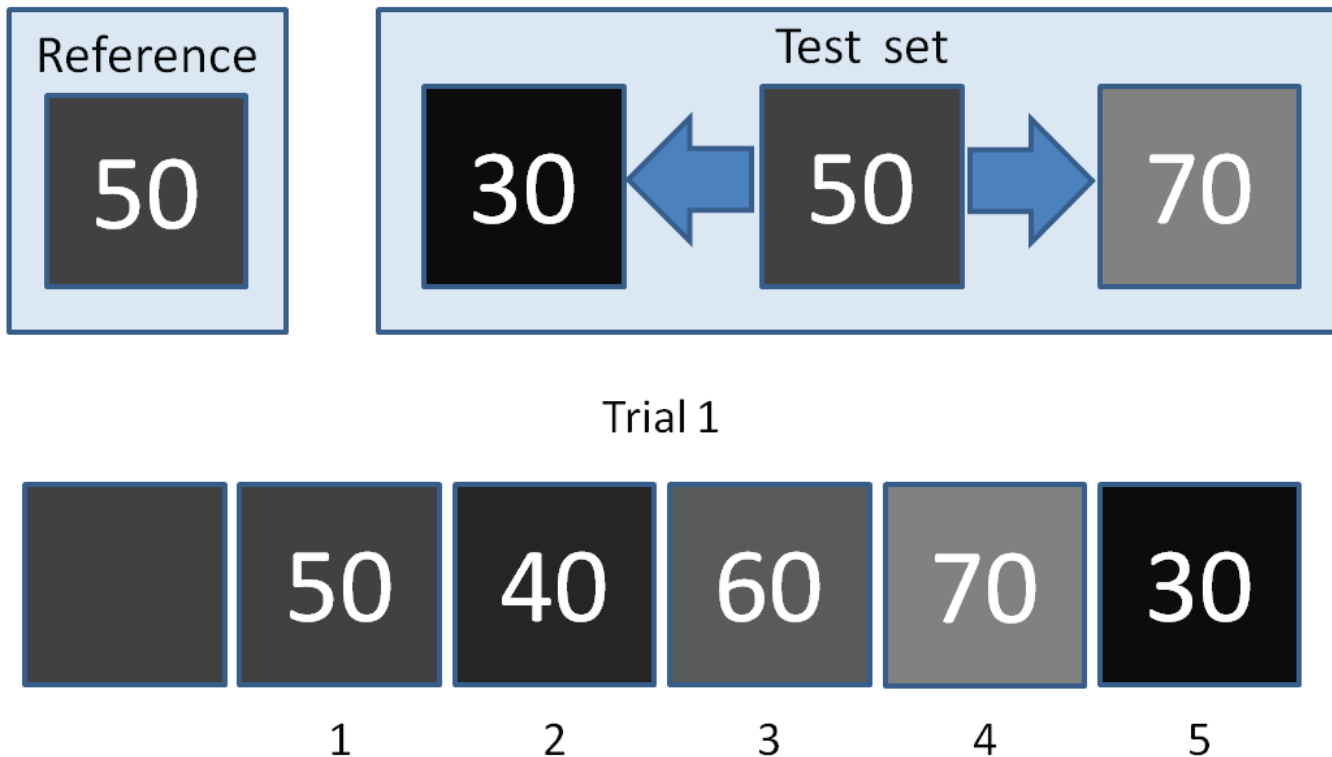
Rank order

- For rank order experiments the observer is presented with a number of images, who is asked to rank them based on a given criterion.
- Rank order can be compared to doing a pair comparison of all Images simultaneously.
- If the number of images is high, the task quickly becomes challenging to the observer.
- However, it is a fast way of judging many images and a simple type of experiment to implement.



Rank order example

- The observer ranks the reproductions from best to worst according to a given criteria.





OTHER METHODS

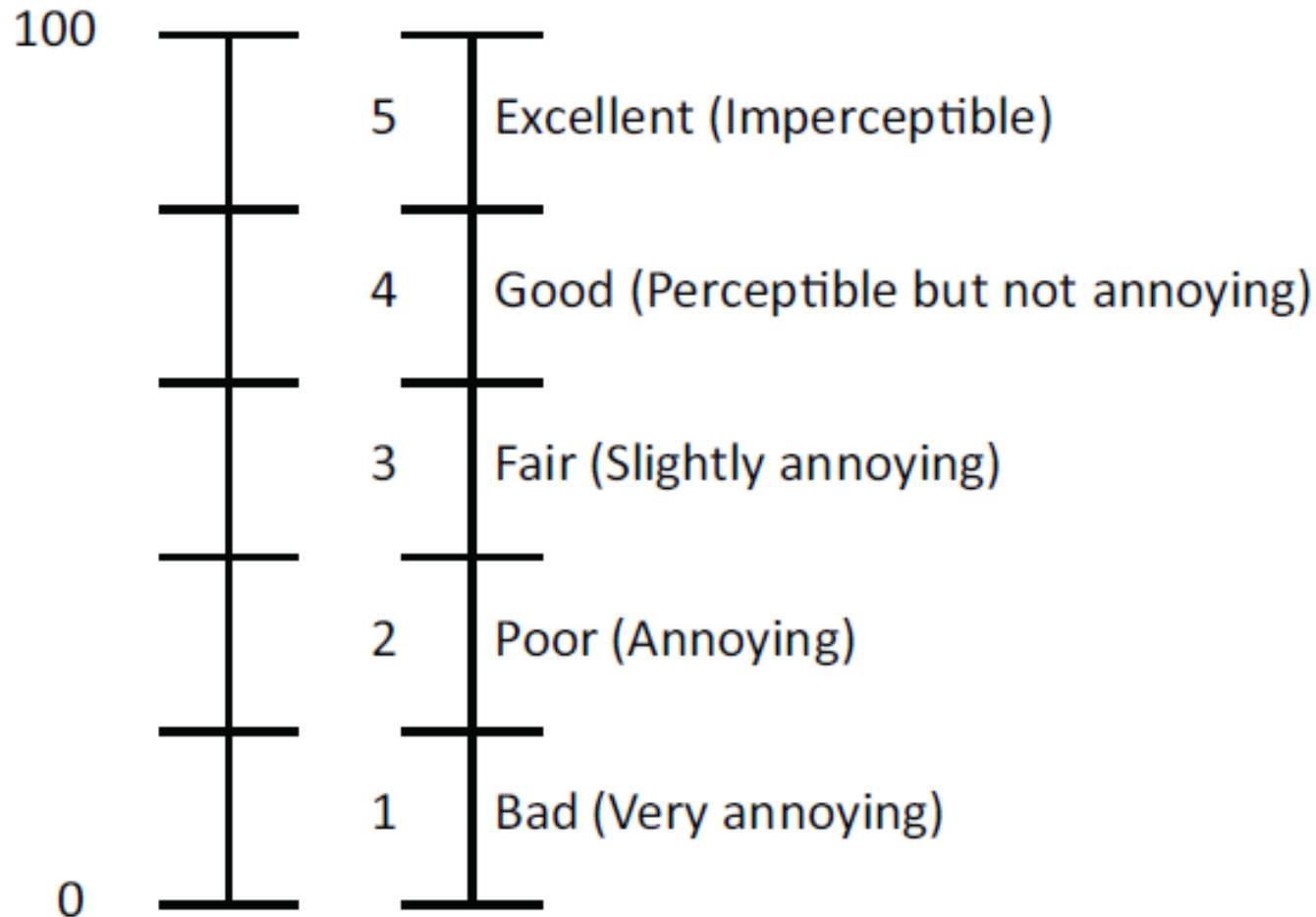


Mean opinion score

- Mean Opinion Score (MOS) is defined by the International Telecommunication Union (ITU) for audio quality, but has also been extensively used for image quality.
- Observers judge the quality from one to five where five is the best quality
 - similar to category judgment.
- Common to have a five point descriptive quality scale (bad, poor, fair, good, excellent)



Quality scale and impairment scale





Mean Opinion Score

- The MOS is the arithmetic mean of the individual scores S

$$MOS = \frac{1}{n} \sum_{i=1}^n S_i$$

- where S_i is the score from one observation, and n is the total number of observations.
- Assumption: data is normally distributed.
- MOS are usually given with a 95% confidence interval.
- MOS can also be used to calculate the Difference Mean Opinion Score (DMOS), where the MOS for the reference is then subtracted from the MOS for the other images.
 - quality of a test image relative to the reference.



Triplet comparison

- ISO 20462-2 presents another way of doing psychophysical experiments, where three images are judged simultaneously.
- Stimuli are presorted in three or more categories, so that only those stimuli falling within certain ranges are subsequently rated against one another.
- ISO proposes the three following categories: "favourable", "acceptable", and "unacceptable".
- The three images in each triplet are not ranked, but rated against a five-category scale.



Number of samples

- The number of sample combinations is fewer than for pair comparison, with the number of comparisons (N) equal to:

$$N = \frac{n(n-1)}{6}$$

- where n is the number of samples. It has been shown that the triplet comparison is almost 50% faster than pair comparison.
- A function is specified for the combinations of samples to be shown:
$$f(i) = 1 + \text{modulo}(i-1, n)$$
 - where modulo indicates the remainder of the division of $(i-1)$ by n .



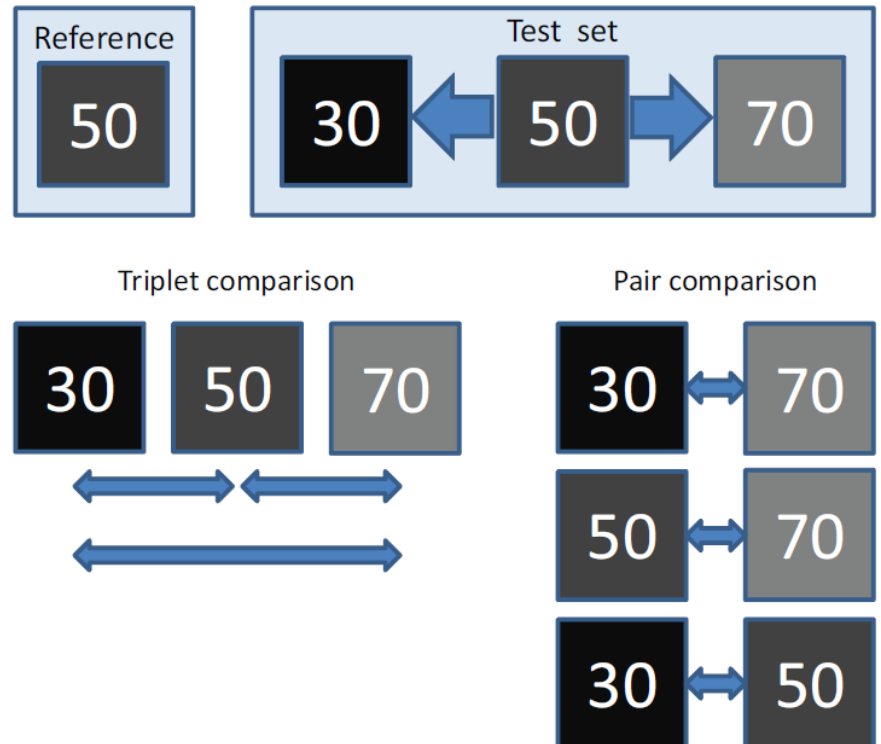
Steps of the triplet comparison method

- 1. Categorical step
- 2. Reduction of the number of samples
- 3. Triplet comparison method step
- 4. Interval scaling



Triplet comparison compared to pair comparison

- One set is needed for triplet comparison, while pair comparison needs three sets.





