POLITECNICO DI TORINO
Repository ISTITUZIONALE

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

Original
Modeling Subject Scoring Behaviors in Subjective Experiments Based on a Discrete Quality Scale / Tiotsop, L.F., Servetti, A., Barkowsky, M., Masala, E.. - In: IEEE TRANSACTIONS ON MULTIMEDIA. - ISSN 1520-9210. - STAMPA. - 26:(2024), pp. 8742-8757. [10.1109/tmm.2024.3382483]

Availability:
This version is available at: 11583/2989052 since: 2024-05-28T07:56:46Z

Publisher:
Institute of Electrical and Electronics Engineers

Published
DOI:10.1109/tmm.2024.3382483

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

26 November 2026

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

Lohic Fotio Tiotsop, Antonio Servetti, Marcus Barkowsky, *Member, IEEE*, and Enrico Masala, *Senior Member, IEEE*

***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.**

***Index Terms*—Subjective quality recovery, subject scoring model, discrete quality scale, subject bias weights, subject inconsistency.**

I. Introduction

SEVERAL quality scales and rating approaches for conducting subjective media quality assessments have been proposed and standardized [1], [2], [3], [4], [5]. In this work, we focus on modeling subject scoring behaviors in a subjective experiment run on a discrete quality scale. Discrete scales are widely used since they allow for an easier interpretation of the rating task.

The raw individual opinion scores are typically affected by noise caused by subject inconsistency and/or experimental context influence factors (IFs). Thus, approaches to analyzing raw individual ratings to identify subjects with peculiar behaviors and to mitigate the effects of noise sources have been investigated [6], [7], [8], [9], [10].

In several related studies [7], [8], [11], [12], the authors argue that the subject behavior can be reasonably captured by two main characteristics, i.e., subject bias and subject inconsistency. Subject bias is defined as the systematic tendency of a subject to assign lower (negative bias) or greater (positive bias) quality scores than the actual quality scores. For instance, a viewer with low visual acuity is likely to be positively biased and vice versa. Subject inconsistency, instead, captures the ability of a subject to provide accurate ratings and to repeat them if asked to rate the same stimulus several times.

In this work, we also rely on these two main characteristics, i.e., subject bias and inconsistency, since from our point of view, this approach possesses a feature that many other existing approaches lack. In particular, it establishes a clear and direct link between two well-defined aspects of subject behavior and the way subjects rate the stimuli. In many other approaches, instead, "peculiarity" indices are introduced that are shown to be accurate in measuring how peculiar a given subject is, but these indices cannot be directly associated with any specific aspects of the subject behavior, making it difficult to interpret the noise sources.

This paper contributes to advancing the state of the art in three ways.

1) We propose an approach called regularized maximum likelihood estimation (RMLE) of subjective quality from noisy individual ratings.
2) A novel probabilistic model to explain the choices of a subject in a subjective test run on a discrete quality scale is proposed.
3) The proposed probabilistic model is not assumed a priori, as in previous works. Rather, our model is derived from the proof of a theorem. This yields an approach with a stronger theoretical foundation.

This work significantly extends our previous one [13] in which we introduced only the RMLE approach. Here, we extend this previous work by better motivating the RMLE approach and by proposing a novel subject scoring model.

In the proposed RMLE approach, we define and estimate the quality of a given stimulus by considering the contribution of each of the opinion scores that can be chosen on the discrete quality scale. This allows us to model the subject behavior by directly investigating how the subject interacts with each opinion score on the quality scale. In fact, we defined the total

Manuscript received 17 December 2022; revised 29 June 2023, 24 October 2023, and 28 January 2024; accepted 15 March 2024. Date of publication 27 March 2024; date of current version 21 August 2024. This work was supported by PIC4SeR (http://pic4ser.polito.it). Some of the computational resources used were provided by HPC@POLITO (http://www.hpc.polito.it). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yong Luo. *(Corresponding author: Lohic Fotio Tiotsop.)*

Lohic Fotio Tiotsop, Antonio Servetti, and Enrico Masala are with the Control and Computer Engineering Department, Politecnico di Torino, 10129 Torino, Italy (e-mail: lohic.fotiotiotsop@polito.it; antonio.servetti@polito.it; enrico.masala@polito.it).

Marcus Barkowsky is with the Deggendorf Institute of Technology, University of Applied Sciences, 94469 Deggendorf, Germany (e-mail: marcus.barkowsky@th-deg.de).

Digital Object Identifier 10.1109/TMM.2024.3382483

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

attractiveness of each opinion score on the quality scale and proposed a probabilistic scoring model that provides an analytical formulation of the choice probabilities of a given subject.

We conducted several numerical experiments to validate the effectiveness of our proposals. The RMLE approach showed greater robustness to noise in individual opinion scores than did the other four state-of-the-art approaches. We also showed that the proposed subject scoring model allows automatic identification of peculiar subject behaviors not directly observable from the output of other approaches. Finally, unlike previous approaches in which subject behavior is modeled with a single bias and inconsistency value for all stimuli regardless of their quality, our proposed discrete-choice probability model can capture the lower inconsistency of the subject at the extremes of the quality scale.

The paper is organized as follows. In Section II, we discuss the related work. The RMLE approach is presented in Section III, followed by Section IV, where we derive the proposed subject scoring model. In Section V, the model parameters are estimated and interpreted. In Section VI, numerical experiments and results are presented; then, in Section VII, conclusions are drawn, and future research directions are discussed.

II. Related Work

Several authors have studied the factors that cause noise to affect raw individual ratings obtained from subjective experiments. Some of these factors include the following: experimental context influence factors [14], subject fatigue [15], and subject misunderstanding of the task that might yield, e.g., inverted ratings [16].

In an attempt to minimize the effect of noise sources on raw opinion scores, several approaches have been investigated to subjectively assess the quality of media content [17]. These methods include i) single stimulus-based approaches, where the subject views and rates the processed signal exclusively; ii) pair comparisons, involving subjects comparing the quality of stimulus A to that of stimulus B then indicating which one has superior quality; and iii) double stimulus-based approaches, which entail showing the source/reference content first then the processed version and asking the subject to rate the quality of the processed content relative to the reference content.

It has been empirically observed that pair comparison-based subjective experiments are likely to yield more accurate results than those of single stimulus-based experiments [18]. Unfortunately, obtaining a full matrix of comparisons is time demanding, making it difficult to conduct pair comparison-based experiments with as many stimuli as can be done when adopting a single stimulus-based approach. Additionally, while double stimulus-based methods yield quality scores with tighter confidence intervals, they require twice the time needed by single stimulus approaches. Thus, methods aiming for higher accuracy in raw ratings impose constraints on the maximum number of stimuli that can be rated.

Beyond the practical limitations imposed by the approaches that are likely to guarantee greater accuracy of the opinion scores, there are noise sources that are not under the control of the researcher when running the subjective test. For example, a subject rating for a specific video sequence may be significantly influenced by their personal preferences for the content, such as liking or disliking the scene being presented [19]. Therefore, several authors have proposed approaches to model subject behavior in subjective tests to "clean" the raw opinion scores from noise effects, regardless of the method used to collect the scores [6], [7], [8], [9], [10], [20], [21].

The most basic approach for addressing noise when subjectively measuring quality is to calculate the mean opinion score (MOS) by averaging individual ratings. However, the mean operator is highly sensitive to outlier ratings, i.e., those from peculiar subjects. To address this MOS limitation, various approaches, including the algorithms recommended by ITU-R BT.500 [6] and ITU-T P.913 [22], have been proposed for identifying and excluding peculiar subjects before calculating the MOS.

However, the subject exclusion-based approach is perceived by several authors [8], [9] as an approach that throws away more data than should be discarded. In fact, it is very unlikely that a subject wrongly evaluated all the stimuli they were asked to rate. Therefore, by removing the subject from the dataset, one can lose some reliable ratings. Recently, a few researchers have proposed relying on advanced statistical methods to measure how peculiar a subject is and to estimate the subjective quality of the stimuli without excluding subjects from the dataset.

In [10], the authors proposed a generalized linear model-based approach to estimate the actual subjective quality of the stimuli from noisy individual opinion scores. The authors in [9] argued that, when rating media quality is considered, any subject has a certain probability of providing an inaccurate score. This probability is considered a measure of subject inaccuracy. The subject rating is then assumed to be sampled from a mixture of two discrete probability distributions. The first model considers accurate ratings, while the second captures unexpected/unreliable ratings. The authors then proposed an approach to estimate the subject inaccuracy and to recover the actual subjective quality. Although the probability value, which is defined as a measure of subject inaccuracy, is shown to be effective, it is not directly linked to particular characteristics of the subject behavior. This lack of a direct relationship makes it difficult to interpret the source of the noise observed in the data.

Conversely, several authors have embraced methods that enhance the interpretability of model results by establishing a direct connection between model parameters and well-defined subject characteristics. The authors in [7], [8], [11] assumed that each raw rating of a subject derives from a normal random variable. The mean of such a normal random variable depends on the actual quality of the stimulus under evaluation and on the subject bias, while the variance is determined by the subject inconsistency and the complexity of the stimulus.

In this paper, we adopt a similar perspective as in [7], [8], [11]; i.e., we assume that the subject behavior can be reasonably modeled by bias and inconsistency concepts. In these previous papers, subject bias is defined as a single real number. A positive (negative) number indicates a systematic tendency of the subject to choose high (low) opinion scores on the quality scale. Unfortunately, this approach to defining bias does not take into

account the fact that a subject might also have a systematic tendency to choose opinion scores that are significantly far apart on the quality scale. Instead, we define subject bias by a vector of weights. Each weight indicates the subject tendency to prefer each opinion score over the others. This enables our approach, as shown in Section VI, to highlight behaviors due to positional bias, e.g., ternary and bimodal annotators, that cannot be identified by previous approaches. Moreover, our approach introduces a per-stimulus measure of inconsistency, while in previous approaches only the overall inconsistency of a subject across a whole dataset can be calculated. Our approach makes it possible to automatically identify specific stimuli for which the ratings of a given subject might be questionable. Finally, unlike previous authors who used bias and inconsistency as the main parameters of a probabilistic scoring model which they assumed a priori, our proposed scoring model is mathematically derived, thus yielding an approach with a stronger theoretical foundation.

We note that our proposed scoring model allows us to estimate the probability that the perceptual quality of a given content appeals to a subject with certain characteristics. In this respect, our work resembles classical recommender systems [23]. With the advent of deep learning, significant progress has been made toward the design of effective recommender systems [24]. However, the problem we are considering remains difficult to address with new deep learning-based recommender systems given that the subjectively annotated datasets in media quality assessments generally have a limited size, which prevents the effective use of deep learning. Our work does consider the question of data denoising, i.e., ground truth quality recovery, which, from our point of view, goes beyond the typical scope of a recommender system.

III. The Proposed RMLE Approach

In this section, we describe our proposed RMLE approach for estimating subjective quality from noisy individual ratings. First, we introduce the notation used in this paper. Then, we motivate our proposal. Finally, the steps for obtaining the estimated subjective quality are summarized.

A. Notation and Motivation

In this paper, we assume that subjective quality is evaluated by using a standard discrete quality scale with a finite number of available opinion scores. For instance, in the case of the five-point absolute category rating (ACR) scale, the subject is offered the following five opinion scores: *Bad*, *Poor*, *Fair*, *Good* and *Excellent*.

