Controlled vs Crowdsourced Subjective Tests: Analyzing Unexpected Scoring Behaviors in Media Quality Assessment

Lohic Fotio Tiotsop

VQEG JEG-Hybrid

Nov 21, 2024

Introduction

- Statistical models for subjective quality recovery and peculiar behavior identification
  - Different approaches have been proposed
  - Subjects are commonly assumed to exhibit bias and inconsistency
- Beyond Bias and Inconsistency: **Scale quantization**
  - To identify aspects to improve in crowdsourcing settings in order to reach standards similar to ones in controlled environments
  - To detect peculiar behaviors overlooked by bias and inconsistency in crowdsourcing tests
  - To investigate the origin of extreme inconsistency observed in crowdsourcing settings

Approach Summary on the 5-point ACR scale

- Quality of stimulus i: $Q_i \rightarrow [w_{i1}, w_{i2}, \ldots, w_{i5}]$
- Bias of subject j: $b_j \rightarrow \sum_{k=1}^{k=5} k \cdot \mu_k^j(w)$
- Inconsistency of subject j: $\sigma_j^2(\mu, w) \rightarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sigma_{ij}^2(\mu, w)$
- **Scale Quantization Index** of subject j:

$$SQI_j \rightarrow \sum_{k=1}^{k=5} |\mu_k^j(w)|$$

On Theoretical Aspects

- Theoretical details on parameter estimation in this paper:

  Modeling Subject Scoring Behaviors in Subjective Experiments Based on a Discrete Quality Scale

Journals & Magazines > IEEE Transactions on Multimedia > Early Access

Modeling Subject Scoring Behaviors in Subjective Experiments Based on a Discrete Quality Scale

Publisher: IEEE | Cite This | PDF

Lohic Fotio Tiotsop ; Antonio Servetti ; Marcus Barkowsky ; Enrico Masala **All Authors**

33 Full Text Views

Open Access

Under a Creative Commons License

Abstract
Authors
Keywords
Metrics

Abstract:
Several approaches have been proposed to estimate quality in subjective experiments while highlighting peculiar subject behaviors. However, there is some room for improvement in existing approaches, both in terms of robustness to noise and the ability to accurately indicate several peculiar subject behaviors in subjective experiments. This work advances the state-of-the-art in three main directions: i) A new approach to estimate the subjective quality from noisy ratings is proposed and is shown to be more robust to noise than are four state-of-the-art approaches; ii) a novel subject scoring model is proposed that makes it possible to highlight several peculiar behaviors typically observed in subjective experiments; and iii) our proposed probabilistic subject scoring model results from the proof of a theorem, whereas in previous approaches a probabilistic scoring model is assumed a priori . This represents an important first step toward models supported by a stronger theoretical foundation. Numerical experiments conducted on several datasets highlight the effectiveness of our proposal.

Published in: IEEE Transactions on Multimedia (Early Access)

- Main focus: the ability of the approach to offer a more comprehensive analysis of scoring behaviors in subjective tests, especially in crowdsourcing.
- Overview of the ideas behind our approach in the following.

Ground Truth Quality Q_i

Overview:

- The weight w_{ki} represents the importance of the rating k in determining the ground truth quality of stimulus i
- For each stimulus, five weights $\mathrm{w} = \{w_1, w_2, w_3, w_4, w_5\}$ are computed.
- The weights sum up to 1 for each stimulus
- The weights are derived by solving an optimization problem

Optimization Problem:

- Objective function: a regularized likelihood function
- Likelihood term:
  - Captures the information contained in the data
- Regularization Term:
  - Penalizes the noise in the data

Scale Quantization Weights μ_k^j

Definition:

- μ_k^j represents the tendency of subject j to:
  - Over-use the rating k ($\mu_k^j > 0$).
  - Use the rating k less frequently than expected ($\mu_k^j < 0$).