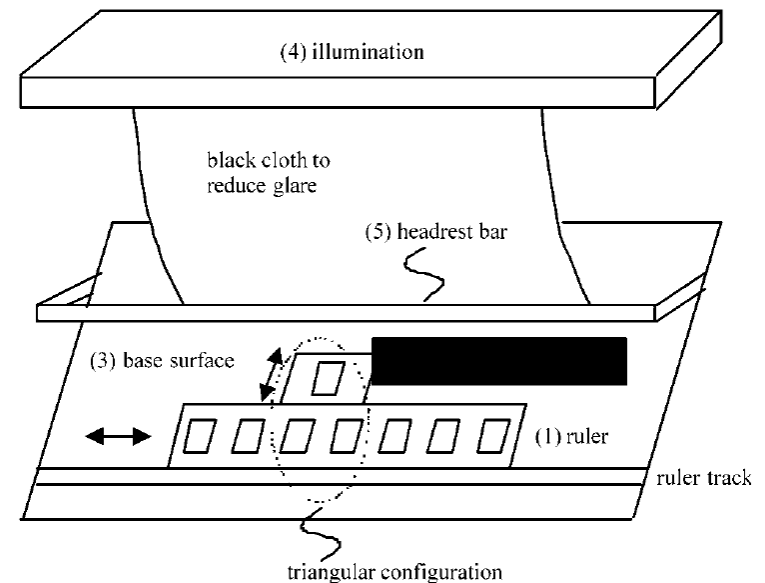
Quality ruler

- The quality ruler uses a Standard Quality Scale (SQS) with a fixed numerical scale so that the stimuli are anchored to a physical standard
 - with one unit corresponding to one JND
 - it also has a zero point.
- The use of reference stimuli results in more reliable results when assessing large sets of stimuli spanning a wide range of quality.



Hardcopy ruler

- The ruler consists of several reference stimuli ordered from highest to lowest quality, spaced at approximately three JNDs, and labeled with an integer.





Process of using the quality ruler

- The observer can slide the ruler back and forth in order to compare the test stimuli with the reference stimuli.
- Since the reference stimuli are labeled 3, 6, 9, and so on, the observer can specify an integer between two reference stimuli, where one integer difference corresponds to approximately one JND.
- The quality ruler method is more suitable for measuring larger quality differences than for example the triplet comparison method.



Advantages and disadvantages

- An average SQS score per stimulus can be easily obtained
 - gives a quality value directly expressed in JNDs.
- Another advantage:
 - scores from different experiments are easily compared.
- However, the process of obtaining SQS scores is complex and delicate, and requires a lot of work in the implementation stage.



Tool for psychometric experiments

- Quickeval www.quickeval.no
- If you want to use it register on the webpage, and send marius.pedersen@ntnu.no an email and I will upgrade your account to «scientist» (allowing you to use all features).



Find experiment
Title, person, type

Test RankOrder version 2
Marius Pedersen

Test RankOrder
Marius Pedersen

Subjective experiment test version 2
Kjetil Grosberghaugen

Subjective experiment test
Kjetil Grosberghaugen

Experiment LMPC
Marius Pedersen

Quality Experiment 3 - SAA&MP version 2
Marius Pedersen

Quality Experiment 4 - SAA&MP version 2
Marius Pedersen

Quality Experiment 2 - SAA&MP
Marius Pedersen

Quality Experiment 1 - SAA&MP
Marius Pedersen

Experiment 2: Translucency Scaling
AKIB JAYED ISLAM

Experiment 1: Transparency scaling
AKIB JAYED ISLAM

Translucency Experiment
AKIB JAYED ISLAM

Experiment 8 version 2
ColourImagingPhD

Experiment 6 version 2
ColourImagingPhD

Experiment 8

Please note that questions are voluntary. By participating you give consent for your participation and on your personal data processing (in case provided).

Please enter your provided user ID

Age

Gender

Do you have confirmed colour vision deficiency?

Do you wear glasses for medical reasons?

Cultural background

Are you working / have you worked in the field of imaging?

START EXPERIMENT



DEMO QUICKEVAL



Link to demos

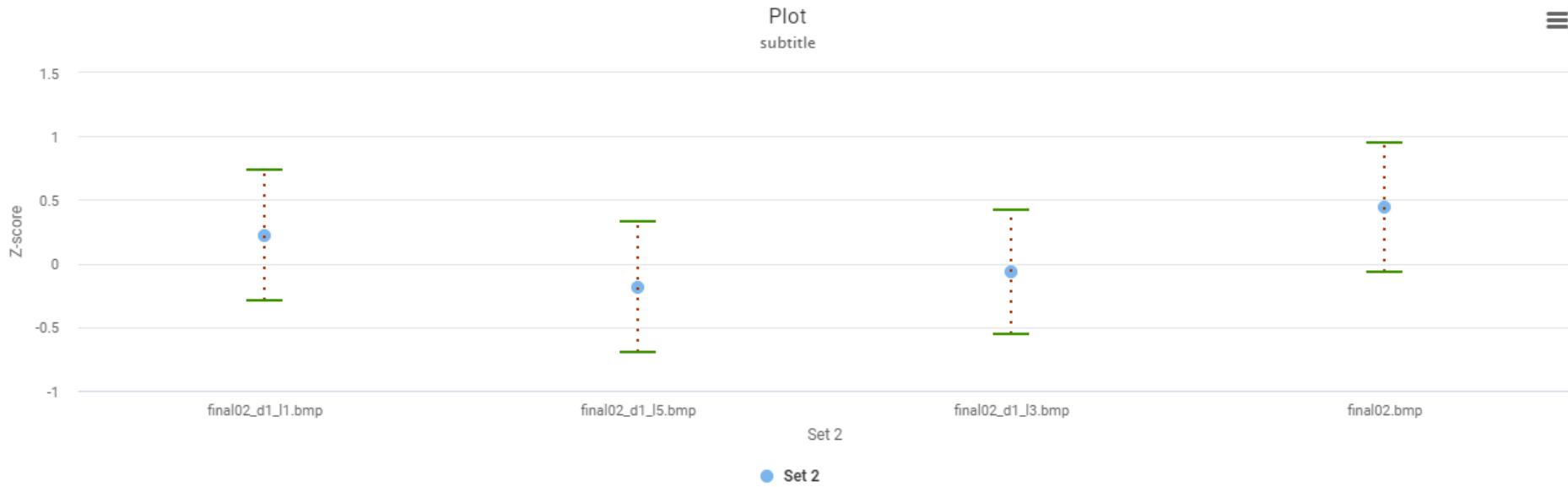
- <https://quickeval.no/observer/159>
- <https://quickeval.no/observer/158>
- <https://quickeval.no/observer/157>



Data analysis

Scatter and errorbar plot


Set 2





Raw data

Set 2

	final02_d1_I1.bmp	final02_d1_I5.bmp	final02_d1_I3.bmp	final02.bmp
final02_d1_I1.bmp		4	6	2
final02_d1_I5.bmp	3		4	2
final02_d1_I3.bmp	2	4		4
final02.bmp	5	5	4	

Z-score

Set 2

Title	Low CI limit	Mean z-score	High CI limit
final02_d1_I1.bmp	-0.289	0.223	0.735
final02_d1_I5.bmp	-0.702	-0.19	0.322
final02_d1_I3.bmp	-0.559	-0.069	0.421
final02.bmp	-0.073	0.439	0.951



EXPERIMENTAL ASPECTS



Number of observers

- The number of observers in subjective experiments is important for
 - the statistical analysis performed on the results.
 - the precision of the statistics.
- When a large number of observers are used
 - The average is more likely to be consistent with "overall quality".
 - the precision of the estimated values increases.
 - In scale values (z-scores) the precision increases with the square root of the number of observers.



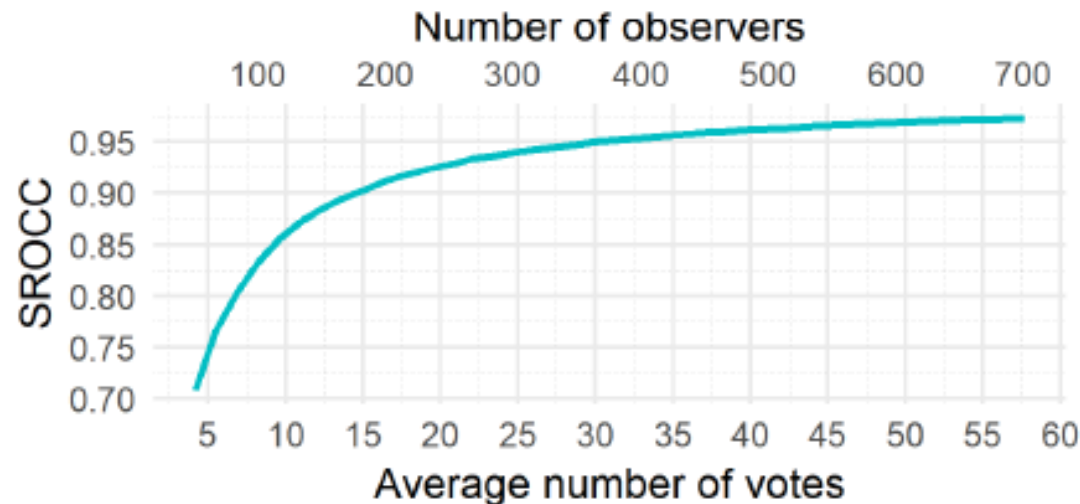
How many observers?

- It is complicated to answer precisely how many observers are required, but guidelines can be found.
 - Engeldrum recommends between 10 and 30 observers are recommended for typical scaling applications.
 - CIE recommends at least 15 observers carrying out pair comparison, category judgment, or ranking experiments with gamut mapped images.
 - 15 observers are also recommended by the ITU for the evaluation of television pictures.
 - Keelan and Urabe recommend a minimum of 10 observers for obtaining relative quality values for JND, and a minimum of 20 observers for obtaining absolute quality scores on the SQS scale.