We introduce the following sets and quantities:

- $\mathcal{I}$: the set of stimuli that have been rated;
- $\mathcal{J}$: the set of subjects that rated the stimuli in $\mathcal{I}$;
- $\mathcal{K}$: the set of opinion scores available on the quality scale;
- $\mathcal{F}$: the set of influencing factors that might affect the ratings of a subject;
- r_i^j: the rating of the subject $j \in \mathcal{J}$ for the stimulus $i \in \mathcal{I}$;
- $\mathcal{R}$: all the ratings collected during the subjective test;
- n_{ik}: the number of subjects in $\mathcal{J}$ for whom the opinion score for stimulus $i \in \mathcal{I}$ is $k \in \mathcal{K}$.

TABLE I
Summary of the Notation

Parameter	Definition
$\mathcal{I}$	Set of stimuli
$\mathcal{J}$	Set of subjects
$\mathcal{K}$	Set of discrete opinion scores
$\mathcal{F}$	Set of all influence factors
r_i^j	Rating of the subject j for stimulus i
$\mathcal{R}$	Set of all ratings $r_i^j, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}$
n_{ik}	Number of subjects in $\mathcal{J}$ that chose the opinion score $k \in \mathcal{K}$ for stimulus $i \in \mathcal{I}$
Q_i	Ground truth quality of the stimulus i
w_{ik}	Weight of the opinion score $k \in \mathcal{K}$ in the determination of the ground truth quality of stimulus $i \in \mathcal{I}$
U_{ik}^j	Total attractiveness of the opinion score $k \in \mathcal{K}$ for the subject $j \in \mathcal{J}$ when rating stimulus $i \in \mathcal{I}$
μ_k^j	Bias weight of the subject $j \in \mathcal{J}$ towards the opinion score $k \in \mathcal{K}$
θ_{ik}^j	Stochastic effect of all the influence factors in $\mathcal{F}$ that might affect the choice of the opinion score $k \in \mathcal{K}$ by the subject $j \in \mathcal{J}$ when rating the stimulus $i \in \mathcal{I}$
β_j	Parameter modeling the effect of influence factors on the inconsistency of subject $j \in \mathcal{J}$
p_{ik}^j	Probability that the subject $j \in \mathcal{J}$ choose the opinion $k \in \mathcal{K}$ when rating the stimulus $i \in \mathcal{I}$
b_j	Overall bias of the subject $j \in \mathcal{J}$
σ_j	Overall inconsistency of the subject $j \in \mathcal{J}$

For the reader's convenience, we summarize the above notation and the definitions of the main parameters considered by the scoring model proposed in this paper in Table I.

The MOS of any stimulus $i \in \mathcal{I}$ can be expressed as:

$$MOS_i = \sum_{j \in \mathcal{J}} \frac{1}{|\mathcal{J}|} \cdot r_i^j = \sum_{k \in \mathcal{K}} \frac{n_{ik}}{|\mathcal{J}|} \cdot k \tag{1}$$

The first equality in (1) indicates the main issue with the MOS: when considering the MOS as the actual quality of a stimulus, all individual ratings have the same importance, i.e., each one of them is weighted with $\frac{1}{|\mathcal{J}|}$. This is problematic because it implies that potentially unreliable ratings have an equal impact on determining the quality of the stimulus as reliable ratings.

From the second equality, we can conclude that by weighting each opinion score $k \in \mathcal{K}$ with the fraction $\frac{n_{ik}}{|\mathcal{J}|}$, one obtains a subjective quality estimator (the MOS) that attributes the same importance to noisy and noiseless ratings.

Here, our main concern is to find a better way to weight the different opinion scores offered by the quality scale to obtain a more robust estimate of the quality. Specifically, our goal is to introduce a weighting scheme that assigns less importance to potentially noisy opinion scores while augmenting the weight of reliable opinion scores, thus enhancing their contribution to quality determination.

We therefore define the quality Q_i of the stimulus $i \in \mathcal{I}$ as follows:

$$Q_i = \sum_{k \in \mathcal{K}} w_{ik} \cdot k \tag{2}$$

in which the weights $w_{ik}, \ k \in \mathcal{K}$ are different from the fractions $\frac{n_{ik}}{|\mathcal{J}|}$ in (1) and are computed in the next section.

B. Mathematical Formulation of the RMLE Approach

Let us assimilate the weight w_{ik} to the unknown probability of choosing the opinion score $k \in \mathcal{K}$ when rating the stimulus $i \in \mathcal{I}$, i.e., the probabilities of the choices that are estimated from a noiseless dataset.

If the raw ratings in $\mathcal{R}$ were noiseless, then the probability of obtaining the observed ratings, also known as the likelihood function, would be expressed as:

$$L(w) = \prod_{i \in \mathcal{I}} \prod_{k \in \mathcal{K}} w_{ik}^{n_{ik}} \tag{3}$$

where w denotes a vector containing all the values $w_{ik}, \quad \forall i \in \mathcal{I}, \quad k \in \mathcal{K}$. The logarithm of $L(w)$, called the log-likelihood function, would be expressed as:

$$LL(w) = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} n_{ik} \cdot \log(w_{ik}). \tag{4}$$

The weights w_{ik} would then be obtained by finding the vector w that maximizes the log-likelihood function $LL(w)$.

Unfortunately, real datasets include noisy ratings. Without any additional input, the MLE framework would consider all ratings in the dataset reliable; hence, the obtained weights would not be robust to noise. In fact, maximizing the $LL(w)$ function would result in estimating each weight w_{ik} as equal to the fraction $\frac{n_{ik}}{|\mathcal{J}|}$, as in the noiseless case. However, we have already noted that this weighting scheme is not particularly robust to noisy ratings.

To incorporate the noisy nature of the dataset into the MLE framework, we introduce a regularization term as an additional input to the estimation process of the weights w_{ik}.

This term is designed to penalize what we refer to as "surprising events" for a particular stimulus, meaning opinion scores on the quality scale that seem to be chosen very infrequently when rating that stimulus. We believe that noisy ratings for a specific stimulus occur only sporadically, while accurate ratings tend to cluster around a set of opinion scores that are commonly selected.

To quantify how surprising the choice of opinion score $k \in \mathcal{K}$ is for stimulus $i \in \mathcal{I}$, we introduce the quantity C_{ik}, which is defined as follows:

$$C_{ik} = -log\left(\frac{n_{ik}}{|\mathcal{J}|}\right). \tag{5}$$

We observe that quantifying the surprise of an event based on the logarithm of its probability is a well-established approach in information theory [25]. The formula in (5) is therefore not considered a peculiarity of this work.

We propose the following regularization term:

$$R(w) = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} C_{ik} \cdot w_{ik} \tag{6}$$

to be subtracted from the log-likelihood function $LL(w)$ to obtain the optimization problem whose solution yields the weights $w_{ik} \ \forall i \in \mathcal{I} \ k \in \mathcal{K}$ that we are seeking. The weights $w_{ik} \ \forall i \in \mathcal{I} \ k \in \mathcal{K}$ are therefore obtained by solving the following optimization problem:

$$\max_{w} \left[LL(w) - \lambda \cdot R(w)\right]$$

$$\text{s.t.} \sum_{k \in \mathcal{K}} w_{ik} = 1 \ \forall i \in \mathcal{I} \tag{7}$$

where λ is the regularization coefficient whose calibration is discussed in Section VI.

Let us interpret the optimization problem in (7) to clarify how the proposed regularization term enables a noise-aware estimation of the weights w.

From the definition in (5), C_{ik} assumes large values if the opinion score k is not frequently selected, i. e., when n_{ik} is close to 0, the logarithm outputs a large number.

By subtracting the regularization term $R(w)$ from the log-likelihood function $LL(w)$, each value C_{ik} is considered by the optimization problem as a virtual cost to be paid by the objective function, depending on the value attributed to the weight w_{ik} of the opinion score k when estimating the quality of the stimulus i. Therefore, to maximize the objective function, for each stimulus i, opinion scores (those with large values of C_{ik}) which are not frequently chosen; hence, potentially noisy scores, receive less weight (lower value of w_{ik}) in the optimal solution so that the total virtual cost to be paid, expressed by the regularization term, is minimized.

Our decision to regularize the likelihood function rather than to utilize other established regularization techniques was primarily driven by empirical findings. The maximum likelihood estimation framework has demonstrated its effectiveness in the development of subjective quality recovery methods [8], [9], [11]. Our approach is to capitalize on this empirical evidence to create a more robust quality recovery method. We needed a new regularization term that aligns with the characteristics of our problem with a similar order of magnitude as our likelihood function to avoid unbalancing the objective function of the optimization problem in (7). As we could not find any existing regularization term meeting these criteria, we opted to design one from scratch.

While the RMLE approach is primarily designed for discrete quality scales, it can also be adapted for analyzing data on continuous scales. This may involve dividing the continuous scale into intervals and using the RMLE to weigh the ratings within each interval. However, evaluating this adaptation is beyond the scope of this paper.

As already mentioned, the weight w_{ik} can be interpreted as the ground truth probability that the opinion score $k \in \mathcal{K}$ is chosen when rating the quality of the stimulus $i \in \mathcal{I}$. Hence, the ground truth standard deviation of the opinion scores on the quality of the stimulus $i \in \mathcal{I}$ can be expressed as:

$$std_i = \sqrt{\left(\sum\nolimits_{k \in \mathcal{K}} k^2 w_{ik} - \left(\sum\nolimits_{k \in \mathcal{K}} k w_{ik}\right)^2\right)} \tag{8}$$

Therefore, the 95% confidence interval (CI) of the recovered quality of the stimulus $i \in \mathcal{I}$ can be estimated as follows:

$$CI_{Q_i} = Q_i \pm 1.96 \cdot \frac{std_i}{\sqrt{|\mathcal{J}|}} \tag{9}$$

IV. A Novel Subject Scoring Model

A. The Attractiveness of Opinion Score

In this paper, we consider that when assessing a stimulus, each subject mentally evaluates the attractiveness of each opinion score on the quality scale. This attractiveness is not directly observable, as some of its aspects are inherently subjective. However, we contend that this attractiveness is influenced by i) the ground truth quality of the stimulus, ii) the subject systematic preference for specific opinion scores over others, and iii) the subject level of inconsistency. Therefore, we introduce:

- U_{ik}^{j}, as the overall attractiveness attributed to the opinion score $k \in \mathcal{K}$ by the subject $j \in \mathcal{J}$ when asked to rate the stimulus $i \in \mathcal{I}$.

To propose an analytical expression of attractiveness U_{ik}^{j}, let us first recall some well-known peculiar subject behaviors observable in subjective tests run with a discrete quality scale. The authors in [16] identified the following eight main types of behaviors:

1) *Positively biased annotators:* subjects who tend to assign high opinion scores;
2) *Negatively biased annotators:* subjects who tend to assign low opinion scores;
3) *Unary annotators:* subjects who tend to assign the same opinion score;
4) *Binary annotators:* subjects who tend to assign only the lowest or the highest opinion score;
5) *Ternary annotators:* subjects who tend to assign the lowest, middle and highest opinion scores;
6) *Adversary annotators:* subjects who assign inverted ratings;
7) *Spammer annotators:* subjects whose data were randomly assigned;
8) *Competent annotators:* subjects who are very accurate.

Clearly, these behaviors are not mutually exclusive, as a subject may exhibit multiple behaviors during the same experiment. Nevertheless, they provide a solid foundation for designing subjective scoring models. Here, we present them to better introduce and to motivate our proposed scoring model. We describe them in more detail in Section VI.

