Updates on Maximum Likelihood Estimation (MLE) Methods for Subject Behavior Modeling

Zhi Li & Christos Bampis, *Netflix, USA*
Lucjan Janowski, *AGH, Poland*
Ioannis Katsavounidis, *Facebook, USA*

VQEG Meeting March 2020
3/10/2020

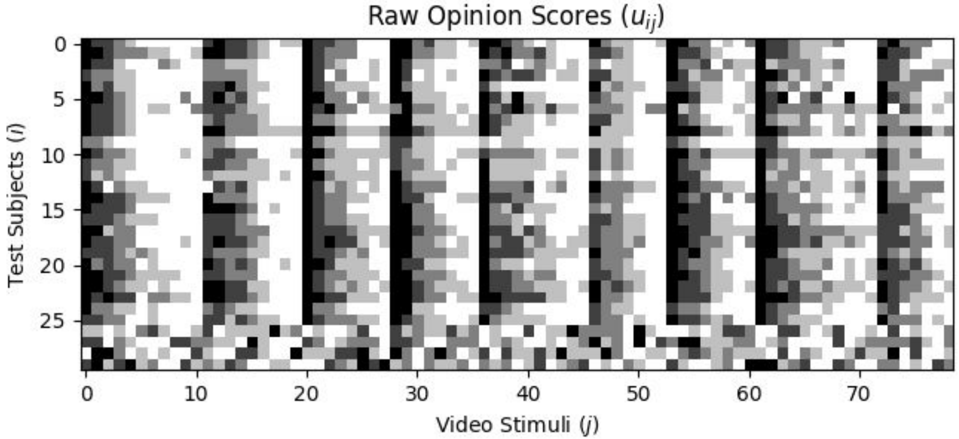
MLE Activities in SAM Study Group

- Motivation
  - Mission: to improve data quality coming from subjective experiment
  - Saw opportunity to improve data cleanup methods currently adopted in ITU-T/R recommendations via statistical methods (e.g. MLE)
- Progress since the last VQEG meeting
  - A comprehensive study with comparison to prior standards ITU-R BT.500 and ITU-T P.913
  - Bayesian Information Criterion (BIC) to validate model fitting with real-world data
  - Besides the original Newton-Raphson method, proposed an alternative solution based on projection, proven to be faster and more intuitive

Subjective Test

Credit: memegenerator.net

Raw opinion scores are noisy and unreliable



Black: lowest score
White: highest score

Subject bias

Last 4 Outliers

- Would MOS or DMOS be good enough?
- Corrective mechanisms
  - Subject outlier rejection
  - Subject bias removal

Prior Art: Subject Outlier Rejection (ITU-R BT.500)

For each test presentation, calculate the mean, $\bar{u}_{jkr}$, standard deviation, S_{jkr}, and kurtosis coefficient, β_{2jkr}, where β_{2jkr} is given by:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \quad \text{with} \quad m_x = \frac{\sum_{i=1}^{N}(u_{ijkr}-\bar{u}_{ijkr})^x}{N} \tag{5}$$

For each observer, i, find P_i and Q_i, i.e.:

for $j, k, r = 1, 1, 1$ to J, K, R

if $2 \le \beta_{2jkr} \le 4$, then:

if $u_{ijkr} \ge \bar{u}_{jkr} + 2\,S_{jkr}$ then $P_i = P_i + 1$

if $u_{ijkr} \le \bar{u}_{jkr} - 2\,S_{jkr}$ then $Q_i = Q_i + 1$

else:

if $u_{ijkr} \ge \bar{u}_{jkr} + \sqrt{20}\,S_{jkr}$ then $P_i = P_i + 1$

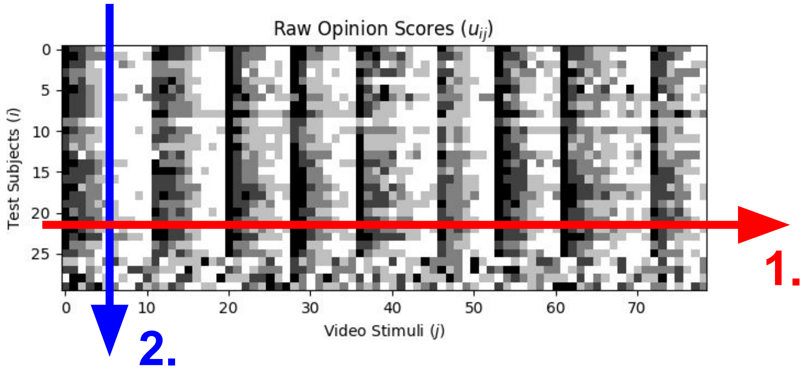
if $u_{ijkr} \le \bar{u}_{jkr} - \sqrt{20}\,S_{jkr}$ then $Q_i = Q_i + 1$

If $\frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05$ and $\left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3$ then reject observer i

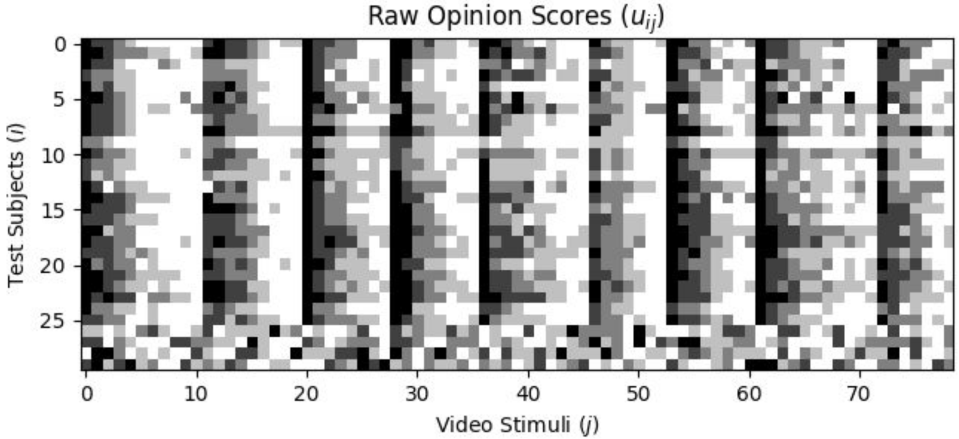
with:

- N: number of observers
- J: number of test conditions including the reference
- K: number of test images or sequences
- R: number of repetitions
- L: number of test presentations (in most cases the number of presentations will be equal to $J \cdot K \cdot R$, however it is noted that some assessments may be conducted with unequal numbers of sequences for each test condition).

1. Video by video, the algorithm counts the number of instances when a subject's opinion score deviates by a few sigmas
2. Subject by subject, if the occurrences are more than a fraction, reject the subject



Limitations of BT.500-Style Subject Outlier Rejection



Black: lowest score
White: highest score

Subject bias

Last 4 Outliers

- All scores corresponding to rejected subjects are discarded - an overkill
- Often only identifies a subset of outliers
    - In the example above, only subjects #26, #28, #29 were rejected
- Hard-coded parameters and heuristic steps lack interpretability

Prior Art: Subject Bias Removal (ITU-T P.913)

First, estimate the MOS for each PVS:

$$\mu_{\psi_j} = \frac{1}{I_j}\sum_{i=1}^{I_j} o_{ij}$$

where:

o_{ij} is the observed rating for subject i and PVS j;

I_j is the number of subjects that rated PVS j;

μ_{ψ_j} estimates the MOS for PVS j, given the source stimuli and subjects in the experiment.

Second, estimate subject bias:

$$\mu_{\Delta_i} = \sum_{j=1}^{J_i}\left(o_{ij} - \mu_{\psi_j}\right)$$

where:

μ_{Δ_i} estimates the overall shift between the ith subject's scores and the true values (i.e., opinion bias)

J_i is the number of PVSs rated by subject i.

Third, calculate the normalized ratings by removing subject bias from each rating:

$$r_{ij} = o_{ij} - \mu_{\Delta_i}$$

where:

r_{ij} is the normalized rating for subject i and PVS j.

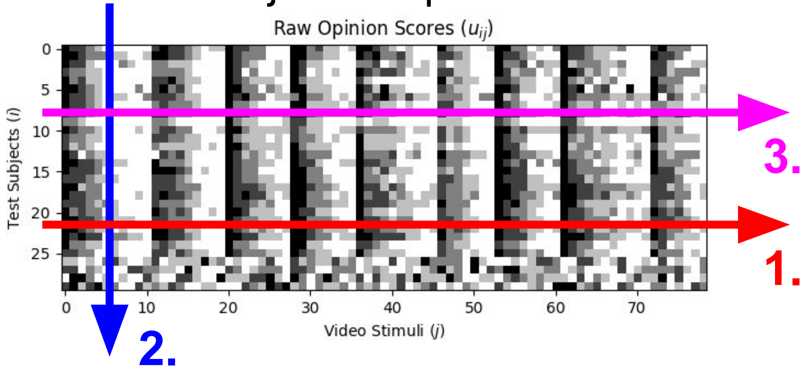
MOS and DMOS are then calculated normally. This normalization does not impact MOS:

$$\mu_{\psi_j} = \frac{1}{I_j}\sum_{i=1}^{I_j} r_{ij} = \frac{1}{I_j}\sum_{i=1}^{I_j} o_{ij}$$

where:

μ_{ψ_i} estimates the MOS of PVS j.

1. Video by video, estimate MOS by averaging over subjects
2. Subject by subject, estimate subject bias by comparing against MOS, and remove bias from opinion scores
3. Video by video, estimate MOS again based on bias-removed outlier-rejected opinion scores



Can we do better?

- Two most dominant effects of test subject inaccuracy:
    - Subject bias
        - Picky viewers tend to be biased toward lower scores
        - Not every subject has “golden eyes” - their sensitivity to impairment varies
    - Subject inconsistency
        - Subjects may not rate consistently throughout a session
        - Outliers - a special case with very large inconsistency
- Our proposal:
    - A simple yet effective model to account for subjective bias and inconsistency
    - Jointly solve the model parameters via maximum likelihood estimation (MLE)
        - Incorporate implicit “subject outlier rejection” and “subject bias removal” during model solving

A Simple Model*

$$\underbrace{\boxed{U_{ijr}}}_{\text{Raw Opinion Score}} = \underbrace{\boxed{\psi_j}}_{\text{True Score}} + \underbrace{\boxed{\Delta_i}}_{\text{Subject Bias}} + \underbrace{\boxed{\upsilon_i}}_{\text{Subject Inconsistency}} X$$

- U_{ijr} - Opinion score of subject i, stimulus j and repetition r
- ψ_j - true quality of stimulus j
- Δ_i - bias of subject i
- υ_i - inconsistency (std) of subject i
- X - i.i.d. normal random variables, $X \sim N(0, 1)$

*The model is a simplified version of [Li&Bampis'17] without considering the ambiguity of content. Compared to the original, the simplified model is more efficient and stable.

Solving the Model via MLE

- Given observations $\{U_{ijr}\}$
- The task is to solve for free parameters $\theta = (\{\psi_j\}, \{\Delta_i\}, \{v_i\})$
- Define log-likelihood function $l(\theta)$

