National differences in image quality assessment: An investigation on three large-scale IQA datasets

Dietmar Saupe*, Simon Hviid Del Pin†
*University of Konstanz, Germany
†Norwegian University of Science and Technology, Gjøvik, Norway

***Abstract*—This paper investigates the potential effects of national differences on image and video quality assessment using discrete rating scales. Drawing on cultural psychology, we hypothesize that observers from different countries may exhibit distinct response styles in interpreting and applying the five-level absolute and degradation category rating scales (ACR, DCR). For our study, we adapt state-of-the-art statistical models for three large-scale image quality datasets (KonIQ-10k, KADID-10k, and NIVD). Our models include country-specific components such as variable rating category thresholds and the probability for extreme ratings on these scales. We found statistically significant differences between ratings collected in different countries. Our results have implications for the analysis and design of current, respectively future datasets and contribute to a more comprehensive understanding of image quality in a global context. We also propose to include lapse rates into statistical models for categorical judgements. Lapse rates model unintentional erroneous responses of subjects in a quality assessment study and provide a regularization mechanism for the scale estimation by maximum likelihood estimation.**

***Index Terms*—Image and video quality assessment, absolute category ratings, degradation category ratings, statistical modeling, maximum likelihood estimation, national differences, category thresholds, extreme ratings**

I. Introduction

Subjective image quality assessment involves observers rating sets of images, but beneath the surface lies a complex interplay of cultural influences on response styles. Our study delves into cross-cultural image quality assessment, exploring whether observers from different countries exhibit distinct tendencies in providing their ratings.Several phenomena may lead to variations in how people interpret and utilize discrete rating scales such as ACR and DCR, which are ordinal scales with five categories ranging from 'bad' to 'excellent' for ACR and from 'imperceptible' to 'very annoying' for DCR.

We develop and apply statistical models to explore national differences in the use of the 5-level ACR and DCR scales for image and video quality assessment. Our study is based on data collected from observers of diverse countries who rated the same images or videos. The objective was to uncover whether cultural nuances play a role in how observers tend to assign stimuli to given quality categories and to what extent extreme ratings are chosen.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – DFG Project ID 251654672 – TRR 161. S.H.D.P. is supported by the project "Quality and Content: understanding the influence of content on subjective and objective image quality assessment" (grant number 324663, approved on 1 October 2021) from the Research Council of Norway. We thank Vlad Hosu and Mirko Dulfer for their help in preparing the KonIQ-10k raw dataset and Shaolin Su for reformatting the NIVD dataset.

Many subjective image and video quality assessment studies were carried out across several countries, either in different labs or on crowdsourcing platforms. The category labels for the responses of the subjects may be uniformly presented in English for participants from all countries, or they may have been adapted to the respective languages. In either case, the interpretation of the category labels may depend on the cultural background of the subjects.

For example, an Italian subject might rate an image on an Italian language ACR scale as 'mediocre', and the same image on an English language scale as 'fair', although the primary meaning of 'mediocre' is 'poor'. Namely, 'mediocre' can also be translated by 'moderate' which means 'average in quality', i.e., something that is neither particularly good nor particularly bad, which is just how 'fair' quality can be defined.

Moreover, subjects from different cultural backgrounds may give different category ratings for the same stimulus, even when the perceived qualities are identical. For example, the chances for an image of very good quality to receive a rating 'excellent' could be much larger when asking subjects from one country than from another one.

Similarly, some people, due to cultural background or personal style, prefer choosing the most extreme options on the scale instead of more moderate middle responses. This is called an extreme response style. It means they are more likely to pick 'bad' or 'excellent' on the 5-point ACR scale rather than a mid-point response like 'fair' [1]–[3].

An extreme response style is not inherently positive or negative. However, it can lead to biases when comparing research findings across different cultures. If a particular group consistently leans towards extreme responses, it could distort the perceived cultural differences, making them appear larger or smaller than they truly are. A thorough understanding of response styles is crucial for the accurate interpretation of results.

By nature, perceived image or video quality is a latent variable. It cannot be measured directly, but must be inferred by a mathematical model from responses of subjects who judge the quality of the stimuli in an experiment. In these models, latent variables commonly are treated as continuous

normally distributed variables. Such models were introduced by Thurstone in 1927 [4] and therefore often are called Thurstonian.

For example, the five ACR categories are commonly interpreted as values 1, 2, ..., 5 on an interval scale, i.e., the categories are not only ordered but also have values that are evenly spaced. Then the mean opinion score (MOS) is the average of the collected ratings for a stimulus. It follows that the MOS is the maximum likelihood estimate (MLE) of the mean of the corresponding normally distributed random variable [5].