An example: KonIQ-10k

- KonIQ-10k: 1.2 million quality ratings from 1,459 crowd workers. (web experiment)





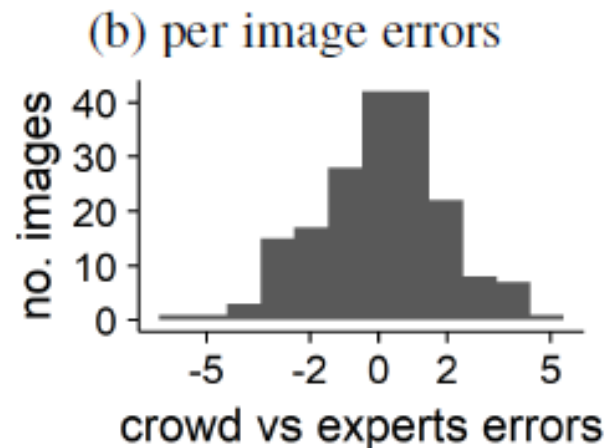
KonIQ-10k: Filtering observers

- Quiz
 - Before starting the actual experiment, workers took a quiz with test questions. Only workers with an accuracy over 70% were eligible to continue.
- Hidden test questions
 - Presented throughout the experiment, below an accuracy were not able to complete.
- Outliers.
 - Workers who had a very low agreement with the preliminary MOS were regarded as outliers.
- Line clickers
 - workers with an unusually high frequency for any single answer choice.



KonIQ-10k: Reliability of the crowd

- To check the reliability of the crowd scores, a comparison was done to scores obtained from 11 experts.
- They regard the expert scores as “ground truth”
- Most images were within the 95% confidence interval of the experts’ scores.





Trade-off

- There is usually a trade-off between the number of stimuli and the number of observers.
- Due to time restrictions, it is often more desirable to have a large number of observers than a large stimuli material.*

* G. Sharma. *Digital Color Imaging Handbook*. CRC Press, Inc., Boca Raton, FL, USA, 2002.



Web-based experiments - recruiting observers

- Larger experiments have been carried out online
 - Qiu and Kheiri <http://hdri.cs.nott.ac.uk/siq/>,
 - Simone et al.
 - Fairchild
 - KonIQ-10k Image Database <http://database.mmsp-kn.de/koniq-10k-database.html>
- Advantage: almost unlimited number of observers can be recruited.
- Disadvantage: If we collect numerous observations even trivial differences can be declared statistically significant.
 - When two stimuli are judged to be different when they are not is commonly referred to as type II errors in statistics, also known as a false negative since it fails to reject a false null hypothesis.
 - Additionally, time and resources will be wasted by collecting too many observations, often for minimal gain.



Observer type

- The expertise and background of the observers will influence the results of experiments.*

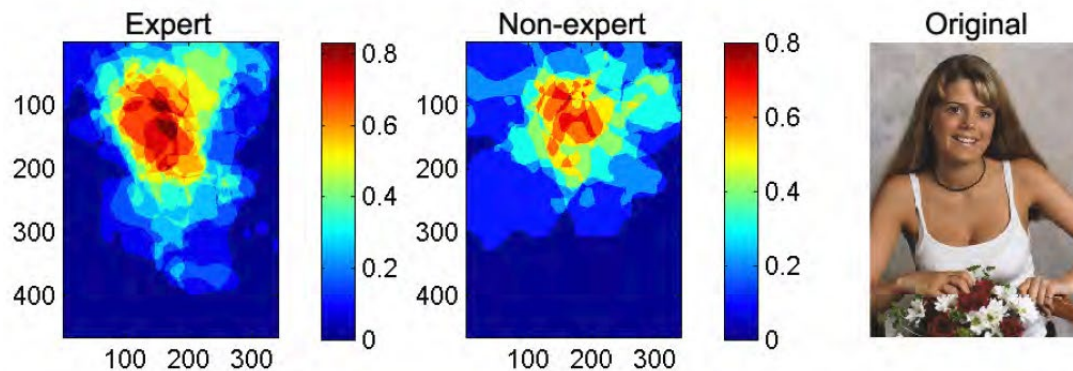
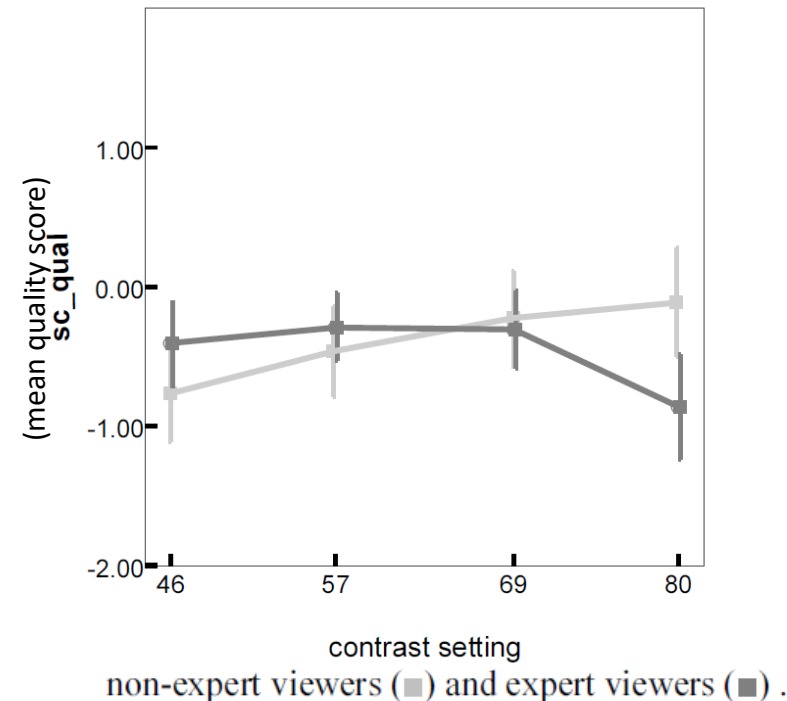


Figure 31: Eye tracker correlation between experts and non-experts in Girl scene.



- G. Deffner, M. Yuasa, and D. Arndt. Evaluation of display image quality: experts vs. non-experts. In *Symp Soc Inf Disp Dig*, volume 25, pages 475–478, 1994.
- Figure on the left from Pedersen, M., 2007. Importance of region-of-interest on image difference metrics (Master's thesis).
- Figure on the right from Ingrid E.J. Heynderickx and Soren Bech "Image quality assessment by expert and non-expert viewers", *Proc. SPIE 4662, Human Vision and Electronic Imaging VII*, (30 May 2002); <https://doi.org/10.1117/12.469509>



Observer type

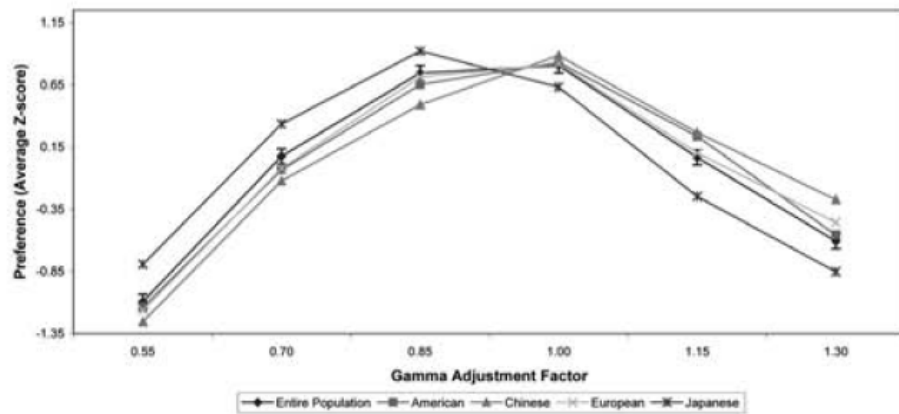
- Observers usually split in two types: experts and naïve.
 - Those considered to be experts usually have experience in judging or evaluating images.
 - The experts can, to a larger degree, distinguish among attributes and they often have a more precise scale than non-experts.
 - In experiments where small differences are needed to be quantified experts are usually more suitable than non-experts.
 - It has been shown that experts have a stronger consensus in their response than non-experts.
 - Experts also look at more regions of smaller and more precise size than non-experts.



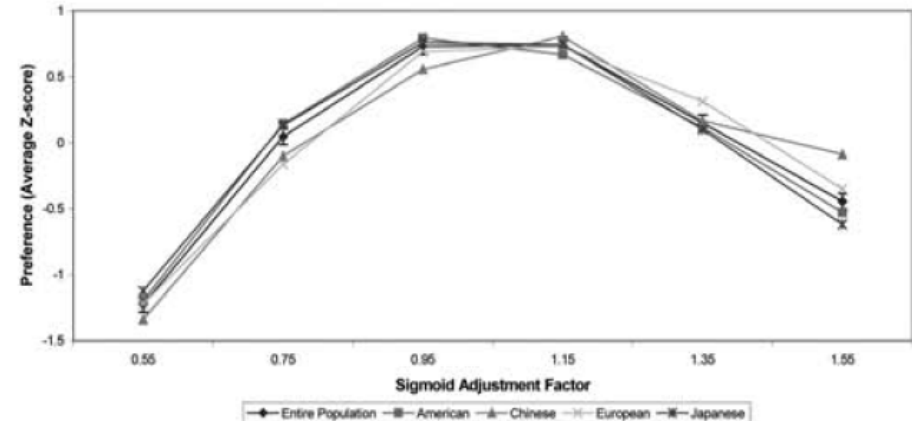
Cultural differences

- Cultural differences have also been found to influence image quality experiments
 - Usually small differences and therefore usually not taken into account.

