SUBJECT BIAS: INTRODUCING A THEORETICAL USER MODEL

Lucjan Janowski *

Department of Telecommunications
AGH University of Science and Technology,
Krakow 30–059, Poland

Margaret Pinson

U.S. Department of Commerce,
National Telecommunications and Information Administration (NTIA),
Boulder, CO, USA

ABSTRACT

We propose a model for rating behavior based on subject bias and subject error. Evidence for subject bias can be found in freely available subjective experiments. When subject bias is removed from ratings, the sensitivity of statistical comparisons between stimuli usually improves. According to our model, subject biases characterize the subject pool. These between-subject differences are important when analyzing and comparing people. On the other hand, it is advantageous to remove subject bias when analyzing mean opinion score. We conclude that bias acts like a random variable within ratings.

Index Terms— design of experiments, mean opinion score, QoE, subjective ratings, video quality assessment

1. INTRODUCTION

Subjective video quality experiments are wildly used to assess opinions about telecommunication services. In a typical experiment, a pool of 24 people rate the perceptual quality of various video sequences. It is important for our analysis that we understand the human factors influencing the subject ratings.

Pinson et al. [1] describes a systematic study of audiovisual subjective quality testing conducted by six laboratories from four countries. The stimuli and scale were held constant across experiments and labs; only the environment of the subjective test was varied. This indicated that after the number of subjects, the most important variable was how opinions differed among people.

Cermak and Fay [2] examined the T1A1 subjective dataset[1] and observed that subjects center their scores around different fulcrum points, resulting in a bias in scores. Cermak and Fay theorized that these biases were meaningful and important, and could be explained if sufficient information were known. They tried to prove this hypothesis using the scores and questionnaire data gathered for the 114 subjects (i.e., gender, experience with video teleconferencing, age, visual acuity, color vision test). Their conclusion was that, "From a practical standpoint, *subject differences are not meaningful in predicting ratings of video quality.*" [2]

Cermak and Fay proposed two different methods to respond to subject bias:

1. Collect data from more subjects, or
2. Remove subject mean from the data

So far, the subjective testing community has embraced the first option. A few researchers have removed both mean and variance from each subject's scores. Ostaszewska and Żebrowska-Łucyk [4] convert each subject's ratings into z-scores with zero mean and unit variance. Also, van Dijk et al. [5] used z-scores to span the same range on the rating scale.

Nevertheless, those methods are not commonly used. One reason could be insufficient proof that that this technique is valid for video quality tests. Also Cermak and Fay [2] do not recommend removing variance. Their theoretical justification for only removing subject bias is that this describes the subject without impacting the subject's relative ranking of processed video sequences (PVS). This can be justified if mean opinion scores (MOS) are relative rather than absolute, which the results presented in [1] support.

This paper proposes a model for subject behavior that includes subject bias. We do not treat subject bias as an error, rather as a natural feature. Our goal is to better understanding scoring behaviors. We will show that subject bias can be observed in a variety of subjective datasets, that subject bias does not depend upon the stimuli, and that it has a consistent behavior. We will then explore the impact of subject bias normalization on data analyses by showing that the obtained results precision can be increased by removing subject bias.

2. DATASETS

We will use four sets of subjective ratings to analyze our subject model.

First, dataset AGH/NTIA is from an experiment con-

*The work of L. Janowski has been supported by the AGH University of Science and Technology under contract no. 28.28.230.7027, C 2013/1-5/MITSU/2/2014.

[1]This dataset is from tests conducted in 1994, as the first independent validation of objective video quality models. For more information, see [3].

This paper was presented at and will appear in the Proceedings of the Sixth International Workshop on Quality of Multimedia Experience (QoMEX), 2014.

structed by Pinson and Janowski for the purposes of exploring this subject model. Dataset AGH/NTIA contains a sparse matrix of 94 source video sequences (SRC) and five hypothetical reference circuits (HRC) for a total of 110 PVSs. Data from 28 subjects was collected using the five-level absolute category rating (ACR) method, in accordance with ITU-T Rec. P.910. For a full description of dataset AGH/NTIA, see [6].

Second is a collection of six high definition television (HDTV) experiments conducted by the video quality experts group (VQEG) to validate HDTV objective quality metrics. These datasets are named vqegHD1, vqegHD2, vqegHD3, vqegHD4, vqegHD5, and vqegHD6. The individual subject ratings are available in the VQEG report [7]. Each of these six experiments was designed according to identical specifications, to contain a full matrix of 9 SRCs by 15 HRCs, plus a common set of 24 PVSs, for a total of 168 PVSs. The subjective data were collected using the same ACR method.

