VQEG Meeting
Dec 18-21, 2023

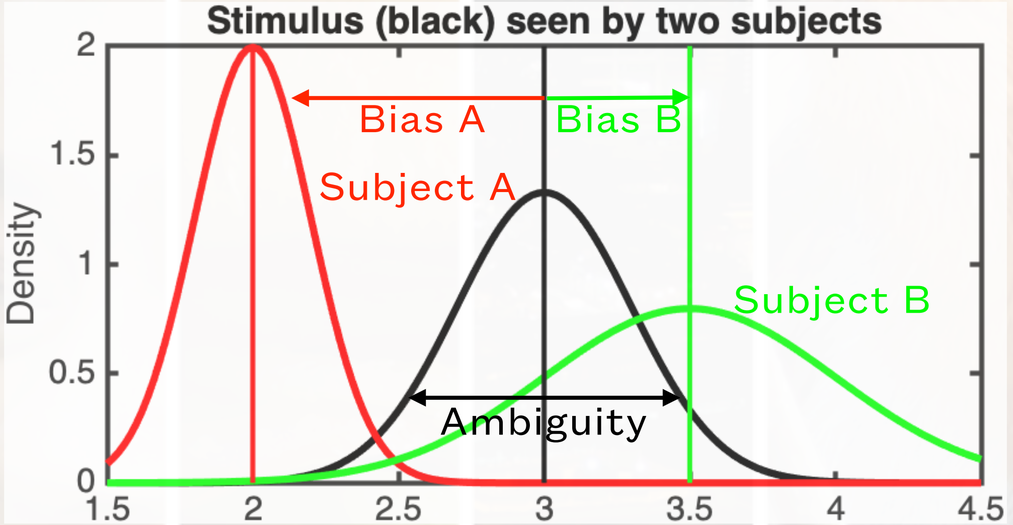
National Differences in IQA
An Investigation on three large-scale IQA datasets

DIETMAR SAUPE, UNIVERSITY OF KONSTANZ, GERMANY – SIMON DEL PIN, NTNU, GJØVIK, NORWAY

SFB-TRR 161
Quantitative Methods for Visual Computing

Overview / Motivation: Statistical models in subjective IQA

- People rate image quality differently. Statistical models in P.913/BT.500 consider subject-specific features:
  - additive bias and variance (inconsistency).
- Cultural psychology found significant national differences in various areas.
- Our hypothesis: Similar differences exist in the perception of image quality.
- Investigation of statistical models with country-specific components for three datasets
  - KonIQ-10k,
    KADID-10k,
    NIVD (=Netflix International Video Dataset).
- National differences could be relevant for
  - design and analysis of crowdsourcing studies for IQA,
  - services to adaptively stream content worldwide.



Previous work

Previous work done on extracting cross-national differences in rating behavior.
None of these presented a model for such differences.

- Bampis, C. G., Krasula, L., Li, Z., & Akhtar, O. **Measuring and Predicting perceptions of video quality across screen sizes with crowdsourcing**. QoMEX 2023.
  - large dataset (1860 videos, 14k subjects), well-balanced over 4 countries
  - biases across nations observed but not analyzed

- Pinson, M. H., Janowski, et al (2012). **The influence of subjects and environment on audiovisual subjective tests: An international study**. IEEE Journal of Selected Topics in Signal Processing, 6(6), 640-651.
  - 1 set of stimuli, 4 countries, 10 datasets of AV quality (ACR)
  - preliminary finding: Datasets „appeared not to be influenced by language or culture“.

- Guntuku, S. C., Scott, M. J., Yang, H., Ghinea, G., & Lin, W. **The CP-QAE-I: A video dataset for exploring the effect of personality and culture on perceived quality and affect in multimedia**. QoMEX 2015.
  - detected significant differences in perceived video quality between subjects from 4 nations

In VQA dominance of direct single stimulus assessment ACR and VAS scales

Absolute Category Rating (ACR)

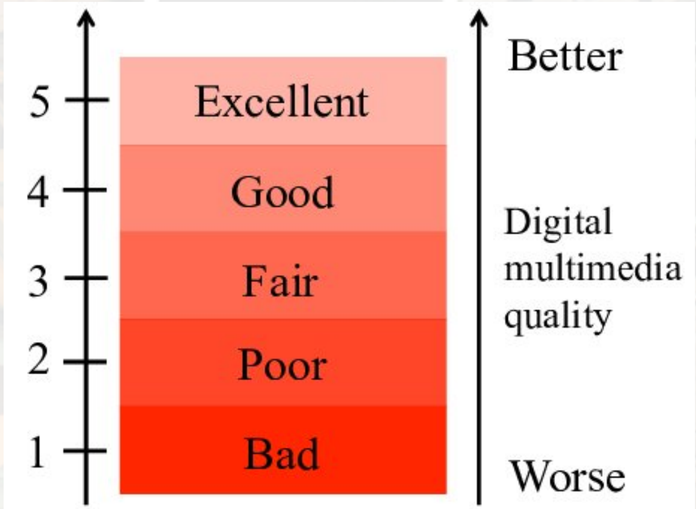
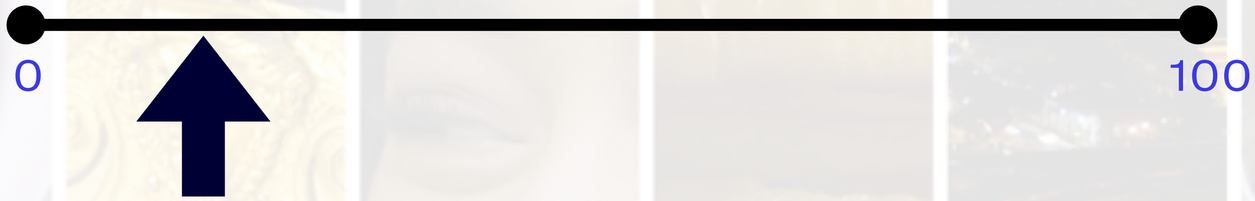