Gamma Adjustment Dimension



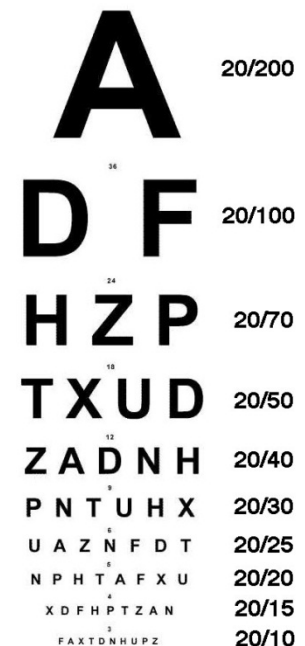
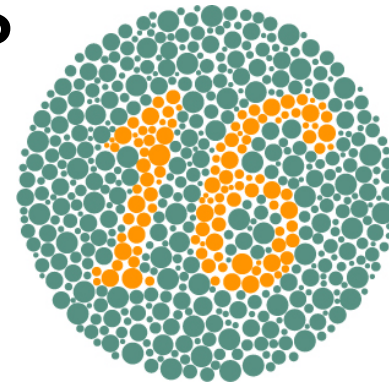
Sigmoid Adjustment Dimension

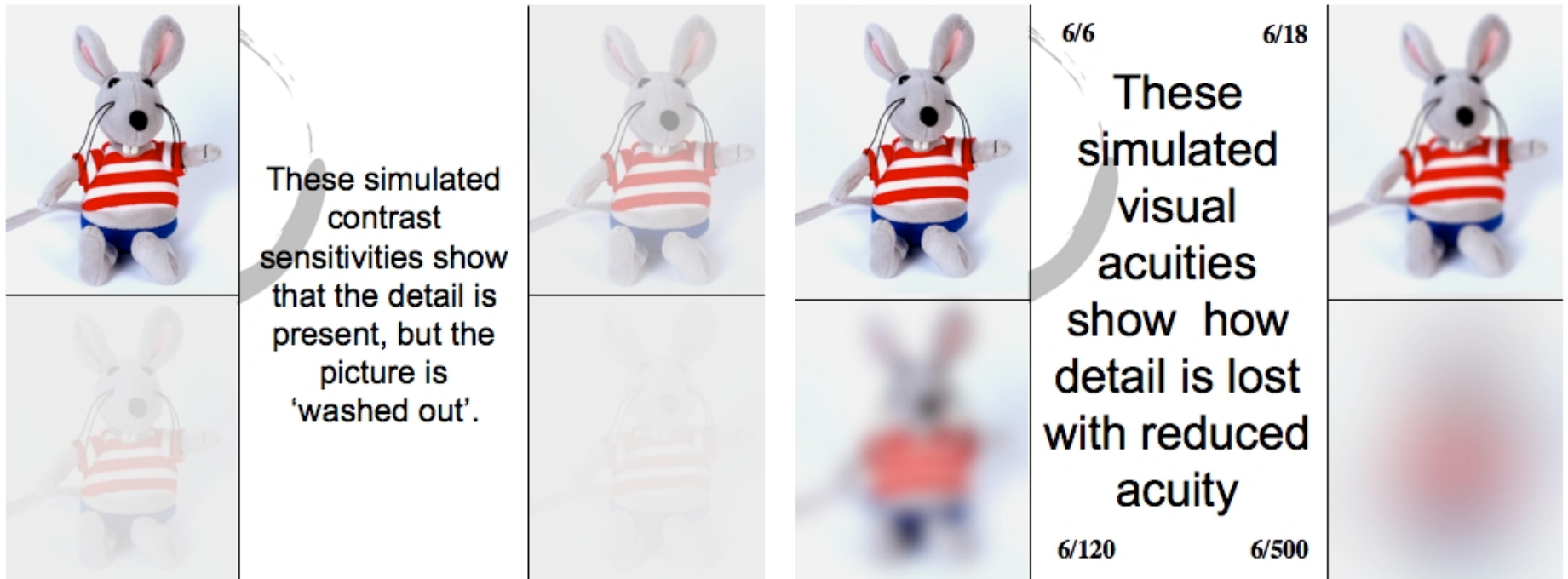




Observer characteristics

- A portion of the population has color vision deficiencies, and thus have a decreased ability to perceive differences between some of the colors that others can distinguish.
 - This will influence how they perceive images, and therefore they are usually not considered as optimal observers
 - Good practice to test all observers for color deficiency using pseudoisochromatic plates, for example with an Ishihara test or Dvorine test.
- The visual acuity of the observers might also influence the results (for example sharpness experiments).
 - Where relevant, conduct visual acuity tests, for example using a Snellen chart.
 - Observers should have 20/20 (or 6/6 vision), normal vision.



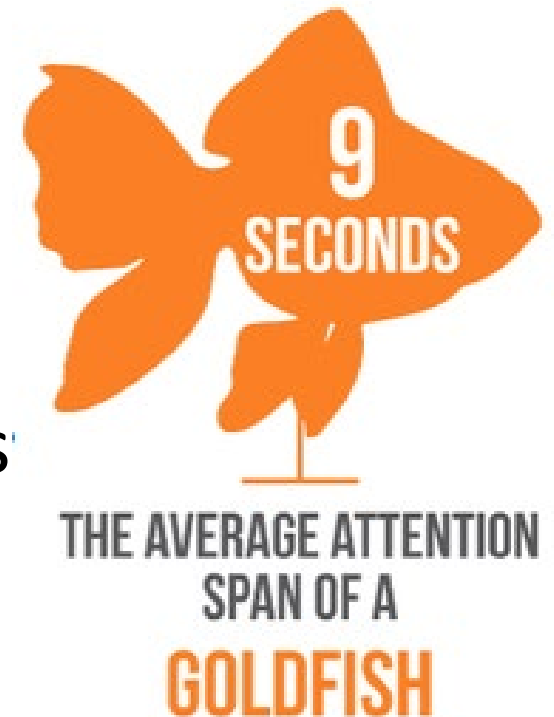






Experiment duration

- Subjective experiments can be long, and they are usually consisting of repetitive tasks.
- *Some* repetitive or monotonous tasks are experienced as boring by *some* people.





Experiment duration

- A high number of stimuli will increase the time spent by observers, and recommendations indicate that the duration of an experiment should be limited to avoid observers fatigue.
 - ITU recommends not more than 30 minutes.
 - Larabi recommends that the median time over the observers should not be more than 45 minutes.
 - In research by Van Der Linde et al. observers showed on average a sustained level of concentration/effort during an eye tracking experiment lasting 1 hour.



Number of stimuli

- Keelan and Urabe state that a minimum of three scenes should be used in order to obtain relative quality values of JND.
- For the SQS scale a minimum of six stimuli is required* .
- ISO 20462-1 reports that the number of test stimuli should be equal to or exceed three scenes, and preferably be equal to or exceed six scenes.
- CIE recommends to use at least one of the obligatory test images specified by CIE, together with at least three additional images, for the evaluation of gamut mapping algorithms.
- Field indicates that between five and ten images are required to evaluate color image quality issues.
- The number of stimuli used is often depending on other aspects, such as the number of observers required, the experimental method, and the precision of the results.

B. W. Keelan and H. Urabe. ISO 20462, a psychophysical image quality measurement standard. In Y. Miyake and D. R. Rasmussen, editors, *Image Quality and System Performance*, volume 5294 of *Proceedings of SPIE*, pages 181–189, San Jose, CA, Jan 2004.

* ISO. ISO 20462-3 photography - psychophysical experimental methods to estimate image quality - part 2: Quality ruler method, jul 2004.

ISO. ISO 20462-1 photography - psychophysical experimental methods to estimate image quality - part 1: Overview of psychophysical elements, jul 2004.

CIE. Guidelines for the evaluation of gamut mapping algorithms. Technical Report ISBN: 3-901-906-26-6, CIE TC8-03, 156:2004.

G. G. Field. Test image design guidelines for color quality evaluation. In *Color Imaging Conference*, pages 194–196, Scottsdale, AZ, Nov 1999. IS&T



Type of stimuli

- There are also recommendations and guidelines for the selection of test stimuli
- Two different types of test stimuli; pictorial images and research images (i.e. test targets).
- Pictorial images most commonly used, since observers are confident in judging them.
 - However, they must be chosen with care since the content might influence the results.
- Research images are artificially created test images, often made to test a specific problem. They have the advantage over pictorial images that they are content free and often have areas that can be read by measuring instruments.
 - Example:
 - Macbeth ColorChecker Color Rendition Chart





Type of stimuli

- There are several guidelines regarding the characteristics of test stimuli.
- They should for example include several levels (low, medium, and high) of several different characteristics (such as tonal distribution, detail level, and saturation).
- It is recommended to test a broad range of images to reveal different quality issues.
- For a complete overview of characteristics of test images see:
 - CIE. Guidelines for the evaluation of gamut mapping algorithms. Technical Report ISBN: 3-901-906-26-6, CIE TC8-03, 156:2004.
 - G. G. Field. Test image design guidelines for color quality evaluation. In *Color Imaging Conference*, pages 194–196, Scottsdale, AZ, Nov 1999. IS&T



Example selecting images

- For a «complete» sheet see «Excel sheet for selecting images for experiments» on Fronter.

Image No.	Attributes Images				Large area of the same color	Neutral gray area	Color transition	Fine details
		Skin color	Sky-blue	Grass				
1	01-picnic-sRGB-16bits-150dpi.tif	1	1	1	1	1	0	1
2	02-WoolBalls-SRGB-16bits-150dpi.tif	0	0	0	0	0	1	0
3	03-Bridge.tif	0	0	0	1	0	0	0
4	04-Sea.tif	0	1	0	1	0	0	0
5	05-Firehydrant.tif	0	0	0	0	1	0	0
6	06-Roses.tif	0	0	0	0	0	0	0,5
7	07-JellyFish.tif	0	0	0	1	0	1	0
8	08-Arctic_Bloom.tif	0	0	0	0	0	0,5	0
9	09-Color-Models.tif	0	0	0	0	0	0	0
10	10-DigiQ_Studio2.tif	1	0	0	1	0	0	1
10	Total	2	2	1	5	2	2,5	2,5



Standard test images

- CIE recommendation for gamut mapping
 - CIE. Guidelines for the evaluation of gamut mapping algorithms. Technical Report ISBN: 3-901-906-26-6, CIE TC8-03, 156:2004.
 - gives one image as obligatory for the evaluation of gamut mapping; the ski image
 - In addition to this image ten graphics (Canon Development Americas computer graphics images) and three other pictorial images (Sony sRGB standard images) are recommended.





Standard test images

- ISO 12640 test images
 - different sets of test images for evaluation of different processes.
- The image set in ISO 12640-1 consists of 8 natural and 10 synthetic images.
 - The image set was developed for comparison of color output systems such as printing, proofing, and color facsimile, and therefore they are in CMYK format.



The eight natural images found in ISO 12640-1



Standard test images

- ISO 12640-2 defines XYZ/sRGB standard color image data.
 - The image set consists of 15 color images, encoded as both 16-bit CIEXYZ and 8-bit RGB digital data, for the evaluation of quality changes.
 - The set has eight natural images and seven synthetic images.
 - This set is optimized for viewing on a reference sRGB display in the reference sRGB viewing environment, and relative to CIE standard illuminant D65.



The eight natural images found in ISO 12640-2



Standard test images

- ISO 12640-3 provides a test image data set with a large color gamut related to illuminant D50.
 - The 18 test images, 8 natural and 10 synthetic, are encoded as 16-bit CIELAB digital data. The natural images are 16 bits per channel, while the synthetic images are 8 bits per channel.



The eight natural images found in ISO 12640-3



Standard test images

- ISO 12640-4 specifies a standard set of wide gamut display-referred color images.
 - These are encoded as 16-bit Adobe RGB digital data. These images compliment the existing images of ISO 12640-2 which are based on the sRGB display gamut.
 - These test images have a larger color gamut than sRGB, and these images will require much less aggressive color re-rendering going to print than sRGB encoded images



Standard test images

- Canon Development Americas computer graphics images
 - ten computer graphics, which are recommended by CIE for the evaluation of gamut mapping algorithms.