Key Observations:

- For each subject j, five weights are computed:

$$[\mu_1^j, \mu_2^j, \mu_3^j, \mu_4^j, \mu_5^j]$$

- These five weights satisfy the constraint:

$$\sum_{k=1}^{5} \mu_k^j = 0.$$

Estimation Procedure:

- Perform one-hot encoding of the opinion scores for subject j
- Subtract the vectors w of quality weights
- Average the resulting vectors over all stimuli to estimate μ_k^j

Bias and Scale Quantization Index

Bias (b_j) of subject j:

- Calculated by weighting each opinion k with the subject's tendency to choose or avoid it, μ_k^j:

$$b_j = \sum_{k=1}^{5} k \cdot \mu_k^j$$

Scale Quantization Index (SQI):

- Measures the extent to which a subject uses the rating scale in an imbalanced manner.

$$\text{SQI}_j = \sum_{k=1}^{5} |\mu_k^j|$$

- A higher SQI indicates tendency to use just a subset of scale items.

Subject Inconsistency σ_j

Overview

- The Rating R_{ij} of subject j for stimulus i is modeled as a realization of a discrete random variables
- A theorem was formulated to derive the probability distribution of R_{ij}
- The variance σ_{ij}^2 of R_{ij} for each stimulus i was estimated using this probability distribution

Estimation of Inconsistency

$$\sigma_j = \frac{1}{N} \sum_{i=1}^{N} \sigma_{ij},$$

where N is the total number of stimuli

Computational Experiments: Datasets

Table: Summary of Datasets Categorized by Experimental Setting

Category	Dataset	Number of Stimuli	Number of Subjects
Controlled Environment	NETFLIX-PUB	70	26
	VQEG-HD1	144	24
	VQEG-HD3	167	24
	VQEG-HD5	168	24
	ITS4S	514	27
Crowdsourcing	KonViD-1k	1200	624
	KonIQ	10076	1261
	KADID	11085	2212
	NIVD	1860	12812
	MovieLens-1M	3706	6040

Note: MovieLens-1M is actually a QoE experiment addressing a more general question than the assessment of the perceived quality

Results: Statistics to Improve in Crowdsourcing Settings

- The average of SQI values
- The variance of inconsistency values

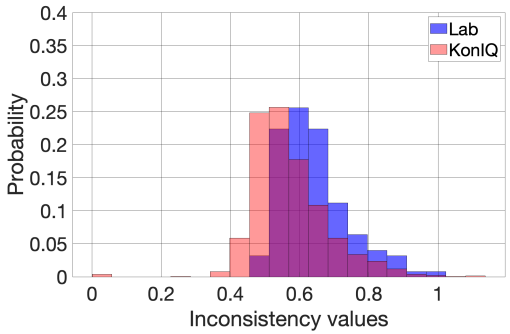
Average SQI: Lab vs Crowdsourcing

Table: Average SQI values for all datasets. For each crowdsourcing experiment, it was also tested whether the average SQI differs significantly from Lab experiments.

Datasets	**Avg SQI**	**Test of Mean Diff**
NETFLIX-PUB	0.2600	–
VQEG-HD1	0.2471	–
VQEG-HD3	0.3093	–
VQEG-HD5	0.3120	–
ITS4S	0.2954	–
KonViD-1k	0.4435	3.0815×10^{-24}
KonIQ	0.3740	4.7898×10^{-11}
KADID	0.4415	1.4214×10^{-26}
NIVD	0.5017	4.6067×10^{-39}
MovieLens-1M	0.3721	7.8599×10^{-12}

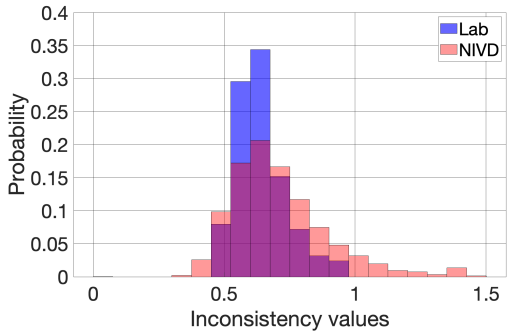
- Scale quantization seems to be more pronounced in crowdsourcing settings

