A Scoring Model Considering the Variability of Subjects' Characteristics in Subjective Experiments

2023 15th International Conference on Quality of Multimedia Experience (QoMEX)

A Scoring Model Considering the Variability of Subjects' Characteristics in Subjective Experiments

Lohic Fotio Tiotsop, Antonio Servetti, Enrico Masala
Control and Computer Engineering Department
Politecnico di Torino
Turin, Italy
lohic.fotiotiotsop@polito.it, antonio.servetti@polito.it, enrico.masala@polito.it

***Abstract*—Many authors argued that the scoring behavior of a subject in a subjective quality evaluation experiment can be modeled by two main characteristics, i.e., the subject's bias and the subject's inconsistency. However, for simplicity's sake, they disregarded the fact that subjects are usually less inconsistent when evaluating stimuli with very low or very high quality. This work addresses this shortcoming by providing an analytical formulation about how to link subjects' bias and inconsistency to the *ground truth subjective quality* of the stimulus under evaluation. By integrating this formulation into a state-of-the-art subject scoring model we obtain a more realistic model to recover the *ground truth subjective quality* of each stimulus. An iterative algorithm able to estimate the model parameters is also provided. Computational experiments show that our proposed model yields more realistic confidence intervals for the recovered *ground truth subjective quality* values and exhibits more robustness to synthetically added noise in several testing conditions.**

***Index Terms*—Subjective quality, Subject's bias, Subject's inconsistency, Quality recovery, SOS hypothesis**

I. Introduction

Raw ratings from subjects in a subjective quality evaluation experiment are usually noisy [1]. This is due, for instance, to potential distractions and/or fatigue of the subject, and also to the complexity of the stimuli under evaluation. Approaches to recover reliable subjective scores from raw ratings are therefore of paramount importance.

Most of the recent approaches make assumptions on the subject behavior in order to recover the ground truth subjective quality from noisy ratings. An assumption adopted by several authors [1]–[3] is that the subject's scoring behavior can be captured by two main characteristics, i.e., the subject's bias and inconsistency.

The bias is a systematic tendency to provide low (negative bias) or high (positive bias) ratings, while the inconsistency is a measure of the inability of a subject to provide an accurate rating and repeat it when rating the same stimulus several times.

In previous works [2], [3], in order to limit complexity, the dependency of the bias and the inconsistency on the actual quality of the stimulus under evaluation has been disregarded. However, there are strong empirical evidences of the fact that the manifestation of the inconsistency as well as the bias of a subject when rating a given stimulus depends on the quality of the stimulus itself.

For instance, at the extremes of the quality scale, i.e., when the quality is particularly low or high, subjects tend to express similar ratings [4]. In such a case, the manifestation of the peculiarities of each individual subject is therefore not noticeable. Thus, the subject bias and inconsistency should assume near-zero values at the extremes of the scale when modeling the subject behavior in that range.

We propose two analytical formulations linking, respectively, the bias and the inconsistency of a subject to the quality of the evaluated stimulus. We then integrate the proposed formulation into a state-of-the-art subject scoring model and provide an iterative procedure to recover the ground truth quality of each stimulus, as well as each subject's bias and inconsistency as defined from our new integrated subject scoring model.

Computational experiments conducted on several datasets show that our proposal yields more realistic confidence intervals (CIs) of the recovered quality. In particular, unlike previous approaches [3] that compute CIs with equal size for all stimuli, our approach computes smaller CIs for very low and very high quality stimuli, thus effectively modeling the lower inconsistency of the subjects at the extremes of the quality scale. Furthermore, the newly proposed model showed higher robustness in several testing conditions to synthetically added noise.

The paper is organized as follows. Section II briefly reviews the related work. In Section III, we propose two formulations linking subjects' bias and inconsistency to the quality of the stimulus under evaluation, and integrate them into a state-of-the-art subject scoring model. Section IV describes the algorithm we propose to estimate the parameters of the integrated scoring model. Section V discusses computational experiments and conclusions are drawn in Section VI.

II. Related Work

The literature focusing specifically on how to recover the subjective quality of a media from noisy raw opinion scores is rather limited. The simplest approach, i.e. the Mean Opinion Score (MOS), is known to be a not so suitable estimator of the subjective quality when the gathered raw individual

979-8-3503-1173-0/23/$31.00 ©2023 IEEE

ratings include outliers. Therefore, approaches to identify outlier subjects and to exclude them from the dataset before computing the MOS have been investigated, such as the quality recovery algorithm proposed in the ITU-T Rec BT.500 [5] and the ITU-T Rec P.913 [6].