Standard test images

- Sony sRGB standard images
 - The Sony standard images are three photographs provided by Sony for CIE for the evaluation of gamut mapping, being a part of the recommended images from CIE.
 - The set consists of two studio scenes, representing a portrait and a party image, and an outdoor image with a picnic theme.

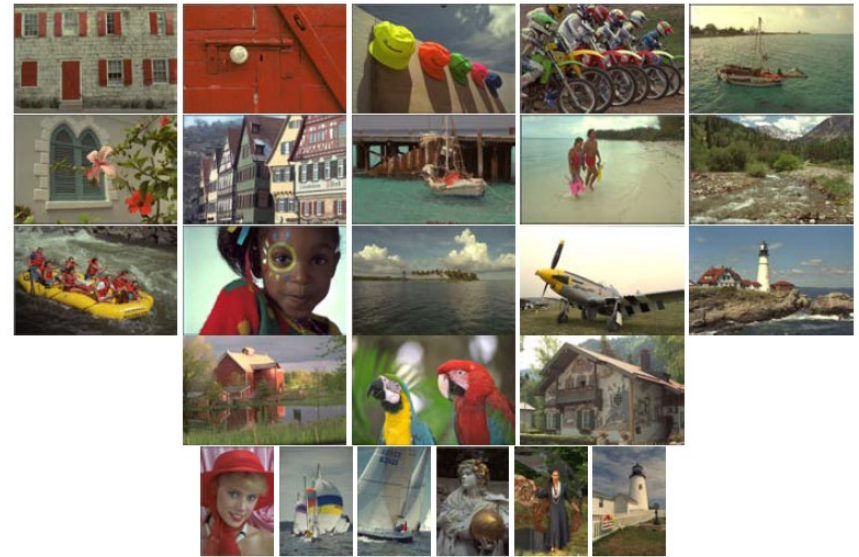


Sony sRGB standard images.



Standard test images

- Kodak lossless true color image suite
 - A set of 24 images (22 outdoor and two studio images)
 - This image suite has become a very popular suite for evaluating different aspects of imaging.
 - However, these images were made when digital images was a new concept. Therefore, their quality is limited, and they are probably not comparable to current digital photos.



Kodak lossless true color image suite.



Standard test images

- DigiQ
 - Halonen et al. proposed the DigiQ image suite, which consists of three test images for print quality evaluation
 - In designing the images, aspects taken into consideration included a recognizable theme, memory colors, shapes, surface materials, detail information and saliency.
 - The images are 16-bit TIFF images in Adobe RGB format, with a resolution of 360 DPI and a print size of 100×150mm.



DigiQ image suite.



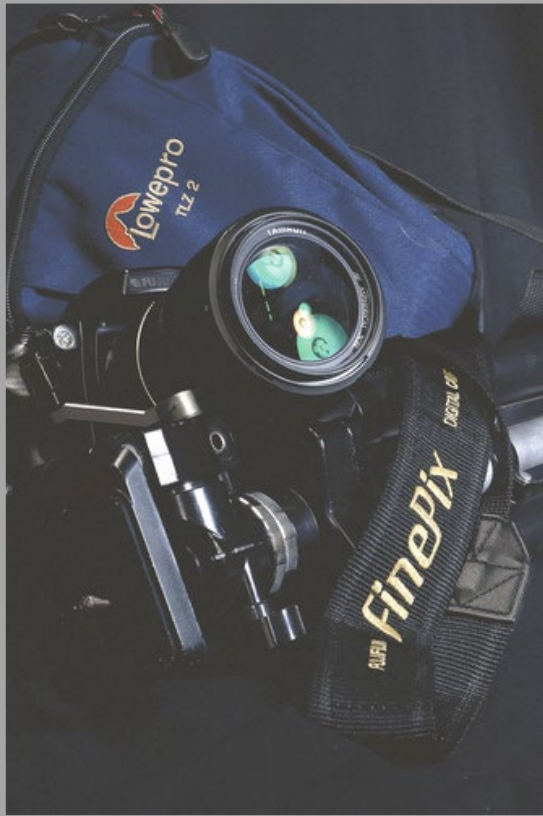
Marking stimuli

- Marking the stimuli is very important.
- One should use «codes» that the observer cannot «interpret».
 - Avoid using A,B,C,D,... And 1,2,3,4,5 since observers can rank/judge them in ascending/descending order.
- Where to mark depends on the experiment (in case of hardcopies).
 - If the experimenter records the data: mark on the back of the image.
 - If the observer records the data: mark on the front.
 - Use for example a combination of letters; BD, TU, IJ, etc...
 - The code identifies the distortion/algorithm: A-E = distortion 1, F-L distortion 3, M-R distortion 4, etc.



Example of marking

Gamut mapping algorithm 1



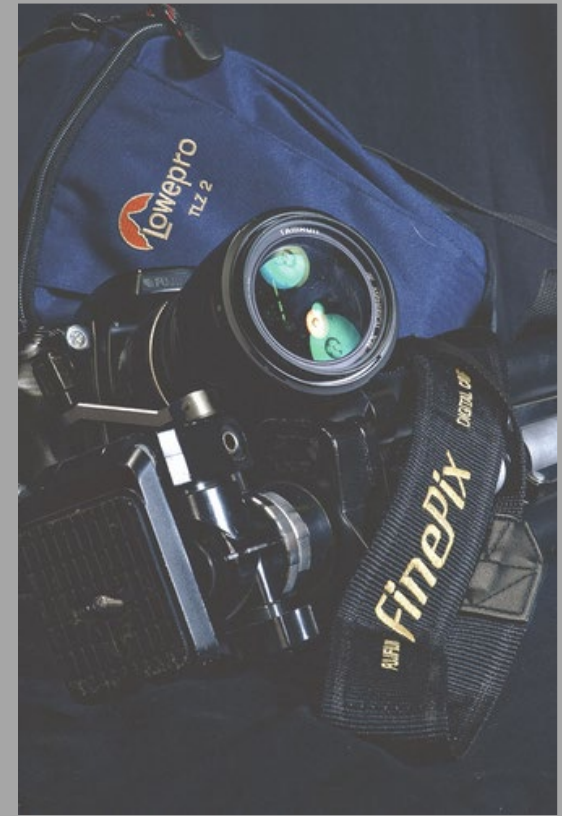
EC

Gamut mapping algorithm 2



HG

Gamut mapping algorithm 3



RO



Viewing conditions

- Controlled and uncontrolled environments
 - Controlled experiments are carried out in a laboratory where the viewing conditions meet standards (such as described by CIE),
 - Uncontrolled experiments can be carried out in the field or on the web.
- The most important reason not to carry out uncontrolled experiments has been that the environments are not standardized
 - However research have shown small differences between controlled and uncontrolled experiments.*

CIE. Guidelines for the evaluation of gamut mapping algorithms. Technical Report ISBN: 3-901-906-26-6, CIE TC8-03, 156:2004

* I. Sparow, Z. Baranczuk, T. Stamm, and P. Zolliker. Web-based psychometric evaluation of image quality. In S. P. Farnand and F. Gaykema, editors, *Image Quality and System Performance VI*, volume 7242 of *Proceedings of SPIE*, page 72420A, San Jose, CA, Jan 2009.



Viewing conditions - distance

- The perceived quality of an image is related to the distance at which the observers view it.
- A change in the viewing distance will change the information that our HVS can perceive or detect, and will therefore also change the quality.
- Thus the viewing distance should be kept constant when conducting experiments
- In experiments where the viewing distance is not controlled, the variability of the results might increase and the results spread.
- In specific experimental methods, such as the quality ruler, a headrest bar is used to fix the distance from the stimuli to the observer.



Viewing conditions - illumination

- The illumination under which the stimulus is viewed has an impact on the perceived quality.
- The intensity of the illumination (illuminance), which is measured in lumens per square meter or lux, influences the responses from human observers.
- Controlling the illuminance is important since an increase in illuminance will increase the colorfulness (Hunt effect), perceived hue can also change with a change in luminance.
- The spectral power distribution of the illumination (Correlated Color Temperature (CCT)) is important. Color appearance changes with the spectral distribution of the illumination.
 - When conducting experiments involving both monitors and printed stimuli the CIE recommends using a monitor with a D65 white point while the prints should be viewed under D50.
- The geometry of the illumination could have an impact on the results, for example when viewing glossy prints. It is recommended to illuminate the test stimuli from an angle of 45° from the normal to the sample, and view the sample normal to the surface*.
 - Commonly achieved by the use of a normalized viewing booth.
- CIE gives guidelines for different conditions depending on the task at hand. For printed samples and transparencies see ISO 3664:2009 and on experiment design ISO 20462-1:2004.



Presentation of stimuli – on a monitor

- The area immediately surrounding the displayed image and its border shall be neutral, preferably dark grey or black to minimize flare, and of approximately the same chromaticity as the white point of the monitor.
- The border should be white if comparisons are to be made to reflective hardcopy and dark grey if comparisons are to be made to transparencies. It should be mid-grey otherwise.
- The monitor shall be situated such that there are no strongly coloured areas (including clothing) directly in the field of view or which may cause reflections in the monitor.
- Ideally all walls, floors and furniture in the field of view should be mid-grey and free of any posters, pictures, or any other object which may affect the vision of the viewer.
- All sources of glare should be avoided since they degrade the quality of the image.
- The monitor shall be situated such that no illumination sources such as unshielded lamps or windows are directly in the field of view or are causing reflections from the surface of the monitor.



Presentation of stimuli – printed samples

- For reflective prints the surround and backing to the sample shall be neutral and matt, and the unprinted substrate should extend beyond the image by 12 mm – 24 mm on all sides.
- Also see ISO 3664:1999 *Viewing conditions - Prints, transparencies and substrates for graphic arts technology and photography*



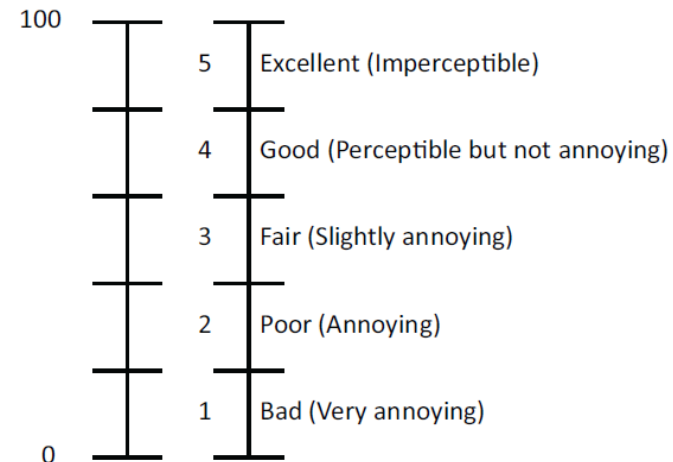
Instructions

- The instructions given to the observers are among of the most important aspects when designing an experiment.
- There are several considerations to make when writing the instructions:
 - What are the observers going to judge?
 - Overall quality, a specific attribute (sharpness, an artifact, others)?
 - in what context
 - Office documents, official letters, private context etc.
 - what are the criteria for the judgment
 - are there any definitions needed for the observers to carry out their task?