When examining the first five behaviors in the list, a crucial observation becomes evident: When modeling the subject scoring behavior, it is essential to acknowledge that subjects may have inherent tendencies to favor specific opinion scores over others, regardless of the stimulus. For example, unary, binary, and ternary annotators prefer only one, two, or three opinion scores from those available on the quality scale. Positively biased subjects lean toward higher opinion scores, while negatively biased subjects tend to prefer lower scores.

To address the fact that a subject might systematically favor certain opinion scores at the expense of others, we introduce the following position bias weights:

- μ_{k}^{j}, i.e., the systematic tendency of subject $j \in \mathcal{J}$ to choose opinion score k rather than another that contributes to determining total attractiveness U_{ik}^{j}.

With the exception of "Spammer annotators," all other types of annotators are supposed to make their choices on the quality scale based on the ground truth subjective quality of the stimuli they are assessing. Consequently, we consider that the attractiveness U_{ik}^{j} also depends on the following:

- w_{ik}, i.e., the quality weight, computed by the RMLE approach, quantifies the importance of the opinion score k in determining the quality of the stimulus $i \in \mathcal{I}$.

Finally, to address subject inconsistency and to encompass possible behaviors resembling "spammer annotators", we assume that the attractiveness U_{ik}^{j} also depends on a random variable:

- θ_{ik}^{j} models the effect of all the influencing factors that might affect the choice of opinion score $k \in \mathcal{K}$ by subject $j \in \mathcal{J}$ when rating the stimulus $i \in \mathcal{I}$.

Summarizing the previous observations in a formula, we express the total attractiveness of the opinion score k for subject j when rating stimulus i as follows:

$$U_{ik}^{j} = w_{ik} + \mu_{k}^{j} + \theta_{ik}^{j}. \tag{10}$$

Let us denote by θ_{ikf}^{j} the random variable representing the relevance of the influence of the specific factor $f \in \mathcal{F}$. In practice, the number of IFs that might affect the choice of the subject is truly large. Moreover, these factors are not expected to have similar impacts on subject choice at all times. More precisely, we believe that, in a given context, an IF might be considered the most relevant. Hence, we assume that the subject choice on the quality scale is mainly determined by the IF with the greatest relevance.

Therefore, the stochastic term θ_{ik}^{j} of the attractiveness in (10) can be written as

$$\theta_{ik}^{j} = \max_{f \in \mathcal{F}} \theta_{ikf}^{j}. \tag{11}$$

The attractiveness of opinion score k for subject j when evaluating stimulus i can be reformulated as follows:

$$U_{ik}^{j} = w_{ik} + \mu_{k}^{j} + \max_{f \in \mathcal{F}} \theta_{ikf}^{j} \tag{12}$$

In practice, the complexity of IFs makes it difficult to hypothesize a specific probability distribution that any of the random variables $\theta_{ikf}^{j},\ f \in \mathcal{F}$ should follow. We therefore assume that such a distribution is unknown. In the next section, under a mild assumption about the shape of this unknown probability distribution, we derive the probability of a particular subject selecting a specific opinion score on the quality scale when evaluating a given stimulus. This derivation forms the basis for modeling the choices of each subject and, in turn, our proposed subject scoring model.

B. Deriving the Proposed Subject Scoring Model

Since the number of IFs that might affect the choices of a subject during a subjective test is truly large, it is reasonable to assume that the cardinality $|\mathcal{F}|$ of the set $\mathcal{F}$ of IFs tends to infinity.

Let us denote by p_{ik}^{j} the probability that subject $j \in \mathcal{J}$ chooses the opinion score $k \in \mathcal{K}$ when asked to rate the stimulus $i \in \mathcal{I}$. The expression of such a probability is the subject scoring model we are looking for.

To derive the probability p_{ik}^{j} and thus our proposed scoring model, we make a mild assumption on the shape of the unknown probability distribution of each random variable θ_{ikf}^{j} to model the effect of the IF $f \in \mathcal{F}$.

In particular, let us denote by $F_{ik}^{j}(x)$ the unknown cumulative probability distribution of any random variable $\theta_{ikf}^{j} \quad f \in \mathcal{F}$. We assume that two constants exist, $\alpha_{|\mathcal{F}|}$ and $\beta_j > 0$, such that $\forall i \in \mathcal{I}, \forall j \in \mathcal{J}, \forall k \in \mathcal{K}$:

$$\lim_{|\mathcal{F}|\to+\infty} F_{ik}^{j}\left(\frac{1}{\beta_j}x + \alpha_{|\mathcal{F}|}\right)^{|\mathcal{F}|} = \exp\left(-e^{-x}\right) \quad \forall x \in \mathbb{R}. \quad (13)$$

At first, the assumption presented in (13) may appear to be restrictive. However, this is not the case, as it holds true for numerous commonly used probability distributions, including the Gaussian, logistic, log-normal, exponential, Laplace, and Gumbel distributions, as shown in [26]. Therefore, by making this assumption, we are not substantially limiting the applicability of our proposed subject scoring model.

Let us note that the constant β is indexed by $j \in \mathcal{J}$. This finding is therefore subject specific. In particular, we later show that this inconsistency is related to the subject. The constant $\alpha_{|\mathcal{F}|}$ has no practical interpretation, as it is introduced only to implement a simple normalization trick that is useful for the proof of Theorem 1, yielding our proposed subject scoring model.

Theorem 1: Under the assumption in (13) and assuming that the random variables θ_{ikf}^{j} $f \in \mathcal{F}$ are independent, as the number of IFs tends to infinity, i.e., $|\mathcal{F}| \to +\infty$, the probability that subject j chooses the opinion score k when rating the stimulus i is:

$$p_{ik}^{j} = \frac{e^{\beta_j(w_{ik}+\mu_k^j)}}{\sum_{k\in\mathcal{K}} e^{\beta_j(w_{ik}+\mu_k^j)}}, \quad k \in \mathcal{K}, \quad j \in \mathcal{J}, \quad i \in \mathcal{I}. \quad (14)$$

Proof: The opinion score k of the quality scale can be chosen by the subject j when rating the quality of the stimulus i if and only if the subject identifies that opinion score as one of those having the greatest attractiveness.

Therefore, for a given stimulus $i \in \mathcal{I}$, by subtracting or adding the same constant to the attractiveness U_{ik}^{j} of each opinion score, the choice probabilities p_{ik}^{j} of the subject j remain unchanged. Hence, without loss of generality, the attractiveness of each opinion score k can be modified by subtracting from it the constant $\alpha_{|\mathcal{F}|}$ introduced in (13) and can be written as a function of $|\mathcal{F}|$ as follows:

$$U_{ik}^{j}(|\mathcal{F}|) = w_{ik} + \mu_k^j + \max_{f\in\mathcal{F}} \theta_{ikf}^{j} - \alpha_{|\mathcal{F}|} \quad (15)$$

The probability p_{ik}^{j} can then be expressed as follows:

$$p_{ik}^{j} = \mathbb{P}\left[U_{ik}^{j}(|\mathcal{F}|) = \max_{k\in\mathcal{K}} U_{ik}^{j}(|\mathcal{F}|)\right] \quad (16)$$

Applying the total probability theorem [27], one can write:

$$p_{ik}^{j} = \mathbb{P}\left[U_{ik}^{j}(|\mathcal{F}|) = \max_{k\in\mathcal{K}} U_{ik}^{j}(|\mathcal{F}|)\right]$$
$$= \int_{-\infty}^{+\infty} \mathbb{P}\left[\bigcap_{h\in\mathcal{K},h\neq k} U_{ih}^{j}(|\mathcal{F}|) \le x\right]$$
$$\left(\mathbb{P}\left[U_{ik}^{j}(|\mathcal{F}|) \le x\right]\right)' dx \quad (17)$$

Now, let us consider the probability $\mathbb{P}[\bigcap_{h\in\mathcal{K},h\neq k} U_{ih}^{j}(|\mathcal{F}|) \le x]$, which holds:

$$\lim_{|\mathcal{F}|\to\infty} \mathbb{P}\left[\bigcap_{h\in\mathcal{K},h\neq k} U_{ih}^{j}(|\mathcal{F}|) \le x\right]$$
$$= \lim_{|\mathcal{F}|\to\infty} \mathbb{P}\left[\bigcap_{h\in\mathcal{K},h\neq k} w_{ik} + \mu_k^j + \max_{f\in\mathcal{F}} \theta_{ikf}^{j} - \alpha_{|\mathcal{F}|} \le x\right]$$
$$= \lim_{|\mathcal{F}|\to\infty} \mathbb{P}\left[\bigcap_{h\in\mathcal{K},h\neq k} \max_{f\in\mathcal{F}} \theta_{ikf}^{j} \le x - w_{ik} - \mu_k^j + \alpha_{|\mathcal{F}|}\right]$$
$$= \lim_{|\mathcal{F}|\to\infty} \prod_{h\in\mathcal{K},h\neq k} \mathbb{P}\left[\max_{f\in\mathcal{F}} \theta_{ikf}^{j} \le x - w_{ik} - \mu_k^j + \alpha_{|\mathcal{F}|}\right] \quad (18)$$
$$= \lim_{|\mathcal{F}|\to\infty} \prod_{h\in\mathcal{K},h\neq k} F_{ih}^{j}((x - q_{ih} - \mu_h^j) + \alpha_{|\mathcal{F}|})^{|\mathcal{F}|} \quad (19)$$
$$= \prod_{h\in\mathcal{K},h\neq k} \exp(-e^{-\beta_j(x-q_{ih}-\mu_h^j)}) \quad (20)$$

where for the equality in (18) and (19), we exploited the independence of the random variables θ_{ikf}^{j}. To obtain (20), we exploit the assumption in (13).

From (18) and (19), it is not difficult to observe that $\mathbb{P}[U_{ik}^{j}(|\mathcal{F}|) \le x] = F_{ik}^{j}((x - w_{ik} - \mu_k^j) + \alpha_{|\mathcal{F}|})^{|\mathcal{F}|}$. Therefore, by using (13), the following limit holds:

$$\lim_{|\mathcal{F}|\to+\infty} \mathbb{P}\left[U_{ik}^{j}(|\mathcal{F}|) \le x\right] = \exp(-e^{-\beta_j(x-w_{ik}-\mu_k^j)}). \quad (21)$$

By inserting (21) and (20) in (17) and by defining $A_i^j = \sum_{k\in\mathcal{K}} e^{\beta_j(w_{ik}+\mu_k^j)}$ as $|\mathcal{F}| \to +\infty$, it follows that

$$p_{ik}^{j} = \mathbb{P}\left[U_{ik}^{j}(|\mathcal{F}|) = \max_{k\in\mathcal{K}} U_{ik}^{j}(|\mathcal{F}|)\right]$$
$$= \int_{-\infty}^{+\infty} \prod_{h\in\mathcal{K},h\neq k} \exp(-e^{-\beta_j(x-q_{ih}-\mu_h^j)})(\beta_j e^{-\beta_j(x-w_{ik}-\mu_k^j)}$$
$$\exp(-e^{-\beta_j(x-w_{ik}-\mu_k^j)}))dx$$
$$= \int_{-\infty}^{+\infty} \beta_j \exp(-A_i^j e^{-\beta_j x}) e^{-\beta_j(x-w_{ik}-\mu_k^j)} dx$$
$$= e^{\beta_j(w_{ik}+\mu_k^j)} \int_{-\infty}^{+\infty} \beta_j \exp(-A_i^j e^{-\beta_j x}) e^{-\beta_j x} dx$$
$$= \frac{e^{\beta_j(w_{ik}+\mu_k^j)}}{A_i^j} \int_{-\infty}^{+\infty} \beta_j A_i^j e^{-\beta_j x} \exp(-A_i^j e^{-\beta_j x}) dx$$
$$= \frac{e^{\beta_j(w_{ik}+\mu_k^j)}}{A_i^j}$$

$$= \frac{e^{\beta_j(w_{ik}+\mu_k^j)}}{\sum_{k\in\mathcal{K}} e^{\beta_j(w_{ik}+\mu_k^j)}}. \tag{22}$$

This proves the Theorem. ■

Therefore, motivated by (14), we argue in this paper that the rating r_i^j of the subject j for the stimulus i is a realization of a discrete random variable that can assume $|\mathcal{K}|$ possible values on the quality scale, i.e.,

$$r_i^j = \text{DRV}\left(p_{ik}^j = \frac{e^{\beta_j(w_{ik}+\mu_k^j)}}{\sum_{k\in\mathcal{K}} e^{\beta_j(w_{ik}+\mu_k^j)}},\ \ k \in \mathcal{K}\right) \tag{23}$$

where DRV represents a discrete random variable.