Several authors, e.g., [3], [7], believe that these approaches exclude more data than needed from the dataset, since it is very unlikely that a subject inaccurately rated all the stimuli. In more recent works [3], [7]–[10], the authors leveraged advanced statistical methods to propose new approaches that avoid subjects' rejection. Among these approaches, an interesting research direction has been to model each subject's scoring behavior with a bias and an inconsistency [1], [2]. The latest advances in this research direction are summarized in [3] and implemented in the Netflix's Sureal software [11].

This work contributes to advance the state-of-the-art in this research direction, i.e., the one assuming that the subject's behavior can be reasonably modeled by a bias and an inconsistency. In particular, it is the first work that considers the fact that the manifestation of the bias and the inconsistency of a subject depends on the quality of the stimulus the subject is asked to rate. This consideration allows, for instance, to model the higher consistency of subjects when rating very low or very high quality stimuli.

III. Linking the Subject Bias and Inconsistency to the Stimulus Quality

Let us denote by $\mathcal{I}$ a set of subjects, $\mathcal{J}$ a set of stimuli, and r_{ij} the rating of the subject $i \in \mathcal{I}$ for the stimulus $j \in \mathcal{J}$.

The work in [3] presents the more recent advances on approaches that recover the subjective quality from noisy individual ratings by modeling the subjects behavior through a bias and an inconsistency. In particular, the authors argue that the rating r_{ij} can be modeled as it follows:

$$r_{ij} = q_j + \beta_i + N(0, \sigma_i) \tag{1}$$

where q_j is the ground truth quality of the stimulus $j \in \mathcal{J}$, β_i is the bias of the subject $i \in \mathcal{I}$, σ_i is the inconsistency of the subject $i \in \mathcal{I}$, and $N(0, \sigma_i)$ is a normal random variable with mean equal to 0 and standard deviation equal to σ_i.

According to the model in Eq (1), the manifestation of the bias and the inconsistency of a subject does not depend on the quality of the stimulus under evaluation. In fact, b_i and σ_i do not depend on j.

This is a shortcoming of such a model, since it is, for instance, empirically known that subjects are less inconsistent when evaluating stimuli of very low or very high quality [12]. In addition, the authors in [4] formulated the so-called Standard deviation of the Opinion Scores (SOS) hypothesis. According to the SOS hypothesis, at the extremes of the quality scale the SOS tends to 0. That means that subjects tend to express similar ratings at the extremes of the quality scale. In other words, there is not a significant diversity among subjects' behaviors at the extreme of the quality scale. Thus any feature, e.g., the bias and inconsistency, aimed at measuring how peculiar is a specific subject with respect to the others, must assume a value close to 0 at the extremes of the quality scale.

Therefore, in this paper, we propose to express the manifestation of the bias and the inconsistency of a subject i when rating a stimulus j as a function of the ground truth quality q_j of that stimulus as it follows:

$$b_{ij} = b_i \cdot \mathbb{1}_{[2,4]}(q_j) \tag{2}$$

$$\sigma_{ij} = \alpha_i(-q_j^2 + 6q_j - 5) \tag{3}$$

where $\mathbb{1}_{[2,4]}(q_j)$ is an indicator function, i.e., it is equal to 1 if the quality q_j of the stimulus j belongs to [2, 4], and 0 otherwise. b_i and α_i are the bias and inconsistency factors of the subject i, respectively.

The formulas in Eq (2) and Eq (3) are designed in such a way as to cancel the bias and the inconsistency of all subjects at the extremes of the scale, and thus to model the subjects similarity in those areas of the scale. For the bias, the cancellation is performed by means of the function $\mathbb{1}_{[2,4]}(q_j)$ which is 0 when the quality is in the intervals [1, 2) and (4, 5] whereas, for the inconsistency, we use the polynomial function $-q_j^2 + 6 \cdot q_j - 5$ that assumes its maximum value when $q_j = 3$ and vanishes in $q_j = 1$ and $q_j = 5$. In this way, we model the fact that we expect larger inconsistency at the center of the scale, while going towards the extremes, it must decrease progressively towards zero.