Number of categories

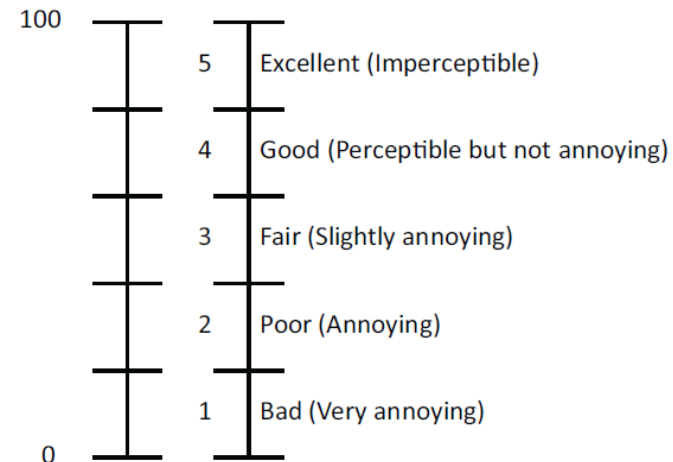
- Category judgement requires a number of categories.
- How many should be used?
 - 2,3,4,5,6,7,8,9,10,11 or more?
- Should you have a mid-point?
 - A mid-point allow observers to be «neutral»
- Are each category equal in «size» and with the same «distance»?
 - Recommendation with equal intervals (Engeldrum, 2000)





Scale

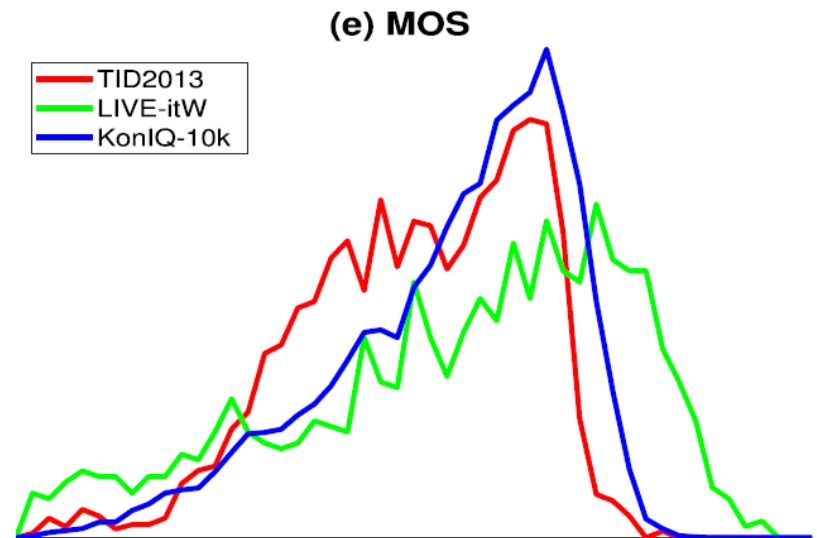
- Some criticism has been made to the ITU-R BT500.7 scale.
 - Not having equal intervals (Zwrick 1984 and Jones and McManus 1986).
 - «Poor» and «bad» being almost the same. (Jones 1986)
 - Inexperienced observers might have problems using the scale.





Category span

- Data analysis require some «confusion» (as Engeldrum puts it). If all observers judge images to be in the same category, one cannot use the law of comparative judgement.
- If all images are judged to be equal, your categories are «too wide» (or no difference the images).



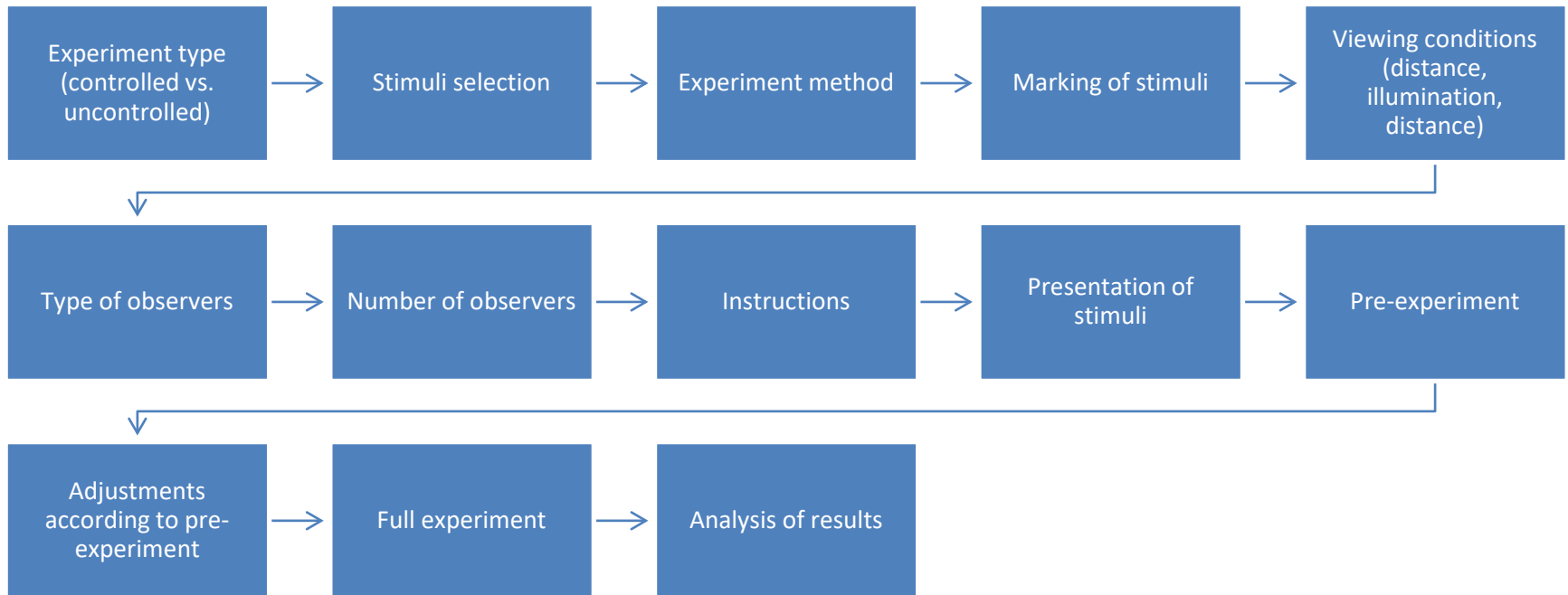


Before commencing the experiment

- Trial experiment
 - 1-2 observers to test the experimental setup
 - Did they understand what to do?
 - Any problems?
- Did you get the data you expected?
- Adapt the experiment to fix any problems, then do a full-scale experiment.



Flow chart



Please note that the order of the boxes can change, and is dependent on the experiment.



IMAGE QUALITY DATABASES



LIVE

- JPEG compresses images (169 images).
- JPEG2000 compressed images (175 images)
- Gaussian blur (145 images)
- White noise (145 images)
- Bit errors in JPEG2000 bit stream (145 images)
- 5 categories: ``Bad'', ``Poor'', ``Fair'', ``Good'' and ``Excellent''.
- 20-29 observers per image



irccyn/ivc

- 10 original images
- 235 distorted images were generated from 4 different processing:
 - JPEG
 - JPEG2000
 - LAR coding
 - Blurring
- Subjective evaluations were made at viewing distance of 6 times the screen height
- DSIS (Double Stimulus Impairment Scale) method with 5 categories and 15 observers.
- Distortions for each processing and each image have been optimised in order to uniformly cover the subjective scale.





TID2008 database

- 25 reference images
- 17 types of distortion over 4 levels
- 1700 images in total
- Subjective scores from 654 observers
- No defined viewing distance!
- Commonly used by many.
- More information in
 - N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, and F. Battisti. Color image database for evaluation of image quality metrics. In *International Workshop on Multimedia Signal Processing*, pages 403–408, Cairns, Queensland, Australia, Oct 2008. 25/09/12:<http://www.ponomarenko.info/tid2008.htm>





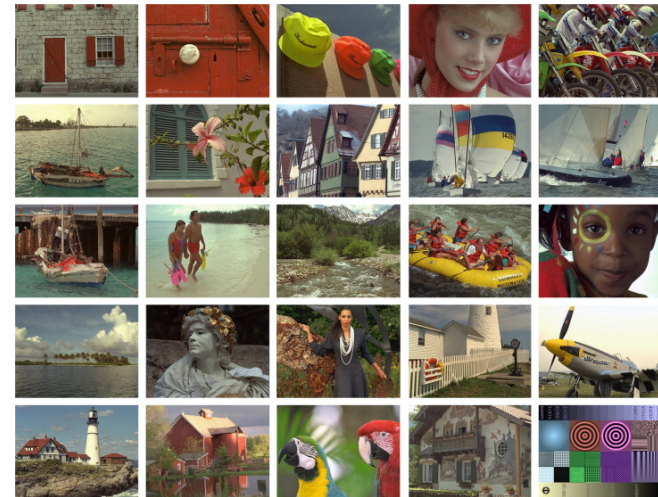
Overview of the distortions

	Type of distortion	Dataset							Full
		Noise	Noise2	Safe	Hard	Simple	Exotic	Exotic2	
1	Additive Gaussian noise	+	+	+	-	+	-	-	+
2	Noise in color components	-	+	-	-	-	-	-	+
3	Spatially correlated noise	+	+	+	+	-	-	-	+
4	Masked noise	-	+	-	+	-	-	-	+
5	High frequency noise	+	+	+	-	-	-	-	+
6	Impulse noise	+	+	+	-	-	-	-	+
7	Quantization noise	+	+	-	+	-	-	-	+
8	Gaussian blur	+	+	+	+	+	-	-	+
9	Image denoising	+	-	-	+	-	-	-	+
10	JPEG compression	-	-	+	-	+	-	-	+
11	JPEG2000 compression	-	-	+	-	+	-	-	+
12	JPEG transmission errors	-	-	-	+	-	-	+	+
13	JPEG2000 transmission errors	-	-	-	+	-	-	+	+
14	Non eccentricity pattern noise	-	-	-	+	-	+	+	+
15	Local block-wise distortion	-	-	-	-	-	+	+	+
16	Mean shift	-	-	-	-	-	+	+	+
17	Contrast change	-	-	-	-	-	+	+	+



TID2013

- 24 distortions
- 5 levels
- 3000 distorted images
- 971 observers (from 5 countries)