Equation (23) represents our proposed subject scoring model. In this model, the choice probability p_{ik}^j considers the impact of the ground truth quality of the stimulus through the weights w_{ik}, the subject bias via the weights μ_k^j, and the subject inconsistency through the parameter β_j, which characterizes the probability distributions of IFs.

V. Estimating and Interpreting the Parameters of the Model

The scoring model presented in (23) incorporates the bias weights μ and the parameters β, both of which need to be estimated for each subject. In this section, we outline our methodology for parameter estimation, and we provide insights into the appropriate interpretation. Additionally, we introduce several indices used by our scoring model to objectively identify peculiar behaviors from individual raw ratings.

A. Bias Weight Estimation

To estimate the bias weights μ_k^j for subject $j \in \mathcal{J}$ and each opinion score $k \in \mathcal{K}$, we represent the rating r_i^j of subject $j \in \mathcal{J}$ for stimulus $i \in \mathcal{I}$ with the array R_i^j containing $|\mathcal{K}|$ values defined as follows:

$$R_i^j(k) = \begin{cases} 1 & \text{if } k = r_i^j \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

For instance, in an experiment using the five-point ACR scale, the representation of the opinion score "Bad" is the array [1 0 0 0 0], while "Poor" is represented as [0 1 0 0 0], and so forth.

We then compute the deviation weights of the rating of subject j from the actual quality of stimulus i for opinion score k as follows:

$$\mu_{ik}^j = R_i^j(k) - w_{ik}. \tag{25}$$

The bias weight μ_k^j is estimated as follows:

$$\mu_k^j = \frac{\sum_{i\in\mathcal{I}} \mu_{ik}^j}{|\mathcal{I}|}. \tag{26}$$

In brief, μ_k^j is estimated as the average deviation between the importance that the subject j attributed to the opinion score k (expressed by $R_i^j(k)$) and the actual importance w_{ik} of that opinion as computed by the RMLE approach.

Let us note that the sum of the bias weights μ_k^j of a given subject j over all the possible opinion scores is equal to 0:

$$\sum_{k\in\mathcal{K}} \mu_k^j = 0 \ \ \forall j \in \mathcal{J} \tag{27}$$

since, by definition, $\sum_{k\in\mathcal{K}} R_i^j(k) = 1$ and $\sum_{k\in\mathcal{K}} w_{ik} = 1$.

Hence, (27) implies that for each subject and each stimulus, certain bias weights are positive, signifying a preference for the associated opinion scores, while others are negative, indicating a tendency to avoid selecting those particular opinion scores.

We define the overall bias of the subject $j \in \mathcal{J}$ as follows:

$$b_j = \sum_{k\in\mathcal{K}} k \cdot \mu_k^j. \tag{28}$$

We argue that, by using the values of the bias weights μ_k^j and the overall bias b_j derived from the raw individual ratings, it is possible to identify the behavioral characteristics of annotators, such as unary, binary, ternary, positively biased, and negatively biased ones. This is shown in more detail in Section VI.

Although the proposed model does not involve parameters that directly and explicitly capture the behavior of adversary annotators, by exploiting the bias weights in (25), we formulate an index that can also identify annotators with adversarial behavior.

More precisely, our idea is to first invert the ratings of all subjects on the quality scale, i.e., to transform all subjects into adversary annotators. By doing so, a subject that was originally an adversary annotator becomes a very accurate subject; hence, his or her ratings deviate less from the actual quality than those of all the other subjects who are now adversary annotators. In other words, the deviation weights μ_{ik}^j from the actual quality weights for that observer become small in absolute value, while those of the other subjects assume larger values in general.

Exploiting the observations made in the previous paragraph, we define the index I_{adv}^j, which establishes whether subject j should be considered an adversary annotator as follows:

$$I_{adv}^j = \left(\frac{1}{|\mathcal{I}||\mathcal{K}|} \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} |\bar{\mu}_{ik}^j|\right)^{-1} \tag{29}$$

where $\bar{\mu}_{ik}^j$ represents the deviation weights computed as in (25) after inverting the ratings of all the subjects in the dataset.

By taking the inverse of the average deviation from the actual quality, we expect that the index I_{adv}^j assumes large values for adversary annotators and a lower value for all the other subjects. This is verified in Section VI.

B. Parameter β and Subject Inconsistency

In this section, we discuss the link between subject inconsistency and parameter β. Subsequently, we derive an analytical expression for subject inconsistency within the framework of our proposed subject scoring model. Finally, we detail our methodology for estimating the parameter β for each subject.

Considering the scoring model in (14), for a given subject $j \in \mathcal{J}$, the following holds:

$$\lim_{\beta_j \to 0} p_{ik}^k = \lim_{\beta_j \to 0} \frac{e^{\beta_j (w_{ik} + \mu_k^j)}}{\sum_{k \in \mathcal{K}} e^{\beta_j (w_{ik} + \mu_k^j)}} = \frac{1}{|\mathcal{K}|}. \tag{30}$$

Hence, if the parameter β_j of the subject j is close to 0, this indicates that the subjects vote by choosing at random one opinion score among the $|\mathcal{K}|$ scores available on the quality scale. In other words, subjects whose β parameters assume low values are likely to be inconsistent. However, if β_j assumes a large value and the subject is not particularly biased toward a specific set of opinion scores, i.e., the bias weights μ_k^j are very close to 0, then the main factors determining the subject choice probabilities are the weights w_{ik}. Hence, subject j would be particularly consistent, as his or her choices are strongly based on the actual quality of the stimulus under evaluation. In any case, if a subject is not inconsistent, the β parameter does not tend to 0; therefore, the probability of choosing a certain opinion score (see (23)) is mainly determined by the total attractiveness (see (12)) of that opinion score for that subject.

For instance, for subjects who tend to provide lower (larger) scores that are accurate, i.e., correlated with the ground truth quality of the stimuli under evaluation, the bias weights corresponding to low (high) opinion scores on the quality scale are significantly greater than those of the others. As a consequence, from (12), we conclude that low (or high) opinion scores are more attractive for this type of subject. Therefore, since this type of subject is not inconsistent, i.e., β does not tend to 0, our scoring model in (23) simply indicates that they have a high probability of choosing a lower (or higher) opinion score when evaluating the quality of any stimulus. Thus, our scoring model can perfectly capture their tendency to provide lower (respectively higher) scores.

Let us note that a subject inconsistency in rating a particular stimulus is influenced not only by the stimulus quality but also potentially by the subject bias. For instance, subjects tend to exhibit lower inconsistency when rating stimuli that are of extremely low or high quality. Additionally, a subject with a strong positive bias may predominantly use the upper part of the quality scale, leading to reduced variance in their choices. Consequently, we should not consider that a single parameter, β_j, can account for all aspects of the inconsistency of subject j.

In fact, we define the inconsistency σ_{ij}^2 of subject $j \in \mathcal{J}$ on stimulus $i \in \mathcal{I}$ as a function of parameter β_j, quality weights w and subject bias weights μ. More precisely, we use the variance of the discrete probability distribution determined by the $|\mathcal{K}|$ probabilities $p_{ik}^j, \quad k \in \mathcal{K}$ as the measure of the inconsistency of the subject j on the stimulus i. The variance is computed as follows:

$$\sigma_{ij}^2(\beta, \mu, w) = \sum_{k \in \mathcal{K}} k^2 \cdot p_{ik}^j - \left(\sum_{k \in \mathcal{K}} k \cdot p_{ik}^j \right)^2 \tag{31}$$

We define the overall inconsistency of the subject $j \in \mathcal{J}$ as the average of the values $\sigma_{ij}^2(\beta, \mu, w), \ i \in \mathcal{I}$ over all the stimuli, i.e.,

$$\sigma_j^2(\beta, \mu, w) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sigma_{ij}^2(\beta, \mu, w) \tag{32}$$

To estimate the parameter β_j for each subject $j \in \mathcal{J}$, we rely on a least squares approach. In particular, we estimate the parameter β_j such that the theoretical overall inconsistency defined in (32) is as close as possible to the variance in the differences between the actual quality scores of the stimuli and the ratings of the subject. Therefore, we first compute:

$$s_j^2 = \mathrm{Var}(Q - R^j) \tag{33}$$

where Var represents the variance of a set of values and Q and R^j are two arrays containing the actual subjective quality of all stimuli computed by the RMLE approach and all the ratings of the subject, respectively j.

We then estimate β_j as the value that minimizes the function $l(\beta_j)$ defined as follows:

$$l(\beta_j) = \left(s_j^2 - \sigma_j^2(\beta_j, \mu, w) \right)^2 \tag{34}$$

Let us note that when estimating the parameter β_j, the quality weights w and the bias weights μ are already known, which is why the function l in (34) depends only on β_j.

VI. Results

In this section, we evaluate the effectiveness of our approach through a series of computational experiments.

A. Experimental Settings

For the experiments, we considered five datasets, namely, the VQEG-HD1, the VQEG-HD3, the VQEG-HD5 [28], the Netflix Public [8] and the ITS4S [29] datasets. Each VQEG dataset comprises ratings from 24 subjects for approximately 160 stimuli. In contrast, the Netflix public dataset contains ratings from 26 subjects for 70 processed video sequences, and the ITS4S dataset includes ratings provided by 27 subjects on the quality of 514 stimuli.

Since there were cases where for a given stimulus, no subject chose a specific opinion score k (for instance, when the quality of a given stimulus was particularly poor, no one chose *Excellent*), we need to compute C_{ik}, which involves the logarithm of zero (see (5)), when the number of subjects was $n_{ik} = 0$. In this case, we set the ratio $\frac{n_{ik}}{|\mathcal{J}|}$ to a very small real number $\epsilon = 10^{-16}$. We experimentally determined that there is no advantage to using a number smaller than that value.

To use our proposed RMLE approach in practice, the regularization weight λ must be estimated. To do so, we considered λ to be i) directly proportional to the number of stimuli rated by each subject to account for noise stemming from subject fatigue; ii) inversely proportional to the number of subjects, as larger subject pools provide more informative datasets; thus, the log-likelihood function $LL(w)$ should carry more weight than the regularization term $R(w)$; and iii) directly proportional to the number of possible opinion scores on the quality scale, which accounts for the expectation that subjects tend to vote more consistently

when they have fewer options on the quality scale, as seen in the greater reliability of subjects in pair comparison-based tests.