Figure: Z. Akhtar

Examples:
- KonIQ-10k
- KADID-10k
- NIVD (converted from VAS)

Visual Analog Scale (VAS)

Example: NIVD



Select quality on the scale

Mean opinion scores (MOS):

- MOS = Mean of all subject ratings for a stimulus
- MOS with subject model (P.913)

Graphic scaling

ACR categories (bad, …, excellent) are **ordinal.**
On a continuous scale they are

- not equidistant like 1,2,3,4,5
- vary between subjects from different countries

Conclusions:
Scaling from ACR/DCR data may benefit from

- **Country-specific models**
- Assigning successive intervals separated by thresholds to ACR/DCR categories:

Italy | US

ACR Categories	Values US N = 37	Values Italy N = 24
excellent ottimio	6.5 +/- 0.6	6.4 +/- 0.6
good buono	4.9 +/- 0.7	5.5 +/- 0.7
fair discreto	3.5 +/- 0.8	4.3 +/- 1.0
poor mediocre	1.4 +/- 0.6	1.9 +/- 1.5
bad cattivo	1.1 +- 0.6	1.5 +- 1.3

On a visual analog scale of 7.1 inch.

Jones, B. L., & McManus, P. R. (1986). Graphic scaling of qualitative terms. *SMPTE Journal*, *95*(11), 1166-1171
Watson, A. O. (2001). Assessing the quality of audio and video components in desktop multimedia conferencing. University of London.

Thurstonian model

- Global unique perceived image quality Q
- Random effect (Gaussian, equal variance)

$$Q \sim N(\mu_j, \sigma^2) \quad \text{with cdf} \quad F_{\mu_j,\sigma^2}$$

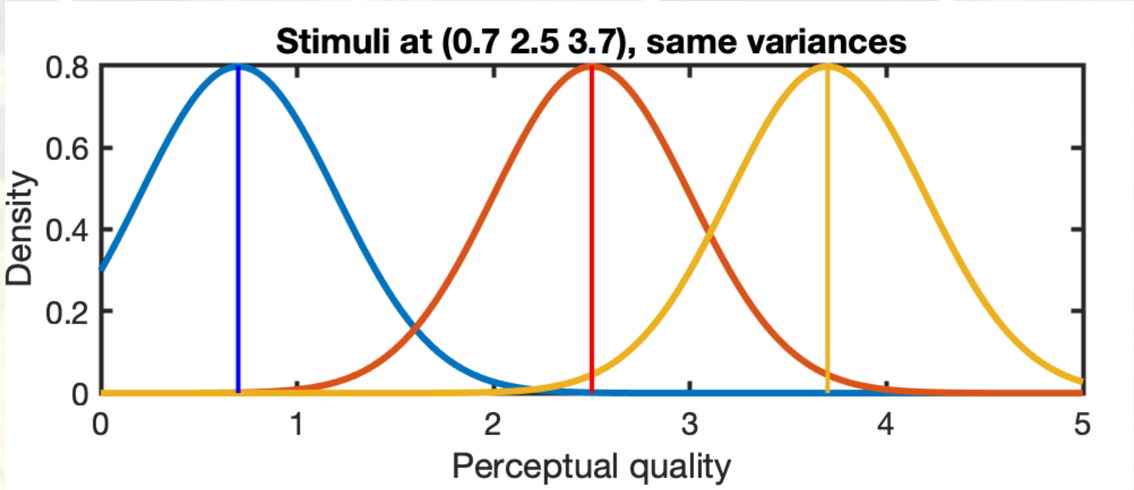
- Country-specific ACR thresholds $\tau_1^k < \cdots < \tau_4^k$
- Global lapse rate $0 < \lambda < 0.2$ (new)
- Probability for ACR rating $m = 1,..,5$

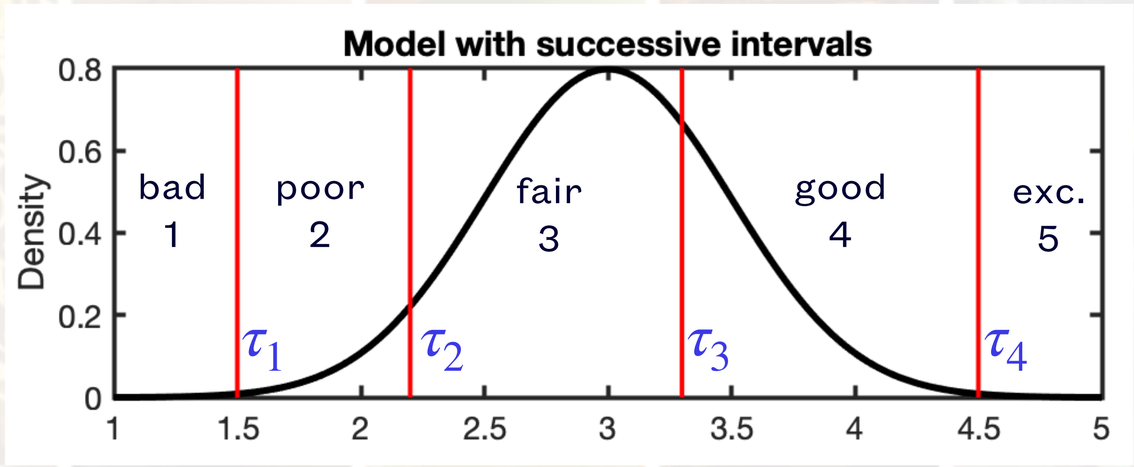
$$\text{Prob}(\text{ACR} = m \mid \text{image } j, \text{country } k) =$$

$$(1-\lambda)(F_{\mu_j,\sigma^2}(\tau_m^k) - F_{\mu_j,\sigma^2}(\tau_{m-1}^k)) + \lambda$$

- To normalize the scale and anchor results

$$\tau_1^k = 1.5 \text{ and } \tau_4^k = 4.5$$





Binomial model

- Extreme response style:
  Some people prefer choosing the most extreme options on a rating scale.
- Rating = 2,3,4 -> extreme = 0
  Rating = 1,5 -> extreme = 1
- Is there a significant difference between nationalities for IQA ratings?
- Generalized linear mixed effects model
  - Family: binomial
    Link function: logit
    Formula: extreme ~ -1 + country + (1 | image)