To account for the effect that the ACR categories may not be equally spaced on the quality scale, we can regard them as intervals of different widths. For the five categories this yields a sequence of five successive intervals that partition the real number line, see Figure 1. For a given number of observers and a set of stimuli, the corresponding statistical model is given by the mean and variance for each stimulus and the list of thresholds that separate the category intervals, i.e., $\tau_1 < \ldots < \tau_4$ in the figure.

This model was discussed by Thurstone in his lectures in the framework of his Law of Categorical Judgement. The standard reference is Torgerson's book [6].[1] The model can also be regarded as a multinomial ordered probit model. A comparison of the metric MOS model with the method of successive intervals was recently presented in [7].

In our work, we apply the method of successive intervals together with maximum likelihood estimation and a generalized linear mixed effects model to study the national differences in rating behavior in terms of the differing interpretations of ACR and DCR categories. More precisely, we consider the following two variations in how people interpret and apply discrete rating scales:

- People from different cultural or national background may associate the rating categories with different intervals on the scale of perceptual quality. This will be checked by introducing independent thresholds $(\tau_1, \ldots, \tau_4)$ per country when scaling image quality from the response data of international studies.
- Extreme response style refers to individuals with a preference for choosing options at the extreme ends of rating scales, influenced by cultural backgrounds and personal styles. To study country-specific extreme response styles, we apply (1) the method of successive intervals, (2) a generalized linear mixed effects model with a parameter for each country accounting for the probability of extreme ratings, and (3) compare with the empirical proportions of extreme ratings.

An additional, technical, contribution is the adaptation of a lapse rate. In Thurstonian reconstruction, a stimulus of high quality may yield a probability for the low category 'bad' that is almost equal to zero. Thus, according to the model a 'bad' rating is extremely unlikely. Yet in practice, such ratings occasionally do occur.

This is because subjects briefly may be inattentive and make the wrong decision, or they may inadvertently press the wrong response button although they had made a correct decision (a 'finger error'). The effect is that in the MLE of the parameters of the model, such outliers receive an undue influence and distort the model parameters, thereby degrading the model quality. A lapse rate introduces a small prior probability for all categories that is then combined with the evidence, i.e., the ratings in the experiment. Such a lapse rate helps to ameliorate the negative effect of outliers. Lapse rates are commonly used in cognitive science to fit models of psychometric functions [8], but have not been considered for Thurstonian reconstructions from ACR/DCR response data before.

[1]The Law of Categorical Judgement is more general by letting the thresholds be random variables instead of fixed numbers. However, this allows the order of the thresholds to vary which complicates theory and algorithms.

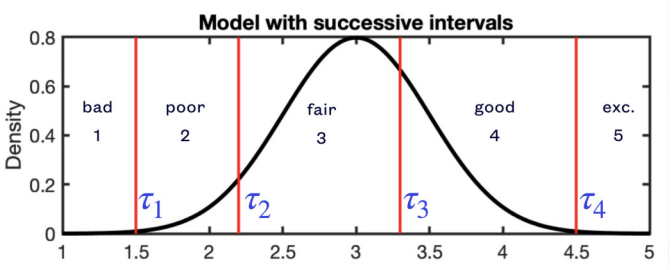


Fig. 1: Probabilities for ACR ratings 'bad' to 'excellent' can be modeled in a two-step process. The latent perceptual quality is taken as a normally distributed random variable, parameterised by its mean and variance. Secondly, ACR categories correspond to successive intervals on the quality scale, separated by thresholds $\tau_1 < \ldots < \tau_4$. The probabilities of an ACR rating is given by the area under the curve in the corresponding interval. Here the mean is 3.0 and the probability for a 'fair' rating (3) is largest.

TABLE I: Graphical scaling for the CCIR quality scale terms in two populations with different languages. Data from [9].

ACR	US		Italy	
Ordinal	Name	Value	Name	Value
5	excellent	6.5 ± 0.6	ottimo	6.4 ± 0.6
4	good	4.9 ± 0.7	buono	5.5 ± 0.7
3	fair	3.5 ± 0.8	discreto	4.3 ± 1.0
2	poor	1.4 ± 0.6	mediocre	1.9 ± 1.5
1	bad	1.1 ± 0.6	cattivo	1.5 ± 1.3

II. Related work

Language and culture influence the interpretation of perceived quality in CCIR terms for ACR and DCR (Consultative Committee on International Radio, founded 1927). Graphic scaling, a technique used in [9], showed non-uniform spacing of ACR categories on an interval scale between US American and Italian citizens, see Table I. This was confirmed in other studies, such as for Dutch-language terms [10].