While our choice to model the link between the inconsistency and the quality with a quadratic function is strongly inspired by the SOS hypothesis in [4], the proposed link between the bias and the quality is rather simple and potentially not the optimal one. However, in the next sections it will be shown to be accurate enough to guarantee the effectiveness of the proposed ground truth quality recovery algorithm.

We propose to integrate the formulas in Eq (2) and Eq (3) into the model in Eq (1), to obtain the following model:

$$\begin{gathered} r_{ij} = q_j + b_{ij} + N(0, \sigma_{ij}) \\ b_{ij} = b_i \cdot \mathbb{1}_{[2,4]}(q_j) \\ \sigma_{ij} = \alpha_i(-q_j^2 + 6q_j - 5) \end{gathered} \tag{4}$$

that we will refer to as the *integrated model* in the rest of the paper.

A. On the Limits of the Proposed Integrated Model

Before describing our approach to estimate the parameters of the proposed integrated model, let us briefly discuss some of its limits. In fact, we believe that, although the proposed model in Eq (4) addresses some of the shortcomings of the model in Eq (1), it still suffers from some limitations.

In particular, the duration of a subjective experiment might influence the bias and inconsistency of a subject. For instance, the subject's fatigue in large-scale subjective experiments can cause larger inconsistency. Our model does not consider the variability of the bias and inconsistency as the subjective experiment goes on.

Moreover, to be able to derive a closed-form analytical formulation of the confidence interval for the recovered quality (see Eq (13)) a Gaussian distribution, i.e., a continuous distribution with an unbounded support, is used to model raw opinion scores that are, instead, very often gathered on a discrete and bounded scale. Thus, in practice, when using the proposed model to simulate subjects' ratings, an approach to truncate the simulated values is required.

Finally, we observe that, although the bias and the inconsistency are fundamental aspects of the scoring behavior of a subject, they do not tell the whole story. The opinion score of a subject for a given stimulus is influenced by many other factors, e.g., the complexity of the stimulus, the interest that the subject has in the content of that stimulus, etc.

IV. An Iterative Algorithm to Estimate the Parameters of the Integrated Model

We propose an iterative algorithm to compute the parameters of the integrated model proposed in Eq (4), inspired by the approach followed by the Sureal software that estimates the parameters of the model in Eq (1). In particular, we will denote by q_j^n, b_i^n and α_i^n respectively, the value of the parameter q_j, b_i and α_i at the n-th iteration of the algorithm. Also, we denote by q_j^0 the initial value of $q_j \;\; j \in \mathcal{J}$, and set it to be equal to the MOS of the stimulus j. Some of the parameters repeatedly used are summarized in Table I for the reader's convenience.

To estimate the inconsistency factor α_i for each subject i, we observe that, according to the model in Eq (4), the overall standard deviation of the difference between the ratings of the subject i and the ground truth quality can be theoretically expressed as:

$$s_i = \left(\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \sigma_{ij}^2 \right)^{\frac{1}{2}} = \alpha_i \left(\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} (-q_j^2 + 6q_j - 5)^2 \right)^{\frac{1}{2}} \tag{5}$$

Algorithm 1 Proposed Algorithm

```
***Inptut:*** $\{r_{ij}\}, \;\; thr, \;\; MaxIter$
$iter \leftarrow 0$
$q_j \leftarrow MOS_j \;\; j \in \mathcal{J}$
$b_i \leftarrow Avg_j(r_{ij} - q_j) \;\; i \in \mathcal{I}$
**while** $target$ **do**
    $q_j^{prev} \leftarrow q_j \;\; j \in \mathcal{J}$
    $e_{ij} \leftarrow r_{ij} - q_j \;\; j \in \mathcal{J} \;\; i \in \mathcal{I}$
    $s_i \leftarrow Std_j(e_{ij}) \;\; i \in \mathcal{I}$
    $\alpha_i \leftarrow \frac{s_i}{\sqrt{\frac{1}{|\mathcal{J}|}\left(\sum_j(-q_j^2+6*q_j+5)^2\right)}} \;\; i \in \mathcal{I}$
    $w_{ij} \leftarrow \frac{e^{-\alpha_i(-q_j^2+6*q_j+5)}}{\sum_i e^{-\alpha_i(-q_j^2+6*q_j+5)}} \;\; i \in \mathcal{I} \;\; j \in \mathcal{J}$
    $b_{ij} \leftarrow b_i \cdot \mathbb{1}_{[2,4]}(q_j) \;\; i \in \mathcal{I} \;\; j \in \mathcal{J}$
    $q_j \leftarrow \sum_i w_{ij} \cdot (r_{ij} - b_{ij}) \;\; j \in \mathcal{J}$
    $b_i \leftarrow Avg_j(u_{ij} - q_j)$
    $target \leftarrow (\|q - q^{prev}\| > thr \;\; and \;\; iter+1 \leq MaxIter)$
**end while**
***Ouptut:*** $\{q_j\}, \; \{\alpha_i\}, \;\; \{b_i\}$
```

TABLE I
Main parameters with the related definitions

Parameter	Definition
$\mathcal{I}$	Set of subjects
$\mathcal{J}$	Set of stimuli
r_{ij}	Rating of the subject i for stimulus j
q_j	Ground truth quality of the stimulus j
b_i	Bias factor of the subject i
b_{ij}	Bias of the subject i when rating stimulus j
α_i	Inconsistency factor of the subject i
σ_{ij}	Inconsistency of the subject i when rating stimulus j
q_j^n	Value of q_j after n iterations of Algorithm 1
b_i^n	Value of b_i after n iterations of Algorithm 1
α_i^n	Value of α_i after n iterations of Algorithm 1

On the other hand, the standard deviation s_i in Eq (5) can be estimated from the ratings gathered from the subject i as it follows:

$$\hat{s}_i = \left(\frac{\sum_{j \in \mathcal{J}} \left((r_{ij} - q_j) - \mu_i \right)^2}{|\mathcal{J}|} \right)^{\frac{1}{2}} \tag{6}$$

where $\mu_i = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} (r_{ij} - q_j)$.

Equating the value of s_i in Eq (5) to that of $\hat{s}_i$ in Eq (6), the value of the inconsistency factor α_i at the n-th iteration is computed as:

$$\alpha_i^n = \frac{\hat{s}_i^n}{\left(\sum_{j \in \mathcal{J}} \left(-(q_j^n)^2 + 6 * (q_j^n) - 5 \right)^2 \right)^{\frac{1}{2}}} \tag{7}$$

where $\hat{s}_i^n$ is the value of $\hat{s}_i$ at the n-th iteration, i.e. when q_j is substituted by q_j^n in Eq (5).

The bias factor b_i of each subject i at the n-th iteration is given by:

$$b_i^n = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \left(r_{ij} - q_j^n \right) \tag{8}$$

that is basically the average deviation of the ratings of the subject j from the ground truth qualities.

The manifestation of the bias of the subject i when rating the stimulus j can then be updated at the n-th iteration as it follows:

$$b_{ij}^n = b_i^n \cdot \mathbb{1}_{[2,4]}(q_j^n)$$

We now focus on the iterative step that updates the value of q_j. This iterative step in the Sureal software was defined as follows:

$$q_j^{n+1} = \sum_{i \in \mathcal{I}} \frac{(\sigma_i^n)^{-2}}{\sum_{i \in \mathcal{I}} (\sigma_i^n)^{-2}} \left(r_{ij} - b_i^n \right) \tag{9}$$

Thus, the contribution, expressed by $(r_{ij} - b_i^n)$, of the subject i to the determination of the ground truth quality of the stimulus j at the n-th iteration is weighted by:

$$w_i^n = \frac{(\sigma_i^n)^{-2}}{\sum_{i \in \mathcal{I}} (\sigma_i^n)^{-2}}. \tag{10}$$

In particular, the higher the inconsistency σ_i of a subject j, the lower is his/her contribution to the computation of the ground truth quality.

Although this weighting schema implemented in Sureal has shown effective performance in many applications, we believe that it suffers a crucial drawback. In fact, the weights defined in Eq (10) for each subject i do not depend on the characteristics of the stimulus under evaluation. As a consequence, if a subject has been particularly inconsistent when rating most of the stimuli but managed to accurately rate a few of them, his/her contribution to the determination of the ground truth quality of the stimuli that have been correctly rated would remain negligible. This is a serious issue if one considers, for instance, that inconsistent subjects at the center of the scale can accurately recognize and score very low or very high-quality stimuli.