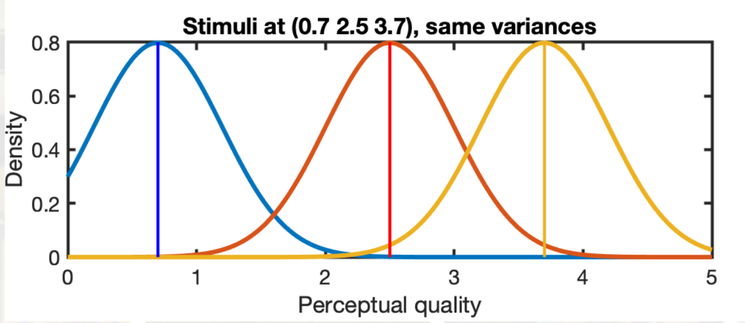
    $\text{Prob}(X_k = \text{extreme} \mid \text{image } j) = \text{logit}^{-1}(\alpha_k + U_j), \quad U_j \sim N(0,\sigma^2)$
  - Fixed effect per country, random effect per image

Clarke III (2000). Extreme response style in cross-cultural research: An empirical investigation. Journal of Social Behavior & Personality, 15, 137–152.

Computational methods

Thurstonian model

- Maximum likelihood estimation (MLE)
- Parameters $(\tau_m^k, \mu_j, \sigma, \lambda)$
- Nonlinear optimization:
  - interior point method, Matlab: `fmincon`
- Confidence intervals:
  - asymptotic Cramer-Rao bounds

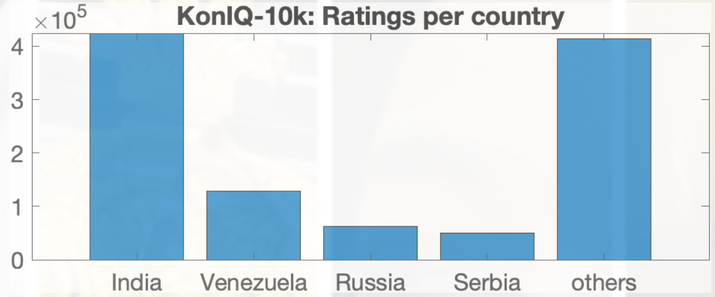


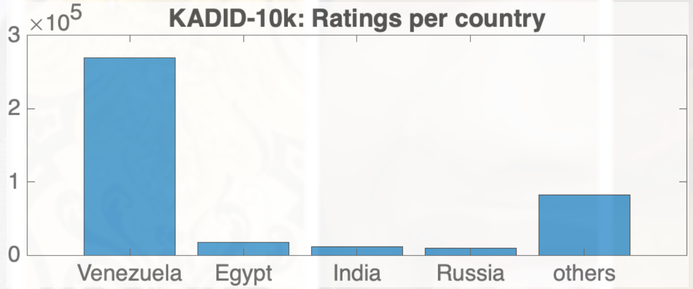
Binomial model

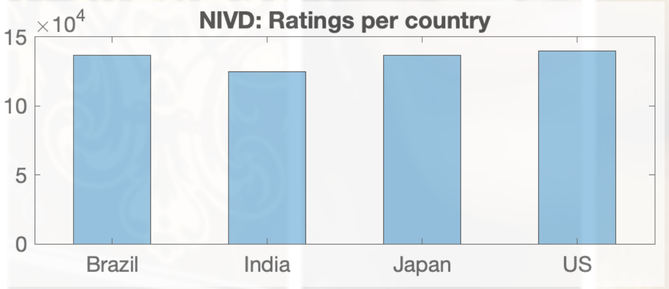
- Bayesian or frequentist analysis possible:
  - R: lme4 library
  - Matlab: `fitglme`
- Parameters (α_k, σ)
- Confidence intervals:
  - Wald method
  - bootstrapping (same results)

International crowdsourced datasets

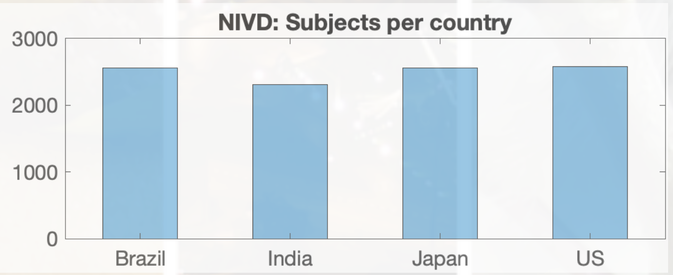
	Images Videos	Ratings	Subjects	Countries
KonIQ	10076	1,078,176 ACR	1261	75
KADID	11085	391,376 DCR	2212	72
NIVD	1860	538,200 VAS	10000	4