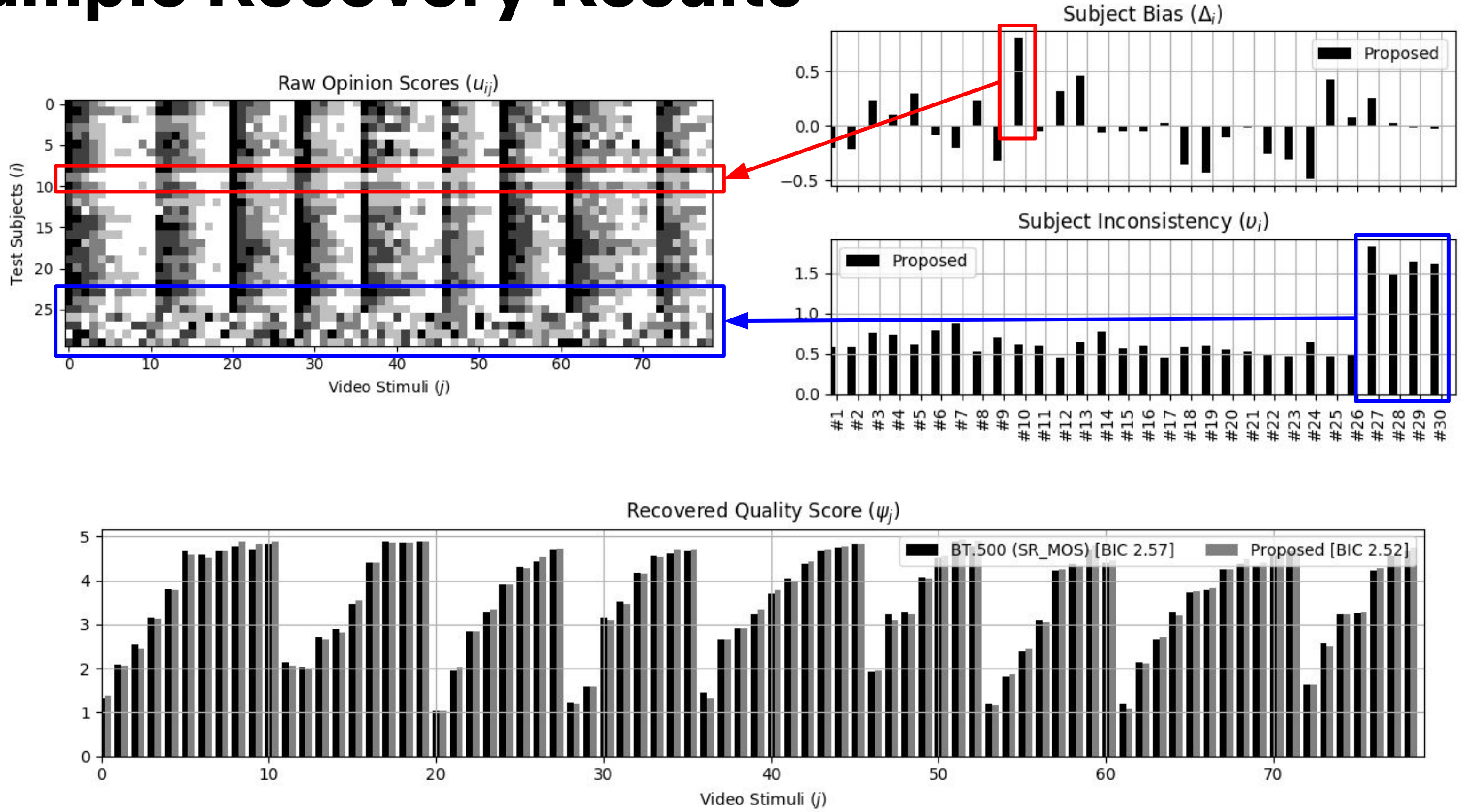
$$l(\theta) = \log P(U_{ijr} | \{\psi_j\}, \{\Delta_i\}, \{v_i\})$$

- Numerically solve to maximize the log-likelihood function

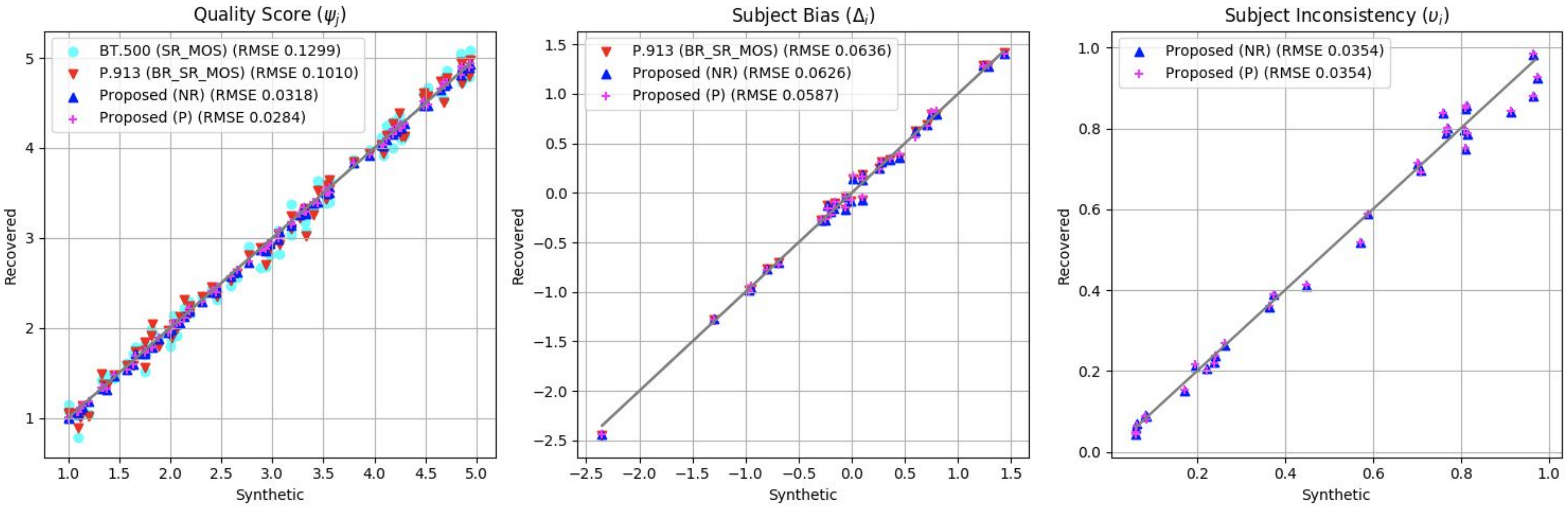
$$\hat{\theta} = \arg\max l(\theta)$$

- Proposed two numerical solutions
  - Newton-Raphson (NR) solution [Li&Bampis’17]
  - Projection-based (P) solution NEW! (thanks to Ioannis!)
    - Faster and strongly intuitive
    - Similar to ITU-T P.913, but 1) iterative 2) the projection is weighted by (sample count)/(residue variance)

Sample Recovery Results



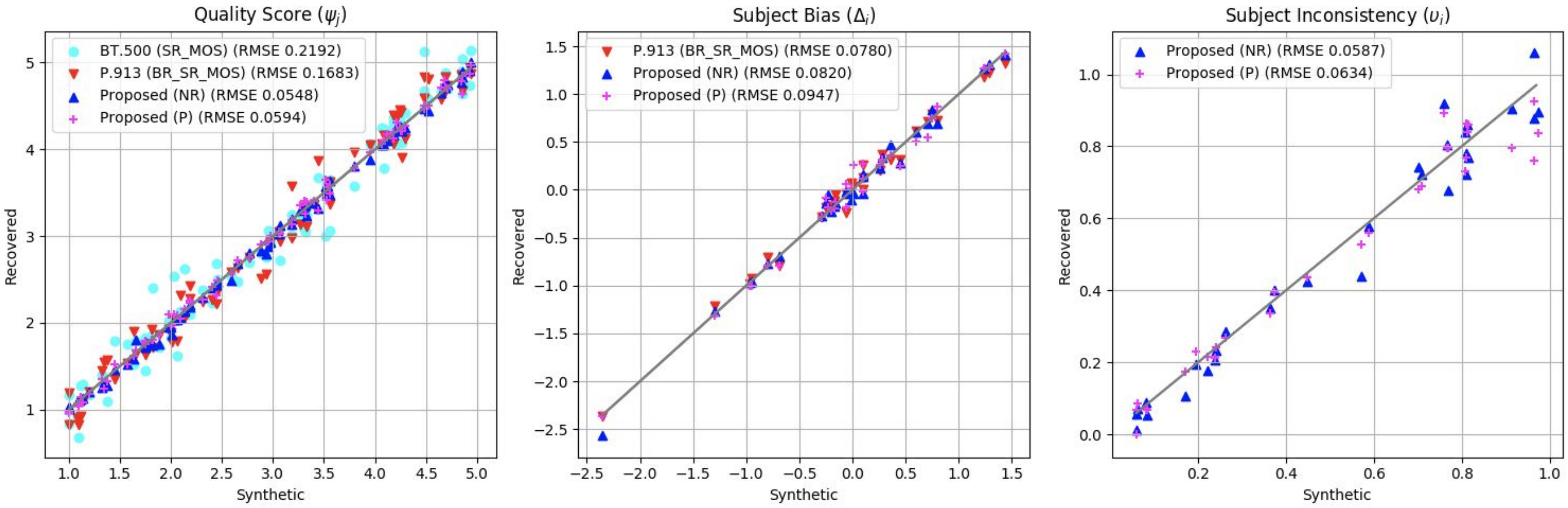
Validation Using Synthetic Data



- Synthetic data generation
  - Randomly generate parameters according to $\psi_j \sim U[1, 5], \Delta_i \sim N(0, 1), \upsilon_i \sim U[0, 1)$
  - Randomly generate observations according to parameters and model

SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

Validation Using Synthetic Data (50% Missing Data)



- Synthetic data generation
  - Randomly generate parameters according to $\psi_j \sim U[1, 5], \Delta_i \sim N(0, 1), \upsilon_i \sim U[0, 1)$
  - Randomly generate observations according to parameters and model
  - Data missing probability 0.5

SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

Validation Using Bayesian Information Criterion

- BIC is a criterion for model fitting, balancing between:
  - The degree of freedom (number of parameters)
  - The goodness of fit (log-likelihood function)

$$\mathrm{BIC} = \frac{\log(n)|\theta| - 2l(\theta)}{n}$$

- $|\theta|$ - the number of model parameters
- n - the number of observations (i.e. raw opinion scores)
- $l(\theta)$ - log-likelihood function

Bayesian Information Criterion (BIC)*

Dataset	MOS	BT.500 (SR_MOS)	P.913 (BR_SR_MOS)	Proposed (NR)	Proposed (P)
NFLX_dataset_public_raw_last4outliers	2.97	2.57	2.55	**2.52**	2.53
VQEGHD3_dataset_raw	2.75	2.74	2.39	**2.30**	2.31
HDTV_Phase_I_Experiment_1	2.45	2.46	2.38	**2.20**	2.22
HDTV_Phase_I_Experiment_2	2.72	2.72	2.52	**2.32**	2.33
HDTV_Phase_I_Experiment_3	2.72	2.71	2.37	**2.29**	**2.29**
HDTV_Phase_I_Experiment_4	2.96	2.96	2.51	**2.27**	**2.27**
HDTV_Phase_I_Experiment_5	2.77	2.77	2.47	**2.33**	**2.33**
HDTV_Phase_I_Experiment_6	2.51	2.49	2.32	**2.16**	**2.16**

*The model with the smallest BIC is preferred.

SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

Bayesian Information Criterion (BIC)*

Dataset	MOS	BT.500 (SR_MOS)	P.913 (BR_SR_MOS)	Proposed (NR)	Proposed (P)
ITU-T_Supp_23_Experiment_1_BNR	2.91	2.91	2.35	**2.31**	2.31
MM2_1	2.80	2.78	2.83	**2.74**	2.75
MM2_2	3.89	3.89	3.52	**3.13**	**3.13**
MM2_3	2.48	2.47	2.45	**2.41**	2.42
MM2_4	2.74	2.73	2.62	**2.47**	**2.47**
MM2_5	2.90	2.82	2.67	**2.64**	**2.64**
MM2_6	2.81	2.74	2.74	**2.72**	2.74
MM2_7	2.73	2.72	2.76	**2.67**	2.71

*The model with the smallest BIC is preferred.

SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

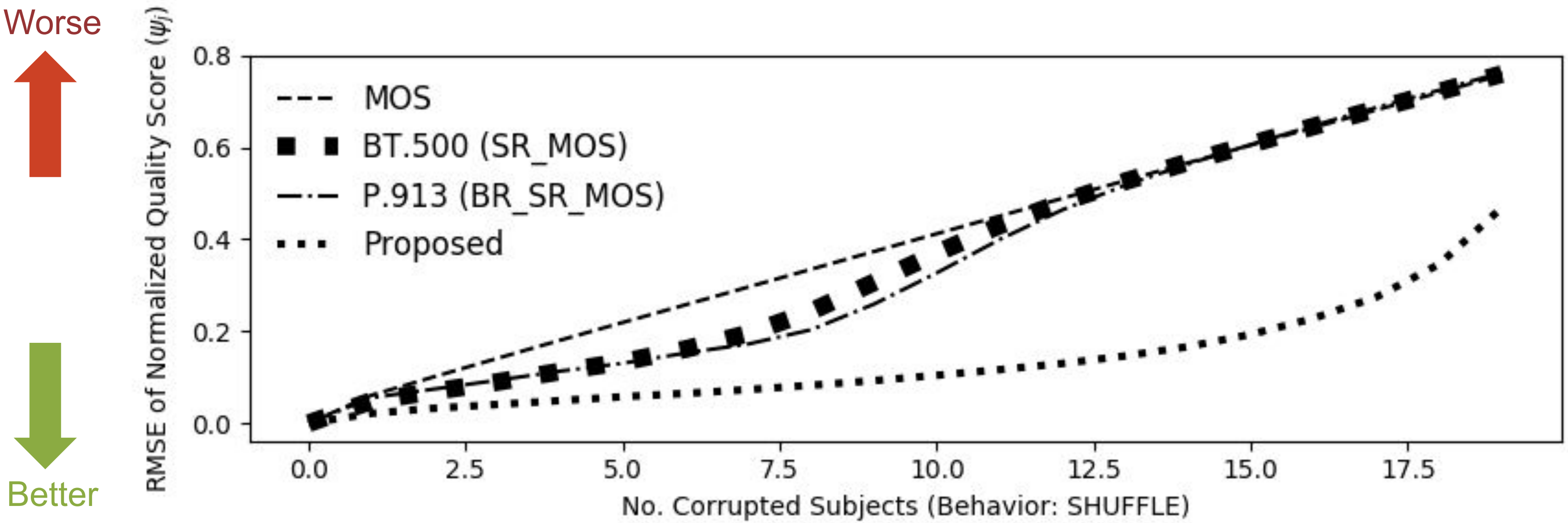
Bayesian Information Criterion (BIC)*

Dataset	MOS	BT.500 (SR_MOS)	P.913 (BR_SR_MOS)	Proposed (NR)	Proposed (P)
MM2_8	3.00	2.92	2.88	**2.70**	2.71
MM2_9	3.27	3.21	2.95	**2.79**	2.80
MM2_10	3.04	3.05	2.98	**2.82**	2.83
its4s2	3.63	3.63	2.96	**2.59**	**2.59**
its4s_AGH	3.14	3.04	2.76	**2.63**	**2.63**
its4s_NTIA	2.92	2.89	2.52	**2.37**	**2.37**

*The model with the smallest BIC is preferred.