Therefore, in our experiments, the value of λ was set to

$$\lambda = \frac{1}{2} \cdot \frac{|\mathcal{I}||\mathcal{K}|}{|\mathcal{J}|} \tag{35}$$

The constant $\frac{1}{2}$ was experimentally determined to be a reasonable proportionality factor for ensuring that our RMLE approach is more robust to noise than are the other quality estimation approaches used in our experiments.

Let us note that this method for estimating λ may not be the optimal choice. Nonetheless, our results, obtained with this straightforward approach to estimating λ, are highly promising, as shown in the following sections.

We compared the proposed RMLE approach to four different state-of-the-art approaches to estimate the subjective quality from noisy raw individual ratings: the MOS, the ITU-T Rec BT.500, the approach proposed in [8], [11] as implemented in the publicly available Netflix SUREAL software [30] and a very recent quality recovery approach called ZREC proposed in [12].

For the sake of completeness, SUREAL software is built upon a subject scoring model in which the rating r_i^j provided by subject j for stimulus i follows a Gaussian distribution, i.e.,

$$r_i^j = q_i + b_j + N(0, \sigma_j) \tag{36}$$

where q_i is the actual quality of the stimulus i and b_j is the bias of the subject j. σ_j is the inconsistency of the subject j, and $N(0, \sigma_j)$ is a realization of a Gaussian random variable with a mean equal to 0 and a standard deviation equal to σ_j.

To estimate the parameters of the above model, the SUREAL software exploits an iterative algorithm called alternating projection (AP). The AP algorithm initializes the ground truth quality values with the MOS values. During each iteration, each subject bias and inconsistency are estimated with the mean and standard deviation of the differences between the subject ratings and the current ground truth quality values, respectively; the ground truth quality values are then updated by performing a weighted sum of the subject ratings after removing their bias. The weight or contribution of each subject in determining the ground truth quality of each stimulus is defined as the inverse of the square of their inconsistency. The iterative procedure continues as long as the Euclidean norm of the difference between the quality values calculated in two successive iterations is greater than 10^{-8}.

It is important to highlight that the AP algorithm utilized by the SUREAL software was endorsed by the ITU in 2021 as the most comprehensive method for subjective quality recovery (as per Section 12.6 of ITU-R P.913 [22]). As the latest standardized approach, the AP algorithm has recently served as the primary benchmark for evaluating newly proposed methods by various authors [12], [31]. Consequently, in the results section, we also consider SUREAL software as the primary benchmarking approach. In particular, in all our numerical experiments, we use the latest version of the SUREAL software, i.e., the one implementing the AP algorithm.

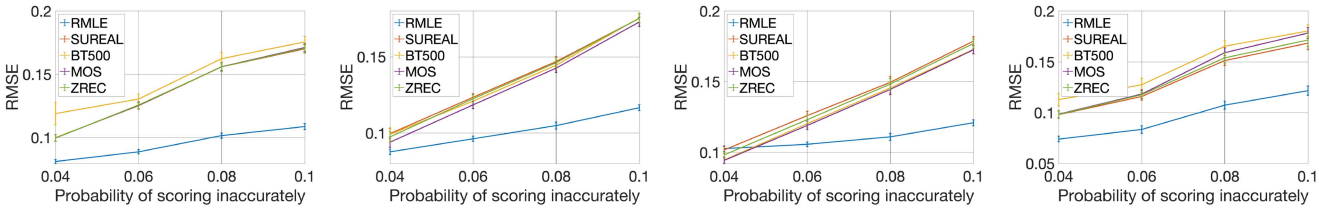
B. *Effectiveness of the Proposed RMLE Approach*

We compared the robustness of each of the considered methods to noise. In practice, following the approach in [9], we added synthetic noise to four datasets and evaluated, for each quality recovery approach, the root mean square error (RMSE) between the ground truth quality (the MOS obtained from the scores in the original dataset without any modifications) and the estimated quality scores from the noisy dataset. The primary objective was to show that the RMLE approach, when applied to noisy ratings, can yield a more accurate estimate of the ground truth quality than can the other methods under consideration. This form of comparison is a standard practice in the literature for assessing the effectiveness of subjective quality recovery methods [8], [9].

As in [8], [9], [11], the robustness of the MOS as a quality recovery method was also tested. In fact, as already mentioned, the MOS computed from the original dataset without adding synthetically simulated noise to the dataset was considered the ground truth or reference quality. Then, noise was added to the dataset. The MOS computed after adding the noise to the dataset was evaluated against the ground truth quality, i.e., the MOS obtained from the noiseless dataset. For instance, looking at Fig. 1(a), it can be said that when replacing 8% (0.08 on the x-axis) of the opinion scores of all subjects in the VQEG-HD1 dataset with a random integer number sampled between 1 and 5, the RMSE between the MOS values computed from the original dataset and those computed on the corrupted dataset is approximately 0.16 (y-axis). By repeating this process with different percentages of random ratings, we obtained the curves for the MOS shown in Figs. 1 and 2. These curves allow us to evaluate the robustness of the MOS to added noise.

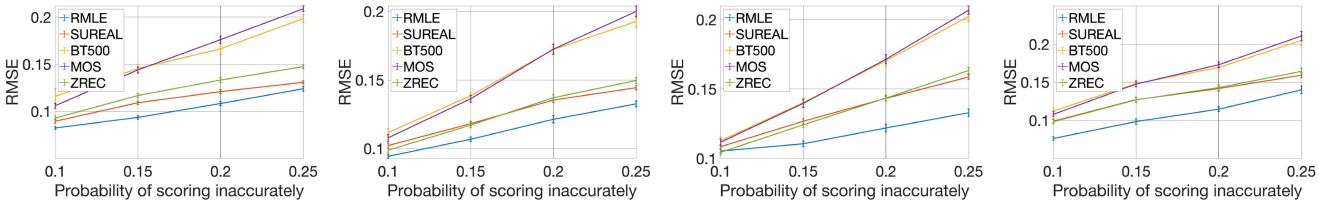
The noise was synthetically added to each dataset by using two different procedures: i) All subjects have a small probability of providing an inaccurate rating when scoring quality; thus, a fraction of the ratings of each subject corresponding to this probability was randomly selected to be replaced with a random integer number between 1 and 5; ii) The ratings of 50% of the subjects were kept unchanged, whereas the ratings of the other subjects were modified as described above. We believe that our first procedure reasonably simulates the introduction of noise, particularly in subjective experiments involving non-expert annotators or those conducted in uncontrolled settings such as crowdsourcing experiments. Our approach also applies well to experiments with a large number of stimuli where subject fatigue may influence the quality of ratings of all subjects. The second approach might be more suitable for simulating noise in subjective experiments that involve both highly competent annotators, such as experts, and naive subjects, who might occasionally provide inaccurate ratings due to the complexity of the stimuli they are assessing.

Figs. 1 and 2 present the obtained results when simulating the noise when using the first and the second procedures, respectively. For the first noise simulation procedure shown in Fig. 1, the less noisy condition consisted of assuming that all subjects provided inaccurate scores with a probability of 0.04, i.e., 4% of the opinion scores of all subjects were converted into random



(a) VQEG-HD1 (b) VQEG-HD3 (c) VQEG-HD5 (d) Netflix Public dataset

Fig. 1. Robustness of the different approaches to synthetically adding noise to individual ratings. All the subjects are assumed to have a certain probability (x-axis) of scoring inaccurately. The experiment was run with 30 different seeds. The average RMSE and the 95% confidence interval computed from the 30 RMSE values are shown. Let us note that the curves related to MOS, ZREC and SUREAL overlap in Fig. 1(a) .



(a) VQEG-HD1 (b) VQEG-HD3 (c) VQEG-HD5 (d) Netflix Public dataset

Fig. 2. Robustness of the different approaches to synthetically adding noise to individual ratings. Fifty percent of the subjects are assumed to have a certain probability (x-axis) of scoring inaccurately. The experiment was run with 30 different seeds. The average RMSE and the 95% confidence interval computed from the 30 RMSE values are shown.

integers between 1 and 5. The probability of all subjects incorrectly scoring a stimulus was then progressively increased to 0.1, i.e., 10% of the ratings of all subjects were modified. A similar interpretation holds for Fig. 2, but in this case, the noise affects only 50% of the subjects. Additionally, higher probabilities of providing inaccurate ratings were considered, up to 0.25.

When noise affects the score of all subjects (see Fig. 1), for almost all considered noise levels, the proposed RMLE approach recovers quality scores with the lowest RMSE with respect to the ground truth. In fact, the RMLE curve lies below all the others, showing that the robustness of our proposal to noise is the greatest among the other approaches. For the case in which only the ratings of half of the subjects in the original dataset are affected by noise (see Fig. 2), our RMLE approach showed the best performance, while the SUREAL software and ZREC outperformed the MOS and the ITU-T Rec. BT.500. In general, the SUREAL software and ZREC showed similar performances. This is not surprising because both approaches use the inverse of the inconsistency to weight the contribution of each subject to the determination of the ground truth quality. The main difference is that SUREAL analyzes the scores as gathered on the quality scale, while ZREC works on Z scores that are obtained by subtracting from the original ratings the MOS and dividing the result by the standard deviation of the ratings.

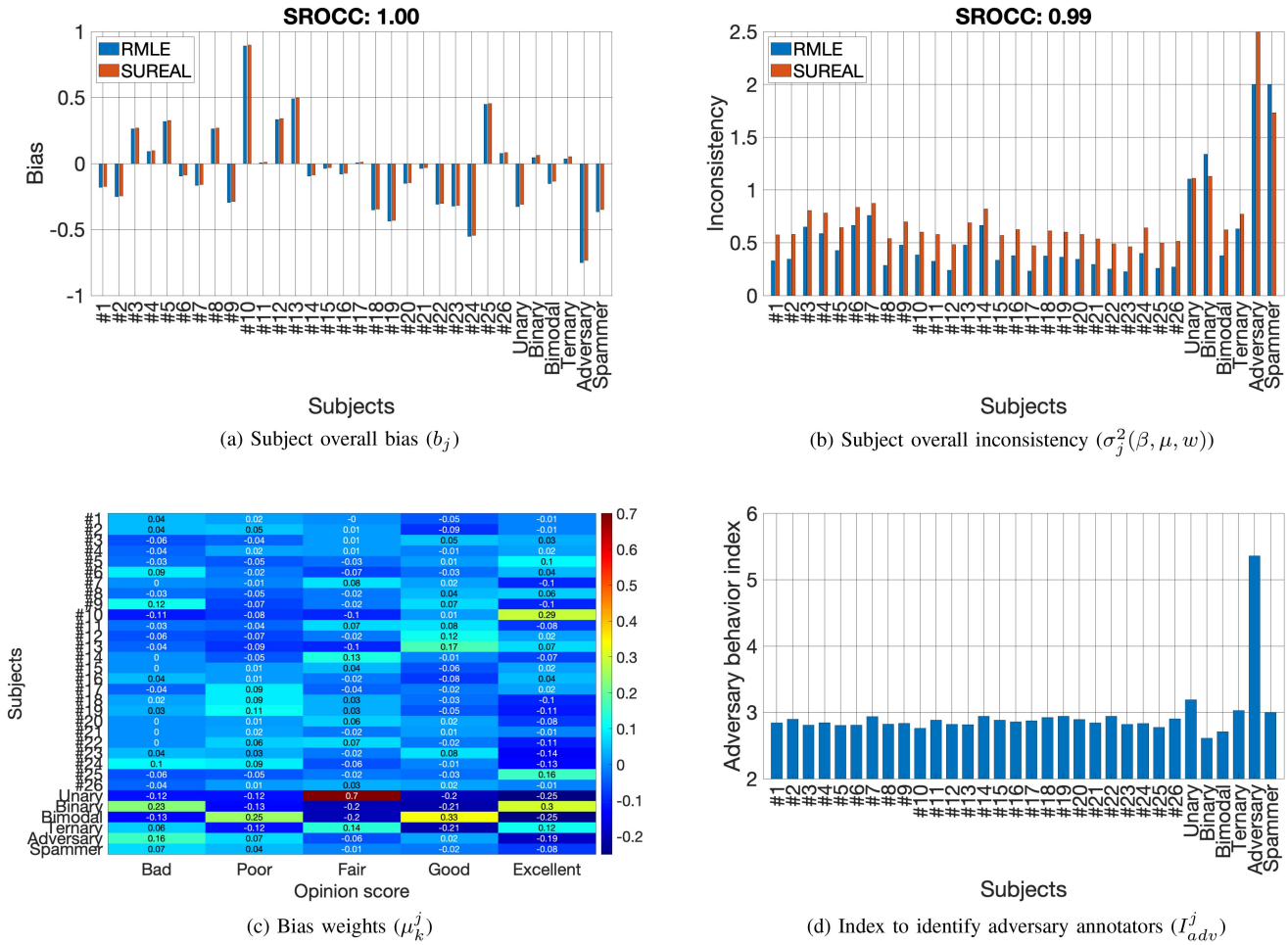
C. Identifying Peculiar Subject Behaviors