Third, dataset vqegMM2 is an audiovisual subjective dataset that contains 60 PVSs. Subjective data was collected at six different labs in ten different environments, for a total of 213 subjects. The subjective data were collected using the ACR method. This dataset was the basis for the Pinson et al. [1] analysis mentioned earlier. For a summary of the experiment and access to the subjective scores, see [8].

Fourth, dataset NTIA/Verizon [9] compares the performance of MPEG-2 and AVC/H.264 on HDTV, both coding only and in the presence of transmission errors. This experiment contains a partial matrix design, drawn from 12 SRC and 9 HRCs, for a total of 144 PVSs. The subjective data were collected using the ACR method.

3. PROPOSED SUBJECT MODEL

We propose that subject rating behavior within an experiment is governed by the following model:

$$o_{ij} = \psi_j + \Delta_i + \epsilon_{ij} \tag{1}$$

where

- o_{ij} is the observed rating for subject i and PVS j
- ψ_j is the true quality value for PVS j;
- Δ_i is the overall shift between the i^{th} subject's scores and the true values (i.e., opinion bias)
- ϵ_{ij} is the error (i.e., scoring imprecision)

We assume that:

- there is an underlying true value ψ_j, despite our inability to measure this value in absolute terms
- random variable Δ_i has a zero mean for all PVSs
- random variable ϵ_{ij} has a zero mean both generally (over all subjects and PVSs) and conditionally (for a particular subject or PVS)
- ψ_j and Δ_i are continuous variables, and thus $\Delta_i + \epsilon_{ij}$ is a deviation from a true value ψ_j

Our goal is to prove that each subject has a bias and that this bias is stable. By stable, we mean that subject bias does not change depending on distortion, source sequence or other factors. Depending on a source sequence would result in scoring better than average one source sequence and worst than average the other source sequence.

Variable ϵ_{ij} is influenced by multiple factors, such as the subject's imprecision and the PVS scoring difficulty (e.g., subjects may have trouble deciding how to rate a PVS with a brief transmission error). Variables o_{ij} and ϵ_{ij} should have a third subscript, r, that denotes the number of times that subject i has rated PVS j. A detailed analysis of ϵ_{ij} and r will be presented in a follow-up paper.

Our estimate of ψ_j is the mean value denoted by μ_{ψ_j} and given by:

$$\mu_{\psi_j} = \frac{1}{I_j} \sum_{i=1}^{I_j} o_{ij} \tag{2}$$

where

- μ denotes estimation from the data
- μ_{ψ_j} estimates ψ_j
- I_j is the total number of subjects for PVS j

By substituting our subject model for o_{ij}, we get:

$$\mu_{\psi_j} = \frac{1}{I_j} \sum_{i=1}^{I_j} \psi_j + \frac{1}{I_j} \sum_{i=1}^{I_j} \Delta_i + \frac{1}{I_j} \sum_{i=1}^{I_j} \epsilon_{ij} \tag{3}$$

Assuming $\frac{1}{I_j} \sum_{i=1}^{I_j} \Delta_i$ and $\frac{1}{I_j} \sum_{i=1}^{I_j} \epsilon_{ij}$ approach zero as I_j approaches infinity, we get:

$$\mu_{\psi_j} \approx \psi_j \tag{4}$$

Our estimate of Δ_i is the mean difference between μ_{ψ_j} and the i^{th} subject's ratings. It is denoted by μ_{Δ_i} and given by:

$$\mu_{\Delta_i} = \frac{1}{J_i} \sum_{j=1}^{J_i} (o_{ij} - \mu_{\psi_j}) \tag{5}$$

where

- μ_{Δ_i} estimates Δ_i
- J_i is the total number of PVSs rated by subject i

Put another way, we take the difference between the average of the i^{th} subject's ratings and the average rating computed over all subjects and all PVSs.