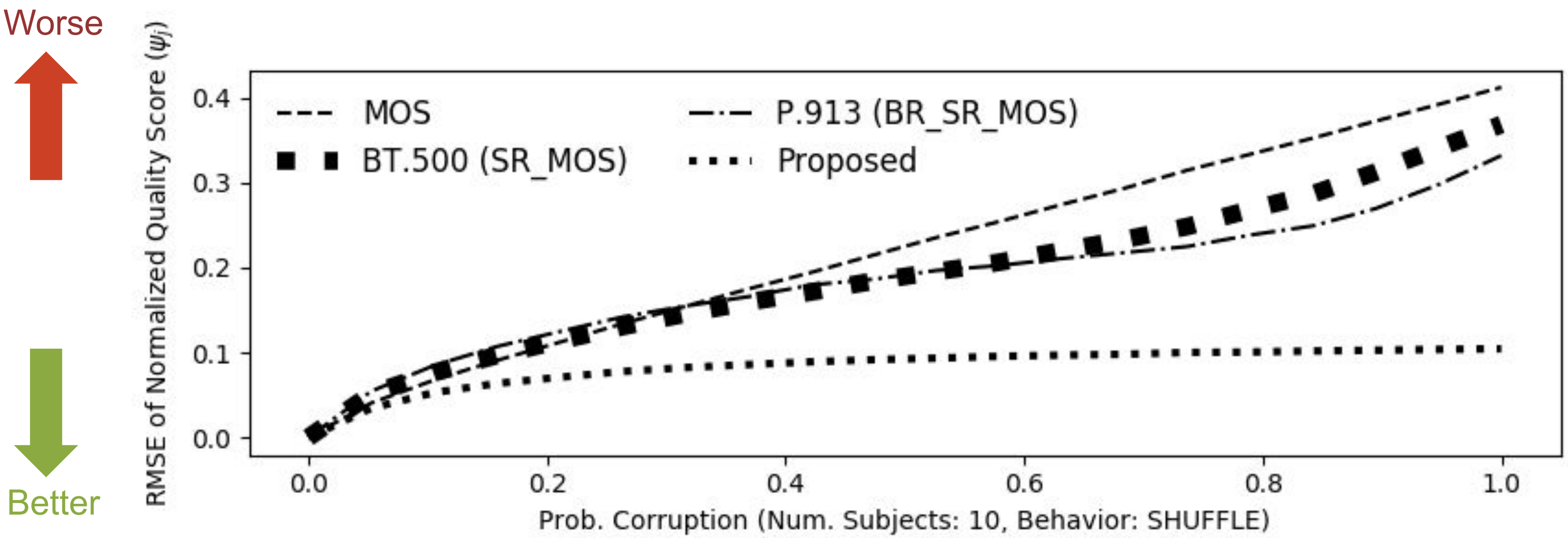
SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

Robustness Against Subjects Giving Random Scores



Random behavior: a subject's scores are shuffled among themselves
Y-axis: RMSE with respect to clean case
SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

Robustness Against Increasing Corruption Probability



10 random subjects are corrupted, with corruption probability varying from 0.0 to 1.0
Y-axis: RMSE with respect to clean case
SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

Conclusions

- Recommendations such as ITU-R BT.500 and ITU-T P.913 standardize the procedure to clean up raw scores from subjective experiments through subject outlier rejection (SR) and subject bias removal (BR)
- We introduce a simple model and the corresponding parameter estimation procedure that implicitly takes into account both SR and BR, with the following advantages:
  - Better model fitting
  - Better robustness in the presence of outlier subjects in terms of recovery accuracy
  - Auxiliary information on test subjects on their bias and consistency, providing guides on subject selection
  - The projection-based solution provides strong intuition
- We propose to update ITU-R BT.500 and ITU-T P.910/P.913 with the new methodology

Source code can be found at https://github.com/Netflix/sureal

Backup Slides

The projection solution is faster and more intuitive

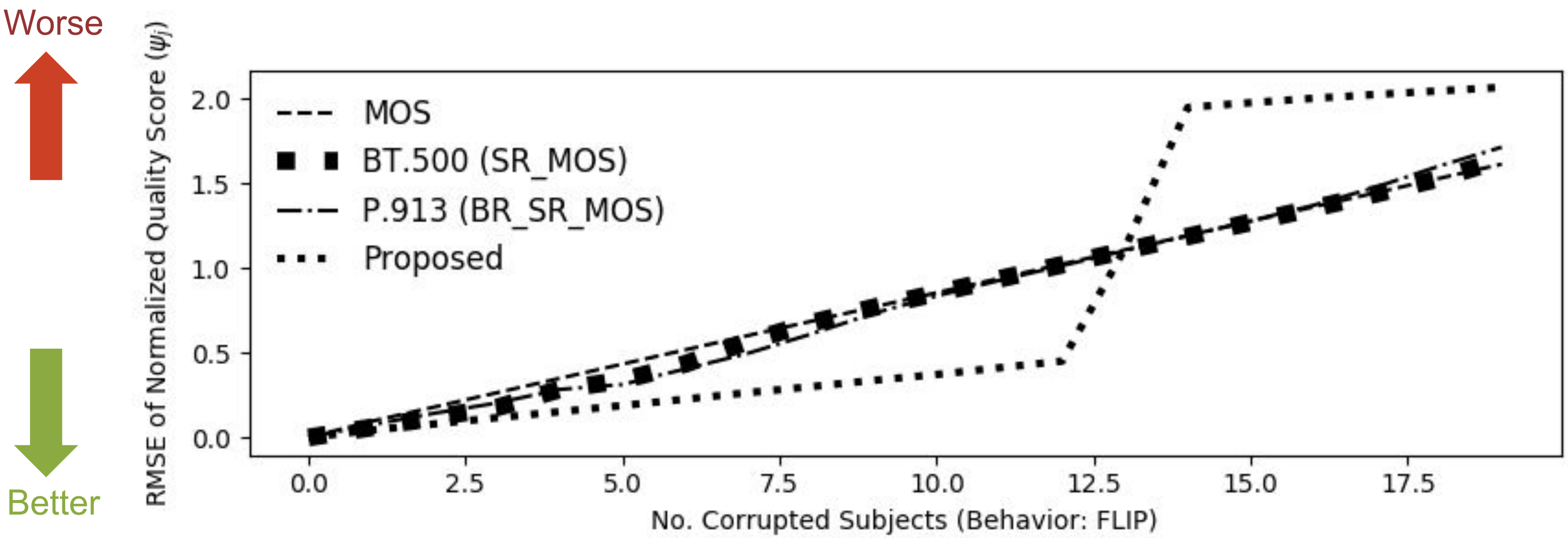
```python
# video by video, estimate MOS by averaging over subjects
s_j = np.nanmean(x_ji, axis=1)  # mean marginalized over i

# subject by subject, estimate subject bias by comparing with MOS
b_ji = x_ji - np.tile(s_j, (I, 1)).T
b_i = np.nanmean(b_ji, axis=0)  # mean marginalized over j

MAX_ITR = 1000
DELTA_THR = 1e-8
EPSILON = 1e-8

itr = 0
while True:

    s_j_prev = s_j

    # calculate residue
    r_ji = x_ji - np.tile(s_j, (I, 1)).T - np.tile(b_i, (J, 1))

    # video by video, estimate MOS by averaging over subjects, inversely weighted by residue variance
    v_i = np.nanstd(r_ji, axis=0)
    s_ji = x_ji - np.tile(b_i, (J, 1))
    w_i = np.divide(cnt_i, v_i ** 2 + EPSILON)
    s_j = weighed_nanmean_2d(s_ji, weights=w_i, axis=1)  # mean marginalized over i

    # subject by subject, estimate subject bias by comparing with MOS, inversely weighted by residue variance
    v_j = np.nanstd(r_ji, axis=1)
    b_ji = x_ji - np.tile(s_j, (I, 1)).T
    w_j = np.divide(cnt_j, v_j ** 2 + EPSILON)
    b_i = weighed_nanmean_2d(b_ji, weights=w_j, axis=0)  # mean marginalized over j

    itr += 1

    delta_s_j = linalg.norm(s_j_prev - s_j)

    if delta_s_j < DELTA_THR:
        break

    if itr >= MAX_ITR:
        break
```

Same as ITU-T P.913

Similar to ITU-T P.913, but weighted

The weight is proportional to the sample count, and inversely proportional to the residue variance

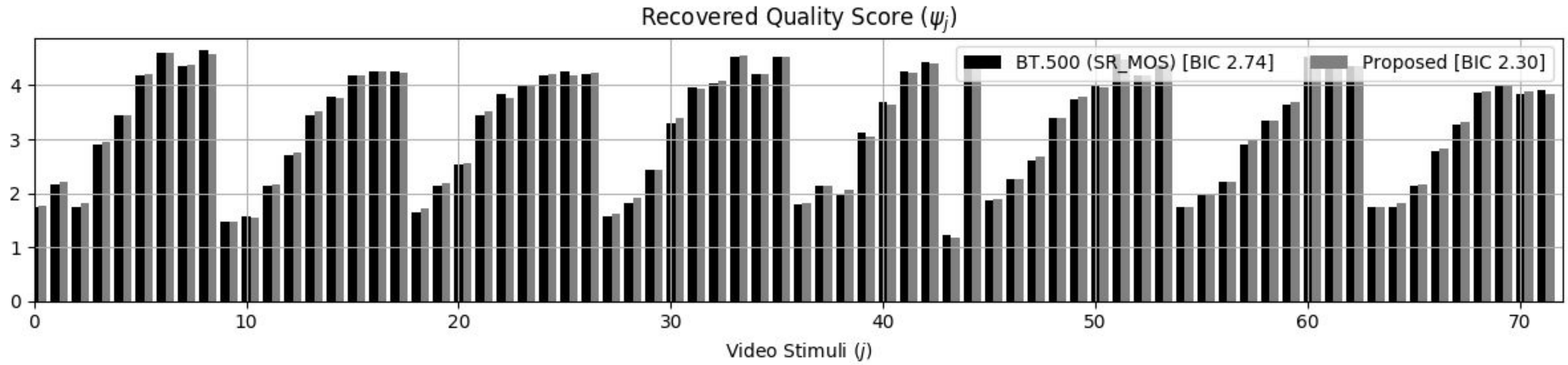
Robustness Against Subjects Giving “Flipped” Scores

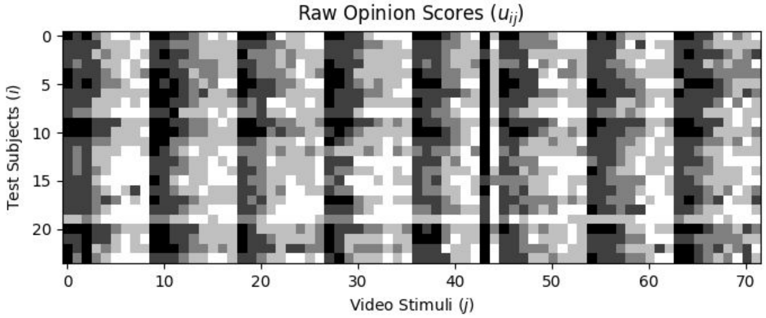


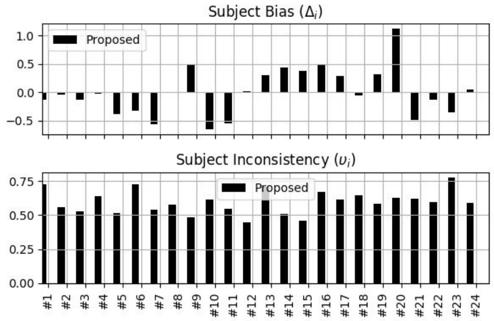
Malicious behavior: scores are “flipped”, for example, 1 for 5, 2 for 4, 2.5 for 3.5, and so on
Y-axis: RMSE with respect to clean case
SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