Variance of Inconsistency: Lab vs Crowdsourcing

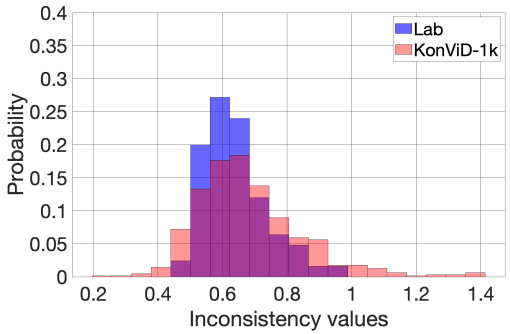


(a) KoniQ

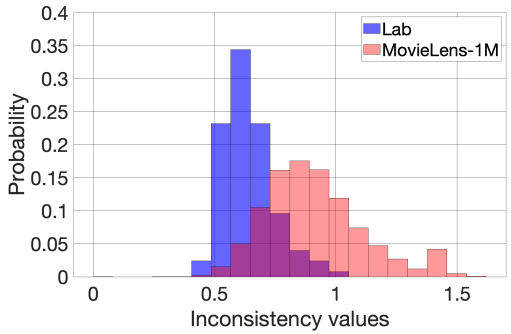
(b) KADID



(c) NIVD



(d) KonViD-1k



(e) MovieLens-1M

Figure: Histogram of inconsistency values from all Lab tests (blue) vs. histogram of inconsistency values from each crowdsourcing test (red).

Variance of Inconsistency: Lab vs Crowdsourcing

Table: P-values of statistical tests to check whether the mean and variance of inconsistency values in crowdsourcing and lab test differ with statistical significance.

Datasets	**Test of Mean Diff**	**Test of Variance Diff**
KonViD-1k	0.4981	1.1428×10^{-10}
KonIQ	0.4772	0.1532×10^{-00}
KADID	**0.0147**	2.6453×10^{-9}
NIVD	0.6818	4.5801×10^{-16}
MovieLens-1M	**0.0038**	5.1548×10^{-19}

- There is in general higher variability of inconsistency values in crowdsourcing tests with statistical significance.

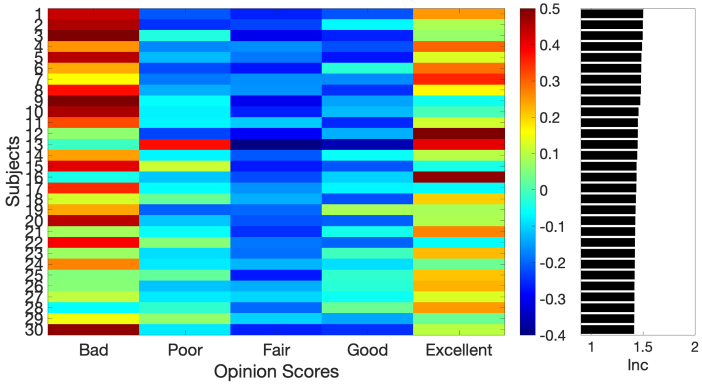
How do Some Highly Inconsistent Subjects Use the Quality Scale?