Therefore, cultural effects are expected in image quality assessment studies. However, little work has been done to extract

TABLE II: Overview of datasets. The average number of ratings per image, subject, and country are given.

	KonIQ-10k [12] 2020	KADID-10k [13] 2019	NIVD [14] 2023
Images or videos	10076	11085	1860
Subjects	1261	2212	12812
Countries	75	72	4
Ratings/image	107.0	35.3	265.3
Ratings/subject	854.8	176.9	38.5
Ratings/country	14372.8	5435.8	123368
Ratings total	1077960	391376	493472
Rating type	ACR	DCR	VAS

TABLE III: The four countries with most ratings per dataset.

Dataset	Country	Subjects	Images	Ratings
KonIQ-10k	India	359	10074	423400
	Venezuela	212	10074	129236
	Russia	66	9871	62077
	Serbia	62	9884	49428
	Other	563	10076	413819
KADID-10k	Venezuela	1332	11085	269923
	Egypt	97	5980	17326
	India	83	5854	11784
	Russia	48	5122	9797
	Other	652	11070	82636
NIVD	US	3287	1860	129244
	Brazil	2963	1860	127720
	Japan	3298	1860	124308
	India	3264	1860	112200

these national differences. An international study [11] found no apparent influence of language or culture on the Mean Opinion Score (MOS) from ACR of audio-visual stimuli.

In [15], nationality was a significant factor for video quality perception. The study used a factorial analysis with personality, culture, and other predictors. 76 university students from four countries rated 144 videos. In [14] the authors created a much larger video quality dataset by collecting ratings from over 12000 subjects from four countries. The work focused on how spatial video resolution and screen size affected perceived quality. Only a few scatter plots revealed that there are national differences.

An international study on the preference of colorimetric adjustments of images showed significant differences due to the cultural background of viewers, but these were considered unimportant for applications [16]. For related international studies showing significant effects of cultural factors on color and visual appeal, see [17], [18].

Extreme response styles can vary across cultures. For example, individuals from individualistic cultures, like the United States, are often more inclined towards extreme responses than those from collectivist cultures, like East Asian cultures [19]–[21]. Even within the US, there are differences in extreme response styles among different ethnic groups [22], [23].

III. Materials and Methods

Here we describe the datasets and the statistical models used to extract country-specific traits in image and video quality assessment. The first model generates country-specific successive intervals for the five ACR/DCR categories, the amount of spread of the ratings, and the lapse rate. The second model extracts the statistical probability for extreme ratings.

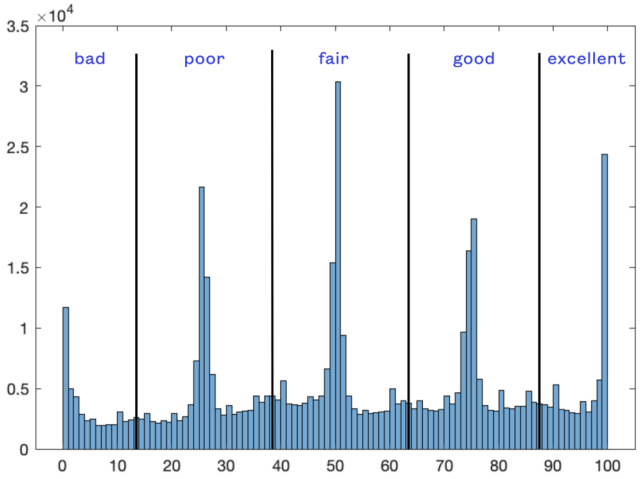


Fig. 2: Histogram of ratings in NIVD, showing the quantization of the percentages of the VAS to the five ACR categories.

A. Datasets

To ensure statistical evidence of our results, we focused on datasets with very large numbers of ratings from several countries. We selected KonIQ-10k [12], KADID-10k [13], and NIVD [14], see their summaries in Tables II and III. The first two are image quality datasets; KonIQ-10k uses no-reference IQA with ACR, and KADID-10k uses full-reference IQA with DCR. NIVD is a video quality dataset assessed on a visual analog scale (VAS).

KonIQ-10k and KADID-10k had been gathered by crowdsourcing without restrictions. That means subjects from any country were accepted, provided they fulfilled the qualification requirements. Moreover, there was no fixed number of stimuli that a subject could rate. Therefore, the resulting ratings are not uniformly distributed across subjects and countries.

