CrowdRMLE: An Efficient Approach to Analyze Crowdsourcing Datasets in Image and Video Quality Assessment

CrowdRMLE: An Efficient Approach to Analyze Crowdsourcing Datasets in Image and Video Quality Assessment

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

***Abstract*—In image and video quality assessment, standardized approaches to analyze subjectively annotated datasets typically consider subject bias and inconsistency. While effective, these methods fail to capture the full complexity of how subjects interpret and use each opinion score on the quality scale. A recently proposed method, Regularized Maximum Likelihood Estimation (RMLE), partially overcomes this limitation by examining how a subject interacts with each opinion score. However, RMLE's high computational cost makes it computationally impractical for large-scale datasets, such as those derived from crowdsourcing experiments. This computational burden stems from solving a nonlinear optimization problem whose complexity grows with the dataset size. In this paper we propose CrowdRMLE, which formulates a different optimization problem and derives an analytical solution, thereby significantly reducing computational cost. Numerical experiments show that CrowdRMLE is not only far more efficient than RMLE but is also more accurate and robust to noise. Using CrowdRMLE we analyze multiple datasets, providing important insights into aspects that need improvement in the design of future methodologies for crowdsourcing tests for image and video quality assessment.**

***Index Terms*—Image and Video Quality Assessment, Crowdsourcing, Subjective Tests, Subject Inconsistency, Scale Quantization, Users' Quality-of-Experience.**

I. Introduction

Accurately measuring the subjectively perceived quality of image and video content is a topic of significant interest, as it contributes to the determination of the end users' overall satisfaction and engagement [1]. Numerous approaches have been proposed in the literature to collect opinion scores from end users [2]. Among these, crowdsourcing has become increasingly popular due to its ability to involve a large number of participants at a relatively low cost [3]–[6]. However, crowdsourcing introduces significant challenges since subjects perform the assessment in a poorly controlled environment, where potential distractions, varying viewing conditions, lack of supervision, misunderstanding of the task, and other external factors may affect both their perception of the stimuli and their use of the quality scale, introducing noise into the gathered data [7].

A substantial body of research has focused on improving the reliability of subjective datasets by modeling and mitigating scorer inaccuracies. In subjective quality assessment experiments, the collected ratings are affected by both unstructured and structured sources of noise. Unstructured noise arises from random rating events that are not driven by a coherent perceptual tendency, such as accidental clicks, momentary inattention, misunderstanding of the task, or software glitches. Since these perturbations are essentially random, they should ideally be mitigated when estimating the underlying perceptual quality of a stimulus.

In contrast, structured noise originates from systematic scoring behaviors exhibited by subsets of subjects. Examples include persistent positive or negative scoring tendencies, central scoring, or preferential use of extreme ratings. Unlike unstructured noise, these behaviors generate coherent and repeatable deviations from the rating patterns observed in the majority of subjects. While such behaviors may appear as noise when aggregating scores across subjects, they reflect structured scoring strategies rather than purely random errors. Therefore, besides mitigating unstructured noise, it is also important to identify and characterize these systematic behaviors when estimating the true perceptual distribution of opinion scores.

Most existing approaches focus on estimating and correcting the scorers' bias (systematic tendency to over- or underscore quality) and inconsistency (erratic or unreliable scoring behavior) [8]–[11]. While effective in many scenarios, these models primarily operate at the aggregate score level and do not explicitly characterize how participants interact with individual rating levels of the quality scale. As a result, structured scoring behaviors related to scale usage remain largely unexplored.

One important source of structured noise is the way subjects interact with the discrete quality scale. In this work, we refer to such behavior as *scale quantization*, defined as a subject's systematic tendency to overuse or underuse specific rating levels of the quality scale. Notably, scale quantization can occur without inducing a significant global bias or inconsistency, yet it alters the empirical distribution of the collected opinion scores and may reduce the discriminatory power of subjective experiments.

This paper addresses two fundamental and interrelated questions in subjective image and video quality assessment: i) How can we accurately estimate the true perceptual distribution of opinion scores for a given stimulus from a collected set of ratings that is affected by noise? ii) How can we explicitly model subject-specific scale quantization behavior in order to distinguish unstructured noise from structured scoring deviations when estimating the true distribution of opinion scores? These two problems were jointly investigated for the first time in [12], where the authors proposed the Regularized Maximum Likelihood Estimation (RMLE) framework. RMLE introduced a mechanism to capture rating-level preferences of raters while estimating the underlying perceptual distribution. In

L. Fotio Tiotsop, A. Servetti and E. Masala are with the Control and Computer Engineering Department, Politecnico di Torino, Torino, Italy. Email: lohic.fotiotiotsop@polito.it, antonio.servetti@polito.it, enrico.masala@polito.it

Manuscript received June 18, 2025.

particular, the framework simultaneously mitigates the effect of unstructured noise and models structured scoring behavior through subject-dependent parameters. However, RMLE is based on a large nonlinear optimization problem that is solved numerically, resulting in a computational burden that restricts its use to small-scale datasets.