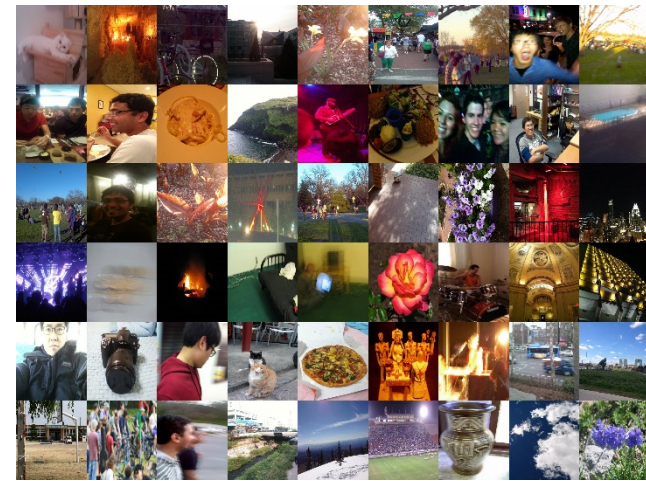
Distortions

No	Type of distortion (four levels for each distortion)
1	Additive Gaussian noise
2	Additive noise in color components is more intensive than additive noise in the luminance component
3	Spatially correlated noise
4	Masked noise
5	High frequency noise
6	Impulse noise
7	Quantization noise
8	Gaussian blur
9	Image denoising
10	JPEG compression
11	JPEG2000 compression
12	JPEG transmission errors
13	JPEG2000 transmission errors
14	Non eccentricity pattern noise
15	Local block-wise distortions of different intensity
16	Mean shift (intensity shift)
17	Contrast change
18	Change of color saturation
19	Multiplicative Gaussian noise
20	Comfort noise
21	Lossy compression of noisy images
22	Image color quantization with dither
23	Chromatic aberrations
24	Sparse sampling and reconstruction



LIVE In the Wild Image Quality Challenge Database

- 1,162 images
- 350,000 opinion scores evaluated by over 8100 unique human observers.
- Each image was viewed and rated on a continuous quality scale by an average of 175 unique subjects.
- Carried out as an online crowdsourcing experiment.





1/7

Bad Poor Fair Good Excellent

Next Image

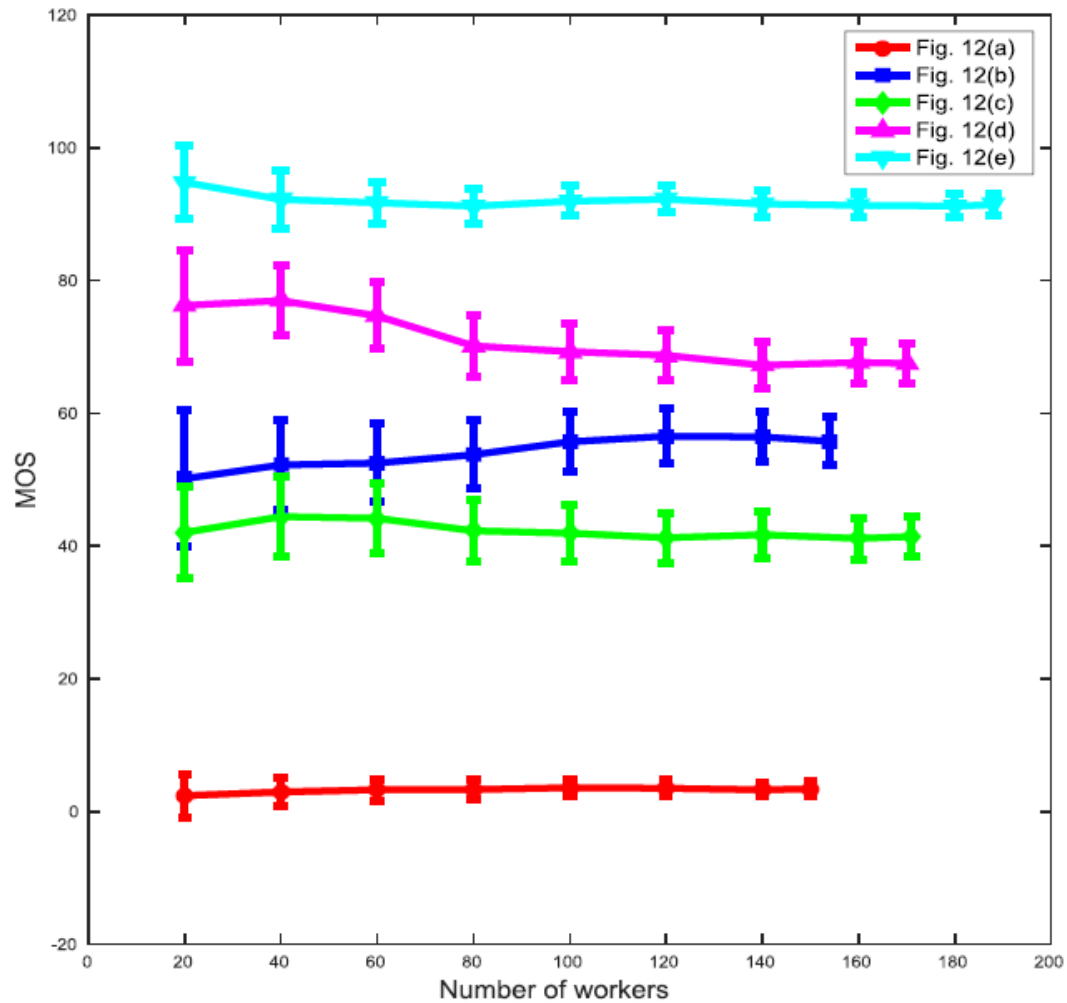


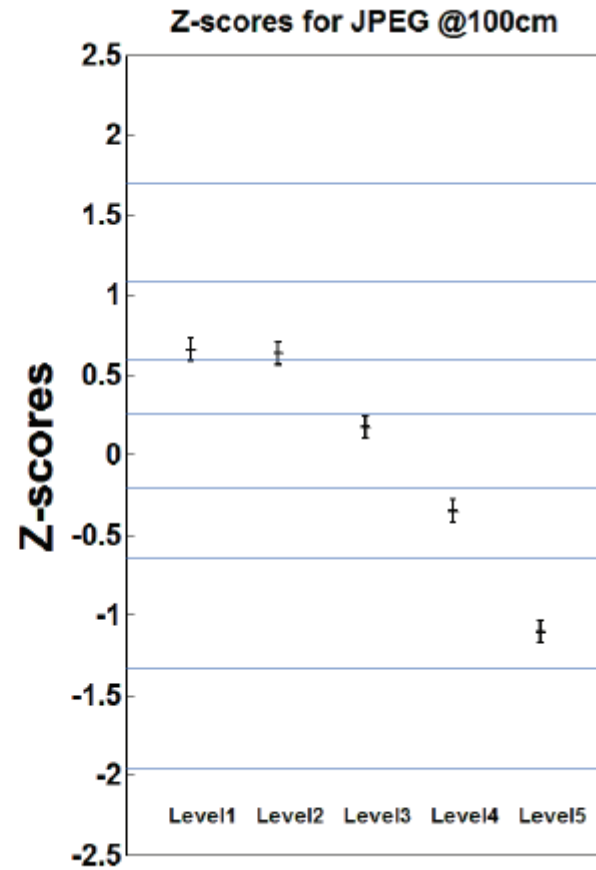
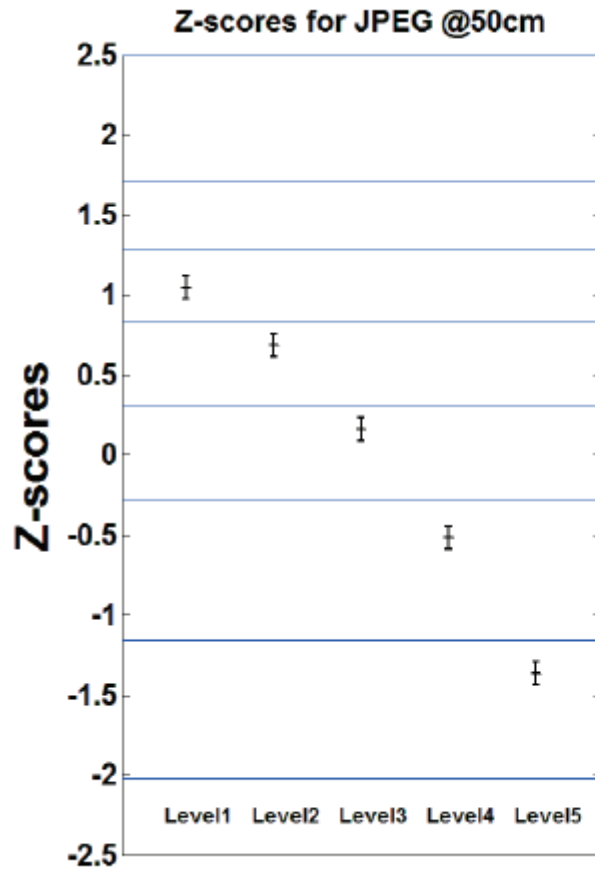
Fig. 15. MOS plotted against the number of workers who viewed and rated the images shown in Fig. 12.



CID:IQ

- Category judgement – 9 point scale
- 5 distortions at 5 levels
 - JPEG, JPEG2000, noise, blurring and gamut mapping
- Experiment done at two different viewing distances.
- Available on <http://www.colourlab.no/cid>.