For this reason, we considered in our analysis of KonIQ-10k and KADID-10k only the four countries that provided the most ratings and pooled the remaining ratings into a fifth one, called ‘other’. The ratings used in our study are a subset of those used in the original publications. We removed data from subjects with unknown nationality.

NIVD was designed to capture country-specific differences, collecting nearly an equal amount of ratings from four selected countries. Ratings were acquired using the SAMVIQ scheme using a visual analog scale (VAS) with tick marks and descriptive ACR labels. For our analysis, we quantized the continuous VAS ratings to discrete ACR ratings, as shown in Figure 2

The three annotated datasets are publicly available [24].

B. The quantized metric model and successive intervals

The common statistical models for the perceived quality of sensory stimuli assume a one-dimensional latent quality scale

TABLE IV: Results with 95% confidence intervals, compare Fig. 3.

Dataset	Country	Std deviation	Lapse rate	Category thresholds			
		σ	λ	τ_1	τ_2	τ_3	τ_4
KonIQ-10k	India	0.5050±0.0016	0.0039±0.0004	1.3867±0.0071	2.3608±0.0028	3.4061±0.0022	4.6590±0.0087
	Venezuela	0.4179±0.0022	0.0078±0.0011	1.6998±0.0086	2.5069±0.0042	3.2330±0.0033	4.1030±0.0064
	Russia	0.3813±0.0030	0.0038±0.0011	1.7161±0.0116	2.5190±0.0058	3.2646±0.0045	4.2292±0.0119
	Serbia	0.3811±0.0035	0.0087±0.0018	1.7089±0.0138	2.5043±0.0066	3.2889±0.0051	4.1533±0.0116
	Other	0.4132±0.0012	0.0053±0.0005	1.6536±0.0050	2.5007±0.0023	3.2752±0.0019	4.2205±0.0044
KADID-10k	Venezuela	0.6372±0.0024	0.0065±0.0009	1.7941±0.0047	2.7047±0.0036	3.2799±0.0036	4.1751±0.0045
	Egypt	0.6910±0.0104	0.0105±0.0042	1.5611±0.0218	2.7732±0.0147	3.2528±0.0147	4.4835±0.0211
	India	0.6442±0.0120	0.0144±0.0056	1.7302±0.0240	2.8174±0.0178	3.3726±0.0179	4.4110±0.0240
	Russia	0.5403±0.0111	0.0058±0.0038	1.8995±0.0221	2.7440±0.0185	3.3349±0.0183	4.1006±0.0208
	Other	0.6013±0.0043	0.0125±0.0019	1.8659±0.0082	2.7664±0.0065	3.3550±0.0065	4.1922±0.0080
NIVD	Japan	0.7028±0.0038	0.0356±0.0026	1.8249±0.0079	2.8243±0.0054	3.7092±0.0056	4.5132±0.0084
	Brazil	0.6343±0.0035	0.0353±0.0027	1.8820±0.0071	2.6355±0.0049	3.3261±0.0049	4.1522±0.0068
	US	0.7603±0.0044	0.0543±0.0036	1.6418±0.0091	2.4355±0.0059	3.1706±0.0055	4.1098±0.0075
	India	0.7467±0.0044	0.0416±0.0033	1.5897±0.0099	2.4910±0.0061	3.2721±0.0058	4.2185±0.0082

of real numbers that is shared by all subjects, but not directly observable. The actual responses in a subjective experiment are also affected by the decisional process that is modulated by personal and cultural influence. In addition, a third layer given by errors in the physical action of communicating the decision by, e.g., a mouse click, may distort the decided rating (so-called finger errors or lapses).

Each stimulus j corresponds to a particular value $\psi_j \in \mathbb{R}$ on the real latent quality scale. The quality as perceived by a subject is modeled by a random variable U_j. In the most basic model, U_j is chosen with a normal distribution centered at the latent quality value ψ_j and a with a global variance σ^2. With this setting, we have

$$U_j = \psi_j + \sigma W \tag{1}$$

where U_j is the random variable producing the observed opinion score for stimulus j. The term σW is the error for the latent variable, which summarizes many random influences, e.g., from differences between test subjects, observation conditions and environmental factors. In regression models, it is common to assume that these errors are normally distributed [25]. Thus, $W \sim N(0,1)$ and $\sigma > 0$ is the standard deviation of U_j, determining the spread of the random variates u_j.

To account for the finite discrete nature of ACR-type data with $K = 5$ categories, we sort the real values of U_j into K successive intervals. For this purpose, we introduce a monotonic sequence of thresholds $\tau = (\tau_0, ..., \tau_K)$,

$$-\infty = \tau_0 \leq \tau_1 \leq \cdots \leq \tau_{K-1} \leq \tau_K = \infty, \tag{2}$$

and define the quantization function $Q_\tau : \mathbb{R} \to \{1, ..., K\}$ by

$$Q_\tau(u) = k \iff \tau_{k-1} \leq u < \tau_k. \tag{3}$$

Given a metric model for the j-th stimulus in the form of a continuous random variable U_j, we define the corresponding quantized metric model by the discrete random variable $Q_\tau(U_j)$. Note that in spite of the normally distributed scores on the latent scale, the resulting ACR distributions may be skewed.

In addition to the quantization, we take into account a lapse rate $\lambda \geq 0$. This yields a discrete random variable V_j that determines the probability of a rating for category k as

$$\Pr[V_j = k] = (1 - \lambda)(G_{\psi_j,\sigma}(\tau_k) - G_{\psi_j,\sigma}(\tau_{k-1})) + \frac{\lambda}{K} \tag{4}$$

where $G_{\psi_j,\sigma}$ denotes the Gaussian cumulative density function with mean ψ_j and variance σ^2. Thus, with zero lapse rate, $\Pr[V = k]$ is just the area under the Gaussian between the thresholds τ_k and τ_{k-1}, as shown in Figure 1.

The above model cannot yet distinguish between ratings from different nationalities. To achieve this, we simply define the following parameters separately for each country: the rating spread σ, the lapse rate λ, and the category thresholds $\tau_1, ..., \tau_4$. Thus, the total number of parameter is equal to the number of stimuli plus six times the number of countries.

The parameters are obtained by MLE. For the optimization we apply an interior point algorithm, implemented in the MATLAB function `fmincon`.

C. Generalized linear mixed-effects model for extreme response styles

Extreme ratings on the 5-level ACR and DCR scales are those at the ends of the scale, i.e., corresponding to values 1 and 5. Various statistical methods can be employed to identify an extreme response style [2], [3], [26]. To assess the probability of extreme ratings per country we apply a generalized linear mixed-effects (GLME) model that incorporates a fixed effect per country and a random effect for the stimuli. Let $p_{j,c}$ be the probability of extreme ratings for the image j from subjects of country c. Then, using the standard link function $\mathrm{logit}(p) = \log \frac{p}{1-p}$ for binomial models, we have

$$p_{j,c} = \mathrm{logit}^{-1}(\alpha_j + \beta_c) \tag{5}$$

where β_c is the parameter for the fixed effect for country c, and α_j is the random effect from the j-th stimulus. The differences between the stimuli (content, color, contrast, etc.) can be assumed to add up to a normal distribution, as is the case with many sums of independent random variables. For this reason

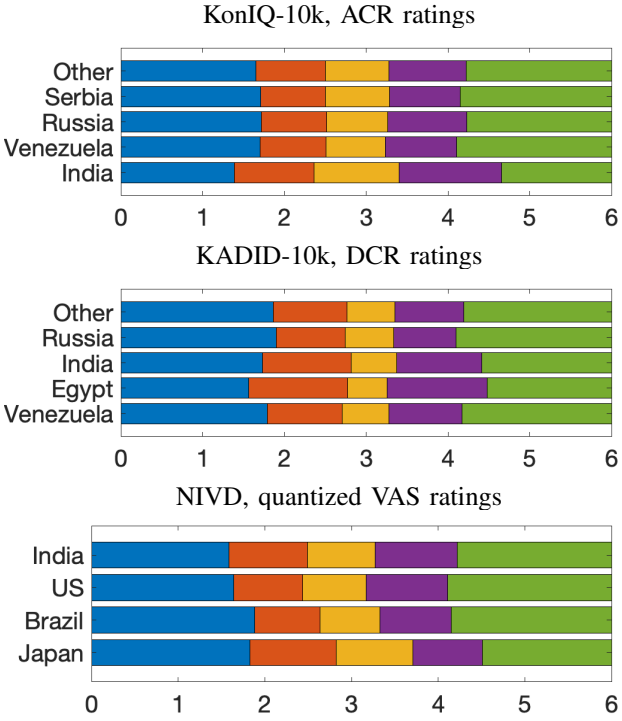


Fig. 3: Results for the country-specific thresholds. For the numerical values and confidence intervals see Table IV.

and for computational and mathematical convenience, random effects are typically assumed to be normally distributed with mean value of zero. Then the estimated probability p_k of extreme ratings for country c is

$$p_c = \text{logit}^{-1}(\beta_c). \tag{6}$$

We used the `lme4` package in R [27] to fit the GLME models and their confidence intervals, as it provides a convenient and efficient way to specify and estimate linear and nonlinear mixed-effects models with various link functions and error distributions. The model can be described in the Wilkinson-Rogers notation as $\texttt{extreme} \sim -1 + \texttt{country} + (1|\texttt{image})$.

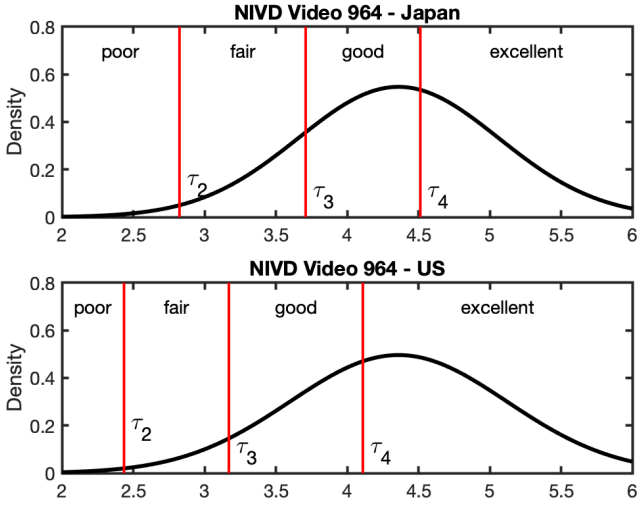
Another estimation of the probabilities p_c for extreme ratings in country c can be based on the quantized metric model with successive intervals by averaging the probability $\Pr[V_j \in \{1, 5\}]$ from Equation (4) over all stimuli j.

IV. Results

The results of the data analysis using the method of successive intervals is shown in Table IV and Figure 3. The scale values for stimuli were also estimated, but are not shown here to keep the focus on the country-specific differences.

Clearly, most thresholds τ_k, and also the standard deviations and lapse rates, significantly differ between countries. For example, in the first two rows of the table for the KonIQ-10k ratings of India and Venezuela, all parameters differ between the countries without overlap of 95% confidence intervals.

To illustrate these results let us consider one example in more detail. In NIVD the video 964 was scaled by the



Country	Model probabilities and MOS 1	2	3	4	5	MOS
Japan	0.0073	0.0209	0.1640	0.4016	0.4062	4.1785
US	0.0110	0.0161	0.0612	0.3060	0.6057	4.4793

Country	ACR fractions and MOS 1	2	3	4	5	MOS
Japan ($N = 69$)	0.0000	0.0290	0.2319	0.2754	0.4638	4.1739
US ($N = 34$)	0.0000	0.0000	0.0000	0.3529	0.6471	4.6471

Fig. 4: Results of the model with successive intervals for the video stimulus numbered 964 in the Netflix International Video Dataset, shown for Japan and US. The category thresholds for the subjective ratings of the perceived quality in Japan are larger than those in the US. In effect, according to the statistical model, the sampled US population generally preferred higher ACR ratings for the video stimuli in NIVD. The table shows the numerical values of the resulting probabilities of the ACR categories for this example. Each of the model probabilities is the corresponding area under the curves plus 1/5 of the lapse rate (0.0356 for Japan and 0.0543 for the US), see Equation (4). For comparison, below we show the fractions of the collected VAS ratings that were quantized to ACR for this study.

statistical model at quality $\mu = 4.360$. The distribution of the latent perceived video quality corresponding to the model parameters for Japan and the US (lines 11 and 13 in Table IV) are pictured in Figure 4. By the assumption of a globally unique perceived quality, we have that for all countries the mean of the distribution is at $\mu = 4.360$. The dispersion of the qualities, the lapse rates, and the ACR category thresholds are different, though. This implies probabilities for the ACR categories that differ between the countries as shown in the table below the plots.

The table also confirms for this example that the model presents an accurate fit to the collected ratings. The corresponding probabilities for the five categories are close to each

TABLE V: Probabilities of extreme ratings with 95% confidence intervals. Results of the GLME model, the Thurstonian successive intervals, and the empirical proportions of ratings in extreme categories 1 and 5 together. Rows are sorted according to their magnitudes.

Dataset	Country	GLME model Prob	GLME model CI	Thurstone Prob	ACR Prop
KonIQ	Venezuela	0.0314	[0.0302,0.0326]	0.0662	0.0674
	Serbia	0.0225	[0.0213,0.0237]	0.0508	0.0505
	Other	0.0201	[0.0194,0.0208]	0.0462	0.0479
	Russia	0.0186	[0.0177,0.0196]	0.0428	0.0462
	India	0.0086	[0.0083,0.0090]	0.0220	0.0201
KADID	Russia	0.313	[0.302,0.324]	0.346	0.375
	Other	0.284	[0.278,0.289]	0.327	0.347
	Venezuela	0.277	[0.272,0.282]	0.322	0.337
	India	0.206	[0.198,0.215]	0.260	0.279
	Egypt	0.172	[0.166,0.179]	0.225	0.240
NIVD	US	0.233	[0.226,0.240]	0.260	0.255
	Brazil	0.218	[0.211,0.224]	0.249	0.238
	India	0.192	[0.186,0.198]	0.223	0.213
	Japan	0.161	[0.156,0.167]	0.191	0.189

other; the measured MOS from the collected ratings differs from the predicted MOS of the model by only about 0.5%.

To study country-specific differences of extreme ratings we computed their occurrences by (a) averaging the probability $\Pr[V_j \in \{1,5\}]$ from the Thurstonian model (4) over all stimuli j per country, (b) the GLME model, and (c) the sum of the empirical proportions of ratings at ACR levels 1 and 5. Table V summarized the results. Clearly, there are significant differences between countries. The largest differences were found for the ACR modality in KonIQ-10k, in which extreme ratings from Venezuela were about three times more likely than those of India.

We note that the three methods of assessment of occurrences of extreme ratings unanimously agree on the ranking of the countries according to the frequencies of extreme ratings. The Thurstonian probabilities are very close to the empirically measured frequencies. However, the probability estimates by the GLME model are smaller, in particular for KonIQ-10k. This effect may be due to several reasons. Firstly, there are differences in the basic setup of the estimation methods. The GLME model assesses the extreme rating probability assuming a normal distribution of the random effects from all stimuli, while the Thurstonian model does not. Moreover, the datasets include test images that were designed to yield extreme ratings, which may incur differing biases.

V. Limitations

We only considered the country of living as a proxy for cultural background, which may not capture the full diversity and complexity of cultural influences on response styles. For example, within-country variations, such as regional, ethnic, or linguistic differences, may also affect how people rate image quality. Other individual factors, such as age, gender, education, or personality may also interact with cultural factors and influence response styles. In addition, technical factors like device characteristics and viewing conditions may have an effect. Future studies should incorporate more fine-grained and multidimensional measures of culture and individual differences to better understand the sources and mechanisms of response styles.

We introduced the lapse rate in the statistical model for ACR/DCR quality assessment. A general analysis on the benefits and limitations of lapse rates in Thurstonian models is outstanding but beyond the scope of this contribution.

A limitation is the large runtime to compute the parameters of model with successive intervals, when the dataset is large. For example, to execute MLE for 10092 parameters for KonIQ-10k took 13h with Matlab on a MacBook Pro. We did not apply any code optimization and did not try alternative solvers like ADAM.

VI. Conclusion

We explored the impact of cultural factors on image quality assessment by adapting statistical models to include country-specific components. We applied our models to three large-scale image quality datasets, KonIQ-10k, KADID-10k and NIVD, and found significant country effects on extreme response styles. For example, Russian observers were more likely to provide extreme ratings than Egyptian ones. These results highlight the importance of considering cultural nuances in image quality assessment, as ignoring them may lead to distorted interpretations of cultural differences in image quality perceptions. Our study contributes to a more comprehensive understanding of image quality in a global context and has implications for the analysis of current datasets. To refine our understanding, we recommend further exploration into the cultural factors contributing to these differences and adjustments to image quality assessment surveys to minimize the impact of extreme response tendencies and ensure more accurate cross-cultural comparisons.

Our study has implications for the field of image quality assessment and cross-cultural research. One implication is that researchers should be aware of the potential effects of extreme response styles on their results and interpretations. Extreme response styles can introduce biases and distortions in the perceived cultural differences in image quality perceptions. Therefore, researchers should consider adjusting their methods and analyses to account for these effects, such as using statistical models that include country-specific parameters or modifying the design of questionnaires to reduce the sensitivity to response styles.

Another implication is that researchers should explore the underlying cultural factors that may contribute to the observed variations in response styles. For example, some possible factors are the degree of individualism or collectivism, the value of moderation or expressiveness, and the preference for direct or indirect communication. Understanding these factors can help to better explain and predict the cross-cultural differences in image quality assessment and to design more culturally appropriate and effective surveys and interventions.

REFERENCES

[1] Irvine Clarke III, "Extreme response style in cross-cultural research: An empirical investigation.," *Journal of Social Behavior & Personality*, vol. 15, no. 1, 2000.