In this section, we assess the ability of our approach to identify annotators with peculiar scoring behaviors by comparing it to the model used in the SUREAL software and the ITU-T Rec BT.500. We simulate the ratings of annotators displaying unary, binary, ternary, adversary, and spammer behaviors, which are five of the eight behaviors outlined in Section IV. The other three behaviors (positively biased, negatively biased and competent annotator) can easily be recognized, as detailed later, from the data gathered during an actual subjective experiment. For this reason, in this experiment, we considered the ratings collected during an actual subjective test, i.e., the Netflix public dataset. We also simulated an additional peculiar scoring behavior typically observed in subjective tests, that we named "bimodal annotator", i.e., subjects that tend to avoid the extremes of the quality scale [2] and provide ratings normally distributed around *Poor* and *Good* depending on whether they judge the quality as not satisfying or satisfying.

For our analysis, we augmented the Netflix Public dataset, which originally included 26 real subjects, by introducing six virtual subjects. These virtual subjects were designed to simulate the behaviors of a unary, binary, bimodal, ternary, adversary, or spammer annotator. We subsequently applied the SUREAL software, ITU-T Rec BT.500, and our proposed approach to the integrated dataset. This method allowed us to assess the ability of the three approaches to accurately identify the simulated peculiar behaviors.

To simulate the ratings of the six virtual subjects, we first identified the most accurate annotator, i.e., the real subject with the lowest bias and inconsistency (subject #17 in Fig. 3(a) and (b)). We refer to this subject as *"gold subject"* in the following. By using the *gold subject* ratings, we generated the following six virtual subjects:

Unary annotator: These subjects tend to score as *Fair* for almost all stimuli; we randomly selected 90% of the ratings of



(a) Subject overall bias (b_j)

(b) Subject overall inconsistency ($\sigma_j^2(\beta, \mu, w)$)

(c) Bias weights (μ_k^j)

(d) Index to identify adversary annotators (I_{adv}^j)

Fig. 3. Comparison of the output of the SUREAL software (red bars in Fig. 3(a) and (b)) to that of our proposed approach on the Netflix public dataset integrated with the simulated ratings of peculiar subjects. In addition to the bias and inconsistency values computed by both approaches, our proposal also outputs the matrix of bias weights in Fig. 3(c) and the index in Fig. 3(d). Such additional output makes our approach more complete than is the SUREAL software, as it enables us to determine the source of the inconsistency of any particular subject.

the *gold subject* and set them equal to 3. The remaining 10% were kept unchanged.

Binary annotator: These subjects tend to choose *Bad* or *Excellent:* we randomly selected 90% of the ratings of the *gold subject*. Selected ratings less than or equal to 2 were set to 1, those equal to 3 were changed to either 1 or 5, and those greater than or equal to 4 were set to 5. The remaining 10% of the ratings were kept unchanged.

Bimodal annotator: These subjects feel that they are not experts and therefore tend to avoid the extremes of the quality scale, i.e., *Bad* and *Excellent*. Instead, they prefer to select *Poor* when judging the quality as not satisfying and *Good* otherwise. To simulate the ratings, we randomly chose 90% of the ratings of the *gold subject*. Any selected rating smaller than 3 was turned into 2, any rating greater than 3 was changed into 4, and any rating equal to 3 was changed into either 2 or 4 with equal probability. The remaining 10% of the ratings were kept unchanged.

Ternary annotator: These subjects tend not to express intermediate opinion scores, i.e., *Poor* and *Good*. To simulate the ratings, we again chose 90% of the ratings of the *gold subject* at random. A rating of 2 was turned into either 1 or 3 with equal probability, and a rating of 4 was turned into either 3 or 5. The remaining 10% of the ratings were kept unchanged.

Adversary annotator: We simply inverted all the ratings of the *gold subject* on the quality scale; e.g., when the *gold subject* rated as 1 (*Bad*), we turned it into 5 (*Excellent*) and vice versa, as well as 2 into 4 and vice versa.

Spammer annotator: The simulated ratings were obtained by substituting 90% of the rating of the *gold subject* with a random integer number uniformly sampled in the range from 1 to 5.

Fig. 3 presents the outcomes achieved by applying both the SUREAL software and our proposed approach to the Netflix Public dataset, which was augmented with the simulated ratings of peculiar virtual subjects as previously described. The ratings of these virtual subjects were generated by using 30 different random seeds. As such, except for the adversary annotator, where randomness was not a factor in the simulation, the statistics shown in Fig. 3 for the virtual subjects represent the average of 30 values.

First, we investigated the overall subject bias and inconsistency values computed by both approaches, as shown in Fig. 3(a) and (b), respectively. The results showed that, for all subjects,

both approaches estimated similar overall subject bias values. Although the overall inconsistency values computed by the two approaches are not equal in absolute terms, the Spearman rank order rank correlation coefficient (SROCC) is 0.99. In other words, given a pair of subjects, both approaches always agree on which one is the most inconsistent. Therefore, if one limits the analysis to overall subject bias and inconsistency, the two approaches can be considered quite similar.

Nevertheless, examining overall bias and inconsistency alone may not provide a comprehensive analysis of subject behavior. In fact, looking only at those values, it is not possible to distinguish between the simulated peculiar behaviors. For instance, from the results in Fig. 3(b), we surmise that both approaches identified the spammer and the adversary annotator as being very inconsistent. If the analysis is limited to the overall bias and inconsistency values, these two subjects are considered equivalent, and their ratings have a very low contribution to the determination of the ground truth quality since they are both considered very inconsistent. Instead, with more information that makes it possible to explain the source of inconsistency between the two subjects, the ratings of the adversary annotator can be easily recovered, and only the ratings of the spanner annotator can receive low consideration.

The previous example compares and illustrates the limits of approaches that rely on only an overall bias and inconsistency value, such as the SUREAL software, to analyze each subject behavior. However, our approach introduces the bias weights shown in Fig. 3(c) and the index in Fig. 3(d), which make it possible to distinguish between peculiar behaviors.

Fig. 3(c) shows the bias weights introduced in (26). Interpreting these weights makes it possible to identify the simulated unary, binary, bimodal, and ternary annotators. It is evident that these subjects assign substantial positive bias weights to the single, double, and triple opinion scores they are predisposed to choose. Hence, after observing the overall inconsistency values in Fig. 3(b), examining the bias weights in Fig. 3(c) may make it possible to determine whether the observed inconsistency derives from one of these four behaviors.

In fact, when looking at the output of the SUREAL software in Fig. 3(a) and in Fig. 3(b), one might erroneously conclude that the *Bimodal* and *Ternary* annotators are not peculiar subjects since their inconsistencies are comparable to those of several other real subjects. However, through the bias weights in Fig. 3(c), our approach clearly highlights the strong tendency of these two annotators to use only 2 and 3 opinion scores, respectively. It is clear that a unary or binary annotator might be more prejudicial than a bimodal or ternary annotator. However, from our point of view, it is still important to have approaches that can automatically highlight bimodal and ternary annotators. Bimodality generates, for example, slight inaccuracies in subject ratings at the extremes of the quality scale. In fact, a bimodal annotator would choose *Poor* as the opinion, with high probability, even when shown a stimulus for which *Bad* would be a better fit. Ternary annotators quantize the quality scale and thus implicitly use a different scale than the one proposed by the test designer.

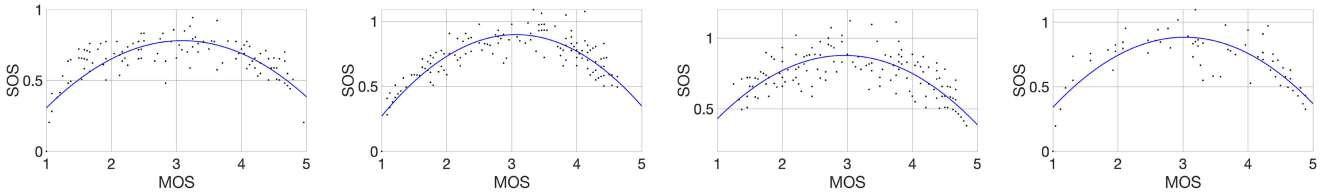
The index for identifying adversary annotators, defined in Section V, is shown in Fig. 3(d). As expected, corresponding to the simulated adversary annotator, the proposed index assumes a very large value compared to those of the other subjects. This shows that the proposed index can effectively determine whether an observed overall subject inconsistency derives from an adversarial behavior.

With regard to the ITU-T Rec BT.500, we calculated the number of times it managed to recognize and to reject each type of peculiar behavior during the 30 repetitions of the experiment. The unary, bimodal and ternary annotators were never rejected. The binary annotator was rejected 24 times out of the 30 repetitions, whereas the adversary and the spammer annotators were rejected 21 and 17 times, respectively. Hence, the ITU-T Rec BT.500 clearly showed lower performance than did our approach, which recognized all the simulated peculiar behaviors.

As mentioned in Section IV, when subjects rate stimuli, it is unlikely that they consistently adopt only one of the six behaviors simulated in this section. Their actual behavior may be a combination of several of these peculiar behaviors. For example, a subject might be competent at the beginning of a test but turn into a spammer annotator toward the end due to fatigue. Consequently, the bias weights of actual subjects in Fig. 3(c) may not be sufficient to entirely characterize a subject's behavior. Nevertheless, they do provide valuable insights in some cases that can be subject to further analysis.

For instance, from the overall bias values shown in Fig. 3(a), let us note that subject #10 is particularly positively biased. Considering the matrix of bias weights in Fig. 3(c), this bias can be explained by the high tendency of the subject to select *Excellent* as the opinion score. In fact, the bias weights of subject #10 for *Bad*, *Poor* and *Fair* are all negative, which indicates that the subject tends not to use the left part of the quality scale. In contrast, positive bias weights are observed for *Good* and, in particular, for *Excellent*, yielding an overall positive bias.