1. For each rowdsourcing test, we selected any subject j for which it holds

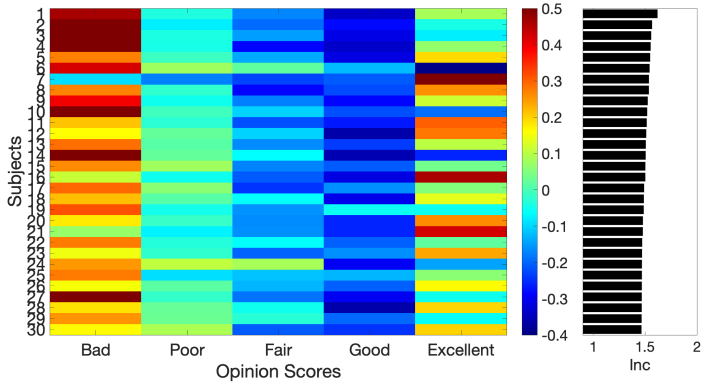
$$\sigma_j > \max(\sigma_{\text{from all lab tests}})$$

2. The selected subjects were ordered by decreasing order of inconsistency
3. The scale quantization weights of the selected subjects were then analyzed

How do Some Highly Inconsistent Subjects Use the Quality Scale?



(a) NIVD



(b) MovieLens-1M

Figure: Scale quantization weights of highly inconsistent subjects.

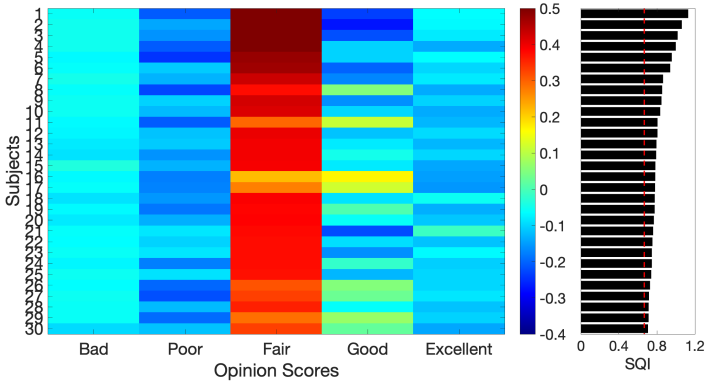
- Highly inconsistent subjects seem to ignore intermediate opinion scores
- Using quality scales with fewer options would potentially help

Detecting Peculiar Subjects that are neither Biased nor Inconsistent

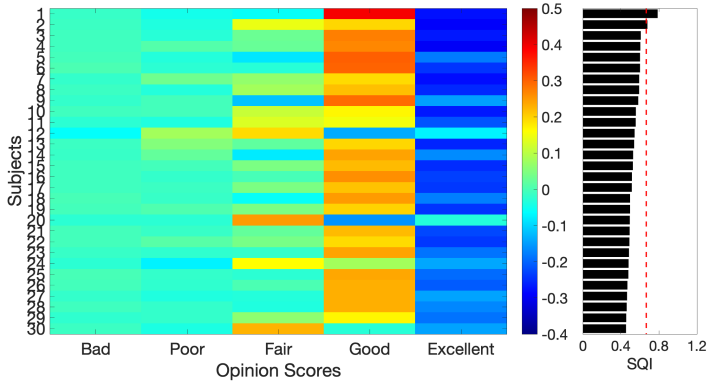
- **Goal**: detecting unbiased and not inconsistent subjects that highly quantized the scale
- These subjects are not penalized by SUREAL and ZREC in the quality recovery process
- Experiments show that BT.500 would not exclude as well

1. For each crowdsourcing test, we selected any subject j for which it holds:
   - $-0.2 < b_j < 0.2$
   - $\sigma_j < Avg(\sigma_{\text{from all lab tests}})$
2. Selected subjects were then ordered in decreasing order of their SQI.

Detecting Peculiar Subjects that are neither Biased nor Inconsistent



(a) NIVD



(b) MovieLens-1M

Figure: Scale quantization weights of the unbiased and not inconsistent subjects ordered by decreasing SQI value.

- Peculiar subjects over-used "Fair" in one case, and ignored "Excellent" in the other.
- A threshold (dashed red line) could be fixed by the test designer to exclude subject that are particularly peculiar. Es:

$$Avg(SQI_{\text{from all lab tests}}) + 3 * Std(SQI_{\text{from all lab tests}})$$

Thanks for your attention