Country	Subjects
IND	359
VEN	212
RUS	66
SRB	62
Other	563

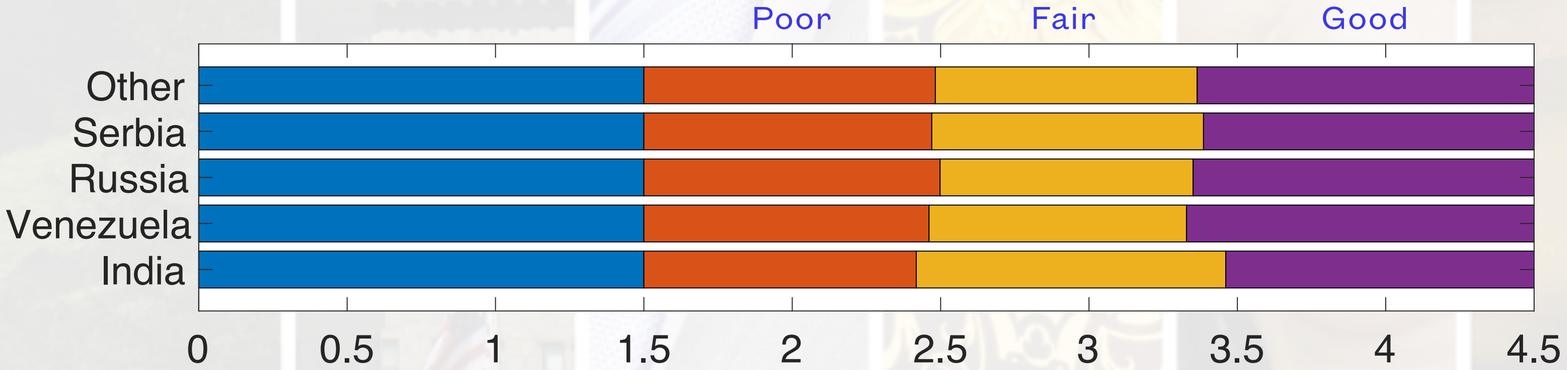


KonIQ-10K: Hosu, Lin, Sziranyi, & Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE TIP 29 (2020) 4041-4056.
KADID-10k: Lin, Hosu, & Saupe. KADID-10k: A large-scale artificially distorted IQA database. QoMEX 2019.
NIVD: Bampis, Krasula, Li & Akhtar, Measuring and predicting perceptions of video quality across screen sizes with crowdsourcing, QoMEX 2023.

Results KonIQ-10k — Thresholds

- Global lapse rate = 0.0051 +/- 0.0003
- Thresholds and 95%-confidence intervals:

	India	Venezuela	Russia	Serbia	Others
tau_3	3.4609+/-0.0021	3.3285+/-0.0041	3.3509+/-0.0053	3.3859+/-0.0062	3.3646+/-0.0022
tau_2	2.4180+/-0.0027	2.4610+/-0.0050	2.4983+/-0.0067	2.4702+/-0.0078	2.4823+/-0.0027
sigma	0.4808+/-0.0015	0.5297+/-0.0028	0.4472+/-0.0036	0.4707+/-0.0042	0.4821+/-0.0015

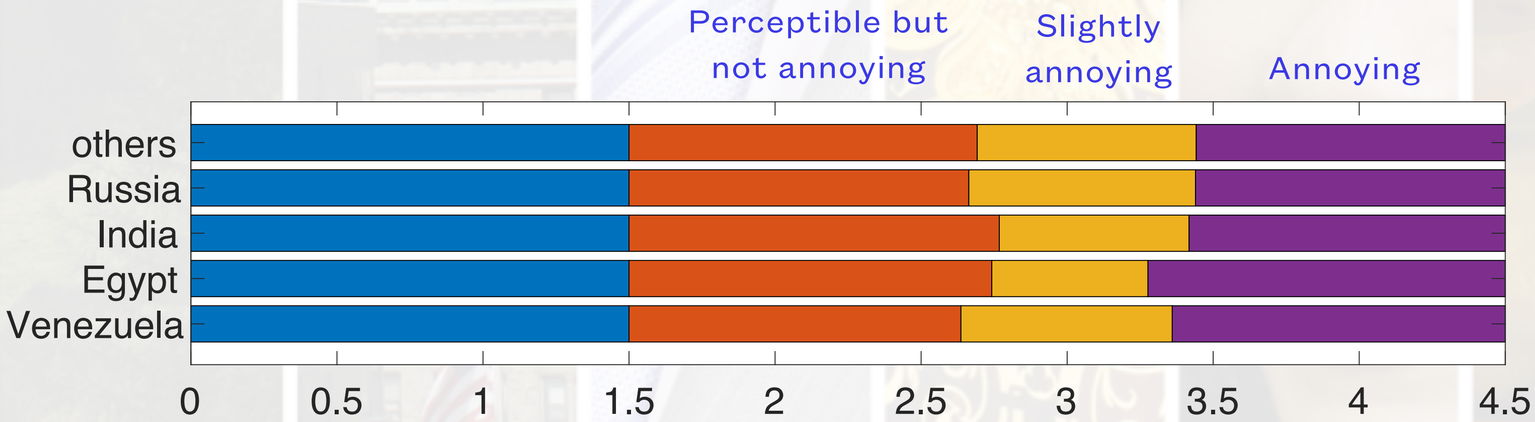


- Findings:
  The ACR interval for „Fair“ is widest for India and smallest for Russia.
  The variance is largest for Venezuela and smallest for Russia.

Results KADID-10k — Thresholds

- Global lapse rate = 0.0078 +/- 0.0008
- Thresholds and 95%-confidence intervals:

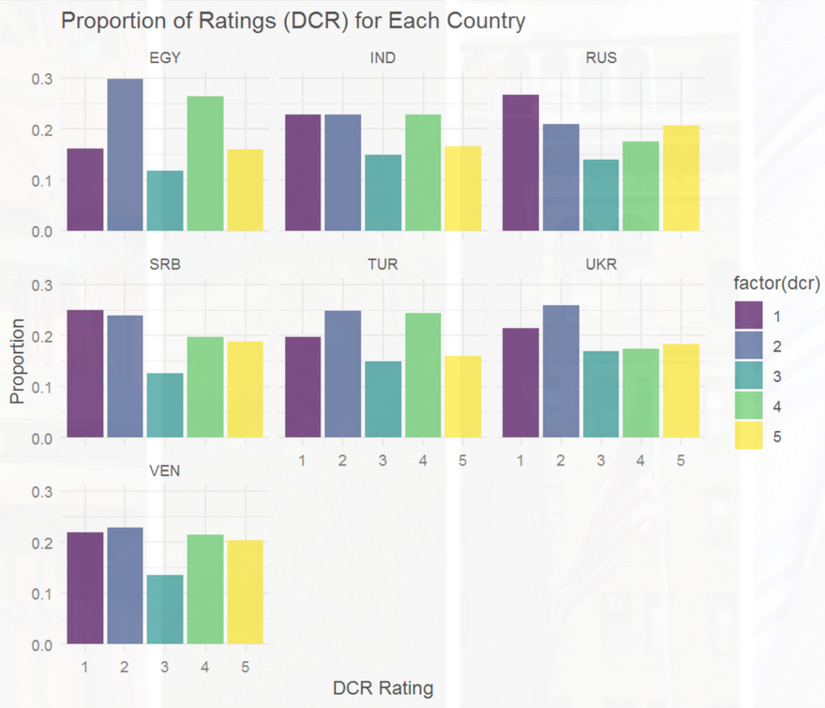
	Venezuela	Egypt	India	Russia	Others
tau_3	3.3598+/-0.0046	3.2768+/-0.0161	3.4174+/-0.0204	3.4395+/-0.0243	3.4420+/-0.0082
tau_2	2.6363+/-0.0046	2.7420+/-0.0162	2.7677+/-0.0208	2.6634+/-0.0245	2.6921+/-0.0084
sigma	0.7997+/-0.0030	0.7651+/-0.0111	0.7609+/-0.0137	0.7154+/-0.0150	0.7765+/-0.0054

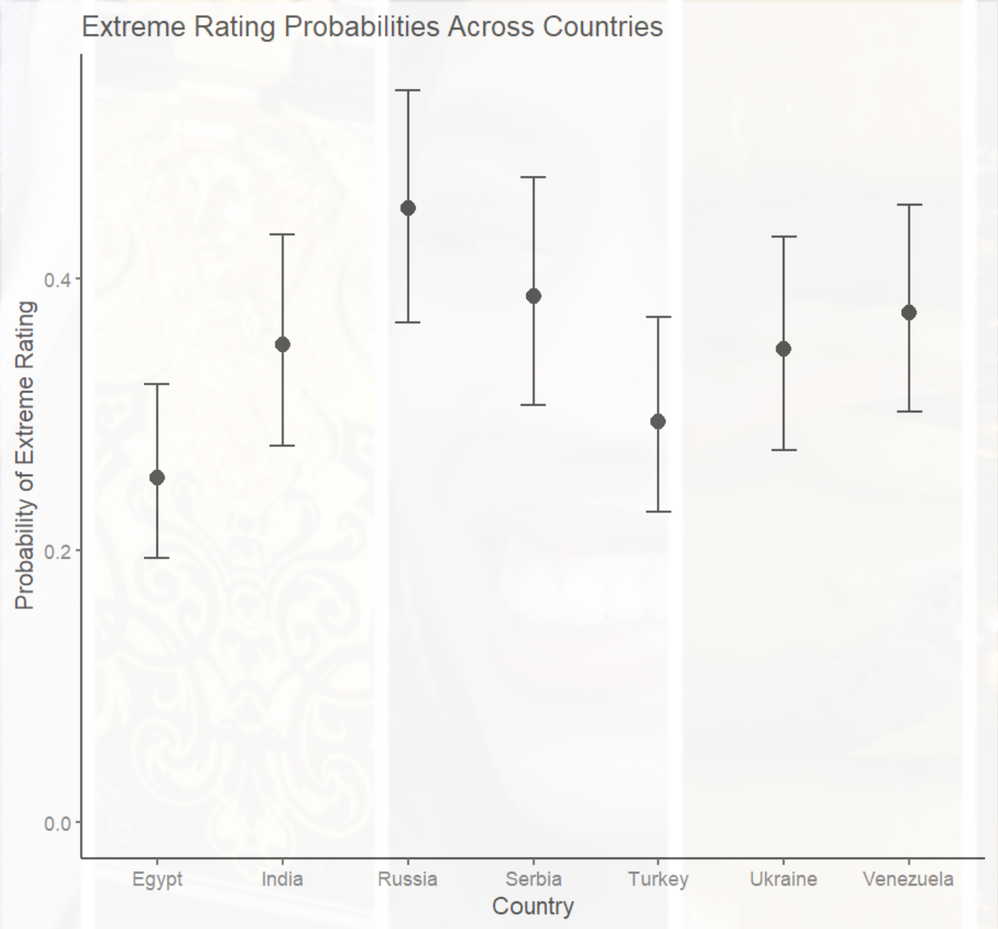


- Findings:
  The DCR interval for „Slightly Annoying“ is smallest for Egypt.
  The variance is largest for Venezuela.

Results KADID-10k — extreme ratings

Select countries and images:
- Each country has ≥ 1000 ratings
- Each image has ≥ 500 ratings
- Result: 7 countries, 67 images





Confidence Intervals

Wald formula

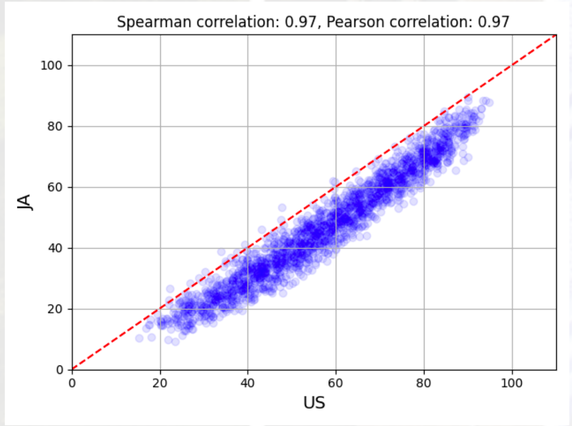
	2.5 %	97.5 %
countryEGY	0.19	0.32
countryIND	0.28	0.43
countryRUS	0.37	0.54
countrySRB	0.31	0.47
countryTUR	0.23	0.37
countryUKR	0.27	0.43
countryVEN	0.30	0.45