We propose in this paper a different weighting scheme. In particular, at the n-th iteration, we weight the contribution of the subject i to the computation of the ground truth quality q_j of the stimuli j by:

$$w_{ij}^{n} = \frac{e^{-\sigma_{ij}^{n}}}{\sum_{i\in\mathcal{I}} e^{-\sigma_{ij}^{n}}} = \frac{e^{-\alpha_i\left(-(q_j^n)^2+6\cdot(q_j^n)-5\right)}}{\sum_{i\in\mathcal{I}} e^{-\alpha_i\left(-(q_j^n)^2+6\cdot(q_j^n)-5\right)}}. \quad (11)$$

For a fixed value of the quality q_j, it can be noticed that the value of the weight w_{ij} decreases as α_i increases. Therefore, similar to the weights in Eq (10) used in the Sureal software, if the subject i is particularly inconsistent, i.e., if α_i assumes a large value, his/her weights $w_{ij}\ \ j \in \mathcal{J}$ are in general smaller than those of a less inconsistent subject.

Unlike the weights in Eq (10), the proposed ones in Eq (11) depend on the quality of the stimulus under evaluation. This allows us to model, for instance, the fact that all subjects are usually accurate at the extremes of the quality scale, and thus none of the subjects is expected to perform particularly better than the others in those areas of the scale. In fact, when $q_j^n \to 1$ or $q_j^n \to 5$, the polynomial function $\left(-(q_j^n)^2 + 6 \cdot (q_j^n) - 5\right) \to 0$. Hence, the weight w_{ij} of any subject i for the determination of the ground truth quality of the stimulus j converges to $\frac{1}{|\mathcal{I}|}$. Therefore, by using the proposed weights, at the extreme of the quality scale, the ratings of all subjects tend to receive the same consideration (one over the total number of subjects) when computing the ground truth quality.

Using the weights proposed in Eq (11), we update the ground truth quality of each stimuli j at the iteration n+1 as it follows:

$$q_j^{n+1} = \sum_{i\in\mathcal{I}} w_{ij}^{n}\left(r_{ij} - b_{ij}^{n}\right). \quad (12)$$

The Algorithm 1 summarizes the iterative steps discussed above for the estimation of the parameter q_j, b_i and α_i. The algorithm takes, as input, a matrix $\{u_{ij}\}$ containing the rating of each subject i for each stimulus j. A threshold, denoted by thr, to monitor the convergence of the algorithm, as well as a maximum number of iterations denoted by $MaxIter$ are also required. The algorithm stops if: i) the convergence has been reached, i.e., the Euclidean distance between the arrays containing the values of the ground truth qualities estimated in two consecutive iterations is smaller than the required threshold; or ii) if the maximum number of iterations has been reached.

We recommend, while using the Algorithm 1, to set $thr = 10^{-8}$ as done in the Sureal software, and $MaxIter = 100$. In fact, we experimentally found that the algorithm converges in general before completing 100 iterations. In the very few cases where more than 100 iterations were necessary, after the iterations number 100, the Euclidean distance between the parameters estimated in two consecutive iterations was never larger that 10^{-4}.

To compute the 95% CI of the estimated ground truth quality q_j of each stimulus j, we observe from Eq (4) that the difference $(r_{ij} - b_{ij})$ is normally distributed with mean q_j and standard deviation σ_{ij}. Hence, the estimator $\sum_{i\in\mathcal{I}} w_{ij}\left(r_{ij} - b_{ij}\right)$ of q_j is a linear combination of normal random variables, thus it is normally distributed with mean q_j and standard deviation $\sqrt{\sum_{i\in\mathcal{I}} w_{ij}^2 \sigma_{ij}^2}$. Therefore, the 95% CI of q_j can be expressed as:

$$CI_{q_j} = q_j \pm 1.96 \cdot \sqrt{\sum_{i\in\mathcal{I}} w_{ij}^2 \sigma_{ij}^2} \quad (13)$$

where 1.96 is the 95% percentile of a normal random variable with mean equal to 0 and standard deviation equal to 1.

V. Results

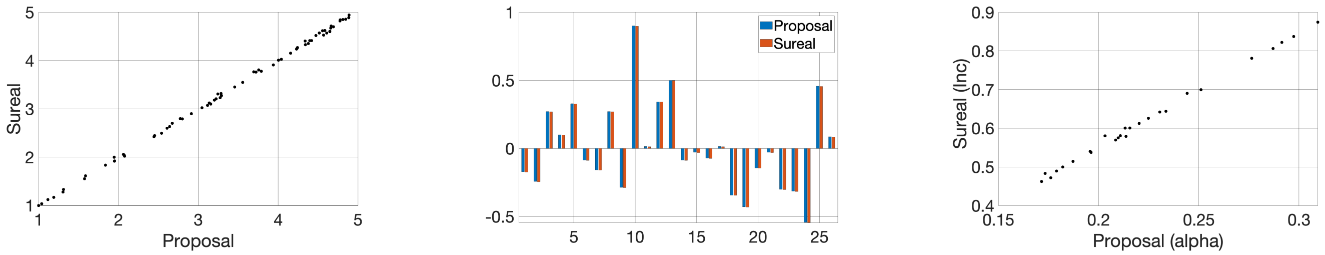
The Sureal software as well as our proposed iterative algorithm were run on four datasets, i.e., the VQEG-HD1, VQEG-HD3, VQEG-HD5 [13] and the Netflix public dataset [2]. We then compared the output of both approaches.

A. Computing more Realistic Confidence Intervals

The Sureal software,e as well as the Algorithm 1 proposed in this paper, were used on the four aforementioned datasets to estimate: i) the ground truth quality of each stimulus; ii) the CI of the estimated ground truth quality; iii) the inconsistency and the bias of each subject in these datasets.

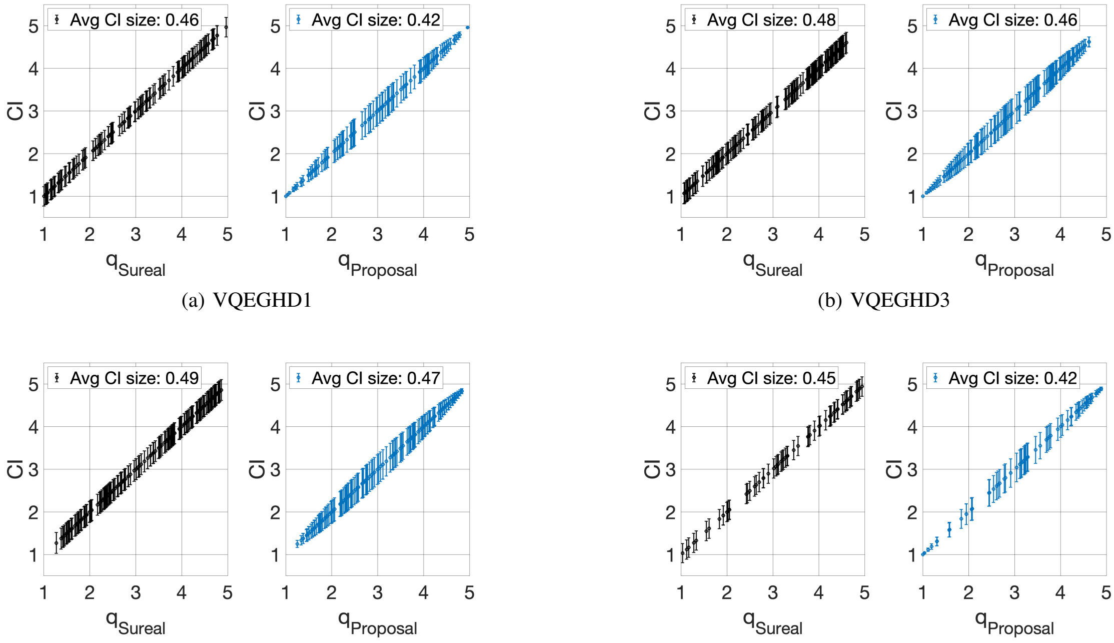
Figure 1 shows a comparison between the recovered ground through quality and the subjects' characteristics as computed by the two approaches on the Netflix Public dataset. For this specific experiment, we discuss only the Netflix public dataset because very similar results were obtained for the other three datasets.

The four datasets were collected during subjective experiments conducted in highly controlled environments under conditions specified by the ITU-T Recommendations, thus they are not particularly noisy. On this type of datasets, i.e., in absence of a significant quantity of noisy ratings, the Sureal software and the proposed Algorithm 1 yielded very similar estimations of the ground truth quality (see Figure 1a). Also, the overall bias estimated by both approaches for each subject was quite similar (see Figure 1b). Finally, the inconsistency estimated by the Sureal software correlated rather well to the parameter α that captures the subject inconsistency in our proposed integrated models (see Figure 1c).



(a) Quality (b) Bias (c) Inconsistency

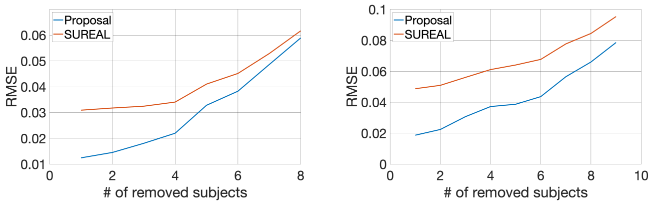
Fig. 1. Comparison of the recovered quality and the Subjects' characteristics computed by the Sureal software and the proposed integrated version.



(a) VQEGHD1 (b) VQEGHD3

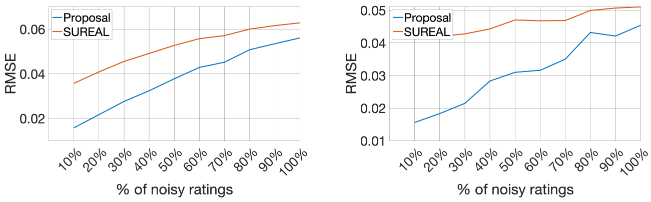
(c) VQEGHD5 (d) NETFLIX PUB

Fig. 2. Confidence Interval of the recovered quality for each stimulus in each dataset. q_{Sureal} and $q_{Proposal}$ are respectively the quality as recovered by the original SUREAL software and the iterative algorithm proposed in this paper.



(a) VQEGHD3 (b) NETFLIX PUB

Fig. 3. Robustness of the approaches to the input data reduction. The quantity of raw individual opinion score available to estimate the ground truth quality is progressively reduced by removing subjects from the dataset.

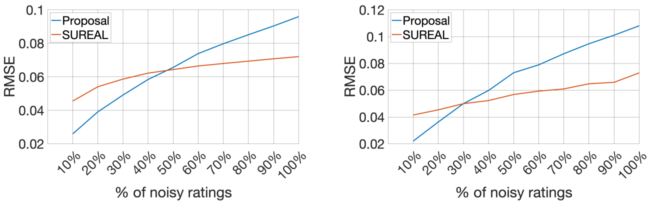


(a) VQEGHD3 (b) NETFLIX PUB

Fig. 4. Robustness of the approaches to synthetic noise added to the ratings of 10% of subjects in each dataset while assuming that even inconsistent subjects can accurately score the quality at the extremes of the scale. The experiment is repeated with 100 different seeds and the average RMSE is reported

Despite both approaches recovered similar ground truth qualities for all stimuli, the corresponding CIs are different as it can be noticed from Figure 2. In fact, the Sureal software computes CIs that have the same size independently of the ground truth quality of the stimulus. On the contrary, the CIs computed by the proposed approach have smaller size at the extreme of the quality scale as expected, since subjects are more reliable at the extremes of the scale. It can also be noted that our proposal yields ground quality estimations with smaller CIs on average (see the legends in Figure 2).

In conclusion, with not particularly noisy datasets, our proposed algorithm and Sureal estimate quite similar Subjects' characteristics and ground truth qualities. However, our approach computes CIs that are more realistic, since they model the high consistency of subjects at the extremes of the scale.



(a) VQEGHD3 (b) NETFLIX PUB

Fig. 5. Robustness of the approaches to synthetic noise added to the ratings of 10% of subjects in each dataset while disregarding the fact that subject are less inconsistent at the extreme of the quality scale. The experiment is repeated with 100 different seeds and the average RMSE is reported.

B. *Comparing the Robustness of the Approaches*

In this section, following the approach of [7], we consider the MOS computed from the ratings in each of the four considered dataset as the reference quality. We then modify each of the four datasets by removing the rating of a few subjects or adding some noise. Finally, we used the Sureal software as well as Algorithm 1 on the modified datasets to estimate the subjective quality and compare it to the reference quality. Due to space constraints, we included the results only for the VQEG-HD3 and the Netflix Public dataset. However, very similar results have been obtained on the other datasets.

First, we focus on the subject removal case whose result is reported in Figure 3. The main point of this experiment is to determine, as in [2], which approach can better estimate the reference quality if the subjective experiment was run with a smaller number of subjects than those that actually participated in the test. As it can be seen from Figure 3, the estimated quality by the proposed algorithm showed a lower Root Mean Square Error (RMSE) with respect to the reference quality in all the testing conditions. This suggests that the proposed algorithm is more robust to the reduction of available individual ratings when estimating the ground truth quality.

We now simulate the presence of a few inconsistent subjects during the test, making the dataset noisy. To this aim, for 10% of the subjects, a certain percentage of their ratings were changed into a random number between 1 and 5. Such a percentage is reported on the x-axis in Figure 4 and 5.

First, in Figure 4, we considered a sub-case in which an inconsistent subject is perfectly able to recognize very low and very high quality. Thus, for the 10% selected subjects, only the scores strictly greater than 1 and strictly smaller than 5 have been changed into a random integer between 1 and 5. Under these settings the proposed algorithm shows higher robustness to noise as it can be seen in Figure 4. In all testing conditions, the proposed approach estimated a quality with lower RMSE with respect to the reference one.

Figure 5 shows the result for the case in which scores are changed regardless of their original value. In this case, the proposed algorithm is better than the Sureal software when the percentage of noisy ratings increases up to about 30%. However, we believe that showing better robustness up to 30% is significant, because in practice subjects are typically accurate at the extreme of the scale, thus conditions such as these artificial experiments can be rarely encountered.

VI. Conclusion

In this paper we focused on the problem of how to accurately model subjects' behavior in subjective experiments and recover the ground truth subjective quality from noisy individual ratings. We assumed that the subject behavior can be reasonably captured by two main characteristics, i.e. the subject's bias and inconsistency. Unlike previous works that disregarded the fact that the manifestation of these characteristics varies with the quality of the stimulus under evaluation, we proposed two analytical formulations to express the link between these subject's characteristics and the quality, and we integrated them into an existing subject scoring model yielding a new, more robust, model. We also proposed an iterative algorithm to estimate the parameters of the new model. The proposed algorithm was compared to a state-of-the-art approach. The results showed that the proposed algorithm can compute more accurate confidence intervals for the recovered ground truth quality, and it is also more robust to synthetically added noise in several testing conditions.

References

[1] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, 2015.

[2] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *2017 Data Compression Conference (DCC)*, April 2017, pp. 52–61.

[3] Z. Li, C. G. Bampis, L. Janowski, and I. Katsavounidis, "A simple model for subject behavior in subjective experiments," *Electronic Imaging*, vol. 2020, no. 11, pp. 131–1, 2020.

[4] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *Proc. 3rd Intl. Workshop on Quality of Multimedia Experience (QoMEX)*, 2011, pp. 131–136.

[5] ITU-T Rec. BT.500, "Methodology for the subjective assessment of the quality of television pictures," Jan. 2012.

[6] ITU-T Rec. P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," Mar. 2016.

[7] J. Li, S. Ling, J. Wang, and P. Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3339–3347.

[8] Q. Xu, M. Yan, C. Huang, J. Xiong, Q. Huang, and Y. Yao, "Exploring outliers in crowdsourced ranking for QoE," in *Proc. 25th ACM Intl. Conf. on Multimedia*, 2017, pp. 1540–1548.

[9] S. Pezzulli, M. G. Martini, and N. Barman, "Estimation of quality scores from subjective tests: beyond subjects' MOS," *IEEE Transactions on Multimedia*, 2020.

[10] L. Fotio Tiotsop, A. Servetti, M. Barkowsky, and E. Masala, "Regularized maximum likelihood estimation of the subjective quality from noisy individual ratings," in *14th Intl. Conf. on Quality of Multimedia Experience (QoMEX)*, 2022.

[11] Netflix, "The sureal software," https://github.com/Netflix/sureal, May 2017.

[12] L. F. Tiotsop, T. Mizdos, M. Barkowsky, P. Pocta, A. Servetti, and E. Masala, "Mimicking individual media quality perception with neural network based artificial observers," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 1, pp. 1–25, 2022.

[13] VQEG, "Report on the validation of video quality models for high definition video content (v. 2.0)," http://bit.ly/2Z7GWDI, Jun. 2010.