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

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

  - large dataset (1860 videos, 14k subjects), well-balanced over 4 countries
  - biases across nations observed but not analyzed

  - 1 set of stimuli, 4 countries, 10 datasets of AV quality (ACR)
  - preliminary finding: Datasets „appeared not to be influenced by language or culture“.

  - detected significant differences in perceived video quality between subjects from 4 nations
Mean opinion scores (MOS):

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

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

<table>
<thead>
<tr>
<th>Thresholds</th>
<th>US</th>
<th>Italy</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N = 24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>excellent</td>
<td>6.5 +/- 0.6</td>
<td>6.4 +/- 0.6</td>
</tr>
<tr>
<td>ottimio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>good</td>
<td>4.9 +/- 0.7</td>
<td>5.5 +/- 0.7</td>
</tr>
<tr>
<td>buono</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fair</td>
<td>3.5 +/- 0.8</td>
<td>4.3 +/- 1.0</td>
</tr>
<tr>
<td>discreto</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poor</td>
<td>1.4 +/- 0.6</td>
<td>1.9 +/- 1.5</td>
</tr>
<tr>
<td>mediocre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bad</td>
<td>1.1 +/- 0.6</td>
<td>1.5 +/- 1.3</td>
</tr>
<tr>
<td>cattivo</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On a visual analog scale of 7.1 inch.
**Thurstonian model**

- Global unique perceived image quality $Q$
- Random effect (Gaussian, equal variance)
  \[ Q \sim N(\mu_j, \sigma^2) \text{ with cdf } F_{\mu_j,\sigma^2} \]
- Country-specific ACR thresholds $\tau^k_1 < \cdots < \tau^k_4$
- Global lapse rate $0 < \lambda < 0.2$ (new)
- Probability for ACR rating $m = 1, \ldots, 5$
  \[ \text{Prob} \ (\text{ACR} = m \mid \text{image } j, \text{country } k) = (1 - \lambda)(F_{\mu_j,\sigma^2}(\tau^k_m) - F_{\mu_j,\sigma^2}(\tau^k_{m-1})) + \lambda \]
- To normalize the scale and anchor results
  $\tau^k_1 = 1.5$ and $\tau^k_4 = 4.5$
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)

    \[
    \text{Prob}(X_k = \text{extreme} \mid \text{image } j) = \logit^{-1}(\alpha_k + U_j), \quad U_j \sim N(0,\sigma^2)
    \]

• Fixed effect per country, random effect per image

Computational methods

Thurstonian model
- Maximum likelihood estimation (MLE)
- Parameters ($\tau^k_m, \mu_j, \sigma, \lambda$)
- 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 ($\alpha_k, \sigma$)
- Confidence intervals:
  - Wald method
  - bootstrapping (same results)
International crowdsourced datasets

<table>
<thead>
<tr>
<th></th>
<th>Images</th>
<th>Videos</th>
<th>Ratings</th>
<th>Subjects</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>KonIQ</td>
<td>10076</td>
<td></td>
<td>1,078,176 ACR</td>
<td>1261</td>
<td>75</td>
</tr>
<tr>
<td>KADID</td>
<td>11085</td>
<td></td>
<td>391,376 DCR</td>
<td>2212</td>
<td>72</td>
</tr>
<tr>
<td>NIVD</td>
<td>1860</td>
<td></td>
<td>538,200 VAS</td>
<td>10000</td>
<td>4</td>
</tr>
</tbody>
</table>


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:

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

- 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:

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

- Findings:
  - The DCR interval for „Slightly Annoying“ is smallest for Egypt.
  - The variance is largest for Venezuela.
Select countries and images:
- Each country has ≥ 1000 ratings
- Each image has ≥ 500 ratings
- Result: 7 countries, 67 images

### Results KADID-10k — extreme ratings

#### Confidence Intervals

**Wald formula**

<table>
<thead>
<tr>
<th>Country</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>countryEGY</td>
<td>0.19</td>
<td>0.32</td>
</tr>
<tr>
<td>countryIND</td>
<td>0.28</td>
<td>0.43</td>
</tr>
<tr>
<td>countryRUS</td>
<td>0.37</td>
<td>0.54</td>
</tr>
<tr>
<td>countrySRB</td>
<td>0.31</td>
<td>0.47</td>
</tr>
<tr>
<td>countryTUR</td>
<td>0.23</td>
<td>0.37</td>
</tr>
<tr>
<td>countryUKR</td>
<td>0.27</td>
<td>0.43</td>
</tr>
<tr>
<td>countryVEN</td>
<td>0.30</td>
<td>0.45</td>
</tr>
</tbody>
</table>

**Bootstrapping**

<table>
<thead>
<tr>
<th>Country</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>countryEGY</td>
<td>0.19</td>
<td>0.32</td>
</tr>
<tr>
<td>countryIND</td>
<td>0.28</td>
<td>0.43</td>
</tr>
<tr>
<td>countryRUS</td>
<td>0.37</td>
<td>0.54</td>
</tr>
<tr>
<td>countrySRB</td>
<td>0.31</td>
<td>0.48</td>
</tr>
<tr>
<td>countryTUR</td>
<td>0.23</td>
<td>0.37</td>
</tr>
<tr>
<td>countryUKR</td>
<td>0.27</td>
<td>0.43</td>
</tr>
<tr>
<td>countryVEN</td>
<td>0.30</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Results NIVD — VAS

- Plain MOS scatterplots country vs. country
- No apparent country-specific differences
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 employed a SAMVIQ scale
• Result: Pseudo ACR
• We quantize VAS to ACR
Results NIVD — Thresholds

- Global lapse rate $0.0471\pm0.0015$

- Thresholds and 95%-confidence intervals:

<table>
<thead>
<tr>
<th></th>
<th>Brazil</th>
<th>India</th>
<th>Japan</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_3$</td>
<td>3.4517$\pm$0.0062</td>
<td>3.4511$\pm$0.0065</td>
<td>3.4535$\pm$0.0062</td>
<td>3.4696$\pm$0.0061</td>
</tr>
<tr>
<td>$\tau_2$</td>
<td>2.5439$\pm$0.0064</td>
<td>2.5417$\pm$0.0067</td>
<td>2.5387$\pm$0.0064</td>
<td>2.5532$\pm$0.0063</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.8709$\pm$0.0047</td>
<td>0.8768$\pm$0.0049</td>
<td>0.8683$\pm$0.0047</td>
<td>0.8713$\pm$0.0046</td>
</tr>
</tbody>
</table>

- 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.
Acknowledgments

- 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.