[2] Eric A Greenleaf, "Measuring extreme response style," *Public Opinion Quarterly*, vol. 56, no. 3, pp. 328–351, 1992.

[3] Martijn G De Jong, Jan-Benedict EM Steenkamp, Jean-Paul Fox, and Hans Baumgartner, "Using item response theory to measure extreme response style in marketing research: A global investigation," *Journal of Marketing Research*, vol. 45, no. 1, pp. 104–115, 2008.

[4] Louis L Thurstone, "A law of comparative judgment," *Psychological Review*, vol. 101, no. 34, 1927.

[5] Zhi Li, Christos G Bampis, Lukáš Krasula, Lucjan Janowski, and Ioannis Katsavounidis, "A simple model for subject behavior in subjective experiments," in *IS&T International Symposium on Electronic Imaging*. IS&T, 2020.

[6] Warren S Torgerson, *Theory and Methods of Scaling*, Wiley, 1958.

[7] Torrin M Liddell and John K Kruschke, "Analyzing ordinal data with metric models: What could possibly go wrong?," *Journal of Experimental Social Psychology*, vol. 79, pp. 328–348, 2018.

[8] Felix A Wichmann and N Jeremy Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit," *Perception & Psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.

[9] Bronwen L Jones and Pamela R McManus, "Graphic scaling of qualitative terms," *SMPTE Journal*, vol. 95, no. 11, pp. 1166–1171, 1986.

[10] Kees Teunissen, "The validity of CCIR quality indicators along a graphical scale," *SMPTE Journal*, vol. 105, no. 3, pp. 144–149, 1996.

[11] Margaret H Pinson, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Christian Schmidmer, Phillip Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 640–651, 2012.

[12] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[13] Hanhe Lin, Vlad Hosu, and Dietmar Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[14] Christos G Bampis, Lukáš Krasula, Zhi Li, and Omair Akhtar, "Measuring and predicting perceptions of video quality across screen sizes with crowdsourcing," in *15th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2023, pp. 13–18.

[15] Sharath Chandra Guntuku, Michael James Scott, Huan Yang, Gheorghita Ghinea, and Weisi Lin, "The CP-QAE-I: A video dataset for exploring the effect of personality and culture on perceived quality and affect in multimedia," in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2015, pp. 1–7.

[16] Scot R Fernandez, Mark D Fairchild, and Karen Braun, "Analysis of observer and cultural variability while generating "preferred" color reproductions of pictorial images," *Journal of Imaging Science and Technology*, vol. 49, no. 1, pp. 96, 2005.

[17] Dianne Cyr, Milena Head, and Hector Larios, "Colour appeal in website design within and across cultures: A multi-method evaluation," *International Journal of Human-Computer Studies*, vol. 68, no. 1-2, pp. 1–21, 2010.

[18] Martin Varela, Toni Mäki, Lea Skorin-Kapov, and Tobias Hoßfeld, "Towards an understanding of visual appeal in website design," in *2013 Fifth international workshop on quality of multimedia experience (QoMEX)*. IEEE, 2013, pp. 70–75.

[19] Chuansheng Chen, Shin-ying Lee, and Harold W Stevenson, "Response style and cross-cultural comparisons of rating scales among East Asian and North American students," *Psychological Science*, vol. 6, no. 3, pp. 170–175, 1995.

[20] Kl-Taek Chun, John B Campbell, and Jong Hae Yoo, "Extreme response style in cross-cultural research: A reminder," *Journal of Cross-Cultural Psychology*, vol. 5, no. 4, pp. 465–480, 1974.

[21] Li Lian Ho, Poh Cheng Loh, and Ai Ling Quah, "A cross-cultural, between-gender study of extreme response style," Nanyang Technological University, 1995.

[22] Irvine Clarke III, "Extreme response style in cross-cultural research," *International Marketing Review*, vol. 18, no. 3, pp. 301–324, 2001.

[23] Allyson L Holbrook, Melanie C Green, and Jon A Krosnick, "Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias," *Public Opinion Quarterly*, vol. 67, no. 1, pp. 79–125, 2003.

[24] Workgroup Multimedia Signal Processing, University of Konstanz, Germany, "The Konstanz Visual Quality Databases," https://database.mmsp-kn.de.

[25] Paul-Christian Bürkner and Matti Vuorre, "Ordinal regression models in psychology: A tutorial," *Advances in Methods and Practices in Psychological Science*, vol. 2, no. 1, pp. 77–101, 2019.

[26] Guy Moors, "Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined," *Quality and Quantity*, vol. 37, pp. 277–302, 2003.

[27] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.