By substituting our subject model for o_{ij}, we get:

$$\mu_{\Delta_i} = \frac{1}{J_i} \sum_{j=1}^{J_i} (\psi_j - \mu_{\psi_j}) + \frac{1}{J_i} \sum_{j=1}^{J_i} \Delta_i + \frac{1}{J_i} \sum_{j=1}^{J_i} \epsilon_{ij} \tag{6}$$

Given (4) and assuming $\frac{1}{J_i} \sum_{j=1}^{J_i} \epsilon_{ij}$ approach zero as J_i approaches infinity, we get:

$$\mu_{\Delta_i} \approx \Delta_i \tag{7}$$

This paper was presented at and will appear in the Proceedings of the Sixth International Workshop on Quality of Multimedia Experience (QoMEX), 2014.

4. ANALYSIS OF Δ_i

By the Central Limit Theorem, we can assume that the measurement error of μ_{Δ_i} is normally distributed. It is not obvious that this assumption is true, especially for both ends of the scale. We validated this assumption on the AGH/NTIA dataset using Kolmogorov-Smirnov test run for each subject separately. For more than 64% of subjects the distribution of $(o_{ij} - \mu_{\psi_j})$ is statistically the same as the normal distribution. Since the visual investigation of the normality plot shows the strong influence of discretization, the obtained results support normality of the samples used to calculated μ_{Δ_i}.

With such an assumption, variable $(o_{ij} - \mu_{\psi_j})$ can be treated as a random variable drawn from a normal distribution with mean μ_{Δ_i} and standard deviation σ_i. The goal of the analysis is to estimate the mean. The confidence interval (CI) for that estimate is:

$$\mu_{\Delta_i} \pm z_{1-\alpha/2} \frac{\sigma_i}{\sqrt{J_i}} \tag{8}$$

where $z_{1-\alpha/2}$ is the inverse of the normalized normal distribution. This is the CI of μ_{Δ_i} for subject i, which indicates how well we have estimated Δ_i.

Table 1 shows the range of μ_{Δ_i} and σ_i for the datasets identified in Section 2. We see that the span of μ_{Δ_i} is similar for all datasets except vqegHD4 and vqegMM2. In the case of vqegMM2, the increased range of μ_{Δ_i} is probably caused by the large pool of 213 subjects i.e. greater chance to draw a strongly bias subject. Dataset vqegHD4 should be investigated in more detail. From Table 1 we can estimate a typical pool of μ_{Δ_i} being between (-0.6, 0.6).

Standard deviation is also similar for all datasets. Again vqegMM2 has the largest pool of subjects and so naturally has the widest range of σ_i. This time, the NTIA/Verizon dataset has a larger range. Similarly to the μ_{Δ_i} analysis, a deeper investigation is needed to understand the reasons or decide that it is caused just by luck.

If the span of μ_{Δ_i} is very small, it could indicate that our subject pool is specific and we should be careful about very general conclusions. On the other hand if the span of μ_{Δ_i} is very wide, a careful analysis of possible outliers is suggested. Dataset vqegHD4 provides an example.

The confidence intervals obtained for different values of μ_{Δ_i} cannot be shown for each experiment. As an example, Fig. 1 shows μ_{Δ_i} for dataset AGH/NTIA with subjects sorted by μ_{Δ_i}.

A detailed analysis of different SRCs and HRSs is difficult, because the number of answers for a single group is small and the discrete nature of o_{ij} becomes an important factor. The influence of SRC and HRC needs further investigation, since some irregularity can be seen for some subjects, especially considering different HRCs.

Nonetheless, we have shown that μ_{Δ_i} has a small standard deviation (see small CI at Fig. 1). This suggests that μ_{Δ_i} does not depend upon j, and so we can reasonably treat this as a variable that depends only upon i. It appears to be reasonable to ignore the underlying complexities that we know occur (e.g., like or dislike SRC, different impact of different HRCs).

Table 1: Range of μ_{Δ_i} and σ_i

Dataset	Subjects	μ_{Δ_i} [min,max]	σ_i [min,max]
AGH/NTIA	28	[-0.59, 0.60]	[0.44, 0.95]
vqegHD1	24	[-0.54, 0.40]	[0.51, 0.68]
vqegHD2	24	[-0.67, 0.49]	[0.49, 0.83]
vqegHD3	24	[-0.56, 0.88]	[0.50, 0.83]
vqegHD4	24	[-1.18, 0.95]	[0.47, 0.80]
vqegHD5	24	[-0.69, 0.70]	[0.50, 0.86]
vqegHD6	24	[-0.79, 0.45]	[0.48, 0.78]
vqegMM2	213	[-0.95, 0.90]	[0.43, 1.18]
NTIA/Verizon	21	[-0.65, 0.60]	[0.52, 1.08]