Fig. 3(b) shows that subjects #6, #7 and #14 had the highest overall inconsistency values. According to the bias weights in Fig. 3(c), subject #6 is slightly more attracted by the opinion scores at the extremes of the quality scale, i.e., *Bad* and *Excellent*, since positive bias weights are observed only in correspondence with these opinion scores. Therefore, one might hypothesize that the observed inconsistency can be partly explained by potential binary annotation behavior. For subject #7, the negative bias weights corresponding to *Excellent* show that the subject tends not to choose that opinion score. Unfortunately, this is not fully compensated for by the choice of the closest opinion score to *Excellent*, i.e., *Good*. Instead, *Fair* is chosen. This may explain the inconsistency observed. Finally, looking at the bias weights of subjects #14, a similar pattern can be observed as in the case of a unary annotator. In particular, the opinion score *Fair* exhibits a positive bias weight, while all the other bias weights are negative. This indicates that subject inconsistency might partially derive from a high tendency to choose *Fair*.



(a) VQEG-HD1 (b) VQEG-HD3 (c) VQEG-HD5 (d) Netflix Public dataset

Fig. 4. Standard deviation of the opinion score (SOS) (seen as a measure of inconsistency) as a function of the MOS. Each point corresponds to one stimulus in the corresponding dataset. The blue curves are a second-order polynomial least square fitting. As expected, subjects showed lower inconsistency at the extremes of the quality scale.

In summary, our proposed approach, which introduces bias weights and allows for analysis at the level of each single opinion score, offers a preliminary means of investigating the sources of overall inconsistency and bias in a subject. This approach goes beyond the output of the SUREAL software, which, by design, does not provide guidance on explaining the observed overall inconsistency so that potential issues in the subjective experiment can be addressed. When using our approach to analyze raw ratings, in addition to estimating the actual quality of the stimuli with the RMLE approach, to gain insights into the scoring behavior of each individual rater, we recommend to also conduct the analysis shown in Fig. 3 on the dataset being examined.

D. Modeling Subject Behavior at the Extremes of the Quality Scale

Empirical observations have shown that subjects tend to exhibit less inconsistency when rating stimuli of very low or very high quality. In [32], the authors argued that a second-order polynomial function is suitable for linking the MOS and the standard deviation of the opinion score (SOS), which is considered a measure of inconsistency between subjects. This is known within the media quality assessment community as the "SOS hypothesis".

In Fig. 4, we analyzed the link between the SOS and the MOS in four datasets. Each point in the figure represents a stimulus. The blue curves are obtained by performing a least square fitting of the MOS values to the SOS values by using a second-order polynomial function.

In accordance with the SOS hypothesis, the shape of the blue curves in Fig. 4 clearly illustrates that lower SOS values are prevalent at the extremes of the quality scale in all four subjective experiments. This implies that subjects tend to provide similar opinion scores when rating stimuli of very high or very low quality.

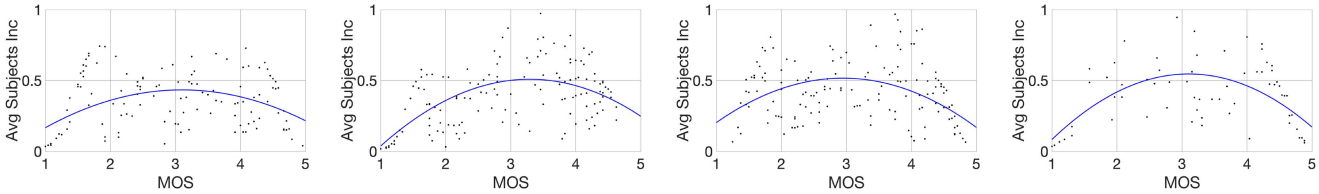
At the individual level, each subject is therefore expected to exhibit lower inconsistency at the extremes of the quality scale. An effective subject scoring model should capture this aspect of subject behavior. However, it is worth noting that, by design, the scoring model employed by the SUREAL software does not account for that aspect of the subject scoring behavior. In fact, the model in (36) assumes that the rating r_i^j of subject j for any stimulus i is affected by the inconsistency (σ_j). Therefore, according to the SUREAL scoring model, given two stimuli obtained from the same source, one with very low quality and the other with a quality score in the middle of the scale, a subject would show the same level of inconsistency when asked to rate these two stimuli. However, this finding contrasts with the observations made from the results shown in Fig. 4 and thus with the SOS hypothesis.

However, our proposed subject scoring model considers subject inconsistency at the level of the single opinion score. Indeed, the random variable θ_{ik}^j, as introduced in (10) to represent subject inconsistency, is defined for each opinion score $k \in \mathcal{K}$. This feature allows our approach to locally model subject inconsistency along the quality scale.

Fig. 5 shows the average inconsistency of each subject as defined in (31), as a function of the quality of the stimulus. On average, the proposed subject scoring model estimated lower inconsistency values for stimuli whose quality is in the range from 1 to 1.5 and from 4.5 to 5, compared to what happens in the middle of the quality scale. Hence, the proposed model captures the lower inconsistency of the subjects at the extremes of the quality scale.

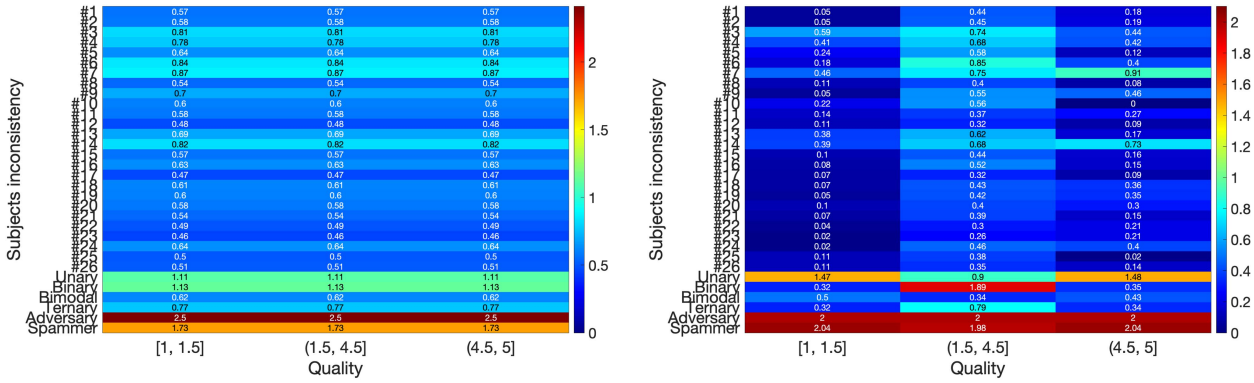
Fig. 6 shows the link between the quality of the stimulus and the estimated inconsistency for each individual subject in the Netflix public dataset, including the peculiar subjects simulated and discussed in the previous section, both for the SUREAL software (Fig. 6(a)) and for our proposed model (Fig. 6(b)). As already mentioned, the SUREAL software outputs a constant inconsistency over the whole quality scale (see Fig. 6(a)). This precludes the possibility of locally analyzing the accuracy of the subject on the quality scale. In Fig. 6(b), instead, one can observe the ability of our approach to predict lower inconsistency for each individual subject at the extremes of the quality scale. For instance, subject #7 is particularly inconsistent when rating stimuli of very high quality. This finding is consistent with the observations made on his or her behavior in the previous section; i.e., the subject tends not to use *Excellent* as an opinion score but does not choose *Good* as the direct alternative.

Interestingly, observing how the analysis in Fig. 6(b) brings to light the sections of the quality scale where the six simulated peculiar subjects are prone to exhibit higher levels of inconsistency. As anticipated, the adversary and spammer annotators display substantial inconsistency across the entire quality scale. Conversely, the unary annotator, by frequently selecting *Fair*, exhibited somewhat lower inconsistency in the middle of the quality scale. The binary annotator shows significant inconsistency in the middle of the quality scale.



(a) VQEG-HD1 (b) VQEG-HD3 (c) VQEG-HD5 (d) Netflix Public dataset

Fig. 5. Average inconsistency of the subjects, computed according to our proposed method, as function of the quality of the stimulus being rated. The blue curves are a second-order polynomial least square fitting. The proposed subject scoring model captures the fact that subjects are expected to rate the stimuli more consistently, with either very low or very high quality.



(a) SUREAL scoring model (b) Proposed scoring model

Fig. 6. Inconsistency of each subject as a function of the quality of the stimulus. SUREAL software computes an overall subject inconsistency value that does not depend on the quality of the stimulus under evaluation. However, our proposed model captures the lower inconsistency of the subjects at the extremes of the quality scale.

E. Assessing the Uncertainty on the Estimated Subjective Quality

In this section, we used our proposed RMLE approach and SUREAL software to estimate the quality of subjectively annotated datasets without adding synthetic noise, as described in Section VI-B. For the experiment, we considered five different datasets, i.e., the four datasets used in Section VI-B plus the ITS4S dataset [29]. Even without the addition of synthetic noise, the ratings in these datasets exhibit some level of noise due to inherent subject inconsistency. For example, in [33], the authors identified a processed video sequence (PVS) in the Netflix public dataset where a subject rated the quality as *Bad*, while the mode of the ratings for that PVS was *Excellent*. With respect to stimuli in the ITS4S dataset, the same authors identified a PVS where subjects uniformly chose opinion scores ranging from *Poor* to *Excellent*. In these cases, the MOS does not provide a suitable estimate of quality. Therefore, these examples also emphasize the importance of applying quality recovery approaches to datasets collected in highly controlled environments.

Following the approach of [11], [12], [34], we benchmarked the performance of our proposed quality recovery approach on real datasets by showing that its estimated subjective quality suffers lower uncertainty than that estimated by the SUREAL software. As in the aforementioned previous papers, here, the level of uncertainty in the estimated subjective quality is measured by the size of the CI. In particular, the larger the CI is, the greater the uncertainty in the estimated quality.

Table II summarizes the results of the experiment. After running the SUREAL software and the proposed software on each dataset, we first evaluated the similarity between the quality recovered by the two methods by computing the Pearson linear correlation coefficient (PLCC), the Spearman rank order correlation coefficient (SROCC), and the RMSE. The results in Table II show that, in general, both approaches recovered very similar subjective qualities. In fact, very large correlation coefficients (> 0.99) and low RMSE values (< 0.14) were observed. This highlights the consistency of our proposal with the prior art on data gathered in highly controlled environments.

We now delve into the results concerning the sizes of the confidence intervals (CIs) for the estimated subjective quality when using both methods. As shown in Table II, it is evident that, in general, the proposed RMLE method results in smaller CIs on average. Thus, although both approaches yield very similar estimates of subjective quality on the analyzed datasets, our method produces estimations that tend to be associated with lower uncertainty.

As in [12], in Table II, we provide the percentage by which the application of each of the two methods reduces the size of the CIs in comparison to what would be obtained from the MOS and the SOS of the raw data. For example, the application of our RMLE approach to the ITS4S dataset yielded subjective quality estimates with CIs whose size was reduced on average by 27%. The percentages in Table II can therefore be considered an indication of how much noise has been removed from the data by the applied quality recovery method. Higher values therefore

TABLE II
COMPARING THE PROPOSED RMLE APPROACH TO THE SUREAL SOFTWARE IN TERMS OF UNCERTAINTY ON THE ESTIMATED SUBJECTIVE QUALITY

Dataset	Recovered Quality Similarity			SUREAL CIs		RMLE CIs	
	PLCC	SROCC	RMSE	Avg CI Size	CI Reduction (%)	Avg CI Size	CI Reduction (%)
VQEG-HD1	1.00	1.00	0.07	0.46	6.93%	**0.42**	**15.58%**
VQEG-HD3	1.00	1.00	0.08	0.48	14.82%	**0.47**	**16.46%**
VQEG-HD5	1.00	1.00	0.11	0.49	14.97%	**0.48**	**16.10%**
NETFLIX PUBLIC	1.00	1.00	0.06	**0.45**	**13.88%**	0.48	9.86%
ITS4S	0.99	0.99	0.14	0.49	13.87%	**0.42**	**27.10%**

The recovered qualities by the two methods are first compared using the PLCC, the SROCC and the RMSE. The uncertainty is measured by the average of the sizes of the CIs of the estimated quality (avg CIs size). We also report by how much (in percentage) each method reduces on average the size of the CIS that can be computed from the raw data, that is, with the MOS and SOS of the ratings (CI reduction (%)).