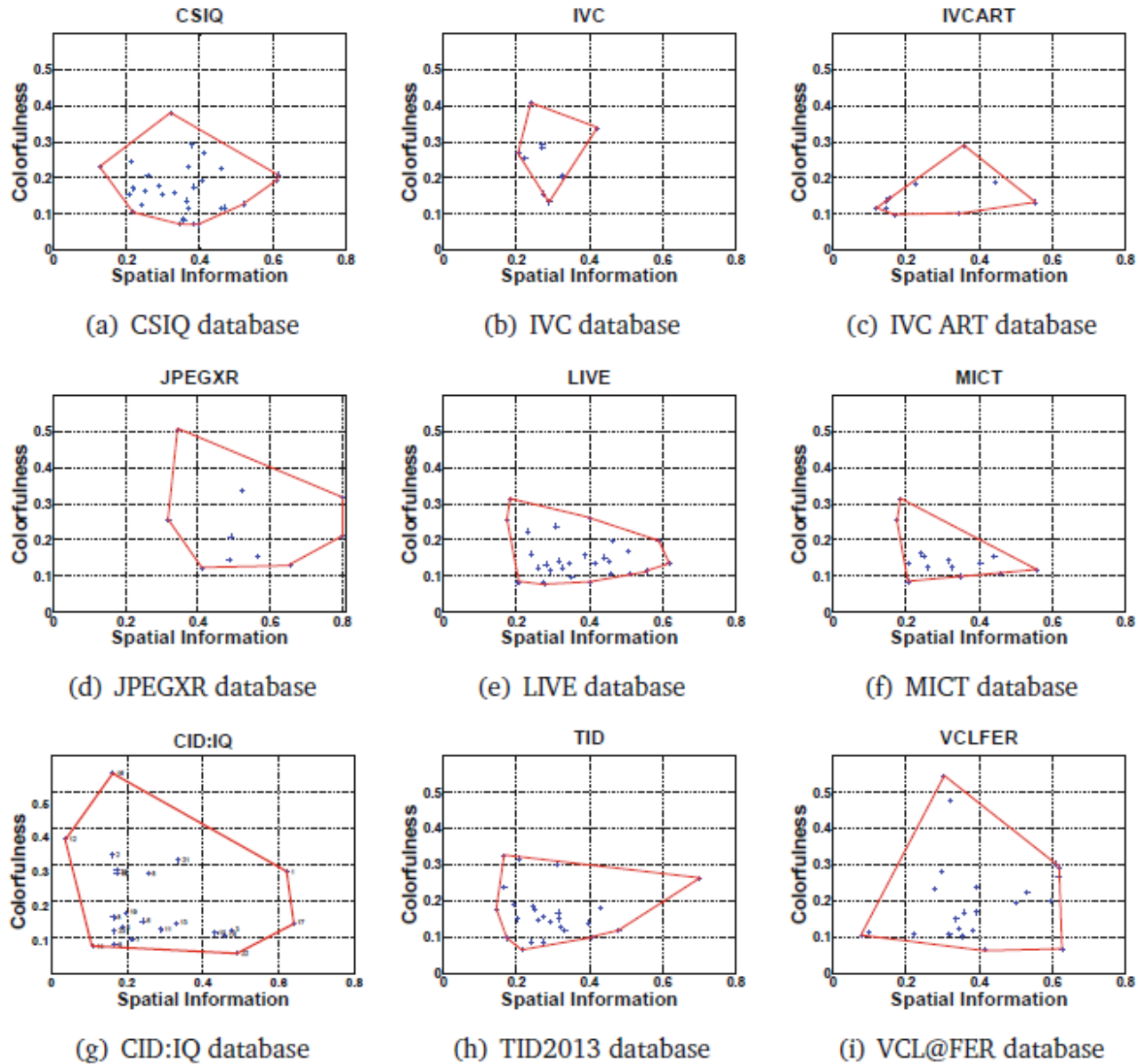


Fig. 2. Comparison of SI vs. CF results between image quality databases



CID2013

- The CID2013 database consists of 480 images captured by 79 imaging devices (mobile phones, DSC, DSLR) in six image sets.
- Complete raw data and background information from the naïve observers
- Link:
<https://zenodo.org/record/2647033#.YC4aJmhnJZc>

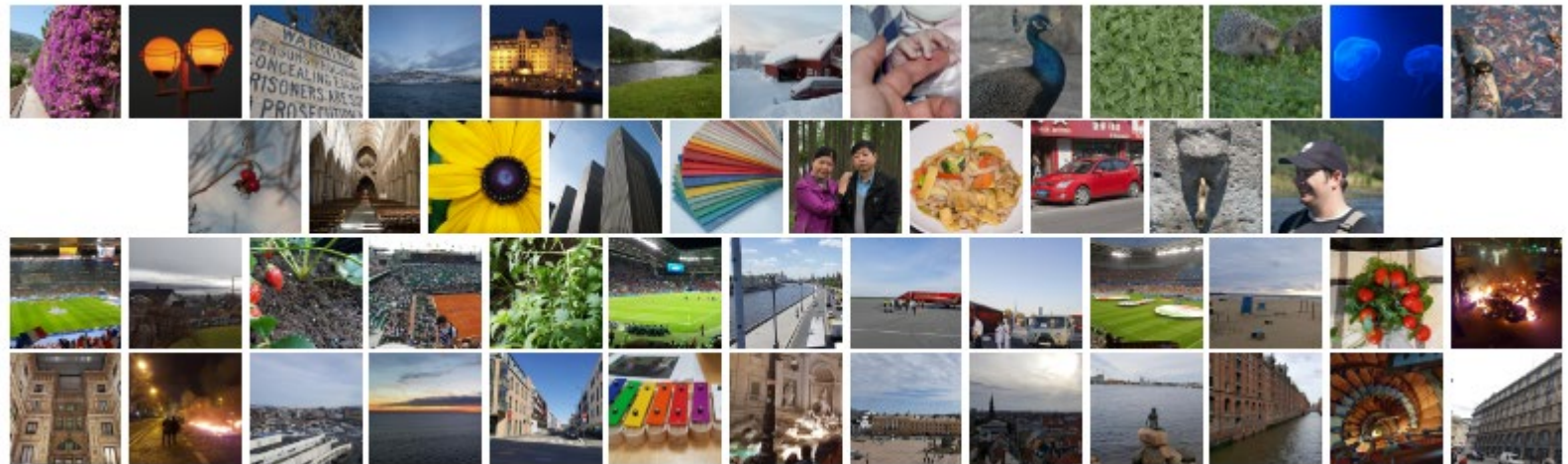






Colourlab Image Database: geometric distortions

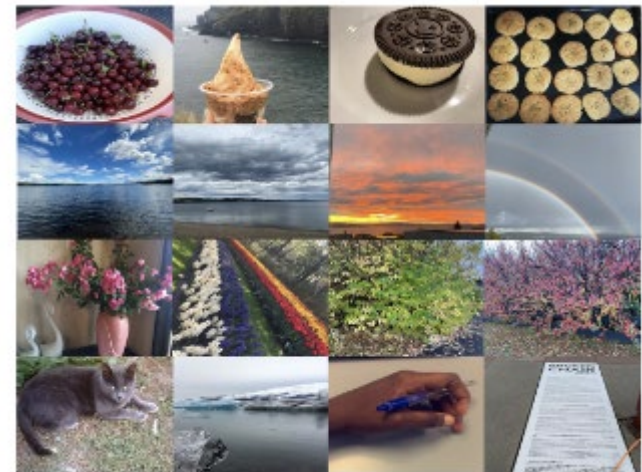
- Geometric distortions: Seam carving, lens distortions and rotation.
- Total of 392 images.
- Category judgement with 5 categories.
- Online with 33 different observers, one average 15 observers per image.





Colourlab Enhanced Image Database

- 16 images
- Enhanced by 5 users through Instagram using brightness, contrast, saturation, sharpness, and warmth.
- 45 unique observers, on average 15 per image.
- Forced-choice pair comparison.





Existing image quality databases

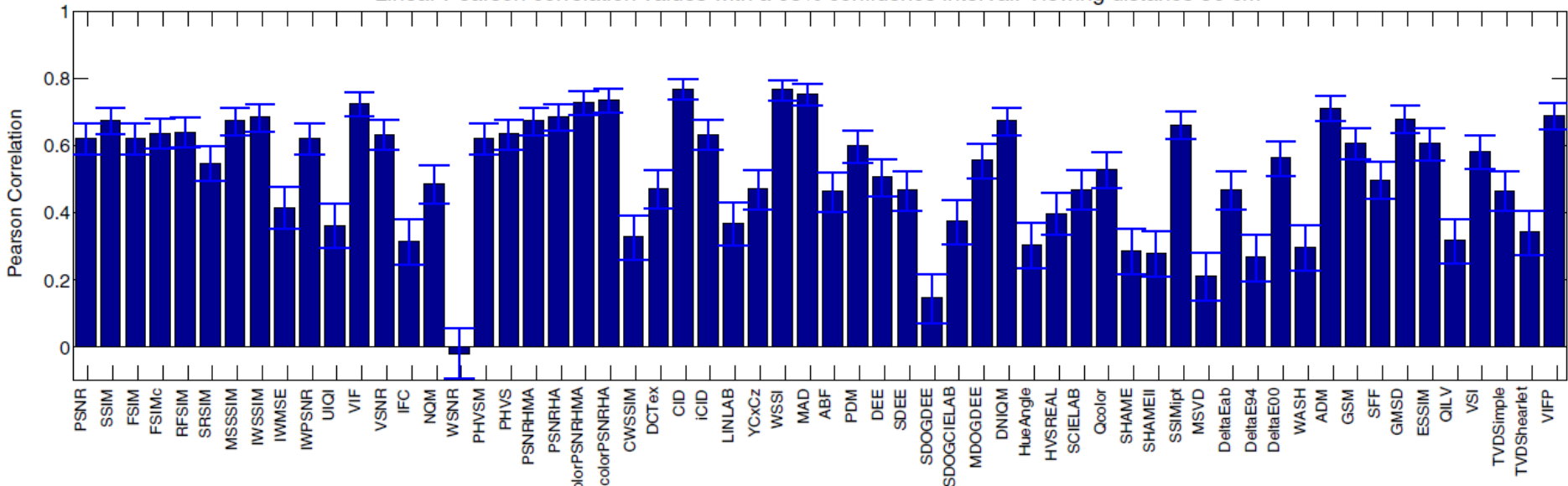
Name	CID:IQ	TID		LIVE (Release 2)	Toyama	CPIQ		IRCCyN/IVC						VCL@FER	VAIQ	TUD		JPEGXR	HTI	IBBI	MMSP 3D	A57	WIQ		
		TID2013	TID2008			CSIQ	DRIQ	IVC	Watermarking							3D image	Art image							TUD1	TUD2
									Enrico	Broken Arrows	Fourier Subband	Meerwald													
Year	2014	2013	2008	2006	2008	2010	2012	2005	2007	2009	2009	2009	2008	2009	2011	2009	2010	2010	2011	2011	2011	2010	2007	2009	
Color or Gray	Color	Color	Color	Color	Color	Color	Color	Color	Gray	Gray	Gray	Gray	Color	Color	Color	Color	Color	Color	Color	Color	Color	Color	Color	Gray	Gray
Number of reference image	23	25	25	29	14	30	26	10	5	10	5	12	6	8	23	42	8	11	10	12	12	9	3	7	
Number of distortion type	6	24	17	5	5	6	3	5	10	2	6	2	15	3	4		1	1	1	1	1		6	1	
Number of distortion level	5	5	4	X	6	5		5	2	6	7	5	1	5	6		2	4	6	5	5		3	X	
Number of image	690	3000	1725	808	196	896	104	195	105	130	315	132	96	120	575	42	16	55	60	60	60	60	54	80	
Number of observer	17	985	838	29	16	35	9	15	16	17	7	14	No Specif y	20	118	15	12	20	No Specif y	18	18	20	7	30	



Image quality metrics

- These databases have been created to evaluate the performance of image quality metrics.

Linear Pearson correlation values with a 95% confidence interval. Viewing distance 50 cm





The Norwegian
Colour and Visual Computing
Laboratory

Thank you for your attention

Contact information:

Marius Pedersen

Office: A208

E-mail: marius.pedersen@ntnu.no

Web: www.colourlab.no

Phone: (+47) 61 13 52 46

Mobile: (+47) 93 63 43 85