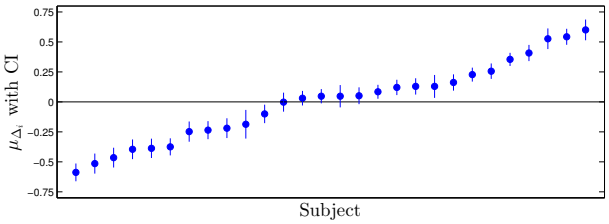


Fig. 1: Observed distribution of subject bias (μ_{Δ_i}) with confidence intervals, for dataset AGH/NTIA.

5. NORMALIZING SUBJECTIVE DATA

Our analysis of μ_{Δ_i} in Section 4 indicates that we should be able to remove the influence of μ_{Δ_i} from any subjective dataset. This normalization will not impact μ_{ψ_j}. However, the standard deviation of scores will change, and should decrease from:

$$s_j = \underset{i}{\sigma}(o_{ij}) \tag{9}$$

to

$$\tilde{s}_j = \underset{i}{\sigma}(r_{ij}) \tag{10}$$

where

$$r_{ij} = o_{ij} - \mu_{\Delta_i} \tag{11}$$

and where

- σ is standard deviation
- s_j is the standard deviation of subject ratings for PVS j
- $\tilde{s}_j$ is this same standard deviation, computed on subject data with the subject bias μ_{Δ_i} removed

We expect to see:

$$\tilde{s}_j < s_j \tag{12}$$

This paper was presented at and will appear in the Proceedings of the Sixth International Workshop on Quality of Multimedia Experience (QoMEX), 2014.

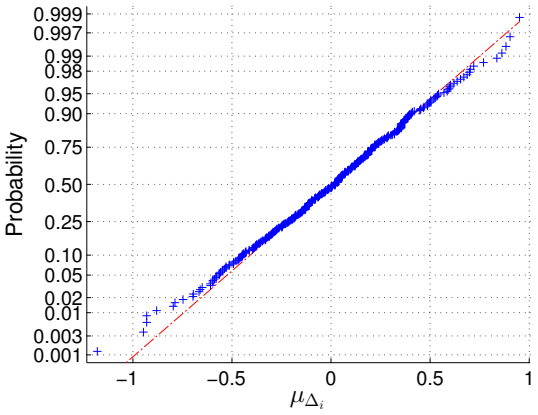


Fig. 2: Normal probability plot of μ_{Δ_i} indicates an approximately normal distribution.

while μ_{ψ_j} remains constant.

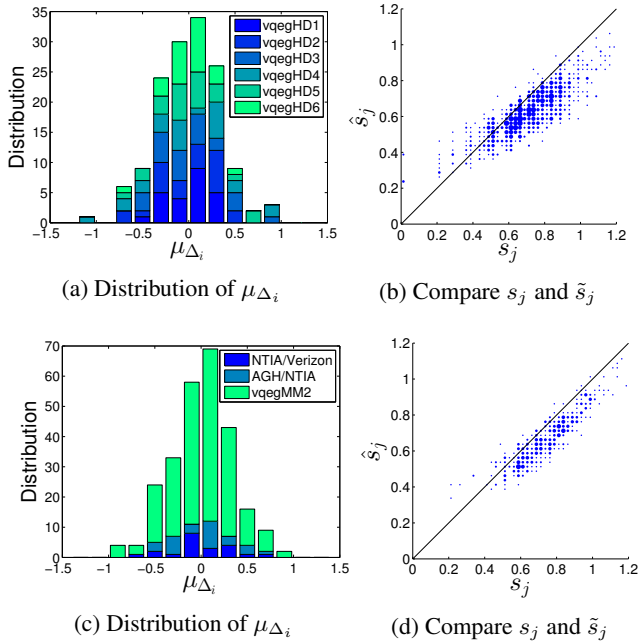
5.1. Distribution of μ_{Δ_i}, s_j and $\tilde{s}_j$

To test this theory, let us examine μ_{Δ_i}, s_j and $\tilde{s}_j$ for several freely available datasets. Fig. 2 shows the normal probability plot of μ_{Δ_i} for all nine datasets mentioned in Section 2. From this, we can see that μ_{Δ_i} has a normal distribution, and spans more than 25% of the rating scale.

Fig. 3a shows the distribution of μ_{Δ_i} values for the six VQEG HDTV datasets. Fig. 3b shows the difference between s_j and $\tilde{s}_j$ for the six VQEG HDTV datasets. The dot size indicates the data density. While $\tilde{s}_j$ is generally less than s_j, $\tilde{s}_j$ is larger for small values of s_j. These smaller standard deviations usually correspond to the saturation of the MOS scale, as we can see in Fig. 4. Further investigation of saturation and its impact on Δ_i estimation will require an improved understanding of ϵ_{ij} and the influence of the discrete, five-level scale.

Fig. 4 shows the relationship between μ_{ψ_j} and σ for the raw data (Fig. 4a) and when μ_{Δ_i} is removed (Fig. 4b). On average, σ decreases by 0.035 to 0.167, depending on the dataset. The red arrow on the bottom of the scatter plot indicates the change to the minimum σ as we move from raw data (Fig. 4a) to normalized data (Fig. 4b). The red arrow at the top of the scatter plot indicates the change in maximum σ. Removing μ_{Δ_i} usually moves the minimum and maximum σ value toward the median, thus eliminating extreme values. It means that for the minimum variance, (12) is not satisfied. We believe that the small variances seen before μ_{Δ_i} is removed are caused by clipping of the MOS scale and not an imperfection of the model given in (1).

The six scatter plots in Fig. 4a show a pattern typical of discrete rating scales like ACR. Variable s_j is larger near the middle of the scale and smaller at both ends—raising concerns of scale compression, although studies fail to indicate



(a) Distribution of μ_{Δ_i}

(b) Compare s_j and $\tilde{s}_j$

(c) Distribution of μ_{Δ_i}

(d) Compare s_j and $\tilde{s}_j$

Fig. 3: Impact of removing μ_{Δ_i} from the six VQEG HDTV datasets (a & b) and datasets vqegMM2, AGH/NTIA, and NTIA/Verizon (c & d). Scatter plot shows bins, where larger dots indicate more data.

an inherent superiority of continuous rating scales [10, 11]. After normalization (Fig. 4b), the $\tilde{s}_j$ values are more stable over the ACR scale.

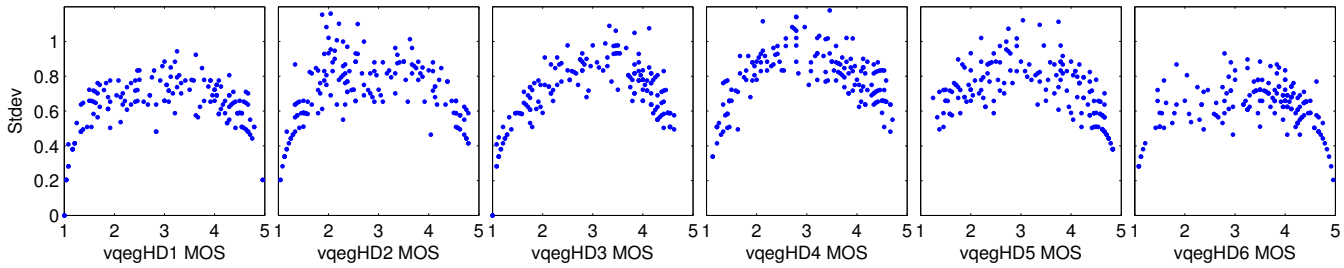
Figs. 3c and 3d show the distribution of μ_{Δ_i} values and the difference between s_j and $\tilde{s}_j$ for datasets vqegMM2, AGH/NTIA and NTIA/Verizon. These figures show the same trends we saw in Figs. 3a and 3b. Variable μ_{Δ_i} spans a similar range. Equation (12) appears to be satisfied except for small values of s_j and a few small increases.

5.2. Impact of Δ_i Removal on MOS Data Analysis

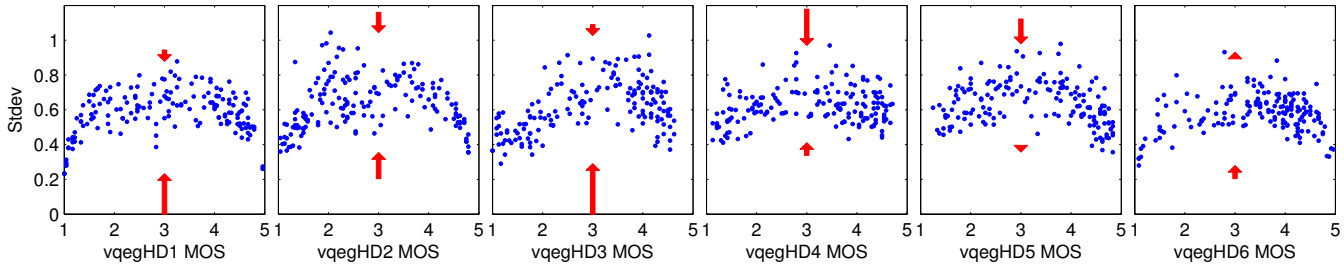
Although removing μ_{Δ_i} does not impact μ_{ψ_j}, the significant difference between PVSs can change. According to our subject model, this change should be beneficial (i.e., removing the part of subjects' differences that are important from the correctness of the subject pool point of view but not important from the MOS point of view). The consequences should be the ability to differentiate between more pairs of PVSs.

For all PVS pairs within each of the six VQEG HDTV datasets, a Student's t-test was used to calculate whether or not the PVSs had equivalent MOS at the 95% significance level. This calculation was repeated on the normalized scores (with μ_{Δ_i} removed). Based on the combined results from all six datasets, removing μ_{Δ_i} had the following impact:

- 97.22% no change
- 2.69% increase in sensitivity (equivalent → different)
- 0.08% decrease in sensitivity (different → equivalent)



(a) Raw subjective data



(b) Subjective data normalized by removing μ_{Δ_i}

Fig. 4: Impact of removing μ_{Δ_i} from the six VQEG HDTV datasets. Long red arrows indicate a large change in the minimum or maximum standard deviation of scores. Short red arrows indicate a small change.

- 0% inversions (opposite conclusions)

Inversions are impossible, because μ_{ψ_j} only changes due to rounding error.

Overall, normalizing each subject's scores by μ_{Δ_i} seems to improve our ability to distinguish between PVS MOS for the VQEG HDTV datasets. This supports the correctness of our subject model and provides corroborating evidence for the term Δ_i.

Now let us consider an applied example. The NTIA/Verizon experiment evaluates the hypothesis that H.264/AVC encoding yields equivalent perceptual quality at half to one third of the bitrate of MPEG-2. This hypothesis is evaluated at four different encoding bitrates. Each comparison includes eight HDTV SRC (see Pinson et al. [9] for details).

Table 2 shows the impact of μ_{Δ_i} removal when a Student's t-test analyzes this hypothesis for specific PVS pairs (e.g., H.264 at 2 Mbps versus MPEG-2 at 6 Mbps, for SRC "NTIA Red Kayak"). The third column (μ_{ψ_j} equivalent using o_{ij}) shows the number of SRC that had statistically equivalent MOS, based on the raw ratings and a 95% significance level. The fourth column (μ_{ψ_j} equivalent using r_{ij}) shows the number of SRC that had statistically equivalent MOS, when μ_{Δ_i} is removed from the subject ratings.

For three of the 24 comparisons, the normalized data allows us to detect a significant difference, while the raw data does not. This is a 12% increase in sensitivity, when the analysis is limited to the question that the experiment was designed to answer.

Table 2: Impact on Student's t-test Sensitivity

H.264	MPEG-2	μ_{ψ_j} equivalent using o_{ij}	μ_{ψ_j} equivalent using r_{ij}
10 Mbps	18 Mbps	8 of 8	7 of 8
6 Mbps	$12\frac{1}{2}$ Mbps	7 of 8	5 of 8
$3\frac{1}{2}$ Mbps	$8\frac{1}{2}$ Mbps	4 of 8	4 of 8
2 Mbps	6 Mbps	3 of 8	3 of 8

6. CONCLUSION

We propose a linear model for subject behavior, which includes three variables: true quality (ψ_j), subject bias (Δ_i), and subject error (ϵ_{ij}). We have shown evidence supporting the existence of the term Δ_i and shown that subjective ratings are influenced by the subject's bias (Δ_i). When subject bias is removed from subject ratings, the sensitivity of statistical comparisons between stimuli usually improves.

Whether or not to remove Δ_i depends upon the type of data analysis.

- When the analysis focuses on MOS comparisons, then Δ_i should be removed. The NTIA/Verizon dataset [9] provides an example. Most subjective tests use this type of MOS analysis, and thus would benefit from removing Δ_i.
- When the analysis compares objective and/or subjective data with user descriptions (e.g., from blogs, forums, or questionnaires), then MOS and subject bias should be taken into consideration.

This paper was presented at and will appear in the Proceedings of the Sixth International Workshop on Quality of Multimedia Experience (QoMEX), 2014.

- When the analysis focuses on subject behavior, then the analaysis could focus only on Δ_i. The vqegMM2 dataset [1] provides an example.

The authors are conducting further investigations into the subject model and subject error (ϵ_{ij}). Another interesting researcher area is finding a typical pool of bias in different groups, like different countries.

7. APPENDIX

The following MATLAB® code implements the equations seen in this paper. Input argument `oij` is o_{ij} from (1). Output argument `rij` is r_{ij} from (11); `deltai` is μ_{Δ_i} from (5); `deltaistd` is σ_i from (8); and array `ni` holds J_i for each subject.

```
% SYNTAX
%   [rij, deltai, deltaistd, ni] = ...
    normalizationByDelta(oij)
% SEMANTICS
% Input: oij matrix with subjects ratings ...
    where rows are different PVSs and ...
    columns are different subjects.
% Output: rij normalized matrix with the ...
    same structure as oij (that is, rij = ...
    oij - deltai).
%         deltai is the subjects bias
%         deltaistd is the standard ...
    deviation of a subject bias
%         ni  number of correct ratings per ...
    subject

function [rij, deltai, deltaistd, ni] = ...
    normalizationByDelta(oij)

psi = nanmean(oij,2);
temp = bsxfun(@minus, oij, psi);
deltai = nanmean(temp);
deltaistd = nanstd(temp);
rij = bsxfun(@minus, oij, deltai);
ni=zeros(1,length(deltai));
for i=1:length(deltai),
    ni(i) = sum(~isnan(rij(:,i)));
end
```

8. REFERENCES

[1] 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.," *J. Sel. Topics Signal Processing*, vol. 6, no. 6, pp. 640–651, 2012.

[2] G. W. Cermak and D. A. Fay, "T1A1.5 video quality project: GTE labs analysis," ATSI T1 contribution T1A1.5/94-148. Sep. 1994. Available: `ftp://vqeg.its.bldrdoc.gov/Documents/OLD_T1A1/`

[3] M.H. Pinson, N. Staelens, and A. Webster, "The history of video quality model validation," in *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, Sept 2013, pp. 458–463.

[4] A. Ostaszewska and S. Żebrowska Łucyk, "The method of increasing the accuracy of mean opinion score estimation in subjective quality evaluation," in *Wearable and Autonomous Biomedical Devices and Systems for Smart Environment*, Aimé Lay-Ekuakille and SubhasChandra Mukhopadhyay, Eds., vol. 75 of *Lecture Notes in Electrical Engineering*, pp. 315–329. Springer Berlin Heidelberg, 2010.

[5] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," in *Advanced Image and Video Communications and Storage Technologies*, N. Ohta, H. U. Lemke, and J. C. Lehureau, Eds., Feb. 1995, vol. 2451 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp. 90–101.

[6] M. Pinson and L. Janowski, "Repeats: A video quality subjective test with repeated sequences," NTIA Technical Memo TM-14-505, June 2014.

[7] M. Pinson, F. Speranza, M. Barkowski, V. Baroncini, R. Bitto, S. Borer, Y. Dhondt, R. Green, L. Janowski, T. Kawano, C. Lee, J. Okamoto, R. Renaud, C. Schmidmer, N. Stalens, A. Takahashi, and Q. Thu, "Report on the validation of video quality models for high definition video content," VQEG, 2010.

[8] M.H. Pinson, C. Schmidmer, L. Janowski, R. Pepion, Quan Huynh-Thu, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "Subjective and objective evaluation of an audiovisual subjective dataset for research and development," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, July 2013, pp. 30–31.

[9] M.H. Pinson, S. Wolf, and G. Cermak, "Hdtv subjective quality of h.264 vs. mpeg-2, with and without packet loss," *Broadcasting, IEEE Transactions on*, vol. 56, no. 1, pp. 86–91, March 2010.

[10] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, June 2010, pp. 82–87.

[11] Quan Huynh-Thu and Mohammed Ghanbari, "Modelling of spatio–temporal interaction for video quality assessment," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 535–546, 2010.