Memory effects in subjective quality assessment of x-ray images

Victor Landre[*], Marius Pedersen, and Dag Waaler

Norwegian University of Science and Technology
Gjøvik, Norway
{marius.pedersen,dag.waaler}@ntnu.no
www.colourlab.no

Abstract. Experiments with human observers is considered as the most precise way for the assessment of image quality. Although widely used, such experiments have its pitfalls and hazards. In this work we investigate if the quality rating of previously viewed images influence the rating given to the current image, which we refer to as the rating memory effect. A subjective experiment with a group of observers rating x-ray images of different radiation dose was used for the basis of the analysis. The results indicate a memory effect, meaning that the rating of an image can be influenced by the ratings given in previously judged images.

Keywords: x-ray, memory, subjective experiments, psychometrics, quality assessment

1 Introduction

Assessment of image quality is very important in many applications, such as medical imaging, printing, image enhancement, etc. Assessment of image quality can be done objectively, through image quality metrics [1], or subjectively, by consulting human observers [2]. Subjective assessment of image quality is considered as the most precise way of assessing quality, and is seen as the ground truth [3].

Various methods are used when assessing quality with human observers. In psychophysics there are traditional methods for measuring thresholds such as method of adjustment, method of limits, and method of constant stimuli [4]. These methods are used in detection or discrimination experiments. Psychometric scaling methods [2] have been proposed to obtain the relationship between physical changes and perceived changes. Common experimental methods include paired comparison, rank order and category judgement [2].

[*] This research has been supported by the Research Council of Norway through project no. 247689 IQ-MED: Image Quality enhancement in MEDical diagnosis, monitoring and treatment. The authors would like to acknowledge the Color in Science and Industry master program (Jean Monnet University, Norwegian University of Science and Technology, University of Granada and University of Eastern Finland).

In most, if not all, psychometric experiments the observers are shown a set of images that are rated. In this work we focus on category judgement experiment, in which the observer is shown an image and is instructed to rate it according to a scale. Then a new image is shown, and the observer is asked to rate it according to the same scale. This is repeated until all images in the set have been rated. We investigate if the rating given to one image is influenced by the ratings of the previously rated images. We refer to this as the memory effect.

Most, if not all, statistical analysis of the results from subjective experiments are based on the assumption that the ratings are independent. If there is a memory effect, then this assumption does not hold, making the standard statistical analysis "useless" [5, 6]. In quality control testing it is common to include visual assessment [7], and if memory effects are present, this might influence the outcome of such tests.

The paper is organized as follows: first relevant background, then we present the experimental setup, before results, and conclusion.

2 Background

We start by introducing work related to still images, then work related to videos, and at last work on decision making.

2.1 Still images

Hoßfeld et al. [8] carried out an experiment with observers to study the impact of quality changes in web browsing. They found that the memory effect is a relevant quality of experience factor.

Short-term memory plays an important role in for example pair comparison experiments, where two images are shown at the time. An observer cannot scrutinise both images at the same time, since viewing one image will automatically place the other in the peripheral field making it substantially less detailed [9]. Because of this, one needs to rely on short-term memory when judging the quality the images, resulting in that a limited quantity of information from the two images can be compared at a time [10].

There has also been work related to long-term memory effects and its influence on image quality [11–13]. However, in this paper we focus on the short-term memory effects. Work has also shown that there is a difference between expert and non-expert observers in psychometric experiments [14, 15]. Le Moan et al. [16] compared two different setups for paired comparison experiments, showing that the results between the two setups were significantly different.

2.2 Video

For video the working memory, or recency, and its impact on video quality has been studied by several researchers [17, 18]. Alridge et al. [17] evaluate this effect on video quality on subjective experiment. Their experimental method consists of showing 30 seconds of a video with increasing or decreasing the quality. They asked the observers to rate the overall quality judgement on a 5 point scale at the end of the video. They evaluated the data using a Mean Difference Score between the ratings and the reference video. The results of this work showed that subjects tended to forgive the bad section by averaging the quality over all the period of the test, and were strongly influenced by what they see in the last section of 10 seconds. In their study they found that the prior information 20 or 30 seconds to the end seem to not influencing the weighting of the final rating.

Seferidis et al. [19] tried to quantify the "forgiveness effect", like the recency memory they evaluated it using videos. The authors introduced a forgiveness factor that adjusted the results obtained from short sessions 10-30 seconds to real-world viewing condition.

Hands and Avons [18] showed 30 seconds sequence films with poor-quality at the beginning or at the end ; and asked the observers to rate continuously the quality during the observation. They found that the ratings changed more slowly following an improvement in quality than following a sudden impairment. They also investigated the duration of the impairment and its effect on observers, and they found that the duration is not cared by the participants, this effect is call "duration neglect".

A study from Pinson and Wolf [20] using videos showed that perception can be affected by the time spending looking the sample. They showed evidence that the human memory effect for quality estimation is limited to about 15 seconds. No significant memory effect occurred after 8 to 9 seconds, there is a low correlation between 1 and 3 sec, and that the correlation were high between 7 and 9 seconds.

Huynh-Thu et al. [21] investigated the difference between discrete and continuous scales. According to their results based on videos they found that most observers tend to align their ratings with the labels, but some observers appear to distribute their ratings more evenly across the scale. This indicates the existence of individual response styles. In conclusion the authors assume that the absolute category rating method, with careful design and proper instruction produce very repeatable subjective results even across different scales.

Ickin et al. [22] studied the challenges in assessing the perceived quality of mobile-phone based video. They found that the extremely good quality videos are remembered better, even if there are intermediate parts with varying qualities.

There is some evidence that viewers have non-symmetrical memory in that they are quick to criticize degradations in video quality but slow to reward improvements [23].

2.3 Decision making

Quality evaluation is a decision making process. Klapproth [24] reviewed the principal literature related to duration of the expected delay for realizing the options, and the time available to reach a decision. They show the relationship between decision making and time in different aspects. They show that physical (objective) time matters and has been related to decision making often and extensively. Also that psychological (subjective) time affect decision. Moreover, they show that anticipating delays in realizing an option not only reduces the value of that option, but also alters its mental representation. The time available for making choices has an impact on the amount of information that is processed and the quality of the final decision.

3 Experimental setup

Double stimulus continuous quality scale (pair comparison) method is claimed to be less sensitive to context (i.e., subjective ratings are less influenced by the severity and ordering of the impairments within the test session). Single stimulus continuous quality evaluation (category judgement) method is claimed to yield more representative quality estimates. In this work we will use the single stimulus continuous quality evaluation method [25].

3.1 Viewing conditions

The experiment was done in a controlled environment and with a medium ambient illumination at around 60 Lux. We used a single monitor, an Eizo ColorEdge CG246 display (24.1" - 61cm), with a resolution of 1920 x 1080, calibrated with the ColorNavigator 6 software for a color temperature of 6500K, a gamma of 2.20, a luminous intensity of 100cd/m^2, and a black level of 0.15cd/m^2. We did not set any restrictions for participants regarding the proximity of the screen to the user during the experiments. The users were allowed to take flexible and comfortable position; the position they are accustomed to take while doing this observation in daily work life. The viewing distance was around 50-60cm from the monitor.

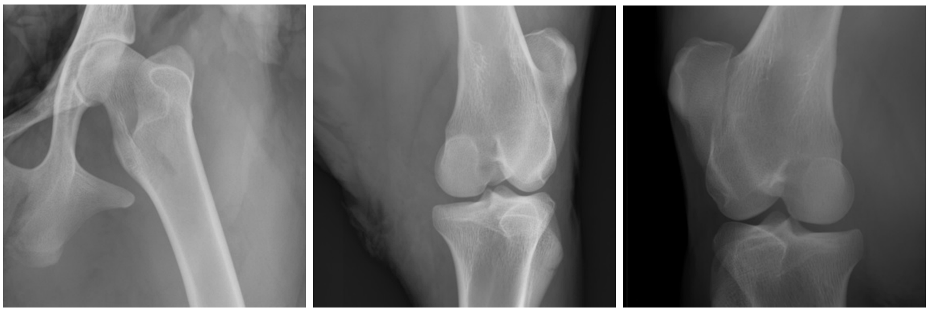

Fig. 1. Experimental X-ray images; image 1 on the left, image 2 in the center, and image 3 on the right.