More Datasets

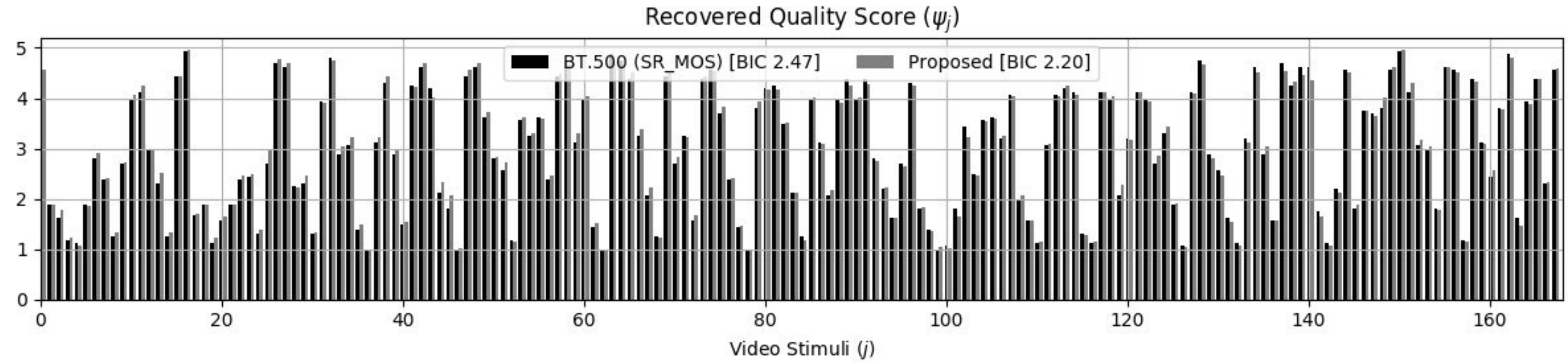
VQEGHD3_dataset_raw

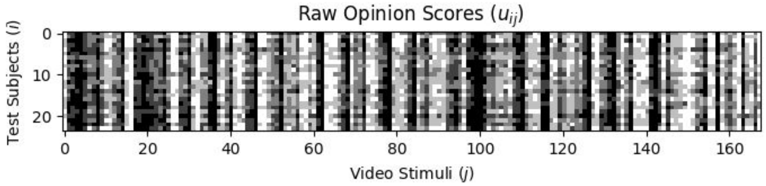


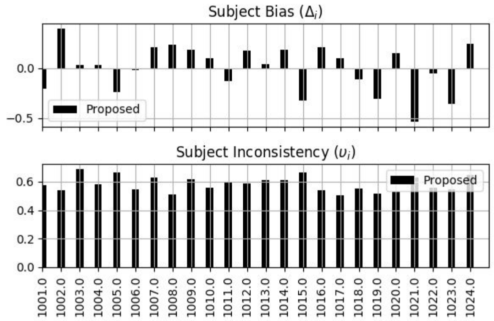




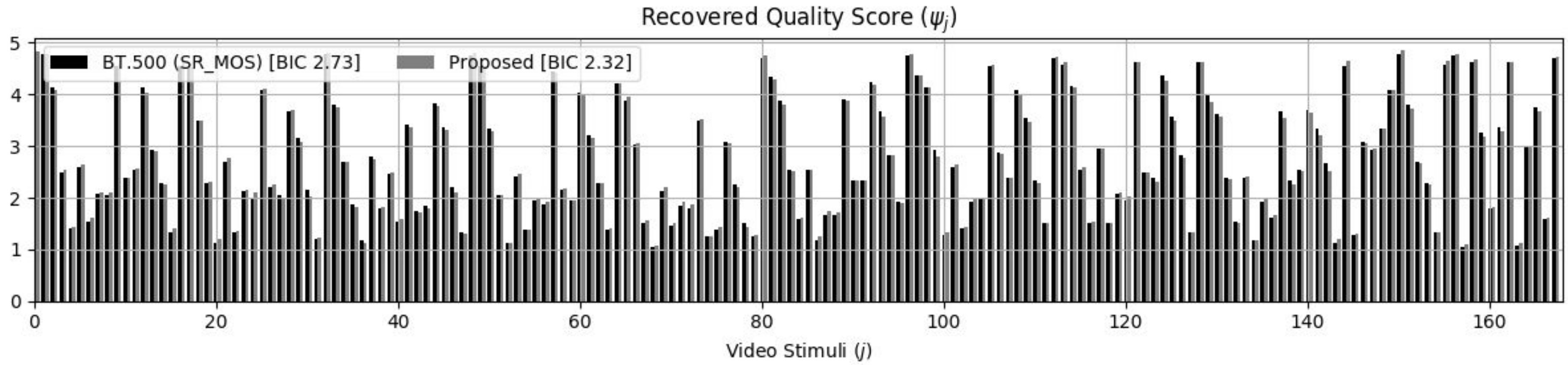
HDTV_Phase_I_Experiment_1

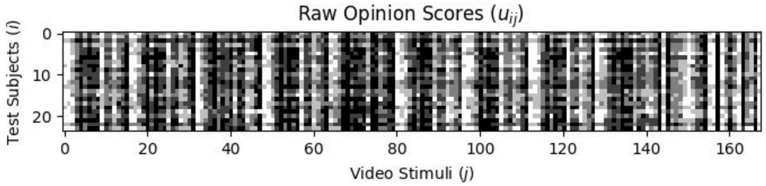


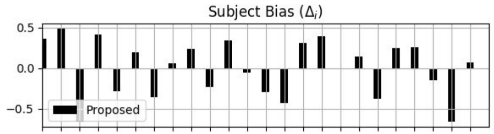


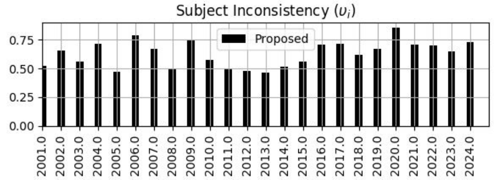


HDTV_Phase_I_Experiment_2

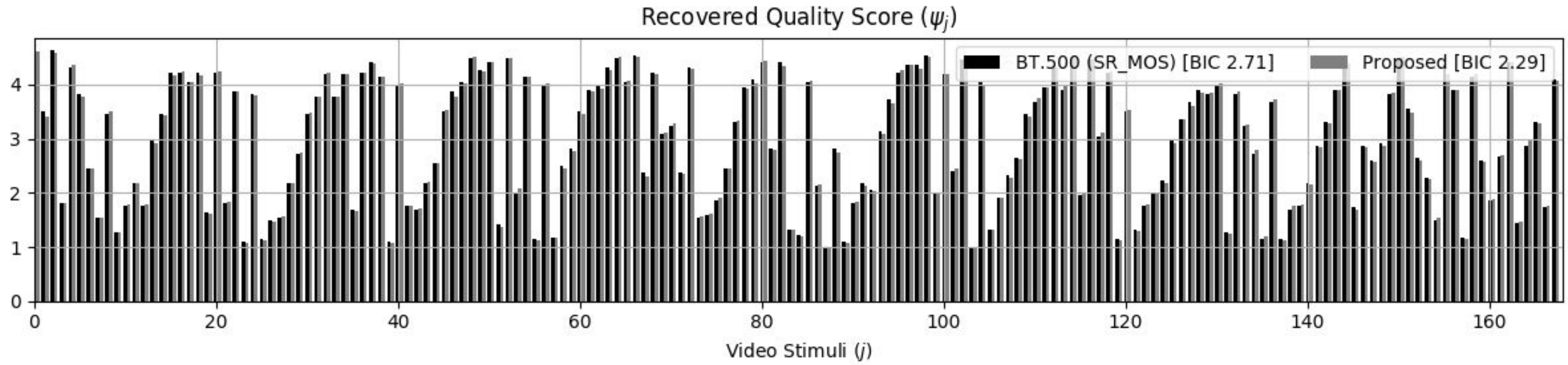


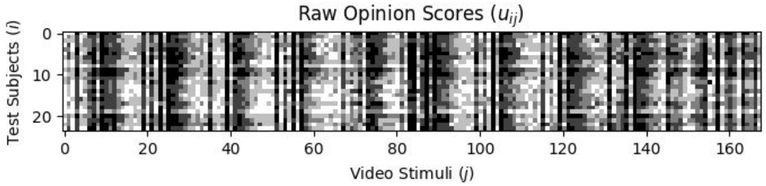


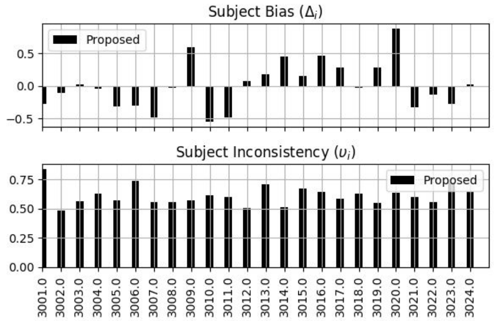




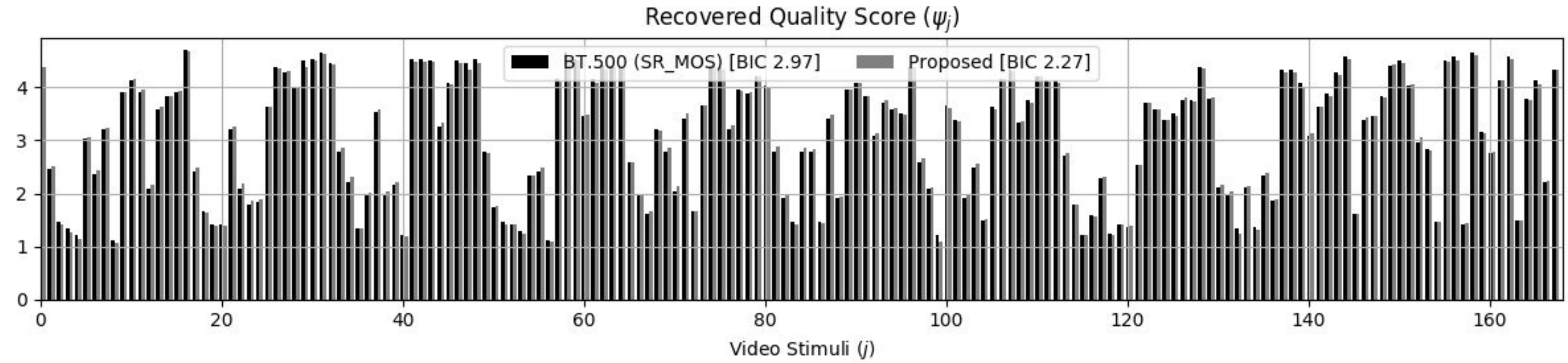
HDTV_Phase_I_Experiment_3

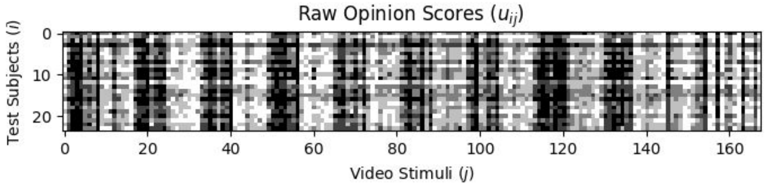


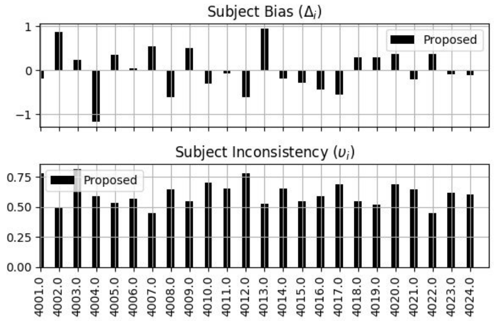




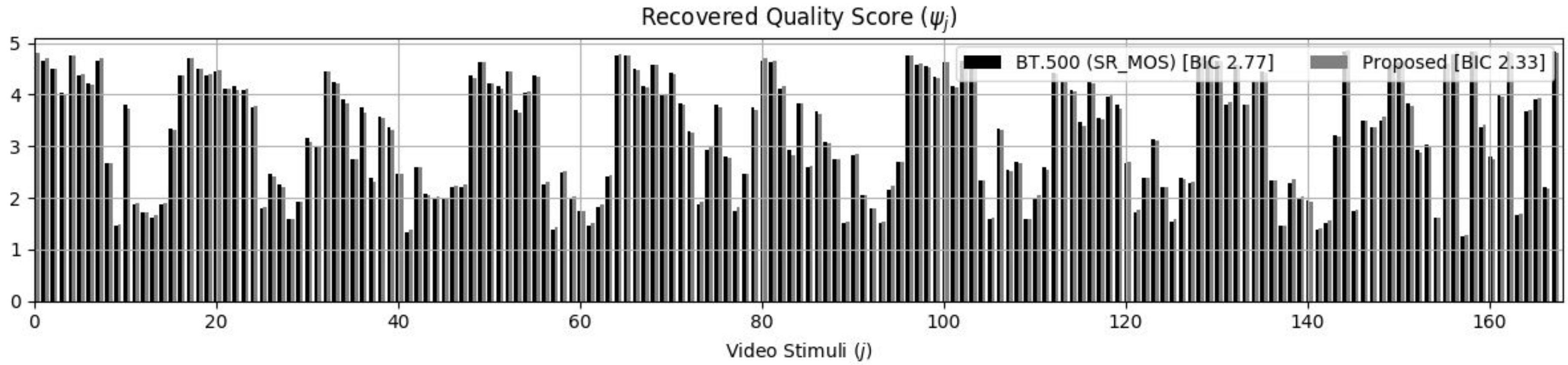
HDTV_Phase_I_Experiment_4

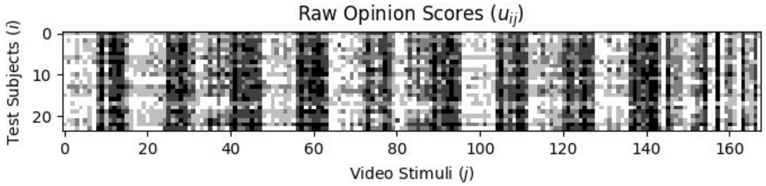


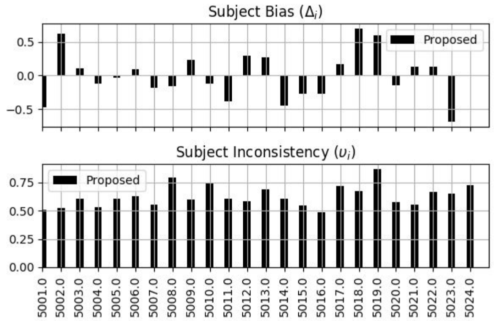




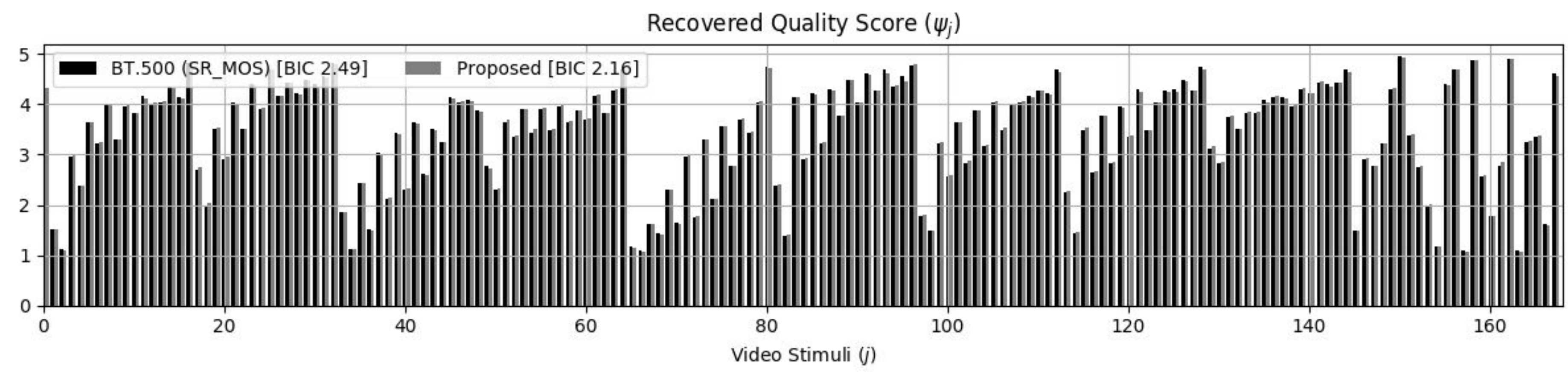
HDTV_Phase_I_Experiment_5

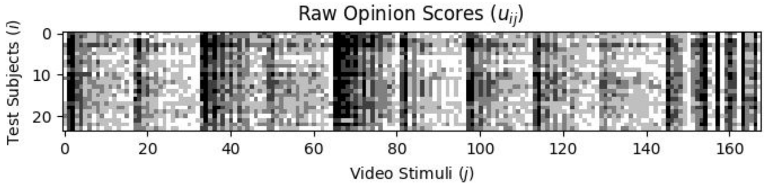


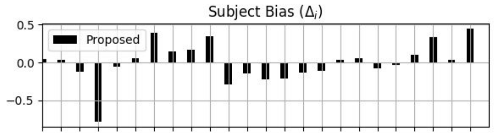


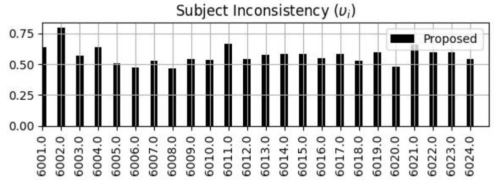


HDTV_Phase_I_Experiment_6

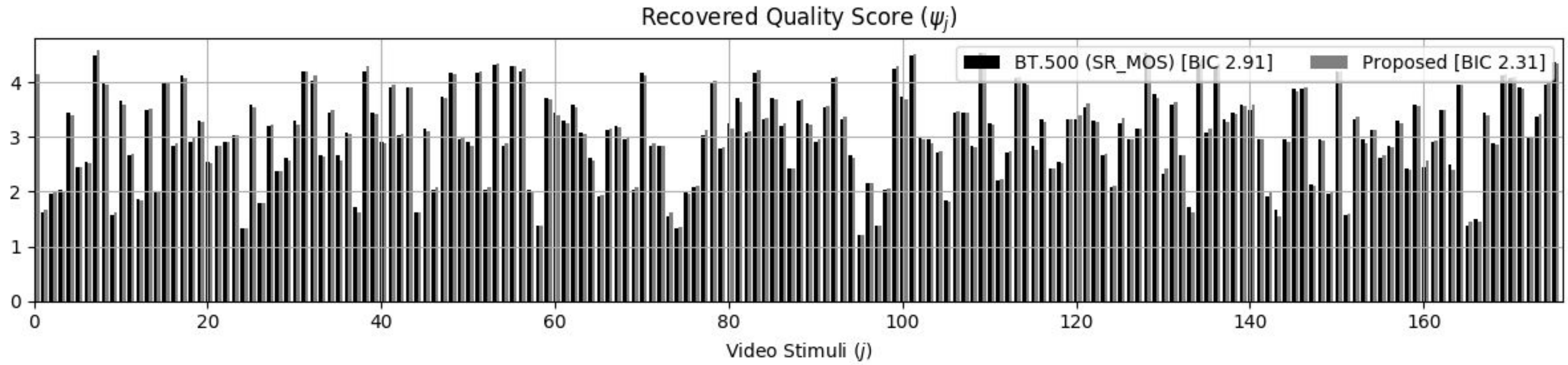


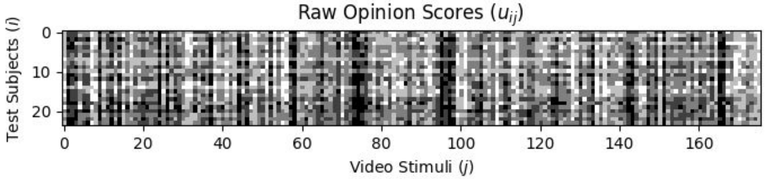


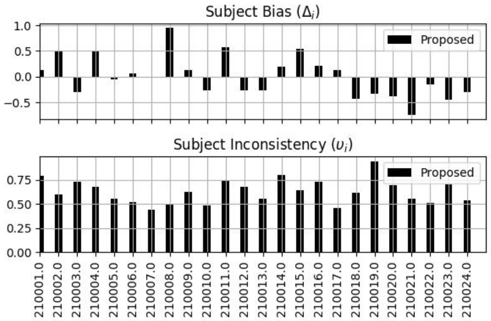




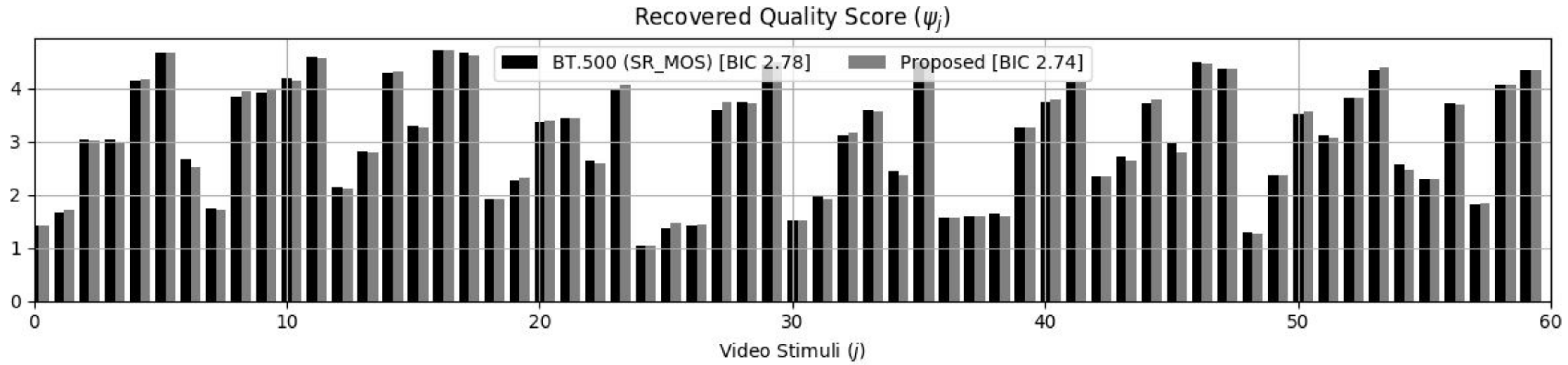
ITU-T_Supp_23_Experiment_1_BNR



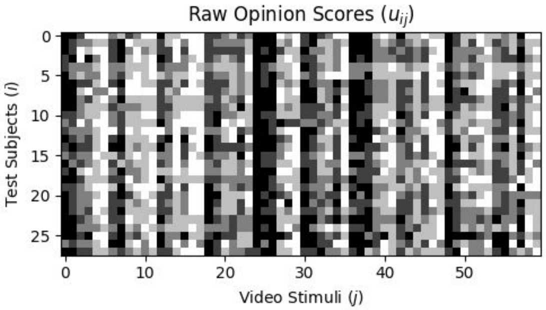




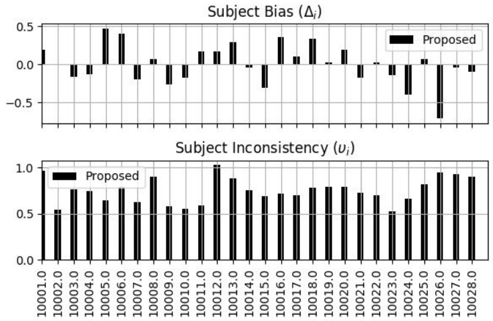
MM2_1



Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.78]
Proposed [BIC 2.74]
Video Stimuli (j)

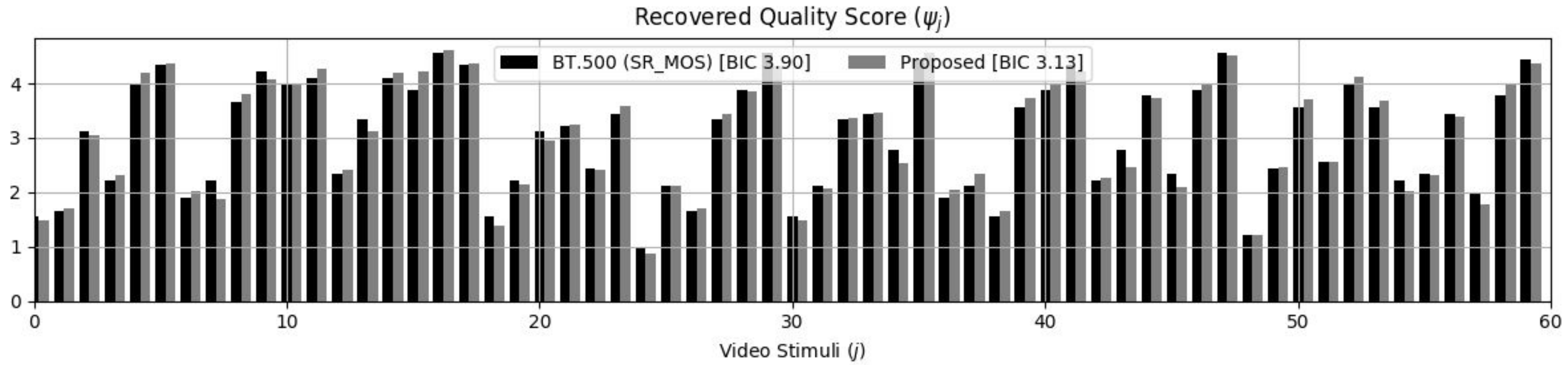


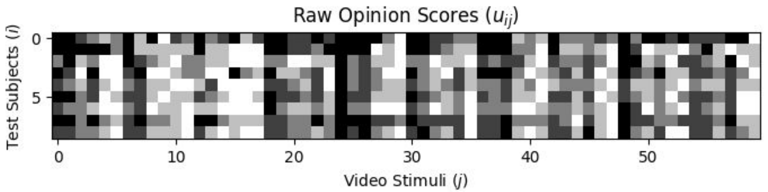
Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

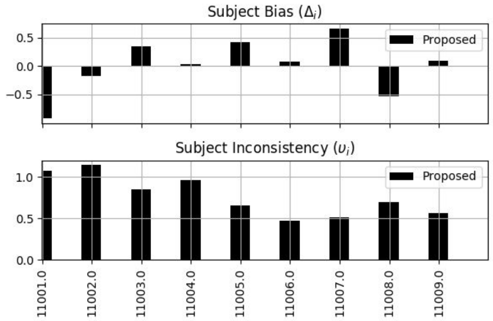


Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed

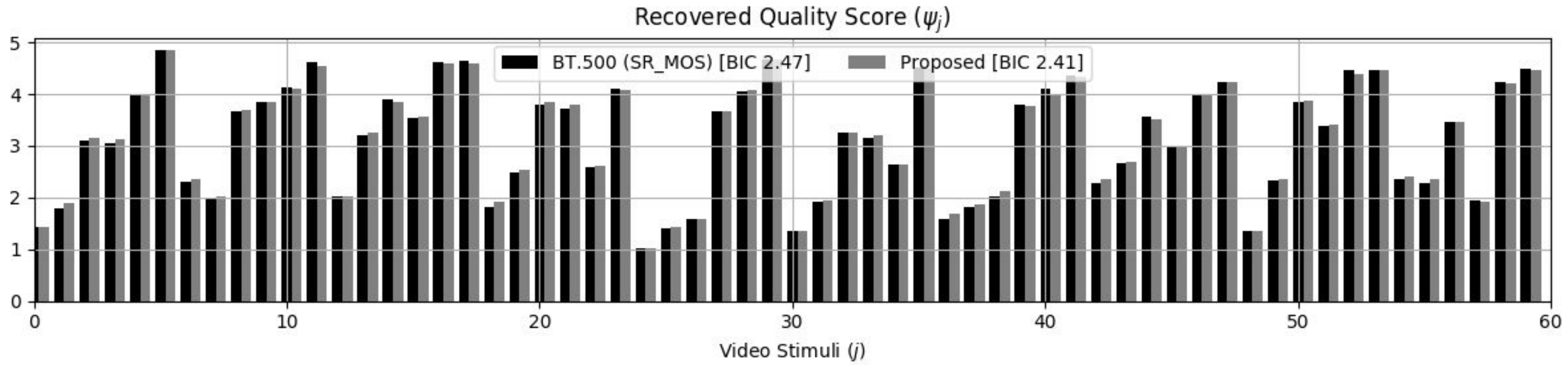
MM2_2

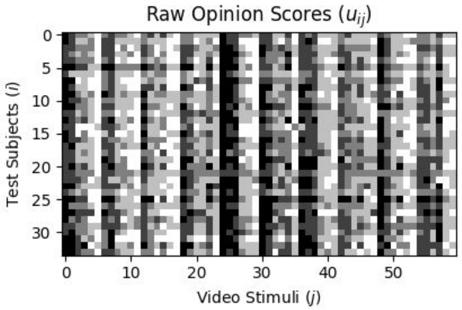

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.90]
Proposed [BIC 3.13]
Video Stimuli (j)

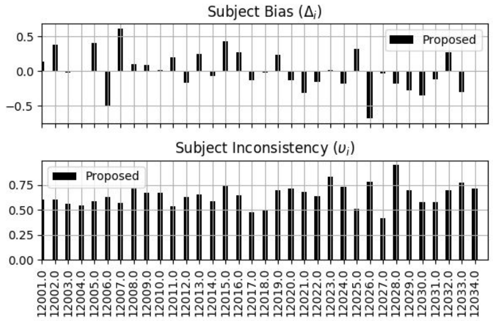

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)


Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed
11001.0
11002.0
11003.0
11004.0
11005.0
11006.0
11007.0
11008.0
11009.0

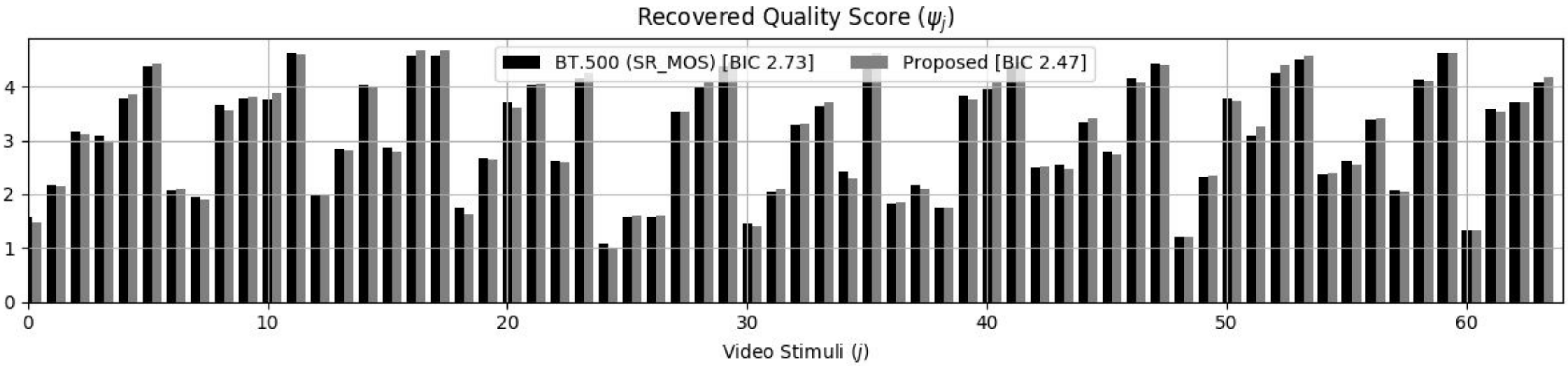
MM2_3

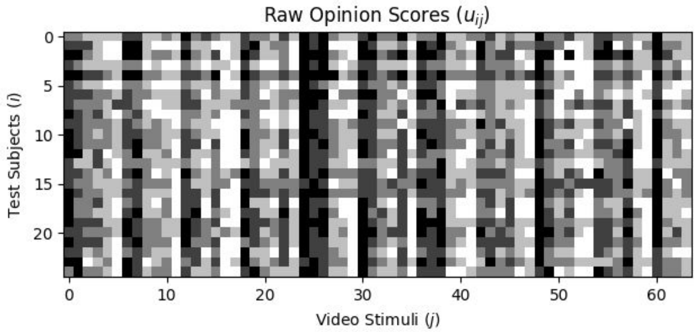

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.47]
Proposed [BIC 2.41]
Video Stimuli (j)

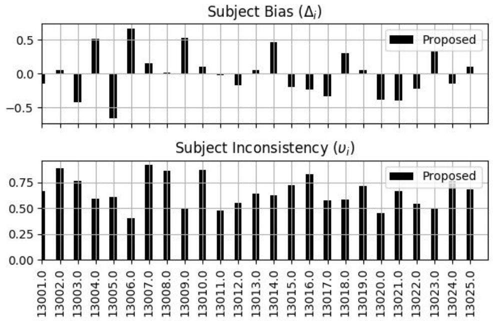

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)


Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed

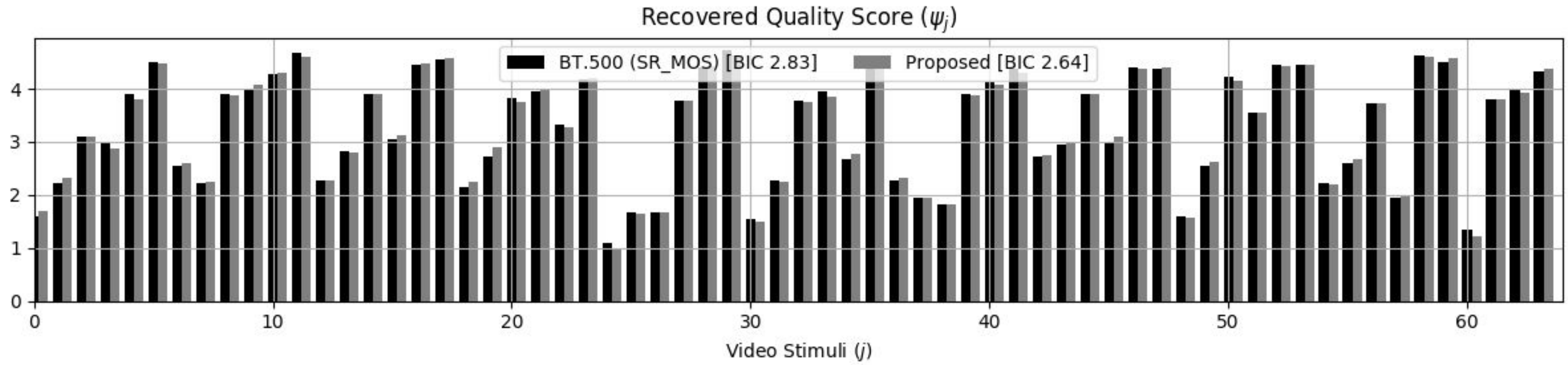
MM2_4


Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.73]
Proposed [BIC 2.47]
0
1
2
3
4
0
10
20
30
40
50
60
Video Stimuli (j)

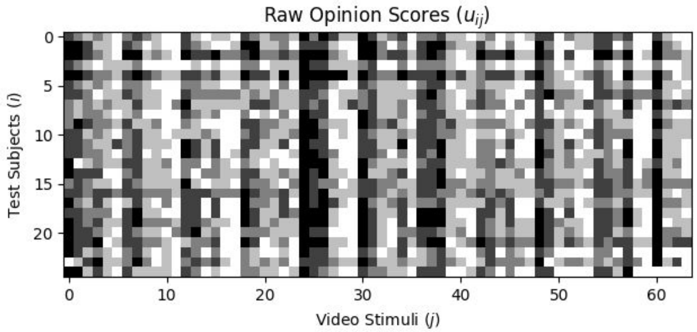

Raw Opinion Scores (u_{ij})
Test Subjects (i)
0
5
10
15
20
0
10
20
30
40
50
60
Video Stimuli (j)


Subject Bias (Δ_i)
Proposed
0.5
0.0
−0.5
Subject Inconsistency (υ_i)
Proposed
0.75
0.50
0.25
0.00
13001.0
13002.0
13003.0
13004.0
13005.0
13006.0
13007.0
13008.0
13009.0
13010.0
13011.0
13012.0
13013.0
13014.0
13015.0
13016.0
13017.0
13018.0
13019.0
13020.0
13021.0
13022.0
13023.0
13024.0
13025.0

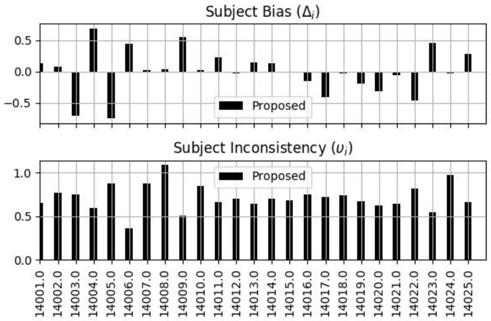
MM2_5



Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.83]
Proposed [BIC 2.64]
Video Stimuli (j)

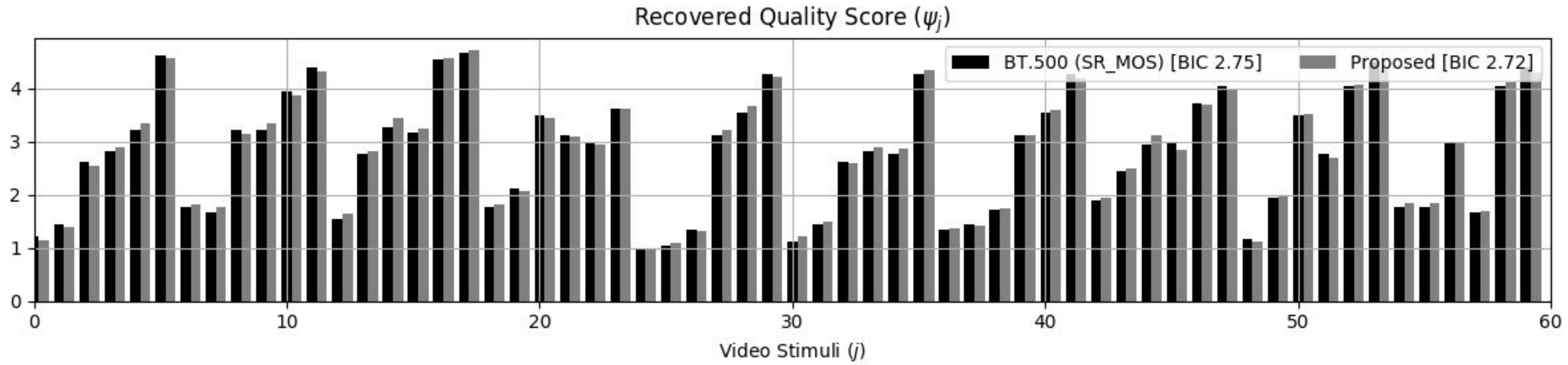


Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

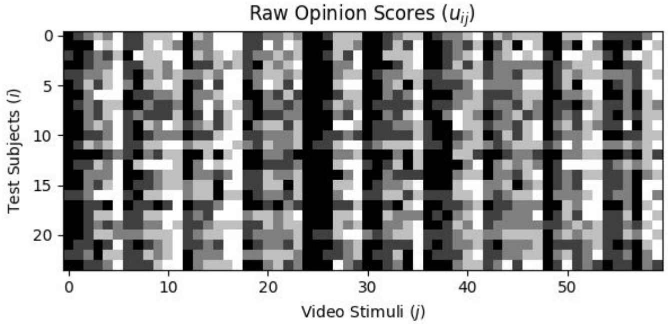


Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed
14001.0
14002.0
14003.0
14004.0
14005.0
14006.0
14007.0
14008.0
14009.0
14010.0
14011.0
14012.0
14013.0
14014.0
14015.0
14016.0
14017.0
14018.0
14019.0
14020.0
14021.0
14022.0
14023.0
14024.0
14025.0

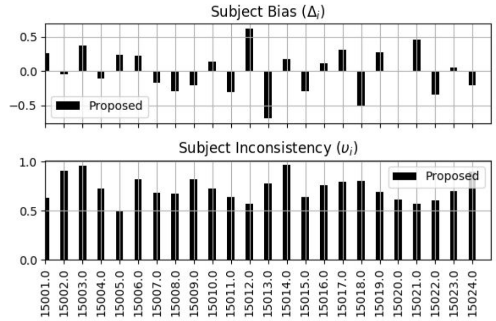
MM2_6



Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.75]
Proposed [BIC 2.72]
Video Stimuli (j)

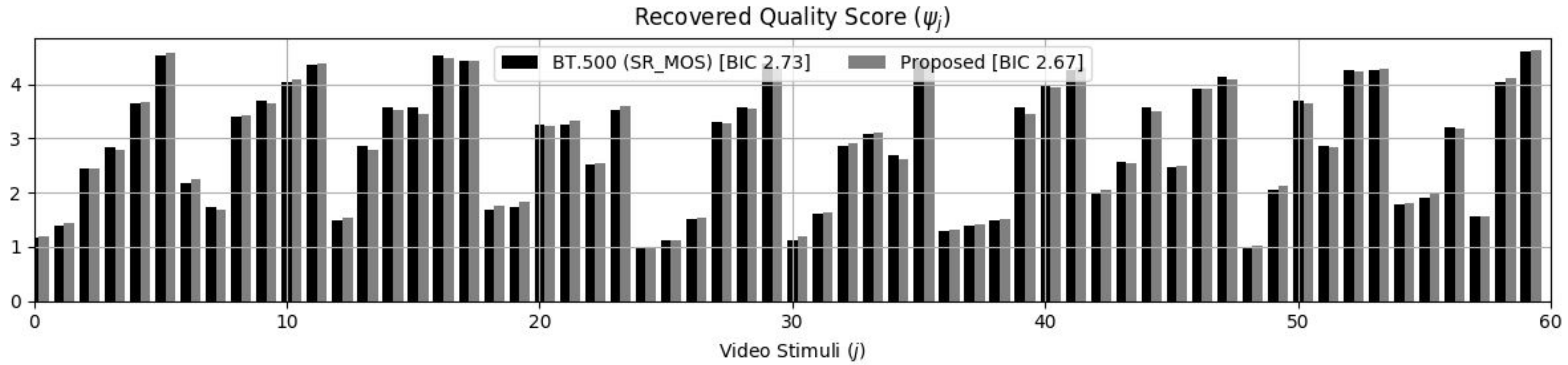


Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

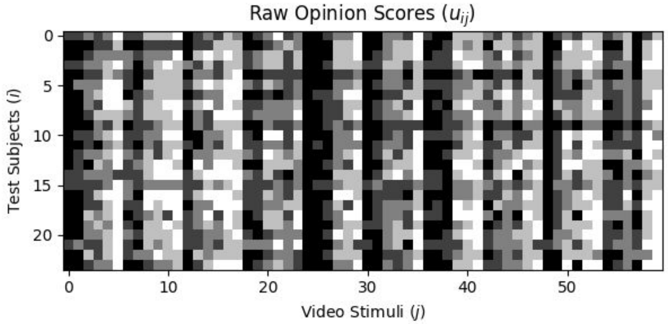


Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed

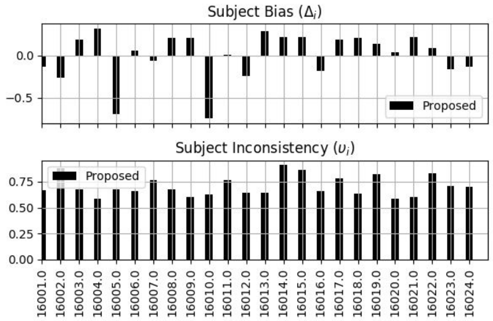
MM2_7



Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.73]
Proposed [BIC 2.67]
0
1
2
3
4
0
10
20
30
40
50
60
Video Stimuli (j)

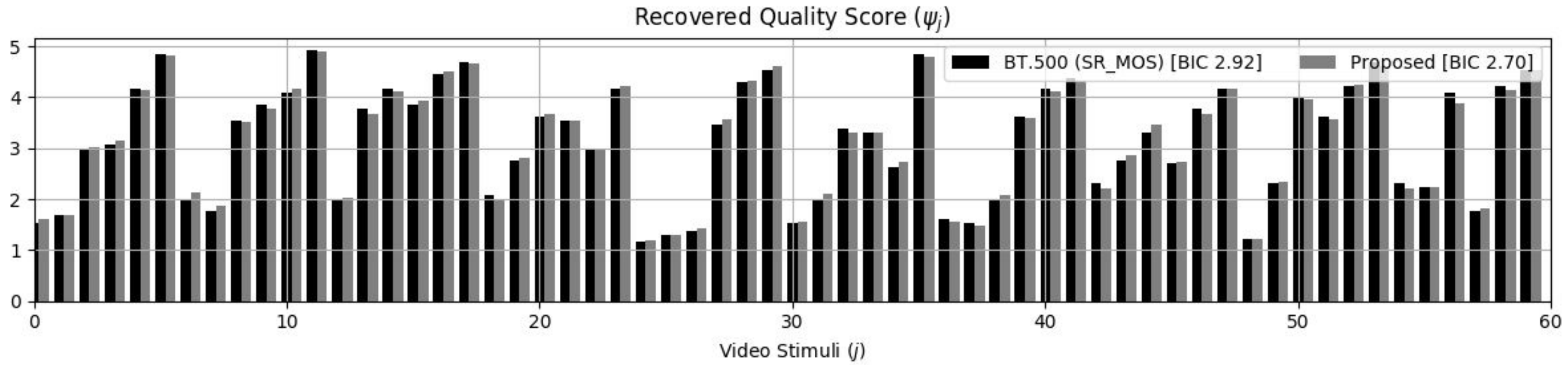


Raw Opinion Scores (u_{ij})
Test Subjects (i)
0
5
10
15
20
0
10
20
30
40
50
Video Stimuli (j)

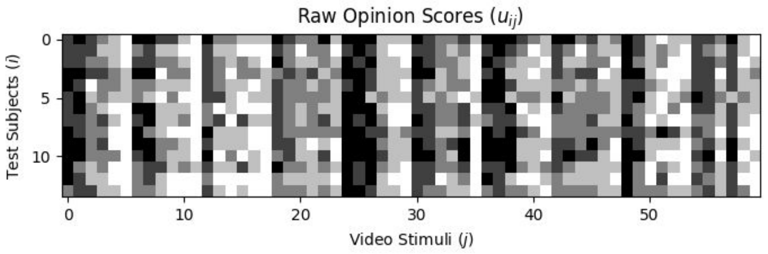


Subject Bias (Δ_i)
0.0
−0.5
Proposed
Subject Inconsistency (υ_i)
Proposed
0.75
0.50
0.25
0.00
16001.0
16002.0
16003.0
16004.0
16005.0
16006.0
16007.0
16008.0
16009.0
16010.0
16011.0
16012.0
16013.0
16014.0
16015.0
16016.0
16017.0
16018.0
16019.0
16020.0
16021.0
16022.0
16023.0
16024.0

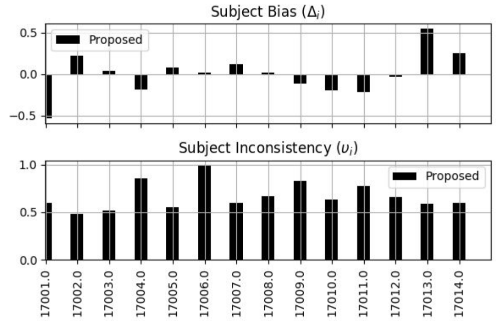
MM2_8



Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.92]
Proposed [BIC 2.70]
Video Stimuli (j)

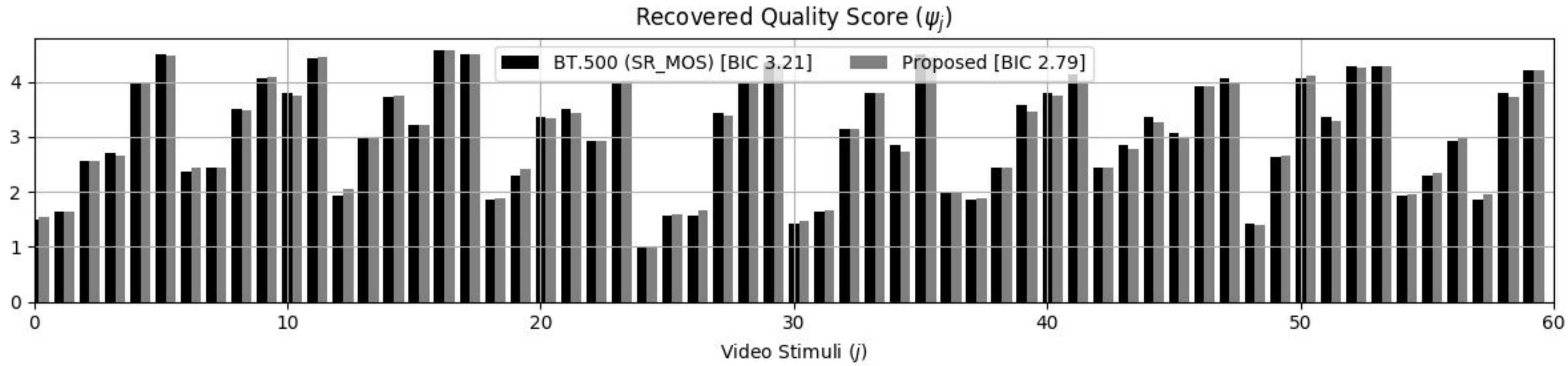


Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

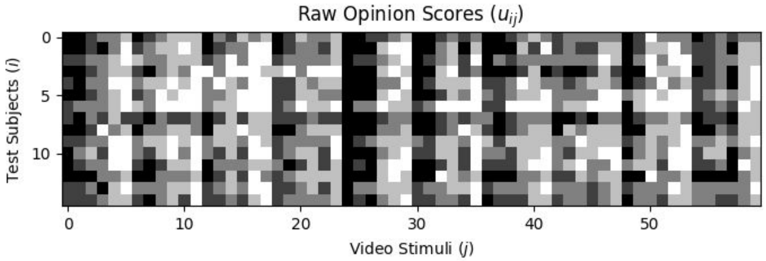


Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed
17001.0
17002.0
17003.0
17004.0
17005.0
17006.0
17007.0
17008.0
17009.0
17010.0
17011.0
17012.0
17013.0
17014.0

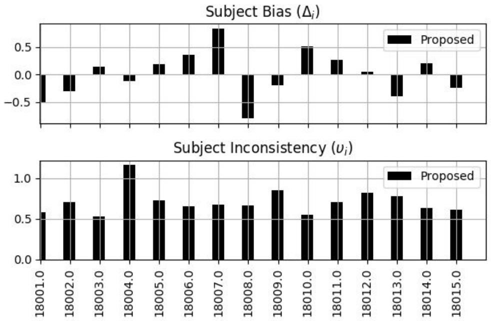
MM2_9



Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.21]
Proposed [BIC 2.79]
0
1
2
3
4
0
10
20
30
40
50
60
Video Stimuli (j)

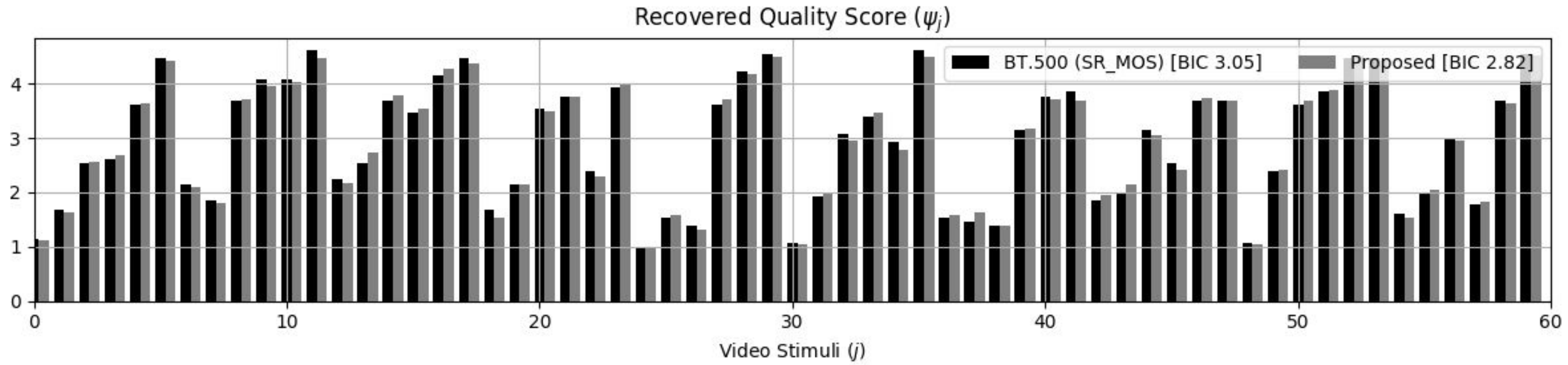


Raw Opinion Scores (u_{ij})
Test Subjects (i)
0
5
10
0
10
20
30
40
50
Video Stimuli (j)

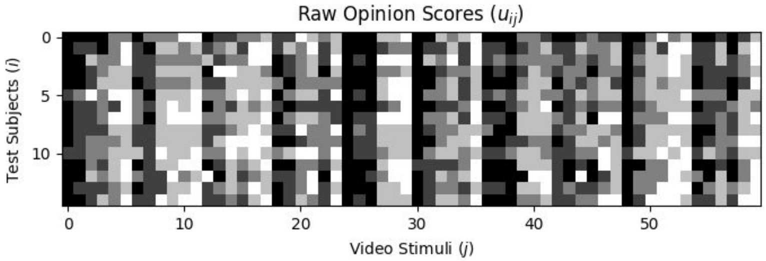


Subject Bias (Δ_i)
Proposed
0.5
0.0
−0.5
Subject Inconsistency (υ_i)
Proposed
1.0
0.5
0.0
18001.0
18002.0
18003.0
18004.0
18005.0
18006.0
18007.0
18008.0
18009.0
18010.0
18011.0
18012.0
18013.0
18014.0
18015.0

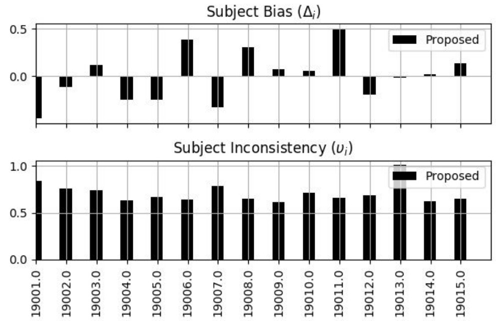
MM2_10



Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.05]
Proposed [BIC 2.82]
Video Stimuli (j)

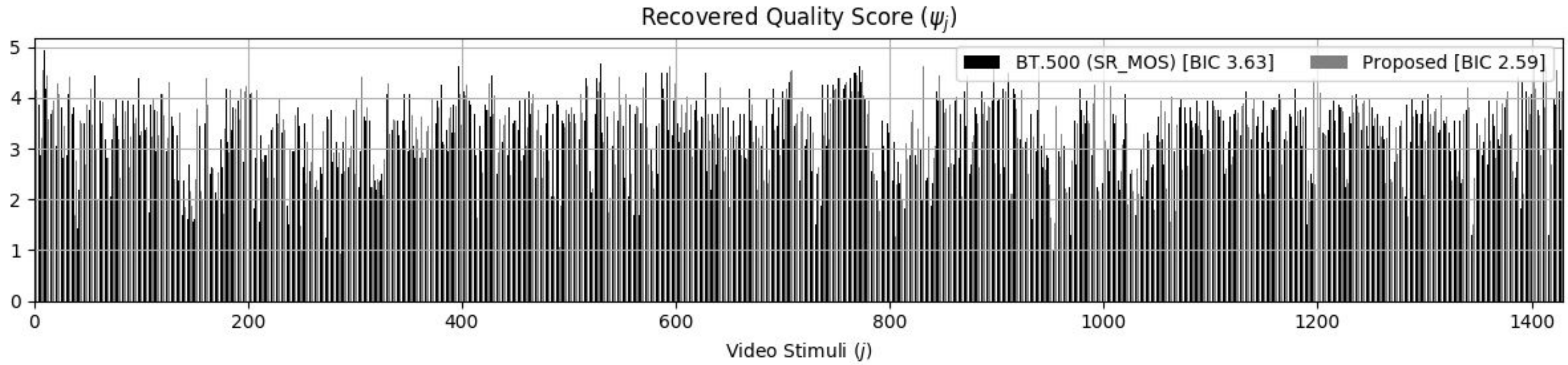


Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)



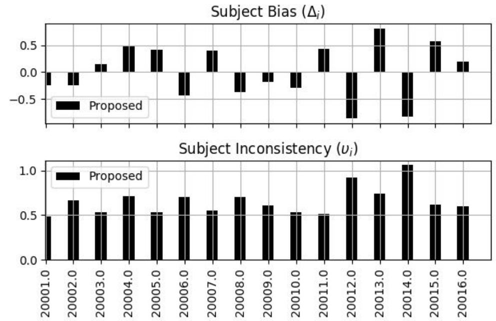
Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed
19001.0
19002.0
19003.0
19004.0
19005.0
19006.0
19007.0
19008.0
19009.0
19010.0
19011.0
19012.0
19013.0
19014.0
19015.0

its4s2



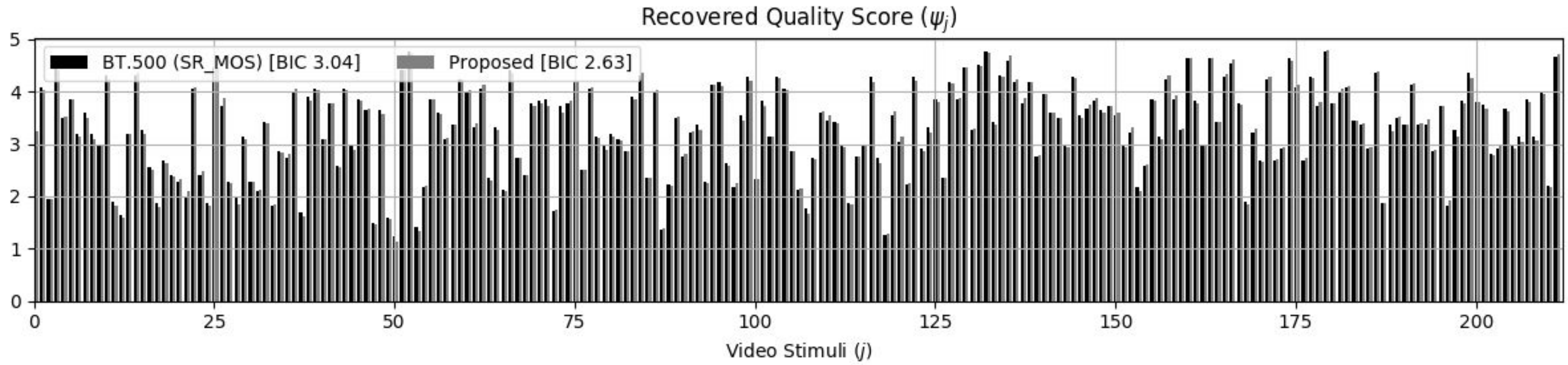
Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.63]
Proposed [BIC 2.59]
Video Stimuli (j)

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

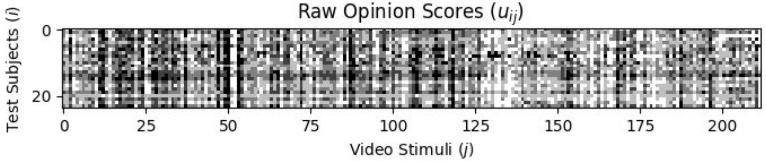


Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed
20001.0
20002.0
20003.0
20004.0
20005.0
20006.0
20007.0
20008.0
20009.0
20010.0
20011.0
20012.0
20013.0
20014.0
20015.0
20016.0

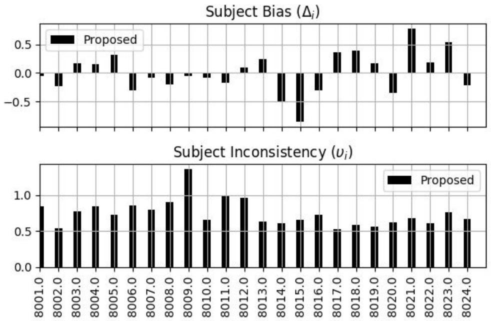
its4s_AGH



Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.04]
Proposed [BIC 2.63]
Video Stimuli (j)

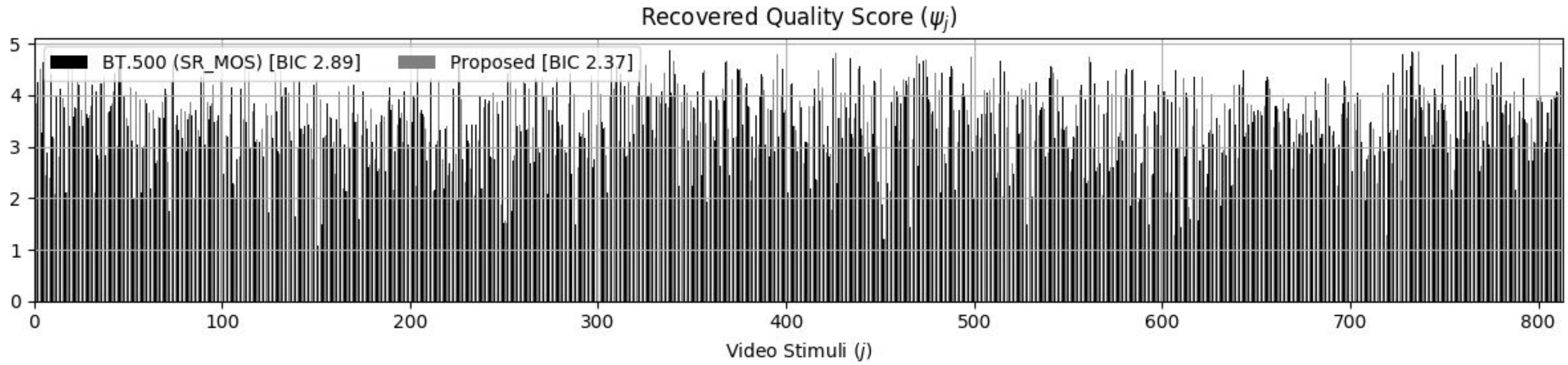


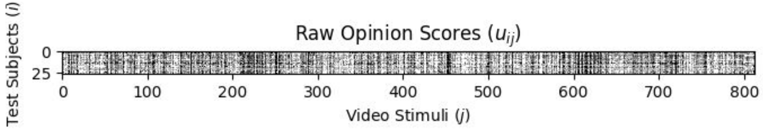
Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

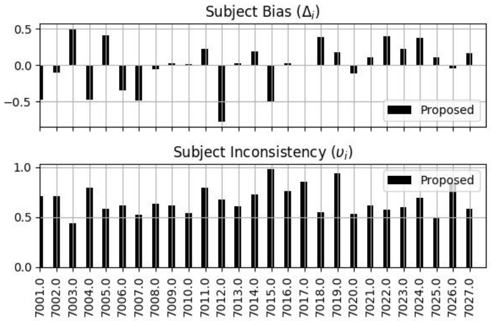


Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed

its4s_NTIA


Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.89]
Proposed [BIC 2.37]
Video Stimuli (j)


Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)


Subject Bias (Δ_i)
Proposed
Subject Inconsistency (υ_i)
Proposed