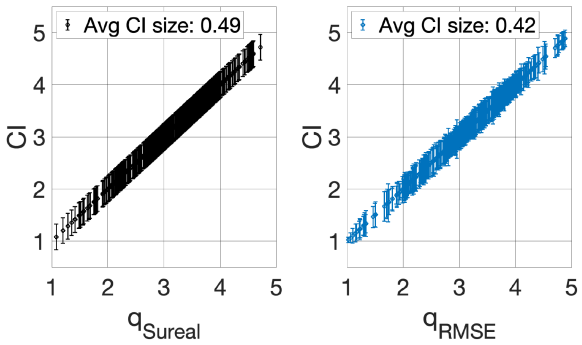


Fig. 7. CIs estimated by the SUREAL software (left) and the proposed RMLE method (right) as a function of the recovered quality of the stimuli in the ITS4S dataset. q_{SUREAL} and q_{RMLE} are the qualities recovered by the SUREAL software and the proposed RMLE approach, respectively.

indicate better performance: for 4 out of 5 datasets, our proposal did better than did the SUREAL software.

Fig. 7 shows the CIs estimated by the SUREAL software and our method as a function of the recovered quality on the ITS4S dataset. Our proposal yielded small CIs on average (see the legends), as already mentioned. An interesting observation that can be drawn from Fig. 7 is that, for extremely low-quality stimuli, the proposed RMLE approach calculates CIs that are smaller than those of other stimuli. This highlights the crucial aspect that, in the computation of CIs, our approach considers the high accuracy of subjects when rating stimuli of very low quality. In contrast, the SUREAL software computes CIs of uniform sizes regardless of the quality of the stimuli being assessed. Consequently, the CIs computed by the SUREAL software may be less realistic than those obtained from the proposed RMLE approach.

VII. CONCLUSION

In this paper, we focused on modeling subject behavior in subjective tests conducted on a discrete quality scale. An approach called regularized maximum likelihood estimation (RMLE) was first proposed to estimate the actual subjective quality from noisy individual ratings. The proposed RMLE approach combines the traditional MLE framework with a regularization term that is meant to attribute less weight to ratings that are potentially noisy in the dataset. The model then outputs the actual contribution/weight of each opinion score to the determination of the actual subjective quality of each stimulus.

An analytical expression of the overall attractiveness of each opinion score for each subject was proposed by using the quality weights estimated by the RMLE approach together with the introduction of subject inconsistency and bias weights. Under the reasonable assumption that the subject select the opinion score with the highest attractiveness, we analytically derived a novel subject scoring model that provides the probability of choosing each opinion score on the quality scale when a subject with specific characteristics is asked to rate a given stimulus.

Computational experiments showed that the proposed RMLE approach is more robust to noise in individual opinion scores than are four state-of-the-art alternative approaches. Moreover, the analysis of bias weights introduced by our proposed approach provides potential insights into the peculiar behavior underlying an observed subject inconsistency. Finally, the proposed subject scoring model effectively captures the typical quadratic link between subject inconsistency and stimulus quality.

Future work includes finding a better theoretical foundation for the estimation of the regularization coefficient. Furthermore, although our model allows for a more detailed analysis of the data, as evidenced by the computational results, it also involves a greater number of parameters than does the model implemented by the SUREAL software. We plan to develop a more parsimonious model in terms of the number of parameters.

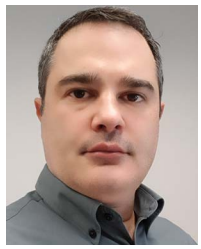
REFERENCES

[1] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *Proc. SPIE*, vol. 5150, pp. 573–582, 2003.
[2] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 1–14, Mar. 2011.
[3] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Syst.*, vol. 22, no. 2, pp. 213–227, Mar. 2016.
[4] ITU-T, "Subjective video quality assessment methods for multimedia applications," ITU-T Rec. P.910, Jul. 2022.
[5] ITU-T, "Subjective evaluation of media quality using a crowdsourcing approach," ITU-T PSTR-CROWDS, May 2018.
[6] ITU-T, "Methodology for the subjective assessment of the quality of television pictures," ITU-T Rec. BT.500, Oct. 2019.
[7] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Trans. Multimedia*, vol. 17, pp. 2210–2224, 2015.
[8] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Proc. Data Compression Conf.*, Snowbird, UT, USA, 2017, pp. 52–61.

[9] J. Li, S. Ling, J. Wang, and P. L. Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proc. 28th Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 3339–3347.
[10] S. Pezzulli, M. G. Martini, and N. Barman, "Estimation of quality scores from subjective tests-beyond subjects' MOS," *IEEE Trans. Multimedia*, vol. 23, pp. 2505–2519, 2021.
[11] Z. Li, C. G. Bampis, L. Janowski, and I. Katsavounidis, "A simple model for subject behavior in subjective experiments," *Proc. Int. Symp. Electron. Imag.*, vol. 32, no. 11, pp. 1–14, 2020.
[12] J. Zhu, A. Ak, P. L. Callet, S. Sethuraman, and K. Rahul, "ZREC: Robust recovery of mean and percentile opinion scores," in *Proc. IEEE Int. Conf. Image Process.* 2023, pp. 2630–2634.
[13] L. F. Tiotsop, A. Servetti, M. Barkowsky, and E. Masala, "Regularized maximum likelihood estimation of the subjective quality from noisy individual ratings," in *Proc. 14th Int. Conf. Qual. Multimedia Experience*, Lippstadt, Germany, 2022, pp. 1–4.
[14] U. Reiter et al., "Factors influencing quality of experience," in *Quality of Experience*, S. Möller and A. Raake, Eds. Cham, Switzerland: Springer, 2014, pp. 55–72.
[15] D. Schwarz, G. Lemaitre, M. Aramaki, and R. Kronland-Martinet, "Effects of test duration in subjective listening tests," in *Proc. Int. Comput. Music Conf.*, Utrecht, Netherlands: Hans Timmermans, 2016, pp. 515–519.
[16] Y. E. Kara, G. Genc, O. Aran, and L. Akarun, "Modeling annotator behaviors for crowd labeling," *Neurocomputing*, vol. 160, pp. 141–156, Jul. 2015.
[17] C. Deng, L. Ma, W. Lin, and K. N. Ngan, Eds., *Visual Signal Quality Assessment: Quality of Experience (QoE)*, 1st ed. Cham, Switzerland: Springer, 2015.
[18] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, Dec. 2012.
[19] P. Kortum and M. Sullivan, "The effect of content desirability on subjective video quality ratings," *Hum. Factors*, vol. 52, no. 1, pp. 105–118, Feb. 2010.
[20] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Online HodgeRank on random graphs for crowdsourceable QoE evaluation," *IEEE Trans. Multimedia*, vol. 16, pp. 373–386, 2014.
[21] Q. Xu et al., "Exploring Outliers in Crowdsourced Ranking for QoE," in *Proc. 25th Int. Conf. Multimedia*, Mountain View, CA, USA, 2017, pp. 1540–1548.
[22] ITU-T, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," ITU-T Rec. P.913, Jun. 2021.
[23] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," *Decis. Support Syst.*, vol. 74, pp. 12–32, Jun. 2015.
[24] R. Mu, "A survey of recommender systems based on deep learning," *IEEE Access*, vol. 6, pp. 69009–69022, 2018.
[25] F. M. Reza, *An Introduction to Information Theory*. New York, NY, USA: McGraw-Hill, 1994.
[26] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*. New York, NY, USA: Wiley, 1978.
[27] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, 2nd ed. Belmont: Athena Scientific, 2008.
[28] VQEG, "Report on the validation of video quality models for high definition video content (v. 2.0)." 2010. Accessed: Dec.16, 2022. [Online]. Available: https://bit.ly/2Z7GWDI
[29] L. F. Tiotsop, A. Servetti, and E. Masala, "Investigating prediction accuracy of full reference objective video quality measures through the ITS4S dataset," in *Proc. Int. Symp. Electron. Imag.*, 2020, vol. 32, pp. 1–6.
[30] Netflix, "The sureal software." 2022. Accessed: Dec. 16, 2022. [Online]. Available: https://github.com/Netflix/sureal
[31] A.-F. Perrin et al., "When is the cleaning of subjective data relevant to train UGC video quality metrics?," in *Proc. 29th Int. Conf. Image Process.*, Bordeaux, France, 2022, pp. 1466–1470.
[32] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!," in *Proc. 3rd Int. Workshop Qual. Multimedia Experience*. Mechelen, Belgium, 2011, pp. 131–136.
[33] L. Fotio Tiotsop et al., "Modeling and estimating the subjects' diversity of opinions in video quality assessment: A neural network based approach," *Multimedia Tools Appl.*, vol. 80, pp. 3469–3487, 2021.
[34] L. F. Tiotsop, A. Servetti, and E. Masala, "A scoring model considering the variability of subjects' characteristics in subjective experiments," in *Proc. IEEE 15th Int. Conf. Qual. Multimedia Experience*, Ghent, Belgium, 2023, pp. 1–6.



Lohic Fotio Tiotsop received the M.Sc. degree in mathematical engineering and the Ph.D. degree in control and computer engineering from the Politecnico di Torino, Turin, Italy, in 2017 and 2021, respectively. He is currently a Postdoctoral Researcher with the Control and Computer Engineering Department, Politecnico di Torino. His primary research interests include statistical models, machine learning, and deep learning-based approaches, specifically delving into media quality assessment.



Antonio Servetti has been an Assistant Professor with the Department of Control and Computer Engineering of the Politecnico di Torino, Turin, Italy, since 2007. His research interests include speech/audio processing, multimedia communications over wired and wireless packet networks, and real-time multimedia network protocols. With the advent of video and audio support in HTML5, his interests also include multimedia Web applications, WebRTC, Web Audio, and HTTP adaptive streaming.



Marcus Barkowsky (Member, IEEE) received the Dr.-Ing. degree from the University of Erlangen-Nuremberg, Erlangen, Germany, in 2009. He joined the University of Nantes, Nantes, France, in 2010, then in 2018 he obtained the professorship on interactive systems and Internet of Things at the Deggendorf Institute of Technology, Deggendorf, Germany, University of Applied Sciences, Germany. His activities range from designing 3-D interaction and measuring visual discomfort using psychometric measurements to computationally modeling spatial and temporal effects of the human perception.



Enrico Masala (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Torino, Italy, in 2004. He is currently an Associate Professor with the Politecnico di Torino. His main research interests include multimedia quality optimisation of communications over packet networks, with special attention to particular scenarios such as remote control applications, 3D video, cloud for multimedia. He is Co-chair of the JEG-Hybrid working group in the Video Quality Expert Group (VQEG) and participates in the activities of the Politecnico di Torino Interdepartmental Center for Service Robotics (PIC4SeR).

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement