

# Subjective Quality Evaluation: What Can be Learnt From Cognitive Science?

Simon Hviid Del Pin<sup>1</sup>, Seyed Ali Amirshahi<sup>1</sup>

<sup>1</sup>Norwegian University of Science and Technology, Gjøvik, Norway

## Abstract

Subjective ratings given by observers are a critical part of research in image and video quality assessment. Like any other field of science, with subjective data collection, researchers may lack the expertise needed to address the different issues they face. In this study, we review different approaches and find potential pitfalls that generally seem overlooked in quality research. To address these issues, we found six relevant pitfalls relating to recruitment, instructions, experimental design, and data analysis that could be addressed by studies done in the field of cognitive science. Combining accessed datasets from quality research with newly collected data, we statistically demonstrated four of the six pitfalls: observers used the scale non-linearly; ratings can change throughout the experiment; features can influence individual observers differently; and allowing observers to decide how many ratings they give can lead to biases. We need additional data to investigate the two pitfalls related to instructions and recruitment. Our findings suggest that pitfalls which might not be initially clear to researchers in the field of image and video processing can still have an empirically demonstrable influence on the data. While this article will not solve every issue, it will try to suggest improvements that researchers can readily employ.


## Keywords

Image quality assessment, subjective data collection, cognitive science

## 1. Introduction

Quality judgments from human observers are crucial to researchers interested in evaluating the quality of images and/or videos. A typical process for such researchers is to ask observers to rate the quality of videos or images. Researchers take these ratings as “ground truth data” which can not only be used to train and test different models to predict observers’ judgement, but can also open new doors for evaluating and understanding different aspects of the human visual system. However, those involved in quality research may lack the expertise in different aspects of subjective data collection, such as instructing human observers or collecting and interpreting the resulting data. Realistically, it is nearly impossible to directly measure the experiences of observers, and no experience can be seen by itself as “right” or “wrong”. Therefore, there are real risks that researchers do not know about possible pitfalls they can face. In this study, our objective is to show some of the relevant pitfalls that are highlighted in cognitive science research. We will aim to demonstrate these pitfalls empirically for quality experiments and provide guidelines on how researchers can avoid them.


---

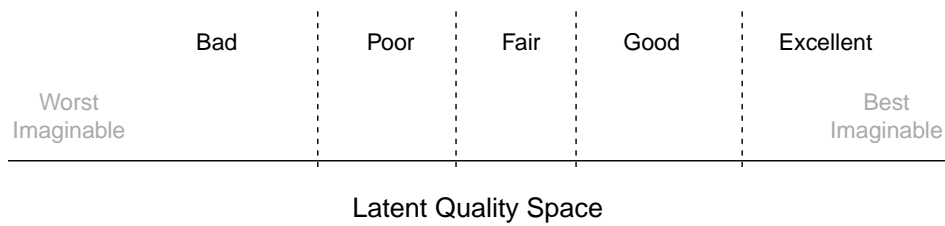
The 11<sup>th</sup> Colour and Visual Computing Symposium 2022, September 08–09, 2022, Gjøvik, Norway 

simon.h.d.pin@ntnu.no (S. H. Del Pin); s.ali.amirshahi@ntnu.no (S. A. Amirshahi)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** A depiction of a latent space representing the experience of quality for an observer. The extremes of this space can be called “Worst Imaginable” and “Best Imaginable” respectively. The observer establishes thresholds for which to rate the quality into one category. The thresholds may not be equally spaced.

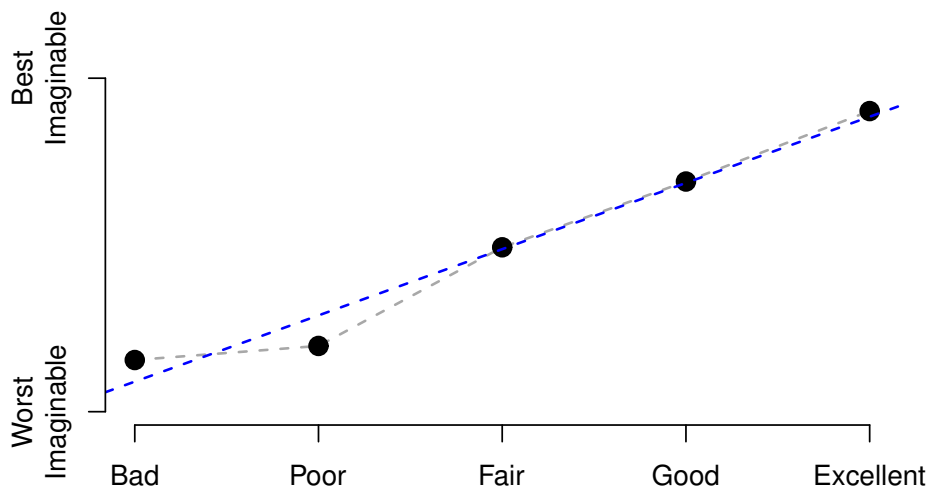
In this paper, we start with an introduction in Section 1. Section 2 is dedicated to introducing the methods used in our study for data collection, followed by Section 3 by a discussion of our findings. Finally, in Section 4 we provide a summary and the conclusion of the work.

### 1.1. Rating quality is a non-objective decision-making process

Researchers often employ scales as a systematic method for observers to report their experiences. When using scales, a tacit (i.e. unspoken) assumption is that there is a latent space of quality that each observer can experience. Two extremes could define this space: on one end is the “Worst Imaginable” and on the other end the “Best Imaginable” quality experiences. The assumption in current studies is that the scale is divided into equally distanced categories. In the field of image and video quality assessment, the typical recommendation refers to 5 or 9 discrete points [1]. For example, when people are asked to use the typical five-point Absolute Category Rating (five-point ACR), we assume that they rate everything below a certain threshold as “Bad” in the given latent space. Anything that is above this threshold but below the threshold of “Fair” will be labeled “Poor”, etc. (Figure 1).

### 1.2. Remember that your scale may be non-linearly understood

In current studies, a common (if not standard) approach to quantifying the quality of an image/video independent of the tasks observers have been given is the use of the Mean Opinion Score (MOS) [2, 3, 4, 5]. MOS takes the numeric average of all subjective scores given to an image by different observers. This practice is tacitly assuming that the different categories introduced to the observers are linear, and so all have the same distance. However, decades of research deem this assumption unreasonable. For example, Jones & McManus [6] investigated how people understood certain terms on a scale from “Worst Imaginable” to “Best Imaginable”. Focusing on the five terms of the five-point ACR, they show that the observers did not see



**Figure 2:** Perceived intervals between terms used for quality category scales. Figure recreated with data presented in Jones & McManus [6]. 37 participants were told to draw in 15 words as points on a line between “Worst Imaginable” and “Best Imaginable”. Results show that compared to an idealised regression (blue line) the five words of the ACR are not equally spaced. “Poor” is a small step above “Bad” whereas “Excellent” is a relatively steep step above “Good”.

these terms as equally spaced (Figure 2). The study from Jones & McManus [6] thus shows that the use of different words and phrases to introduce the different categories could influence the judgement of observers. For example, they show that there is a small perceived difference between “bad” and “poor”. An observer we tested for this article explicitly agreed with this sentiment: (“[I]t was difficult to choose between Bad and Poor as they are so arbitrary[.]” [Observer in our post-experiment questionnaire]).

It is plausible that research can lead to a scale with terms that are closer to equidistant. By using a scale that is closer to the terms that observers themselves would prefer and see as linear, we could address two critical issues. First, the current non-linear nature of the scale and, second, any confusion that using specific terms could cause the observers. Researchers have constructed such scales in cognitive research on, for example, the clarity of briefly flashed figures and the sense of control [7, 8]. In addition to being closer to the participants’ experience, a benefit of such a scale could be that the terms are less arbitrary to the observers. This could lead to more stable thresholds in the latent space. Thus, investigating which words to include in scales could be a fruitful endeavor for future studies.

### 1.3. Viewing ratings as an active decision process

Studies in cognitive science have shown that scale usage is a complex decision process influenced by even minor details. A convincing line of research comes from Siedlecka et al. [9, 10]. In one study, they investigated anagrams, i.e. scrambled letters that may make up a certain word. For instance, the letters ASRONTE can correctly be unscrambled to SENATOR or incorrectly to TOASTER. In their experiment, the participants rated their confidence in the accuracy with which they had unscrambled the letters from “1: I am guessing” to “4: I am very confident”. They also saw a word and judged whether that word represented a correct unscramble. Importantly, the researchers varied the order of these events, meaning that people could see the word before or after giving their input. The order of pressing yes/no and the rating on the scale was also manipulated. The results showed that the procedural order was related to how the observers used the scale. For instance, observers would use the extreme parts of the scale more if they saw the proposed word first.

In follow-up experiments, the researchers showed that pressing even an unrelated button before using the scale could influence how the scale was used [9]. In another paper, people chose a color displayed on a color wheel and rated how clearly they saw it [11]. People used the scale differently if they first tried to match the color on a wheel compared to when the task was absent. Although understanding the cognitive decision process is an ongoing field of research [10] and beyond this article, it may be worth remembering that observers are not merely instruments that output measurements. They are people who make complex decisions. This means that procedural details, instructions, and scale definitions are crucial.

### 1.4. Instructions matter - Importance of sharing instructions

When looking through the literature in the field, it is often ambiguous what exact instructions observers were given before the experiment began and even what specific question they were answering when giving their rating. Merely writing that observers rated the quality on a five-point ACR is not sufficient and directly against the recommendations of the International Telecommunication Union (ITU) [12]. Without in-depth investigation, we cannot know if small variations in instructions or questions matter, but we may again draw on publications from cognitive research. For example, Sandberg et al. [13] tested whether asking three very similar questions influenced how observers used the ratings. They had observers select the object (e.g. a triangle or circle), which was briefly displayed. They then asked the observers “how clearly they saw the object” or “how sure they were of their choice”. The questions yielded different responses, but if one simply analyzed the responses as one to four, this nuance could easily be missed. Similarly, we cannot know if asking “What was the technical quality of this image?” may lead to different responses than “What was the overall quality?”. We strongly encourage researchers to share this aspect of their methods and reviewers to demand such sections before accepting papers. To practice what we preach, we of course have made our specific instructions available in the repository together with our statistical code<sup>1</sup>

---

<sup>1</sup>The raw data, statistical codes and instructions can be found at <https://osf.io/6qvwm/>

**Table 1**

An overview of the datasets analyzed in this paper. Name indicates how we will refer to them throughout the paper.

Name	# observers	# reference images	Local \Online	Reference
Dataset 1	24	10	Local	[14]
Experiment 1	40	235	Online	This Study
Experiment 2	31	10	Online	This Study

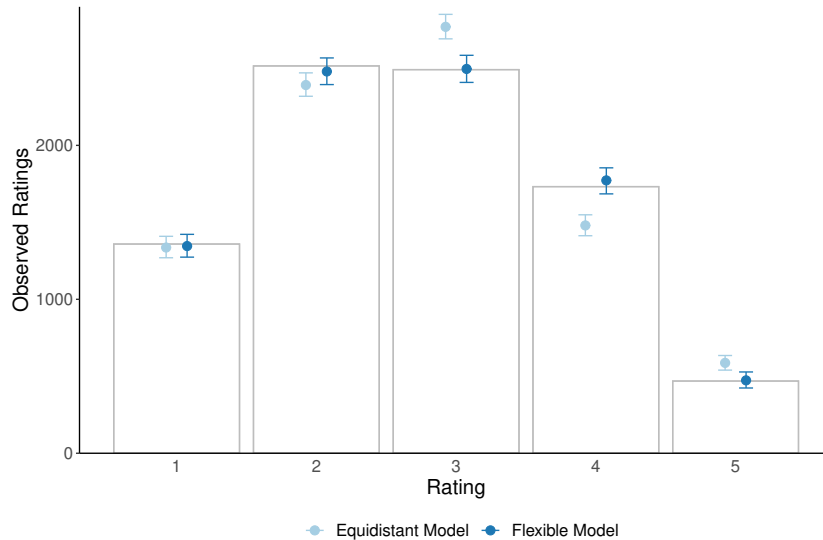
## 2. Methods and Data

To further investigate the issues raised, we used a publicly available dataset in the field of image quality assessment [14] (Dataset 1 in Table 1). The dataset was collected through an experiment conducted locally under controlled conditions and used 10 distorted reference images. As one of our experiments in this study, we recreated an online experiment using the same reference images (Experiment 2 in Table 1). The other experiment had 235 reference images from the KonIQ-10k IQA database [15]. Using Pavlovia, we then performed another set of subjective experiments with 250 trials per observer on the mentioned images (Experiment 1 in Table 1). In addition to using different sets of stimuli, our experiments had identical instructions and experimental paradigms.

### 2.1. Testing if the scale is non-linearly used in quality experiments

We first analyzed whether observers in the three experiments used the scale equidistantly or with flexible thresholds. If you take the means of the ratings, you should (tacitly) expect equidistant usage. Flexible thresholds assume that the distance between each scale point is not equal. This model will by definition have more degrees of freedom and we, therefore, wish to investigate if it also yields correspondingly to better predictions. We created all models in BRMS [16] a package for The R Project for Statistical Computing. We constructed equidistant and flexible models in BRMS (for more information on this process, see [17]). We then compared the two models to see if the flexible model had better predictive power. We used the R package LOO which uses PSIS-LOO to approximate a leave-one-out cross-validation. PSIS-LOO has been shown to be a robust and computationally efficient method for picking models [18]. This type of comparison considers not only the absolute outcome of the model prediction but also gives an estimate of how likely it is that the same model would perform better on future samples from the same population. If two models are within two standard errors (SEs), we cannot be sure which is the better one. In this case, the parsimonious choice would be to choose the simpler model. Using two SEs roughly corresponds to having a 95% probability that the complex model is better (for a more thorough argument on this, see [19]). Comparing the equidistant and flexible models showed that the flexible model was about five SEs better. Therefore, we conclude that a flexible model is best at describing the data (Figure 3).

Our results, along with what we have already emphasized in the literature, indicate that observers neither understand nor use the five ACR categories as equally spaced. Recent research



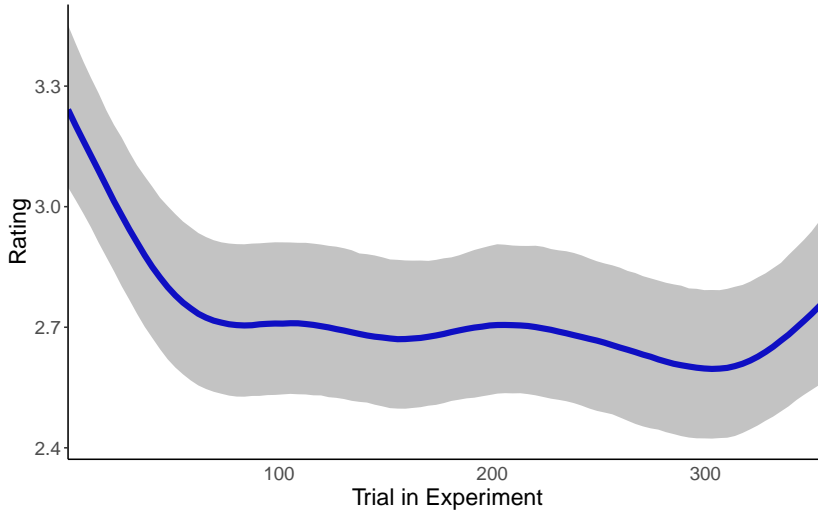
**Figure 3:** Use of ratings modelled for a flexible or an equidistant scale usage. The bars represent the observed ratings and the points represent the estimates from both models. We see that the flexible model accurately captures the usage of all ratings whereas the equidistant model overestimates the usage of 3 and 5 but underestimate the usage of rating 2 and 4.

shows that treating nonlinear data as linear can lead to false conclusions and increase the risk of Type I and Type II errors [20]. This means that using a metric model both increases your risk of missing a real effect and falsely concluding that a non-existent effect is real.

We present a code for a statistical method which does not require ratings to be quality scores but merely ordered (that is, knowing that “Good” is above “Fair” but not to what extent). The method requires more computational power, especially as the size of the dataset increases and may thus not be practical for all situations. We also point out that there are currently a significant number of research studies done on statistical models of ordinal data [17, 20, 21]). The code we present in this paper may therefore not be the same as what we would recommend a few years from now. Nevertheless, we believe that the current methods are mature enough to be widely applied.

## 2.2. How many trials do observers need to learn the task?

Cognitive researchers encourage to begin the experiments with 40-50 trials used purely to let participants learn the task and the scale [22]. We rarely see this practice in image quality experiments conducted in the field of image processing and computer vision. In the few cases that this is done, there is no golden number used and depending on the size of the datasets, the time they have estimated for the experiment etc. This could simply range from a handful [14] to a large but randomly selected number of images [4] as warm-up trials. Additionally, the warm-up trials could range from having the observer evaluate the quality of an image to simply showing randomly selected images from the dataset for a few seconds. Therefore, we investigated if ratings change throughout the experiment, and if so, would it be reasonable to



**Figure 4:** Expected ratings as a function of trial number for Dataset 1. We see that ratings drop over the first roughly 50 trials and then remain relatively stable for the rest of the experiment.

follow the practice from cognitive science?

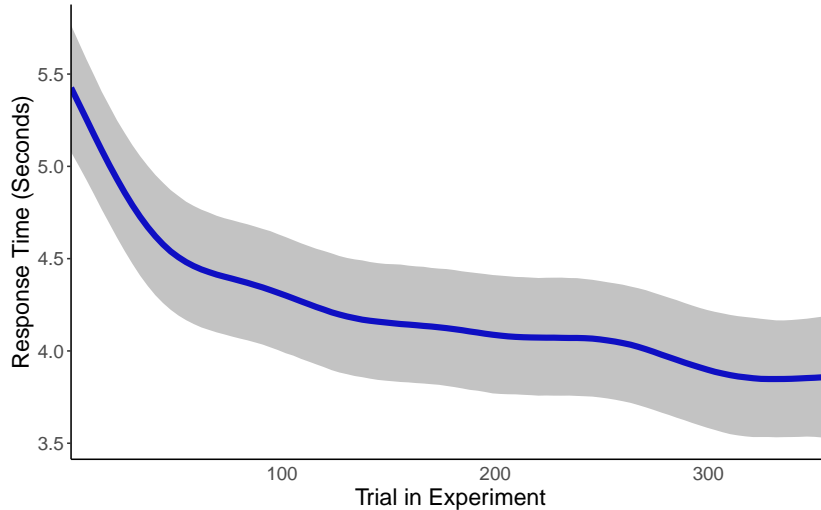
### 2.3. Analysis of scale usage throughout the experiment

To investigate whether scale usage changes throughout the experiments, we modeled the ratings and response time as a function of the trial number. We defined the first model we created as

$$rating \sim s(trials) + (1|id), \quad (1)$$

where  $s(trials)$  represents a spline during the trials and  $1|id$  allows for an intercept per participant. This models the effect of trials on the ratings on a population level and that each individual has their intercept. We thus assume each observer is similarly influenced throughout the experiment but differ in how they rate. This allows for cases where one observer typically rates 4 and another typically rates 3 but their ratings both fall similarly throughout the experiment. In addition, we model the ratings as a spline [23] which allows for a nonlinear effect. This is beneficial if the ratings are not linearly influenced throughout the experiment. We confirmed the model was over five SEs better than a null model not including trials to predict ratings. Due to space limitations, we only present the model calculated for Dataset 1 (Figure 4).

Our model showed that after roughly 50 trials observers in general converge to an average quality value slightly lower than what they normally start with. When we performed the same analysis on Experiment 1 that used the same number of reference images we again saw that ratings dropped throughout the experiment. In the case of Experiment 2 with 235 references, the spline model was less than two SEs better than our null model. This indicates that there was no consistent change in ratings throughout the experiment.



**Figure 5:** Response times throughout the experiment for Dataset 1. We see a large drop over the first roughly 50 trials and a gradual drop over the rest of the experiment.

#### 2.4. Analysis of response time throughout the experiment

We created similar models to capture the response time throughout the experiment. The first model we created was defined

$$response-time \sim s(trials) + (1|id) \quad (2)$$

were  $s(trials)$  representing a spline during the trials and  $1|id$  allows for an intercept per participant. We thus assume there is an effect of trials on the response time on a population level while each individual can be faster or slower. In the case of Dataset 1 (Figure 5) our analysis shows that response time decreased with more trials, especially in roughly the first 50 trials. We found similar effects for Experiment 1 and Experiment 2. Taken together, our analysis shows that ratings decreased similarly to response times. An exception was Experiment 1 with 235 reference images. These results indicate that significant learning occurs, particularly in the first part of the experiments. It seems that the participants “learn” repeated reference images, raising concerns about current practices in the field. As a concrete example, one could speculate that the observer learns they should focus on a particular flower in a reference image to discern if it is of good quality.

#### 2.5. The importance of counterbalancing your conditions

The fact that ratings systematically differed throughout the experiments that had repeating reference images highlights the importance of randomizing all possible aspects of the experiment. An experiment in which conditions are not properly balanced could erroneously show differences in image quality simply because of the order in which they are shown to participants. Imagine, for instance, an observer taking part in an experiment in which they are first shown 40 images compressed with a novel algorithm and then 40 images compressed with the current benchmark.



Without a warm-up phase, the observer most probably will rate the novel algorithm higher - even if it was no better than the benchmark approach. Even in less obvious cases, we recommend counterbalancing. As Brooks [24] puts it, “Reactions of neural, psychological, and social systems are rarely, if ever, independent of previous inputs and states”.

## 2.6. Modelling Individual differences

As argued before, there may be individual differences in how observers understand the scale. This problem may be compounded if there are also personal differences in how image features influence ratings. For instance, you could imagine that both colourfulness and sharpness positively affect ratings. However, it may be the case that there are individual differences in how much these features influence every individual’s evaluation. Sharpness may be more important to one observer, whereas colourfulness is more important to another. Or maybe beyond a certain level of colorfulness, more colors do not matter. Again, this threshold may differ from one observer to the other. To investigate this, we analysed how sharpness influenced the ratings of individual observers in Experiment 1. Preliminary analysis showed that ratings rise with sharpness, but level off or even fall with higher values. We approximated this as a second-order polynomial. The reason we didn’t use splines is that they are not computationally feasible to estimate for each individual. We defined the simple model to assume that each individual can rate higher or lower, but that sharpness will equally influence the observers as

$$rating \sim sharpness + I(sharpness^2) + (1|id) \quad (3)$$

were  $sharpness + I(sharpness^2)$  representing a second-order polynomial over sharpness and  $(1|id)$  allows for an intercept per participant. We thus assume there is an effect of sharpness on the ratings on a population level while each individual can be rate higher or lower. The complex model was defined to assume that each individual is influenced in their own way by sharpness

$$rating \sim sharpness + I(sharpness^2) + (sharpness + I(sharpness^2)|id) \quad (4)$$

were a second-order polynomial over sharpness and  $sharpness + I(sharpness^2)|id$  additionally allows for sharpness to have a unique effect per participant.

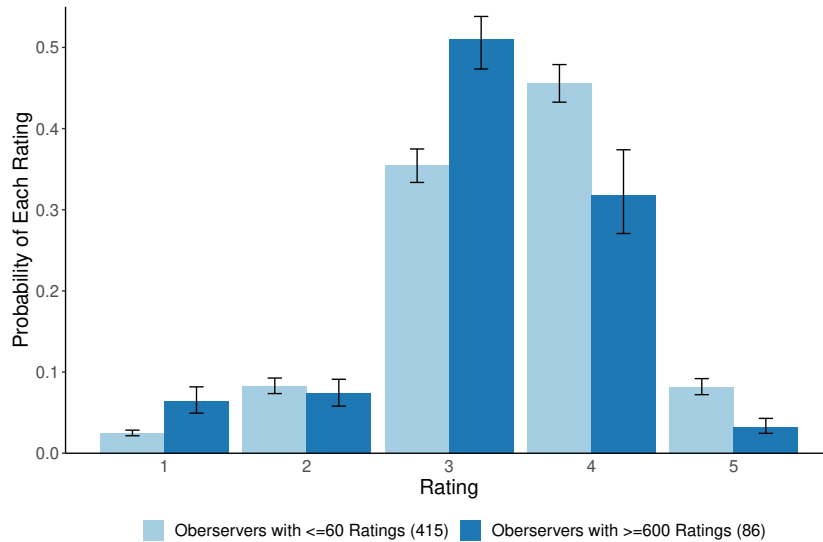
Our analysis showed that the complex model was over 5 SEs better than the simple model. Our observers thus had differences in how they are influenced by sharpness. This result is highly relevant whenever individual ratings are important, but may also be relevant when, for instance, our goal is to model how observers would generally rate an image. Take, for instance, the research from Götz-Hahn et al. [25] where they find that to use the maximum predictive power, in a large image dataset it is optimal for each image to be rated by just five observers. In other words, they would rather have many images rated a few times rather than fewer images rated many times. Though speculative, maybe they would gain even more predictive power, if the rating profiles of the five observers were further investigated. With only five ratings per image, it may, for instance, apply to know if a specific type of distortion or even specific content particularly influences observers.

## 2.7. Recruitment and external validity

Before running an experiment, consider who should be recruited as an observer. Naturally, different observers could represent one or multiple groups of people and so their subjective scores would naturally represent that group(s). Researchers rarely state this tacitly in image/video quality research. While in theory most researchers aim to have observers which ideally represent “all internet/computer users” or some similarly wide group, however, in most cases observers can better be described as “the ones available at campus” or “the first 100 people that responded on the online platform”. In the previous section, we showed that there are indeed individual differences in how people understand and use the scales. When pollsters are conducting surveys, they often use considerable energy addressing the degree to which their respondents represent the entire population of voters. If you, for instance, want to know who will win the next US election it may be more valuable to ask 100 people from a wide range of backgrounds than to ask 500 from the liberal arts college. Likewise, the preferences of young, educated observers who in most cases are working in the field of computer science (if not image processing and computer vision) are over represented in current studies. As the collection of subjective experiments using online platforms has increased, it may be relevant to not only focus on the number of people but also on who these people represent. As yet, the magnitude of this problem seems unknown. We simply do not know how much bias we have in our data.

## 2.8. Observers picking the number of trials themselves

Platforms such as Mechanical Turk and Appen allow participants to decide how many trials to complete. This may inflate variance in ratings because of some observers stopping before they “learn” the task (warm-up) and others contributing many trials after that fact. To investigate how observers empirically behave, we investigated the publicly available KONVID-150k dataset [25]. The dataset represents an experiment on Appen in which observers could choose to quit after each block of 15 videos (one to three of the videos being tests of the observer’s attention). We see the dataset contains 1257 observers for a total of 233,168 observations with a great variance in how many trials each observer completed. The median number of completed trials was just 84, and the max completed trials were 1596. The 640 observers that had given 84 or fewer ratings made up about half the observers, but only made up 17.7% of the total observations. Likewise, the 47 observers that had made over 1000 observations made up 3.7% of the total group but they had made 26.5% of the total observations. We thus see that a minority makes up a disproportionately large portion of the total ratings. The question, however, is how problematic that is? We performed an exploratory analysis (Figure 6) which arbitrarily compared those that completed 60 or fewer trials (60 trials was the 1st Quartile of total ratings) with those that had rated 600 or more images (an order of magnitude more ratings). To avoid differences in learning, we only tested the first 60 trials for all observers. The model with a difference between the two groups was more than eight SEs better than the model which did not include groups. We thus see a difference in how the two groups rated, but it is not clear why. Perhaps observers with certain preferences or understanding of the scale are more likely to continue? We cannot say what makes some observers complete more than a thousand trials whereas others complete less than 60.



**Figure 6:** Response distribution for the 415 observers that completed 60 or fewer trials (left) and the 86 observers that completed 600 or more trials (right). We see the observers that stopped before 61 trials rated four the most whereas the observers that completed 600 trials rated three the most. We also see that the observers that endured 600 trials were more likely to use rating one and less likely to use rating five. This analysis only includes the first 60 trials for all observers.

To address this issue we recommend giving each observer the same number of trials. Not doing so, lets a minority of raters leave a large influence on the entire dataset. In the present example, we see that the observers that endured 600 ratings rated lower than the ones that left before trial 61. If one was training an algorithm to predict how the average observer would rate a video/image, one could therefore end with lower estimates than the population as a whole would rate. Whereas this section could be read as a critique of the KONVID-150k dataset, we wish to commend the researchers for making their datasets with individual ratings publicly available. There could easily be similar issues with other datasets, but way too often such datasets only release the MOS rendering this type of analysis impossible.

### 3. Discussion

#### 3.1. Evidence for the pitfalls presented in this paper

This paper has presented several pitfalls in how subjective datasets are collected in the field of image/video quality assessment and has tried to address them through research performed in the field of cognitive science. We aim to demonstrate these pitfalls empirically either by performing novel statistical work on existing datasets or by additionally collecting new data to analyze (Table 2). Looking through the cognitive literature, we have found six potential pitfalls related to subjective qualitative ratings. We could demonstrate three pitfalls statistically in both accessed and novel collected data. One of them, voluntary number of trials, could only be demonstrated in accessed data as both our experiments had a fixed number of trials.

**Table 2**

An overview of the pitfalls we have listed in this paper and the degree to which we have demonstrated them having relevance to quality research.

Pitfall/Evidence from	Cognitive literature	Accessed data	Collected data
Non-Linear ratings	X	X	X
Warm-up	X	X	X
Individual effects	X	X	X
Voluntary number of trials	X	X	N/A
Recruitment /external validity	X	N/A	N/A
Influence from instructions	X	N/A	N/A

Finally, recruitment /external validity and influence from instructions remain pitfalls that at present are not demonstrated empirically. Both of these effects could be further investigated in future research. Recruitment/external validity could require a relatively high number of observers to be demonstrated, especially since we do not know to what extent, say, a group of college students rate differently from a representative sample of YouTube users. Thus, it may well be that this effect is most relevant to those conducting large-scale research. However, it would not be impossible to compare a convenience sample (such as the first 200 people who volunteer to participate) to a representative sample of people that closely matches a target demographic. Such research needs to be conducted before we can know if it has practical relevance or not. The influence of instructions seems more tangible to demonstrate empirically. After all, Sandberg et al [13] only needed 36 observers in their demonstration. A bigger problem is that the instructions are not always available. This leaves the field in a situation where it would be relatively trivial to test different instructions but with no direct way to access them. Once again, we can only recommend that we share our direct instructions and expect the same from our colleagues.

### 3.2. No current scale is without controversy

Although this paper has focused on five-point ACR, do not assume that simply shifting to another scale will absolve the current issues. For instance, one could be tempted to use a slider to avoid using terms which observers understand differently. However, this is also not without problems. In a review of different response scale characteristics, DeCastellarnau [26] shows the overwhelming options in building a scale. Relating to a slider, there are issues in that the scale takes longer time to use and that observers often will divide it into sections of five and thus still use it more discrete rather than linearly. Moreover, recent research has shown that some cultures understand Fair as average, whereas other cultures understand it as less than average [27]. Therefore, it may be problematic if a scale is developed or tested primarily in a certain cultural context. Taken together, the only way we can know that a scale is useful is when it has been thoroughly tested under different experimental conditions and even cultures. A simple “hunch” to overcome the issues of the scale which we have presented in this paper will probably be insufficient.

## 4. Conclusions

Taken together, this paper has demonstrated several pitfalls empirically and further highlighted some that future research could investigate. Whereas such studies are relevant in themselves, we also hope that this paper is directly useful to researchers in the field, and we, therefore, end with recommendations focusing on the pitfalls that we have demonstrated empirically. Note that these are general recommendations to remove confounding information from future studies, but not necessarily hard rules that must be followed in all cases.

We recommend all observers rate the same number of images or videos, if possible. Allowing observers to select the number of trials for themselves allows people with certain traits to comprise a large part of the collected data. Allowing observers to give only a few ratings also makes it harder to estimate their individual rating profiles. We also recommend that experiments either have at least 35 warm-up trials that are discarded or that a statistical model be used to allow for warm-up effects. This seems particularly relevant if the stimuli consist of a few references that are repeated. We appreciate it may not be possible to discard 35 trials in all cases and therefore share code for a model which can apply to future experiments. Keep in mind to properly balance your experiment. This should be the case whenever possible, but particularly if you cannot follow the previous recommendations. Omitting to do so may lead to false conclusions.

Finally, we recommend that researchers consider whether they are interested in scale ratings themselves or rather what they are supposed to represent. Depending on your specific research question, using the means of ratings may be sufficient. In other cases, remind yourself that ratings represent a nonlinear decision process. We provide code that can test if the data contain non-linear ratings and take that into account while modeling other aspects. We note such models are more computationally heavy and may not apply for very large datasets.

We hope that this paper has not only pointed out the methodological issues that are often seen in the field today, but also shown the relevance of cognitive research to measuring quality. We believe that future research in this overlap between the fields can lead to more robust data that represents the quality that the observers are actually experiencing.

## References

- [1] P. ITU-T, 910: Subjective video quality assessment methods for multimedia applications. geneva, switzerland, International Telecommunication Union (2021).
- [2] S. A. Amirshahi, J. Denzler, C. Redies, Jenaesthetics—a public dataset of paintings for aesthetic research, in: Poster workshop at the european conference on computer vision, 2013.
- [3] S. A. Amirshahi, G. U. Hayn-Leichsenring, J. Denzler, C. Redies, Evaluating the rule of thirds in photographs and paintings, *Art & Perception* 2 (2014) 163–182.
- [4] S. A. Amirshahi, G. U. Hayn-Leichsenring, J. Denzler, C. Redies, Jenaesthetics subjective dataset: Analyzing paintings by subjective scores, *Lecture Notes in Computer Science* 8925 (2015) 3–19.
- [5] M. Pedersen, S. Ali Amirshahi, Colourlab image database: Geometric distortions, in: *Color*

- and Imaging Conference, volume 2021, Society for Imaging Science and Technology, 2021, pp. 258–263.
- [6] B. L. Jones, P. R. McManus, Graphic scaling of qualitative terms, *SMPTE journal* 95 (1986) 1166–1171.
  - [7] M. Y. Dong, K. Sandberg, B. M. Bibby, M. N. Pedersen, M. Overgaard, The development of a sense of control scale, *Frontiers in psychology* 6 (2015) 1733.
  - [8] T. Z. Ramsøy, M. Overgaard, Introspection and subliminal perception, *Phenomenology and the cognitive sciences* 3 (2004) 1–23.
  - [9] M. Siedlecka, J. Hobot, Z. Skóra, B. Paulewicz, B. Timmermans, M. Wierzchoń, Motor response influences perceptual awareness judgements, *Consciousness and cognition* 75 (2019) 102804.
  - [10] M. Siedlecka, M. Koculak, B. Paulewicz, Confidence in action: Differences between perceived accuracy of decision and motor response, *Psychonomic Bulletin & Review* 28 (2021) 1698–1706.
  - [11] Z. Skóra, K. Ciupińska, S. H. Del Pin, M. Overgaard, M. Wierzchoń, Investigating the validity of the perceptual awareness scale—the effect of task-related difficulty on subjective rating, *Consciousness and Cognition* 95 (2021) 103197.
  - [12] I. T. Union, P. 800.1: Mean opinion score (mos) terminology, 2006.
  - [13] K. Sandberg, B. Timmermans, M. Overgaard, A. Cleeremans, Measuring consciousness: is one measure better than the other?, *Consciousness and cognition* 19 (2010) 1069–1078.
  - [14] O. Cherepkova, A. A. Seyed, M. Pedersen, Analyzing the variability of subjective image quality ratings for different distortions, in: *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2022.
  - [15] V. Hosu, H. Lin, T. Sziranyi, D. Saupe, Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment, *IEEE Transactions on Image Processing* 29 (2020) 4041–4056.
  - [16] P.-C. Bürkner, brms: An r package for bayesian multilevel models using stan, *Journal of statistical software* 80 (2017) 1–28.
  - [17] P.-C. Bürkner, M. Vuorre, Ordinal regression models in psychology: A tutorial, *Advances in Methods and Practices in Psychological Science* 2 (2019) 77–101.
  - [18] A. Vehtari, A. Gelman, J. Gabry, Practical bayesian model evaluation using leave-one-out cross-validation and waic, *Statistics and computing* 27 (2017) 1413–1432.
  - [19] S. H. Del Pin, Z. Skóra, K. Sandberg, M. Overgaard, M. Wierzchoń, Comparing theories of consciousness: object position, not probe modality, reliably influences experience and accuracy in object recognition tasks, *Consciousness and Cognition* 84 (2020) 102990.
  - [20] T. M. Liddell, J. K. Kruschke, Analyzing ordinal data with metric models: What could possibly go wrong?, *Journal of Experimental Social Psychology* 79 (2018) 328–348.
  - [21] B. Paulewicz, A. Blaut, The general causal cumulative model of ordinal response, *PsyArXiv preprint* (2022).
  - [22] M. Overgaard, K. Sandberg, The perceptual awareness scale—recent controversies and debates, *Neuroscience of Consciousness* 2021 (2021) niab044.
  - [23] P.-C. Bürkner, Advanced bayesian multilevel modeling with the r package brms, *arXiv preprint arXiv:1705.11123* (2017).
  - [24] J. L. Brooks, Counterbalancing for serial order carryover effects in experimental condition

- orders., *Psychological methods* 17 (2012) 600.
- [25] F. Götz-Hahn, V. Hosu, H. Lin, D. Saupe, Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild, *IEEE Access* 9 (2021) 72139–72160.
  - [26] A. DeCastellarnau, A classification of response scale characteristics that affect data quality: a literature review, *Quality & quantity* 52 (2018) 1523–1559.
  - [27] T. Yan, M. Hu, Examining translation and respondents' use of response scales in 3mc surveys, *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (2018) 501–518.