3.2 Data

The images used for the experiment were three X-ray images of a lamb femur phantom (Figure 1), image set 1 and image set 2 from Precht et al. [26] and image set 3 from Precht et al. [27]. The images in a set are different according to the variation of the X-ray dose and software optimization (Table 1). This resulted in three versions of image 1, three version of image 2, and four versions of image 3.

Table 1. Details of the images used in the experiment. The table shows the version of the images, x-ray dose in milliampere-seconds (mAs), and software optimization for the three different set of images.

Image set	Version	Dose (mAs)	Software optimization
1	1	16	Canon MLT(M)
1	2	6.3	Canon MLT(S)
1	3	2	Canon MLT(S)
2	1	16	Canon MLT(M)
2	2	6.3	Canon MLT(S)
2	3	2	Canon MLT(S)
3	1	8	Canon MLT(S)
3	2	3.2	Canon MLT(S)
3	3	0.5	Canon MLT(S)
3	4	8	Canon MLT(M)

3.3 Observers and task

A total of 20 radiography students were used as observers. All of observers were familiar with medical images, and especially with x-ray images. Ages ranged between 19 and 25, and all observers were Norwegian.

The visual grading assessment was implemented using the "QuickEval" software [28]. Identical instructions were given to all users prior to the experiments with a special training session where we calibrated the observers with examples of best quality image and worst quality image. The scale used was a 5 point scale as recommended by ITU [29]: -2 (Bad), -1 (Poor), 0 (Fair), +1 (Good), +2 (Excellent). For the experiment one image was shown at a time and the instructions was "Rate the images on a 5 points scale according to the sharpness of the trabeculae", the trabeculae are small small elements in the form of beams, struts or rods, that appear in the interior of the bone [30]. Image set 1 was shown first (three versions), then image set 2 (three versions), and at last image set 3 (4 versions). Within each set the images were shown in a random order to the observers. Each version of the image was shown twice to the observers, resulting in a total of 20 images shown to each of the 20 observers, which in total gave 400 ratings for all images.

4 Results and discussion

4.1 Memory effect

Figure 2 shows a histogram of the ratings given by the observers, and we can notice that all categories have been used, but that observers have mostly used categories -1, 0 and 1. To study and evaluate the data and the memory effect, the autocorrelation function [31] is an often used mathematical tool for finding repeating patterns. We define the autocorrelation function by the division of the covariance by the variance [32]:

$$r_k = \frac{c_k}{c_0}, \tag{1}$$

where k is the lag, c_0 is the sample variance and

$$c_k = \frac{1}{T-1} \sum_{t=1}^{T-k} (y_t - \bar{y})(t_{t+k} - \bar{y}). \tag{2}$$

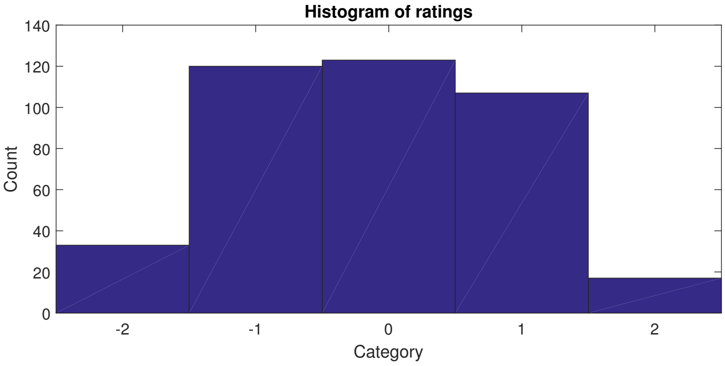

Fig. 2. Histogram of ratings given by the observers.

For the further analysis, we subtracted the average rating given to an image by all observers to the ratings given an observer to this image, and divided it by the standard deviation for all observers for that image:

$$RC_{ij} = \frac{R_{ij} - \frac{1}{N}\sum_{j=1}^{N} R_{ij}}{\sqrt{\frac{1}{N-1}\sum_{j=1}^{N} \left| R_{ij} - \left(\frac{1}{N}\sum_j R_{ij}\right)\right|^2}}, \tag{3}$$

where RC is the standard score rating for image i given by observer j, N the total number of observers, and R the raw rating given by an observer.

The autocorrelation function was calculated for all observers for 20 lags, since the observers each gave 20 ratings. To reduce the impact of the order in which in the observers carried out the experiment 1000 permutations of the observer order were carried out, and the average of the autocorrelation for these 1000 permutations were taken. In a theoretical case, where there is no presence of a memory effect, i.e. that the current rating given by an observer is influenced by the previous rating(s), the autocorrelation function should be close to zero all lags except 0 (which should be 1). The sample distribution can also influence the autocorrelation, for example if most ratings were identical, but as shown in Figure 2 there is more or less an even distribution between category −1, 0 and 1, and some ratings in categories −2 and 2.

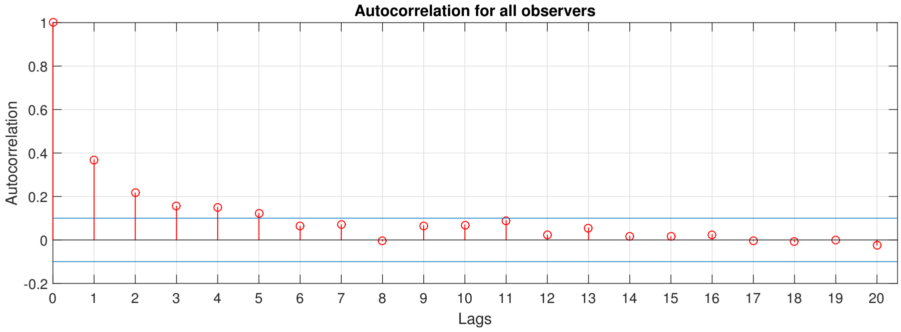

Fig. 3. Autocorrelation for all observers for 20 lags. The horizontal blue lines represents the 95% confidence bounds.

The autocorrelation function is shown in Figure 3. We can notice that there is a drop from 1 to just below 0.4 at lag 1, then there is a steady decrease in the following lags. The blue lines indicate a 95% confidence bounds. There is a fairly high correlation between lag 0 and lag 1, which indicates that the rating given for the current image is influenced by previous ratings when analyzing the results for all observers.

Figure 4 shows the difference in score (raw score) between the duplicates of images (since each version of the image was shown twice). We notice that for most images the original version and its duplicate is rated with the same score. However, there are cases with a difference between the original version and its duplicate. If there are memory effects there would differences in the ratings given by the observers between the original version and its duplicate.

We also analyze the results also for individual observers, as there might be individual differences. Figure 5 shows the results for every single observer in the experiment. The confidence bounds are of course larger as the number of data points are fewer. However, we can notice that there is a large variation between the observers, but that the autocorrelation is reduced with increasing lag and that it is converging.

The content of an image can also influence the ratings given by observers, and therefore we also analyzed the autocorrelation per image set. We also carried out

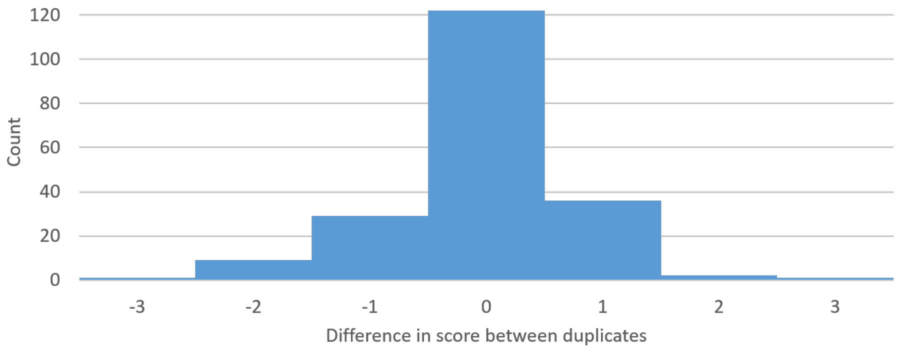

Fig. 4. Difference in score between original version and duplicate for all observers.

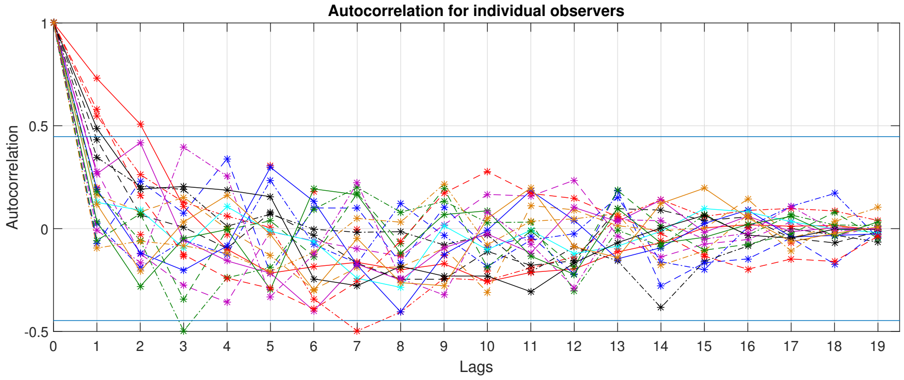

Fig. 5. Autocorrelation for individual observers. Each curve indicate a single observer. The horizontal blue lines represents the 95% confidence bounds.

1000 permutations as described above. Figure 6 shows the results for the three different images, and we can notice that for image 2 there seems to be a higher autocorrelation than for the two others. Given the reduced number of images and number of observers, it is clear that further investigation with regards to the content of images should be done.

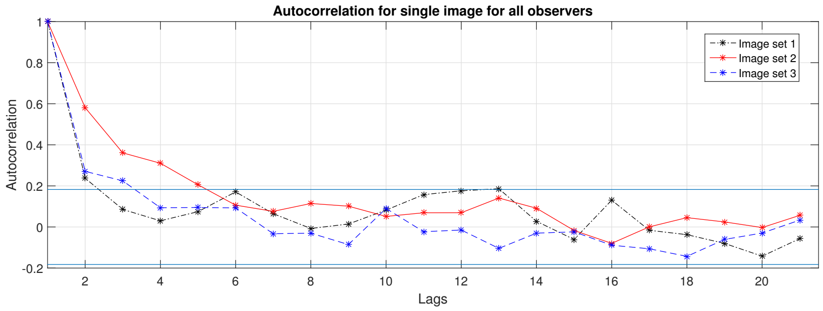

Fig. 6. Autocorrelation per image in the experiment.

One aspect that was not investigated in this work was whether the observers recognized that the same image was shown twice and tried to remember what they scored earlier. If this occurred it might influence the results. This aspect should be investigated in the future.

4.2 Time

Time is also an interesting parameter in psychometric experiments. Figure 7 shows a boxplot of the time the observers spent on the three image sets. We can notice that the average and median time spent is reduced as the observers rated the different image sets.

Furthermore, we also analyze the time spent by the observers when giving the different ratings for each category (-2,-1,0,1, and 2). Figure 8 shows the average time in seconds for each category in the experiment. We can notice that the observers used less time when they gave a high score compared to giving a low score. The effect is not very dominant though: on average the observers use approximately 20% less time than average to assign a top score , and approx. 20% more time than average to assign the lowest score. This result is similar to [22], who found observers to respond faster when giving a high score compared to low scores.

5 Conclusion and Future Work

In this study we want to evaluate the potential presence of memory effects on in subjective experiments. A set of x-ray images were shown to observers, and the

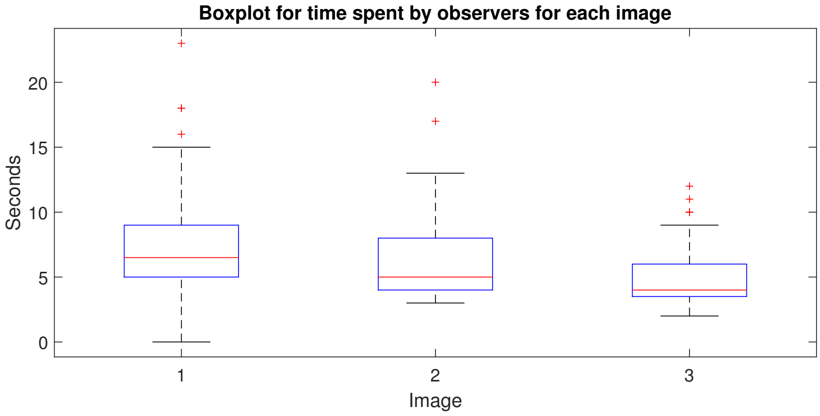

Fig. 7. Boxplot representing the time spent by the observers evaluating the images for the three image sets.

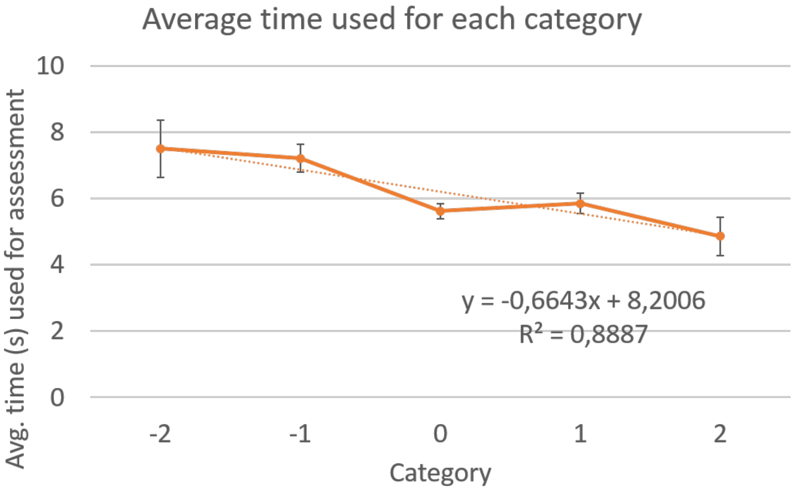

Fig. 8. Average time in seconds for each category represented with a 95% confidence interval.

autocorrelation function was used to analyze the influence of previous ratings on the current rating. The results indicate that there is a correlation between previous ratings and the rating currently given by an observers.

Additional experiments with more images and more observers should be carried out to verify the results found. One should also carry out experiments where the order of the image sets are randomized to see if this influences the results.

6 Acknowledgments

We would like to thank Dr. Helle Precht, who provided the different images for the experiment.

References

1. Marius Pedersen and Jon Yngve Hardeberg. Full-reference image quality metrics: Classification and evaluation. *Foundations and Trends® in Computer Graphics and Vision*, 7(1):1–80, 2012.
2. Peter G Engeldrum. *Psychometric scaling: a toolkit for imaging systems development.* Imcotek press, 2000.
3. Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006.
4. Walter H Ehrenstein and Addie Ehrenstein. Psychophysical methods. In *Modern techniques in neuroscience research*, pages 1211–1241. Springer, 1999.
5. Peter G Engeldrum. Psychometric scaling: avoiding the pitfalls and hazards. In *PICS*, pages 101–107, 2001.
6. Peter G Engeldrum. Image quality modeling: Where are we? In *PICS*, pages 251–255, 1999.
7. International Atomic Energy Agency. *Quality Assurance Programme for Computed Tomography: Diagnostic and Therapy Applications.* Number 19 in IAEA Human Health Series. 2012.
8. Tobias Hoßfeld, Sebastian Biedermann, Raimund Schatz, Alexander Platzer, Sebastian Egger, and Markus Fiedler. The memory effect and its implications on web qoe modeling. In *Proceedings of the 23rd International Teletraffic Congress*, pages 103–110. International Teletraffic Congress, 2011.
9. Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.
10. Michael A Cohen, Daniel C Dennett, and Nancy Kanwisher. What is the bandwidth of perceptual experience? *Trends in cognitive sciences*, 20(5):324–335, 2016.
11. Uwa O. Aideyan, Kevin Berbaum, and Wilbur L. Smith. Influence of prior radiologic information on the interpretation of radiographic examinations. *Academic Radiology*, 2(3):205 – 208, 1995.
12. Lara A. Hardesty, Marie A. Ganott, Christiane M. Hakim, Cathy S. Cohen, Ronald J. Clearfield, and David Gur. memory effect in observer performance studies of mammograms1. *Academic Radiology*, 12(3):286 – 290, 2005.
13. Tamara M. Haygood, Mary A. Qing Liu, Eva M. Galvan, Roland Bassett, Catherine Devine, Elizabeth Lano, Chitra Viswanathan, and Edith M. Marom. Memory for previously viewed radiographs and the effect of prior knowledge of memory task. *Academic Radiology*, 20(12):1598 – 1603, 2013.

14. Fabienne Dugay, Ivar Farup, and Jon Y Hardeberg. Perceptual evaluation of color gamut mapping algorithms. *Color Research & Application*, 33(6):470–476, 2008.
15. Nicolas Bonnier, Francis Schmitt, Hans Brettel, and Stephane Berche. Evaluation of spatial gamut mapping algorithms. In *Color and Imaging Conference*, volume 2006, pages 56–61. Society for Imaging Science and Technology, 2006.
16. S. Le Moan, M. Pedersen, I. Farup, and J. Blahov. The influence of short-term memory in subjective image quality assessment. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 91–95, Sept 2016.
17. R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson. Recency effect in the subjective assessment of digitally-coded television pictures. In *5th International Conference on Image Processing and its Applications.*, pages 336–339, July 1995.
18. David S. Hands and S. E. Avons. Recency and duration neglect in subjective assessment of television picture quality. *Applied Cognitive Psychology*, 15(6):639–657, 2001.
19. V. Seferidis, M. Ghanbari, and D. E. Pearson. Forgiveness effect in subjective assessment of packet video. *Electronics Letters*, 28(21):2013–2014, Oct 1992.
20. Margaret H Pinson and Stephen Wolf. Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing 2003*, pages 573–582. International Society for Optics and Photonics, 2003.
21. Quan Huynh-Thu, Marie-Neige Garcia, Filippo Speranza, Philip Corriveau, and Alexander Raake. Study of rating scales for subjective quality assessment of high-definition video. *IEEE Transactions on Broadcasting*, 57(1):1–14, 2011.
22. Selim Ickin, Lucjan Janowski, Katarzyna Wac, and Markus Fiedler. Studying the challenges in assessing the perceived quality of mobile-phone based video. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 164–169. IEEE, 2012.
23. R. Hamberg and H. de Ridder. Time-varying image quality: Modeling the relation between instantaneous and overall quality. *SMPTE journal*, 108(11):802–811, 1999.
24. Florian Klapproth. Time and decision making in humans. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):509–524, 2008.
25. CIE. Guidelines for the evaluation of gamut mapping algorithms. *Commision Internationale De L'eclairage*, 153:D8–6, 2003.
26. Helle Precht, Ole Gerke, Karin Rosendahl, Anders Tingberg, and Dag Waaler. Digital radiography: optimization of image quality and dose using multi-frequency software. *Pediatric radiology*, 42(9):1112–1118, 2012.
27. Helle Precht, Oke Gerke, Karen Rosendahl, Anders Tingberg, and Dag Waaler. Large dose reduction by optimization of multifrequency processing software in digital radiography at follow-up examinations of the pediatric femur. *Pediatric radiology*, 44(2):239, 2014.
28. Khai Van Ngo, Jehans Jr. Storvik, Christopher Andre Dokkeberg, Ivar Farup, and Marius Pedersen. Quickeval: A web application for psychometric scaling experiments. *Image Quality and System Performance XI], Larabi, M.-C. and Triantaphillidou, S., eds*, 9396:9396–24, 2015.
29. ITUR Rec. Bt. 500-11 (2002). methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*, 39, 2009.
30. Kathy McQuillen Martensen. *Radiographic image analysis.* Elsevier Health Sciences, 2013.
31. George EP Box and Gwilym M Jenkins. *Time series analysis: forecasting and control, revised ed.* Holden-Day, 1976.
32. George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control.* John Wiley & Sons, 2015.