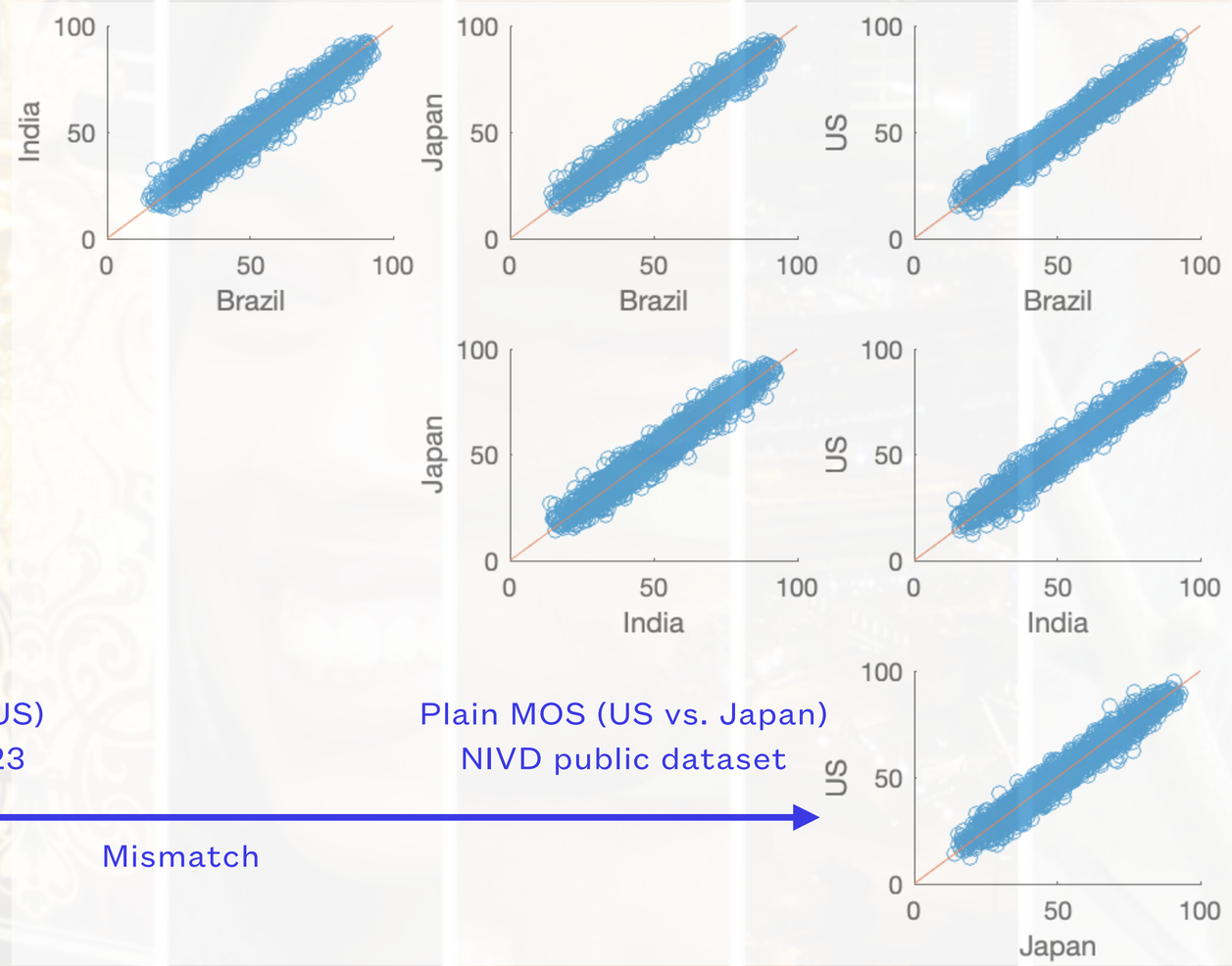
Bootstrapping

	2.5 %	97.5 %
countryEGY	0.19	0.32
countryIND	0.28	0.43
countryRUS	0.37	0.54
countrySRB	0.31	0.48
countryTUR	0.23	0.37
countryUKR	0.27	0.43
countryVEN	0.30	0.46

Results NIVD — VAS

- Plain MOS scatterplots country vs. country
- No apparent country-specific differences





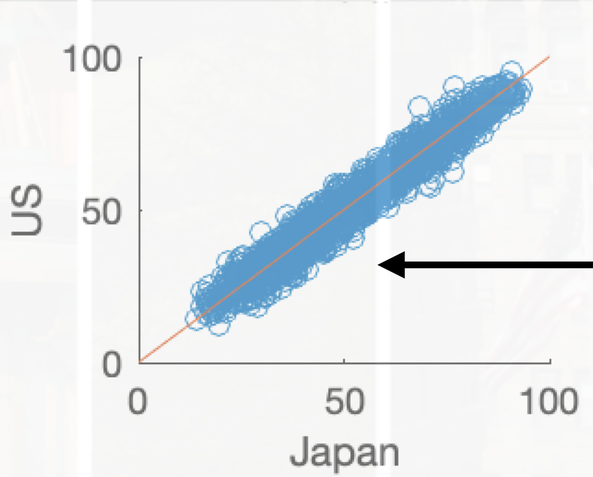
Plain MOS (Japan vs. US)
NIVD paper QoMEX'23

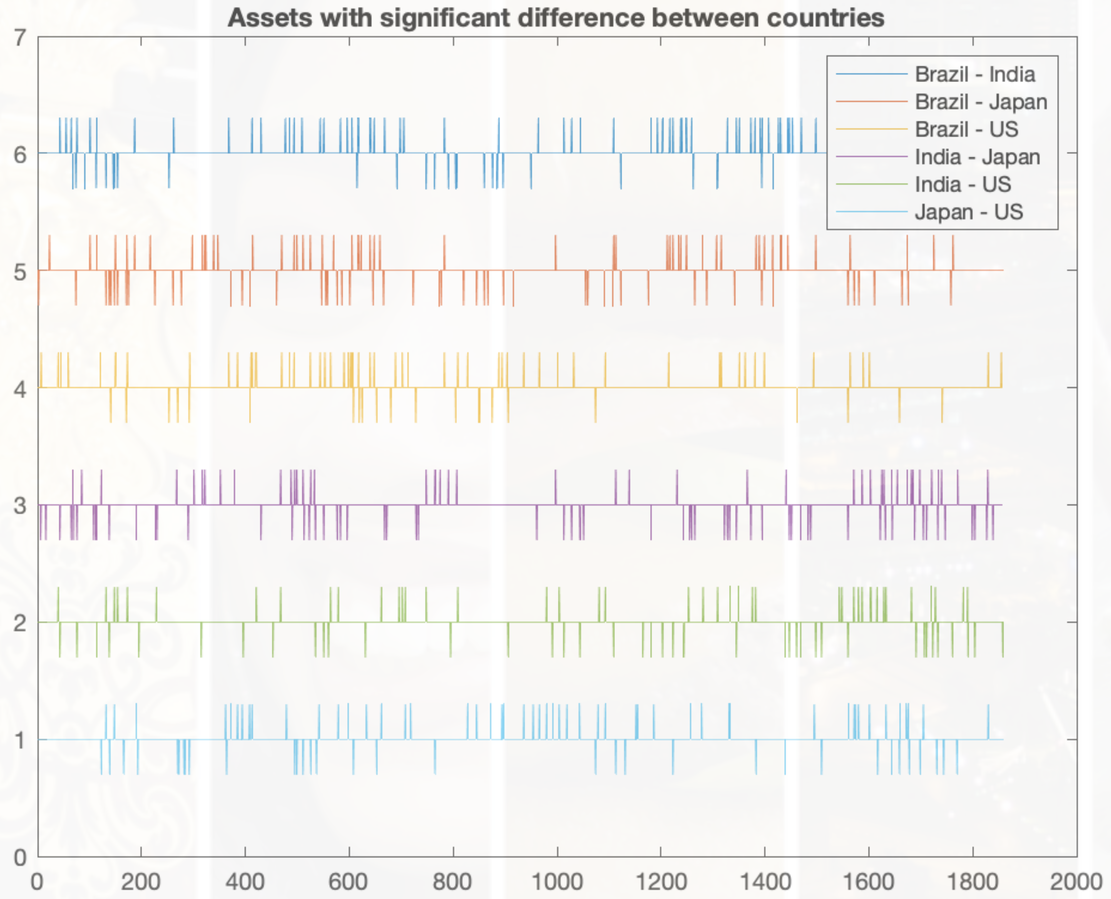
Plain MOS (US vs. Japan)
NIVD public dataset

Mismatch

Statistical analysis for ratings of each video

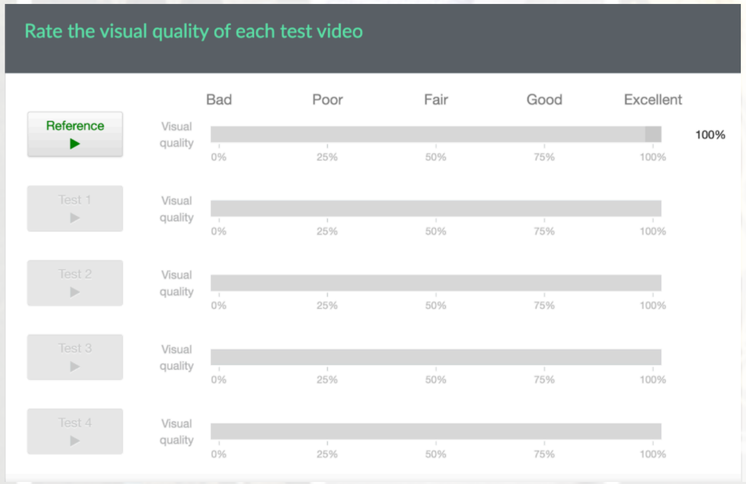
- Two-sample t-test
- Null hypothesis = ratings from two countries are from the same normal distributions (equal mean and variance)

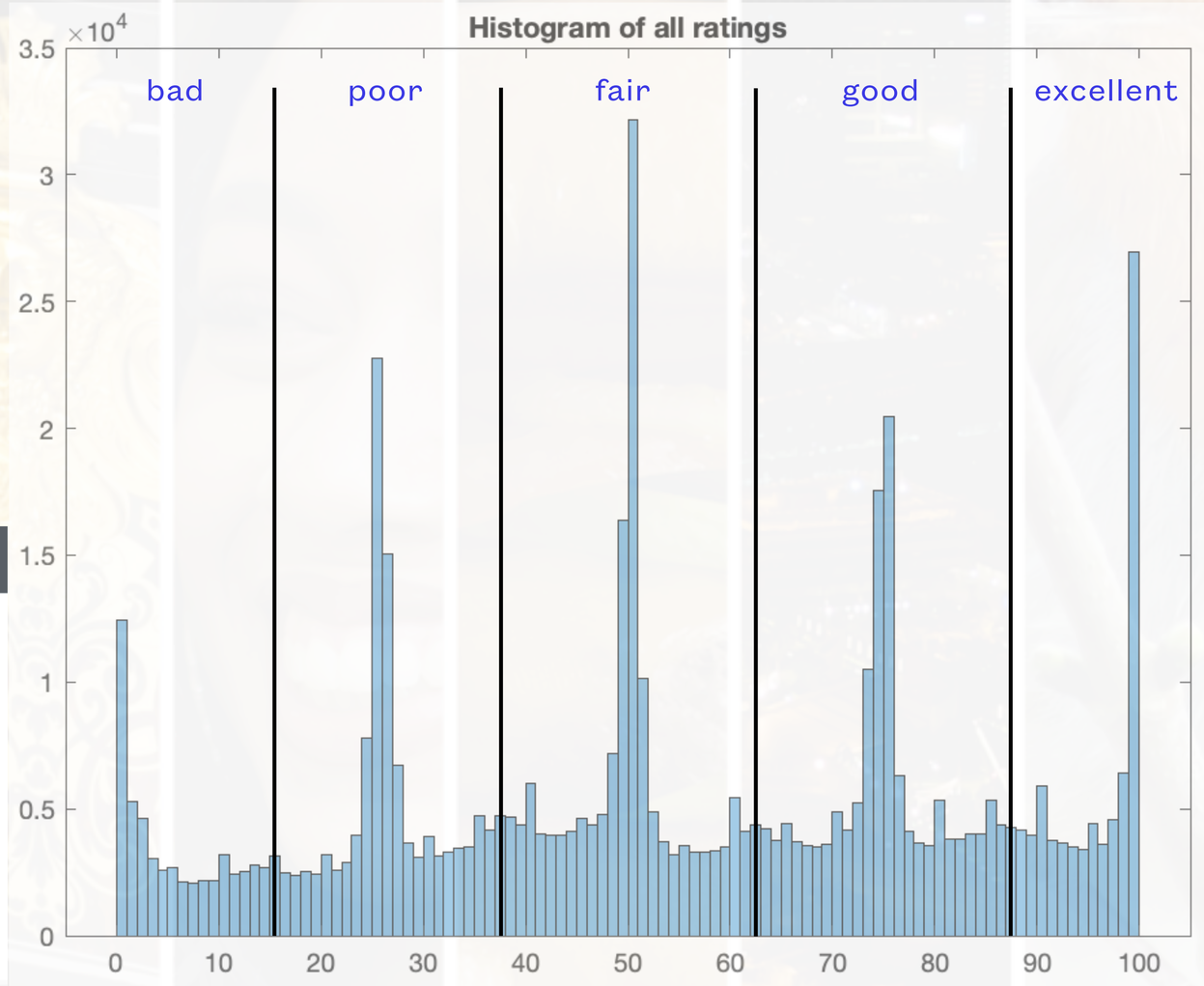




NIVD — Conversion VAS -> ACR

- NIVD employed a SAMVIQ scale
- Result: Pseudo ACR
- We quantize VAS to ACR

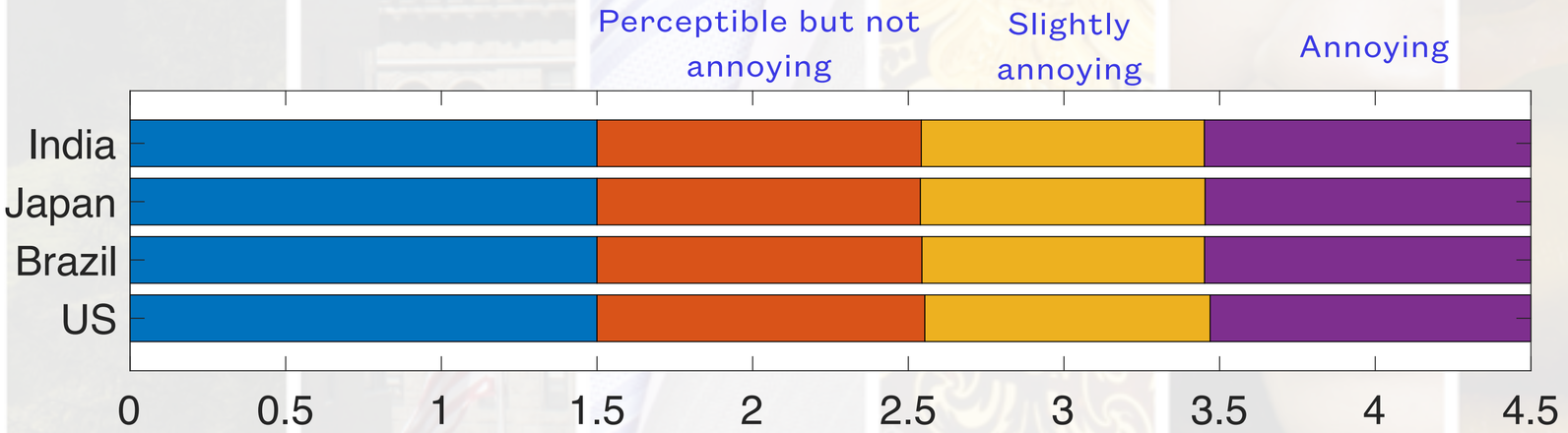




Results NIVD — Thresholds

- Global lapse rate 0.0471+/-0.0015
- Thresholds and 95%-confidence intervals:

	Brazil	India	Japan	US
tau_3	3.4517+/-0.0062	3.4511+/-0.0065	3.4535+/-0.0062	3.4696+/-0.0061
tau_2	2.5439+/-0.0064	2.5417+/-0.0067	2.5387+/-0.0064	2.5532+/-0.0063
sigma	0.8709+/-0.0047	0.8768+/-0.0049	0.8683+/-0.0047	0.8713+/-0.0046



- Findings:
  There are hardly any significant differences in the thresholds.

Limitations / Future work

- The normalization in the models was done by fixing the first and last thresholds to 1.5 and 4.5.
  - It is more informative to also let these be country-specific (normalize by z-scoring).
- The lapse rate was global.
  - Country-specific lapse rates may give better models.
  - A general analysis on the benefits of including lapse rates in Thurstonian models is outstanding.
- MLE for 10092 parameters for KonIQ-10k took 13h with Matlab on a MacBookPro.
  - A reduction of run-time may be achieved by lumping 10076 images into a single random effect.
- Subjective models to our country-specific analysis can be added.
- Binomial model regards influence by images as iid random effects.
  - Add a constant effect per image.
- NIVD public dataset is inconsistent with its QoMEX'23 paper:
  - many more subjects and ratings in paper,
  - country-specific differences shown in the paper are not in the dataset.

Conclusions

1. The hypothesis of significant differences in IQA rating behavior between countries is confirmed:
   - The thresholds of ACR/DCR categories on the latent perceptual scale differ between countries.
   - The likelihood for extreme ratings differ between countries.
   - A more detailed analysis should be carried out.
   - The open questions w.r.t. NIVD should be settled.
2. Future IQA datasets should take country-specific differences into account when
   - selecting subjects from different countries,
   - reconstructing IQA scale values from the responses.
3. Lapse rates have potential to improve Thurstonian models.

Acknowlegdments

- Vlad Hosu and Mirko Dulfer
  - Help regarding the curation of the raw KonIQ-10k dataset.
- Christos Bampis; Lukáš Krasula; Zhi Li; Omair Akhtar (Netflix)
  - Making the NIVD available.
- Shaolin Su
  - Reshaping the format of NIVD
- Michela Testolina
  - Slides background art made from JPEG AIC-3 image dataset.