To overcome this limitation, we propose CrowdRMLE, a computationally efficient framework based on a regularized likelihood formulation that admits a closed-form analytical solution. CrowdRMLE preserves the ability to model scale quantization while substantially reducing computational complexity, making it suitable for large-scale subjective datasets, particularly those collected via crowdsourcing. Although scale quantization effects are often more pronounced in crowdsourcing settings due to scorers' heterogeneity, note that the CrowdRMLE formulation is general and can be applied to both crowdsourcing and controlled laboratory experiments.

The main contributions of this paper can be summarized as follows:

1) We introduce CrowdRMLE, an efficient framework to estimate the true perceptual distribution of opinion scores for a given stimulus.
2) We define and compute the Scale Quantization Weights (SQWs) and the Scale Quantization Index (SQI) to model subject-level scale quantization behavior.

Extensive numerical experiments show that CrowdRMLE is very fast and achieves higher robustness to noise and superior accuracy in subjective quality estimation across both laboratory-controlled and large-scale crowdsourcing settings. Moreover, these experiments also highlight the usefulness of the proposed SQI and SQWs to identify and characterize peculiar subjects' scoring behaviors in subjective experiments.

The rest of the paper is organized as follows. Section II reviews the related work, Section III introduces CrowdRMLE and its analytical solution, Section IV evaluates its efficiency, accuracy, and robustness to noise. In Section V multiple datasets are analyzed with CrowdRMLE, and Section VI concludes the paper.

II. Related Work

Subject opinion scores are fundamental for advancing the research on multimedia signal processing. They intervene in several applications, including the comparison of multimedia processing systems [13], [14], the training and benchmarking of the performance of objective media quality measures [15]–[20] and the training of models to predict individual perception of media quality [21]–[25].

Collecting opinion scores in media quality assessment using the crowdsourcing paradigm is becoming increasingly widespread [26]–[28]. While crowdsourcing settings offer a huge advantage in terms of the quantity of data collected, diversity of the population, and overall cost, the uncontrolled experiment conditions put doubts on the trustworthiness of the collected data [7].

It is, therefore, fundamental to develop approaches that can identify and potentially remove scorers who provided inaccurate ratings or at least mitigate the impact of their noisy ratings on the estimation of the subjective quality. Existing approaches to address this issue can be divided into two main classes.

The first class includes methods that monitor subjects' accuracy directly during the test. These approaches often require subjects to answer a questionnaire designed to evaluate whether they are actually focusing on the rating task or they are just cheating [7], [29], [30]. Another interesting approach to identify peculiar raters involves measuring the average time spent on each stimulus [7].

The second class of approaches focuses on post-test processing, when the collected data are analyzed to identify peculiar behaviors. This paper belongs to this class of approaches.

Early post-test analysis methods focused only on identifying subsets of peculiar annotators to exclude all their ratings from the dataset. One of the first algorithms to implement this subject exclusion idea is described in the ITU-R BT.500 [8]. It recommends to exclude a subject based on the position of their opinion scores with respect to the 95% confidence interval of the Mean Opinion Score (MOS). The authors of [9] introduced CrowdMOS, that suggests to exclude subjects whose ratings poorly correlate to the MOS. In [10], an approach based on a chi-square statistical test was proposed to identify subjects that scored stimuli randomly so that they can be excluded. Finally, in [31] the fraction of triplets of stimuli for which a subject respected the transitivity property (if stimulus A has lower quality than B, and B has lower quality than C, then A has lower quality than C) was used to decide whether to exclude or retain the subject.

All these subject exclusion algorithms suffer from two major drawbacks: i) Excluding all the opinion scores of a subject may discard some reliable ratings, since, some subjects might score the quality accurately at the beginning of the test, but then start cheating in order to conclude the experiment as soon as possible. ii) Algorithms recommending subject exclusion often rely on thresholds whose values are based on empirical rather than theoretical evidence.

Additionally, the authors of [7] observed that subject-screening algorithms alone are insufficient to accurately remove noise from crowdsourcing datasets, concluding that alternative approaches are needed.

In response, recent literature has focused on statistical models that avoid subject exclusion altogether. These models instead use the entire dataset to derive a noiseless estimate of quality while simultaneously estimating parameters that inform how peculiar each subject is. For instance, the authors of [32] proposed a non-parametric statistical approach to measure the reliability of each individual opinion score. Scores can then be weighted based on their reliability to produce a noiseless quality estimate, while the average reliability of the scores of a subject provides a measure of how peculiar the subject is. Similarly, the authors of [33] proposed a statistical model that assumes a subject gives accurate opinion scores with probability $1-\epsilon$, while providing inaccurate scores with probability ϵ due to contextual factors or distractions. Lower values of ϵ indicate greater reliability, and an approach to estimate ϵ for each subject was presented.

These methods, which assess subject reliability and estimate quality without excluding subjects, have shown promising results. However, they lack interpretability. For instance, when a subject's estimated ϵ value is high, the model in [33] does not explain why the subject is unreliable. Similarly, the reliability measure in [32] provides no insight into the causes of unreliability.

An approach that combines theoretical rigor with interpretability involves modeling subjects using the concepts of bias and inconsistency. These characteristics have been extensively studied in recent years [11], [34]–[36]. In fact, the most recent algorithm recommended for subjective quality recovery relies on bias and inconsistency modeling as per clause 13.6 of ITU-T P.910 [37].

The possibility to easily interpret models based on bias and inconsistency allowed the research community to advance significantly. For example, the authors of [38] used simulations guided by bias and inconsistency to propose a measure of the overall quality of crowdsourcing datasets. In [39] it was shown how models based on bias and inconsistency can be used to improve the discriminatory power of a subjectively annotated dataset, while [40] used them in simulations to study the suitability of deep neural networks to model individual scoring behaviors.

Despite their usefulness and interpretability, bias and inconsistency do not capture all types of particularities in scoring behavior. For example, the work in [12] demonstrated through simulation that a bimodal subject, i.e., someone who genuinely scores the quality of the stimuli but overuses "Poor" when dissatisfied and "Good" when satisfied, is neither biased nor inconsistent. Thus, models based solely on bias and inconsistency fail to account for certain unbalanced uses of the quality scale.

To address this limitation, [12] proposed RMLE, an approach that computes not only a subject's bias and inconsistency but also a system of weights to determine whether the subject misused the quality scale. This additional information makes RMLE particularly suitable for analyzing crowdsourcing data, which is collected in uncontrolled environments with people from diverse cultures and backgrounds, that interpret scale items differently [41].

However, directly applying RMLE to crowdsourcing datasets is impractical due to its high computational demands. In the original RMLE paper, its effectiveness was demonstrated only on small datasets with no more than 160 stimuli and 26 subjects. In contrast, crowdsourcing datasets often exceed these dimensions by orders of magnitude, making RMLE unsuitable for such applications.

This paper proposes a different optimization problem than the one underlying RMLE. We prove that the newly proposed formulation admits an analytical solution that can be directly calculated, eliminating the need for intensive numerical optimization. This approach yields *CrowdRMLE*, which is shown to be more efficient, accurate, and robust to noise than RMLE and also highly scalable.

III. The Proposed Approach: CrowdRMLE

In this section we introduce the optimization problem underlying CrowdRMLE and derive the analytical expression of its optimal solution. We then use the derived solution to define the Scale Quantization Weights (SQWs) and the Scale Quantization Index (SQI).

A. The Optimization Problem Underlying CrowdRMLE

Let us introduce the following notation for a given subjectively annotated dataset to be analyzed:

- $\mathcal{I}$: set of stimuli;
- $\mathcal{J}$: set of subjects;
- $\mathcal{I}_j$: subset of stimuli rated by subject j;
- N_i: number of ratings collected for stimulus i;
- $\mathcal{K}$: used discrete quality scale;
- n_{ik}: number of times score $k \in \mathcal{K}$ was assigned to stimulus i;

For each stimulus $i \in \mathcal{I}$ and subject $j \in \mathcal{J}$, we define p^j_{ik} as the ratio between the number of times subject j assigned score k to stimulus i and the total number of times subject j rated stimulus i. Thus, p^j_{ik} represents the empirical choice frequency of score k by subject j for stimulus i. This definition naturally accommodates the most general setting in which the same subject may rate the same stimulus multiple times. In the special case of a single rating per subject–stimulus pair, p^j_{ik} reduces to a binary indicator.

The fundamental assumption underlying CrowdRMLE and, more generally, any subjective quality recovery method is that: for each stimulus in the dataset, there exists a latent distribution of reliable opinion scores representing the genuine perceptual evaluations of that stimulus. The observed ratings are regarded as realizations of this latent distribution, possibly corrupted by noise. It is also assumed that subjects evaluate the quality genuinely much more frequently than they score at random, so that meaningful ratings significantly dominate over purely random responses for each stimulus. Without these assumptions, the notion of recovering a true perceptual distribution would be ill-posed. We note however that these assumptions are not overly restrictive, as subjective quality assessment protocols, both in laboratory and crowdsourcing settings, typically include training phases, qualification tests, and attention checks designed to ensure that genuine evaluations prevail over random behavior.

To estimate the latent true distribution of the opinion score for each stimulus and quantify scale quantization behavior for each subject, we decompose subject-level choice frequencies as:

$$p^j_{ik} = f_{ik} + \nu^j_k + \epsilon^j_{ik}, \tag{1}$$

subject to

$$\sum_{k \in \mathcal{K}} f_{ik} = 1 \quad \forall i \in \mathcal{I}, \tag{2}$$

$$\sum_{k \in \mathcal{K}} \nu^j_k = 0 \quad \forall j \in \mathcal{J}. \tag{3}$$

Here, f_{ik} represents the latent true probability that score k is selected when rating stimulus i, and ν^j_k is a scale quantization

weight that captures the systematic tendency of the subject j to overuse ($\nu_k^j > 0$) or underuse ($\nu_k^j < 0$) opinion score k when assessing quality. Finally, ϵ_{ik}^j is a zero-mean stochastic error term accounting for residual subject-level variability that cannot be explained by systematic scale quantization (e.g., completely random scoring).

We focus mainly on modeling scale quantization behavior, as it has not yet been formally characterized and measured in image and video quality assessment. Other sources of scoring inaccuracy, captured by the stochastic error term ϵ_{ik}^j, are typically associated with general subject inconsistency and are already quantified using established inconsistency indices. Since extensive prior work [12], [35], [36], [42] has addressed inconsistency estimation, we do not introduce a new estimator for this aspect. Instead, standardized inconsistency measures are used as benchmarks in our experiments, where we also analyze their relationship with the explicitly modeled scale quantization behavior.

Although continuous quality scales are sometimes employed in image and video quality assessment, as it can be seen from the model in Eq. (1)-(3), the contribution of the present work is limited to discrete scales.

It is worth noting that the scale quantization weight ν_k^j in Eq. (1) should be interpreted as an *overall measure* of the tendency of the subject j to overuse or underuse the quality level k throughout the *whole experiment*. This formulation assumes that scale-usage behavior remains approximately stable across stimuli. This assumption is reasonable in many practical settings and it is consistent with the stimulus-independent definition of subject bias commonly adopted in subjective quality recovery [11], [34], [36]. Nevertheless, this assumption may not hold for the few subjects that, during the training phase, did not fully understand the difference between perceptual quality and content attractiveness. This may lead the subject to overuse or underuse certain regions of the rating scale when evaluating the stimuli related to the content that they may like or dislike. Such content-dependent interactions between subjects and specific stimuli are not explicitly captured by the proposed stimulus-independent quantization weights.

Recovering the true perceptual distribution of opinion scores for a given stimulus means eliminating or at least attenuating the effect of unstructured noise on the observed empirical distribution of opinion scores. Because ratings deriving from unstructured noise are not systematically replicated across scorers for a given stimulus, they spread across multiple opinion scores on the quality scale. From an information-theoretic perspective, we expect them to increase the entropy of the observed empirical distribution relative to the underlying true perceptual distribution. We therefore propose to estimate the true perceptual distribution through an optimization process that performs a mild entropy reduction while remaining faithful to the observed data.

Notably, no parametric assumptions are imposed on the shape of the true distribution $\{f_{ik}\}_{k\in\mathcal{K}}$. In particular, the proposed approach does not assume unimodality, symmetry, or concentration around a single peak. For instance, if the empirical distribution is bimodal suggesting the presence of two perceptual groups within the population, this structure will be preserved.

Under the probabilistic model defined by $\{f_{ik}\}$, the likelihood of the observed data is:

$$L(f) = \prod_{i\in\mathcal{I}} \prod_{k\in\mathcal{K}} f_{ik}^{n_{ik}},$$

with log-likelihood

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

Under the assumption that genuinely expressed opinion scores dominate significantly over unstructured noise (otherwise the entire experiment would be compromised) and considering that unstructured noise is typically uniformly distributed across all rating levels of the quality scale, if an opinion score k is selected solely due to unstructured noise, its empirical frequency n_{ik}/N_i will likely be a very small value. To attenuate the influence of this entropy-increasing noise, we introduce the empirical surprise measure:

$$S_{ik} = -\log\left(\frac{n_{ik}}{N_i}\right). \tag{5}$$

Thus, if an opinion score k has a very low empirical choice frequency for a given stimulus i, the associated S_{ik} values will be very large by definition, whereas if k is a consolidated opinion score (repeatedly and consistently selected by a subgroup of scorers), this will yield a moderate value of S_{ik}.

We define the CrowdRMLE estimator as the solution to

$$\begin{aligned} \max_{f} \quad & H(f) = \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} S_{ik}^{-1} n_{ik} \log(f_{ik}) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} f_{ik} = 1 \quad \forall i \in \mathcal{I}. \\ & f_{ik} \geq 0 \quad \forall i \in \mathcal{I}, \quad k \in \mathcal{K}. \end{aligned} \tag{6}$$

By down-weighting highly surprising (weakly supported) rating levels, the optimization performs a controlled entropy reduction of the observed empirical distribution in order to obtain an estimate of the latent true distribution of opinion scores.

A potential limitation of the proposed estimation procedure originates from the fact that the regularization weight depends only on the empirical frequency n_{ik}/N_i of each opinion score k for a stimulus i. Consequently, when only one or very few subjects select a particular score for a stimulus, the corresponding frequency becomes very small. Since the regularization mechanism down-weights low-frequency scores, isolated yet genuine perceptual judgments may be attenuated in the estimated perceptual distribution.

To understand when this effect becomes relevant, consider a stimulus i rated by a large number of subjects, and suppose that a structured minority subgroup representing $\alpha\%$ of the population would genuinely assign opinion score k. In this case, the empirical frequency satisfies $\frac{n_{ik}}{N_i} \approx \alpha$, and the corresponding surprise term becomes $S_{ik} \approx -\log(\alpha)$, yielding a weight proportional to $1/(-\log\alpha)$. Therefore, the contribution of score k to the estimated perceptual distribution increases monotonically with the subgroup's representation in the population.

Now consider an unstructured noise affecting $\beta\%$ of the ratings. Since such noise is expected to distribute approximately uniformly across the $|\mathcal{K}|$ rating levels, its empirical contribution to any specific score k is about $\beta/|\mathcal{K}|$. For noise to dominate a structured subgroup selecting score k, it must hold that

$$\frac{\beta}{|\mathcal{K}|} \geq \alpha, \quad \text{i.e.,} \quad \beta \geq |\mathcal{K}|\,\alpha.$$

This means that, on a five-point scale ($|\mathcal{K}| = 5$), at least 25% of the ratings would need to be purely noisy to dominate the contribution of a rating k chosen by a minority subgroup representing only 5% of the population in our quality estimation process. This analysis shows that ratings from structured minority subgroups are preserved and their contribution is proportional to the representation of these subgroups within the population, whereas unstructured noise is even more penalized as the number of rating levels on the quality scale increases.

The limitation arises when α becomes extremely small (i.e., in the case of isolated but genuine outliers). In such situations, the empirical frequency n_{ik}/N_i approaches zero, and the corresponding rating is down-weighted by the regularization mechanism. However, the associated quantization weights $\{\nu_k^j\}_{k\in\mathcal{K}}$ will still explicitly highlight the peculiarity of these subjects. While the current true distribution estimation process may penalize such isolated yet authentic scorers, an iterative refinement procedure could be employed: after identifying structured peculiarities through the quantization weights, ratings could be adjusted (e.g., via bias correction) and the perceptual distribution re-estimated. A detailed investigation of this iterative extension is beyond the scope of the present work.

B. Analytical Solution of CrowdRMLE

The following proposition, together with the subsequent remark, provides the analytical formula of the true choice probabilities as estimated by CrowdRMLE for each stimulus.

Proposition 1. *Assume that*

$$0 < n_{ik} < N_i, \qquad \forall i \in \mathcal{I},\ k \in \mathcal{K}.$$

The optimization problem

$$\begin{aligned} &\max_f H(f) \\ &s.t. \quad \sum_{k\in\mathcal{K}} f_{ik} = 1 \quad \forall i \in \mathcal{I}, \\ &f_{ik} \geq 0 \quad \forall i \in \mathcal{I}, \quad k \in \mathcal{K} \end{aligned} \tag{7}$$

underlying CrowdRMLE admits a unique optimal solution given by

$$\hat{f}_{ik} = \frac{S_{ik}^{-1} \cdot n_{ik}}{\sum_{l\in\mathcal{K}} S_{il}^{-1} \cdot n_{il}}. \tag{8}$$

Proof. We observe that the first-order derivatives of $H(f)$ are:

$$\frac{\partial H}{\partial f_{ik}} = \frac{S_{ik}^{-1} n_{ik}}{f_{ik}},$$

and the second-order derivatives are:

$$\frac{\partial^2 H}{\partial f_{ik}^2} = -\frac{S_{ik}^{-1} n_{ik}}{f_{ik}^2}.$$

For $(ik) \neq (jl)$, the terms are independent, thus:

$$\frac{\partial^2 H}{\partial f_{ik}\partial f_{jl}} = 0.$$

The Hessian matrix $\mathbf{M}$ of the function H is therefore a diagonal matrix with entries:

$$\mathbf{M}_{(ik),(jl)} = \begin{cases} -\dfrac{S_{ik}^{-1} n_{ik}}{f_{ik}^2}, & \text{if } (ik) = (jl), \\ 0, & \text{if } (ik) \neq (jl). \end{cases}$$

Since $0 < n_{ik} < N_i$, it follows that $S_{ik} > 0$. Hence,

$$-\frac{S_{ik}^{-1} n_{ik}}{f_{ik}^2} < 0.$$

Therefore, all diagonal entries of $\mathbf{M}$ are strictly negative. Thus, $\mathbf{M}$ is negative definite and the objective function $H(f)$ is strictly concave. Furthermore, the optimization problem has only linear constraints.

In other words, we are solving a strictly concave maximization problem over a convex feasible set. This implies that any point in the feasible set is an optimal solution if and only if it satisfies the Karush-Kuhn-Tucker (KKT) conditions (see Chapter 10 of [43]).

The Lagrangian of the considered optimization problem is given by

$$\begin{aligned} \mathcal{L}(f,\lambda,\mu) = &\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} S_{ik}^{-1} \cdot n_{ik} \cdot \log(f_{ik}) \\ &+ \sum_{i\in\mathcal{I}} \lambda_i \left(1 - \sum_{k\in\mathcal{K}} f_{ik}\right) + \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} \mu_{ik} f_{ik}, \end{aligned} \tag{9}$$

where $\lambda_i,\ i \in \mathcal{I}$, and $\mu_{ik} \geq 0,\ i \in \mathcal{I},\ k \in \mathcal{K}$, are the Lagrangian multipliers associated with the equality and inequality constraints, respectively.

In the case of our problem, the KKT conditions can be written as:

$$\frac{\partial \mathcal{L}}{\partial f_{ik}} = \frac{S_{ik}^{-1} \cdot n_{ik}}{f_{ik}} - \lambda_i + \mu_{ik} = 0, \tag{10}$$

$$\sum_{k\in\mathcal{K}} f_{ik} - 1 = 0, \tag{11}$$

$$f_{ik} \geq 0, \tag{12}$$

$$\mu_{ik} \geq 0, \tag{13}$$

and

$$\mu_{ik} f_{ik} = 0. \tag{14}$$

Since $0 < n_{ik} < N_i$, it follows that

$$S_{ik}^{-1} n_{ik} > 0, \qquad \forall i \in \mathcal{I},\ k \in \mathcal{K}.$$

Therefore, the objective function $H(f)$ is defined only for

$$f_{ik} > 0, \qquad \forall i \in \mathcal{I},\ k \in \mathcal{K}.$$

Moreover,

$$\lim_{f_{ik}\to 0^+} S_{ik}^{-1} n_{ik} \log(f_{ik}) = -\infty.$$

Hence, whenever any component f_{ik} approaches zero, the objective function tends to $-\infty$. Therefore, no optimal solution can occur on the boundary of the feasible set, and any optimal solution necessarily satisfies

$$f_{ik} > 0, \qquad \forall i \in \mathcal{I},\ k \in \mathcal{K}.$$

Using Eq. (14), it follows that

$$\mu_{ik} = 0.$$

Substituting $\mu_{ik} = 0$ into Eq. (10) yields

$$\frac{S_{ik}^{-1} \cdot n_{ik}}{f_{ik}} - \lambda_i = 0. \tag{15}$$

From Eq. (15), it holds that

$$f_{ik} = \frac{S_{ik}^{-1} \cdot n_{ik}}{\lambda_i}. \tag{16}$$

Eq. (11) implies:

$$\sum_{k\in\mathcal{K}} \frac{S_{ik}^{-1} \cdot n_{ik}}{\lambda_i} = \frac{1}{\lambda_i} \sum_{k\in\mathcal{K}} S_{ik}^{-1} \cdot n_{ik} = 1, \qquad \forall i \in \mathcal{I}.$$

Thus,

$$\lambda_i = \sum_{k\in\mathcal{K}} S_{ik}^{-1} \cdot n_{ik}.$$

Finally, substituting the value of λ_i into the expression of f_{ik} in Eq. (16) yields

$$f_{ik} = \frac{S_{ik}^{-1} \cdot n_{ik}}{\sum_{l\in\mathcal{K}} S_{il}^{-1} \cdot n_{il}}, \tag{17}$$

which is the unique solution that satisfies the KKT conditions and thus the unique optimal solution of the problem. This concludes the proof. □

In practice it is common to encounter situations where opinion score k is not selected by any subject during the evaluation of a stimulus i, i.e., $n_{ik} = 0$, or where opinion score k is selected by all subjects when rating stimulus i, i.e., $n_{ik} = N_i$. Since Proposition 1 assumes $0 < n_{ik} < N_i$, these boundary cases are not covered. Nevertheless, the optimal solution derived in Proposition 1 naturally extends to these boundary cases by continuity, as shown in the following remark.

Remark 1. *Let*

$$w_{ik} = n_{ik} S_{ik}^{-1} = \frac{n_{ik}}{-\log(n_{ik}/N_i)}.$$

When $n_{ik} \to 0$*,*

$$w_{ik} = \frac{n_{ik}}{-\log(n_{ik}/N_i)} \to 0,$$

which implies, from Eq. (17)*, that:*

$$f_{ik} \to 0.$$

Moreover, when $n_{ik} \to N_i$*, it follows from*

$$N_i = \sum_{k\in\mathcal{K}} n_{ik}$$

that $n_{ij} \to 0$ *for all* $j \neq k$*. Furthermore,*

$$w_{ik} = \frac{n_{ik}}{-\log(n_{ik}/N_i)} \to +\infty.$$

Consequently,

$$f_{ik} \to 1, \qquad f_{il} \to 0, \quad l \neq k.$$

Hence, the solution obtained in Proposition 1 admits a natural continuous extension to the boundary cases $n_{ik} = 0$ *and* $n_{ik} = N_i$*.*

Accordingly, CrowdRMLE computes the true choice frequencies in practice as:

$$f_{ik} = \begin{cases} 0, & n_{ik} = 0, \\ \dfrac{S_{ik}^{-1} n_{ik}}{\sum\limits_{l\in\mathcal{K}} S_{il}^{-1} n_{il}}, & 0 < n_{ik} < N_i, \\ 1, & n_{ik} = N_i. \end{cases} \tag{18}$$

Therefore, regarding boundary cases, opinion scores that are never selected receive zero weight, whereas unanimously selected opinion scores receive unit weight.

C. Subjective Quality Recovery and Measures of Scale Quantization

Once the true choice probabilities f_{ik} are calculated using the formula in Eq. (18), we can recover the true subjective quality Q_i of each stimulus $i \in \mathcal{I}$ as:

$$Q_i = \sum_{k\in\mathcal{K}} f_{ik} \cdot k. \tag{19}$$

Each scale quantization weight is computed using the following formula:

$$\nu_k^j = \frac{\sum_{i\in\mathcal{I}_j} \left(p_{ik}^j - f_{ik}\right)}{|\mathcal{I}_j|}. \tag{20}$$

Thus, CrowdRMLE measures the discrepancies between the subject's usage of the opinion score k and the expected true usage for each stimulus. It then averages these discrepancies across all stimuli in the set $\mathcal{I}_j$ to provide an overall assessment of whether subject j overuses ($\nu_k^j > 0$) or underuses ($\nu_k^j < 0$) the score k.

The closer all quantization weights ν_k^j are to 0, the less the subject quantized the quality scale. Hence, we introduce the following statistic, which we name subject Scale Quantization Index (SQI):

$$SQI_j = \sum_{k\in\mathcal{K}} |\nu_k^j|, \tag{21}$$

that we use together with the scale quantization weights (SQWs) of a subject j to measure the extent to which that subject quantized the quality scale during the experiment.

TABLE I
COMPUTATION TIMES REQUIRED BY RMLE AND CROWDRMLE.

Type	Dataset	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{J}\rvert$	RMLE (s)	CrowdRMLE (s)
Laboratory	NETFLIX-PUB	70	26	13	**0.0002**
	VQEG-HD1	144	24	88	**0.0003**
	VQEG-HD3	167	24	173	**0.0003**
	VQEG-HD5	168	24	180	**0.0004**
	ITS4S	514	27	14843	**0.0008**
Crowdsource	KonViD-1k	1200	624	513216	**0.0145**
	KonIQ	10076	1261	–	**0.2628**
	KADID	11085	2212	–	**1.7808**
	NIVD	1860	12812	–	**0.4086**
	MovieLens-1M	3706	6040	–	**0.5197**

IV. EFFICIENCY, ROBUSTNESS AND ACCURACY OF CROWDRMLE

We conducted numerical experiments to compare CrowdRMLE to RMLE in terms of efficiency. We also compared the accuracy and noise robustness of CrowdRMLE to that of several state-of-the-art approaches for subjective quality recovery from a noisy dataset.

A. Datasets overview

The experiments were conducted on ten datasets, five of which were collected in controlled laboratory environments and the remaining five via crowdsourcing. The laboratory datasets include VQEG-HD1, VQEG-HD3, VQEG-HD5 [44], the Netflix Public Dataset [42], and the ITS4S dataset [45]. These datasets were gathered under tightly controlled conditions, ensuring in general high reliability and consistency in subjective quality assessments. In contrast, the crowdsourcing datasets include KonViD-1k [46], KonIQ [41], KADID [41], the Netflix International Video Dataset (NIVD) [47], and the MovieLens-1M dataset [48]. The number of subjects $|\mathcal{J}|$ participating in each experiment and the stimuli $|\mathcal{I}|$ used in these datasets are detailed in Table I.

In all datasets except NIVD participants evaluated the quality of stimuli using a five-point quality scale. In contrast, the original NIVD dataset employed a continuous quality scale ranging from 1 to 100. The authors of [41] demonstrated that these continuous scores can be normalized to the five-point absolute category rating scale without loss of consistency and released this five-point scale version of the NIVD dataset, which is the one we used for our analysis.

Unlike the nine other datasets considered in our analysis, which focus specifically on image and video quality assessment, the MovieLens-1M dataset required participants to score their overall Quality of Experience (QoE) with respect to a given movie. QoE is influenced not only by the perceptual quality of the movie but also by other subjective factors, such as liking/disliking the movie's topic, which can vary significantly based on individual preferences. By including the MovieLens-1M dataset, we therefore aim to evaluate the extent to which the conclusions drawn by CrowdRMLE when analyzing crowdsourcing datasets for image and video quality assessment can be generalized to the overall QoE assessment.

B. The Efficiency of CrowdRMLE

In this section we show that CrowdRMLE provides a highly efficient alternative to RMLE, which is impractical for crowdsourcing datasets.

TABLE II
CORRELATION AND ROOT MEAN SQUARE ERROR BETWEEN THE OUTPUT OF RMLE AND CROWDRMLE

Dataset	PLCC	SROCC	RMSE
NETFLIX-PUB	0.9984	0.9977	0.1072
VQEG-HD1	0.9994	0.9992	0.0529
VQEG-HD3	0.9996	0.9994	0.0383
VQEG-HD5	0.9995	0.9992	0.0519
ITS4S	0.9994	0.9992	0.0271
KonViD-1k	0.9996	0.9998	0.0209

Table I shows the computational time (in seconds) required by RMLE [49] and the proposed CrowdRMLE software for each dataset. The experiments were done on a computer having an Intel Core i9-10900X CPU with a clock speed of 3.7 GHz and 64 GB of RAM.

For the five small-scale datasets from laboratory environments, CrowdRMLE required on average less than a millisecond, whereas RMLE took 13 seconds for the smallest dataset (70 stimuli, 26 subjects) and up to 14,843 seconds (approximately 4 hours) for the largest dataset (514 stimuli, 27 subjects).

On crowdsourcing datasets, CrowdRMLE processed the largest dataset (11,085 stimuli and 2,212 subjects) in less than 2 seconds. In contrast, RMLE required 513,216 seconds (6 days) to process the smallest dataset (1,200 stimuli and 624 subjects). We did not apply RMLE to the other crowdsourcing datasets, as it would have required significantly more than 6 days.

The efficiency gains offered by CrowdRMLE come with no significant discrepancy in the estimated quality scores compared to RMLE on datasets that are not excessively noisy. Table II demonstrates that the outputs of RMLE and CrowdRMLE exhibit an extremely high Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC), along with a very small Root Mean Square Error (RMSE), across all datasets where both methods were applied. Thus, both RMLE and CrowdRMLE provide very similar estimates of the ground-truth quality on the original datasets considered in this study. However, as we will show in the next section, when dataset integrity is compromised by additional noise, CrowdRMLE demonstrates greater robustness than RMLE and other state-of-the-art approaches.

C. CrowdRMLE Robustness to Structured and Unstructured Noise

Since no ground-truth subjective quality scores are available for real datasets, robustness to controlled synthetic perturbations remains the standard methodology for comparing subjective quality recovery approaches [11], [33], [34], [42]. We therefore evaluated the robustness of CrowdRMLE against state-of-the-art recovery methods, namely the most recently recommended algorithm in ITU-T P.910 [37], ZREC [35], and RMLE [12], under progressively increasing synthetic noise. In the simulation study we considered both structured and unstructured noise. In particular, two experimental scenarios were considered:

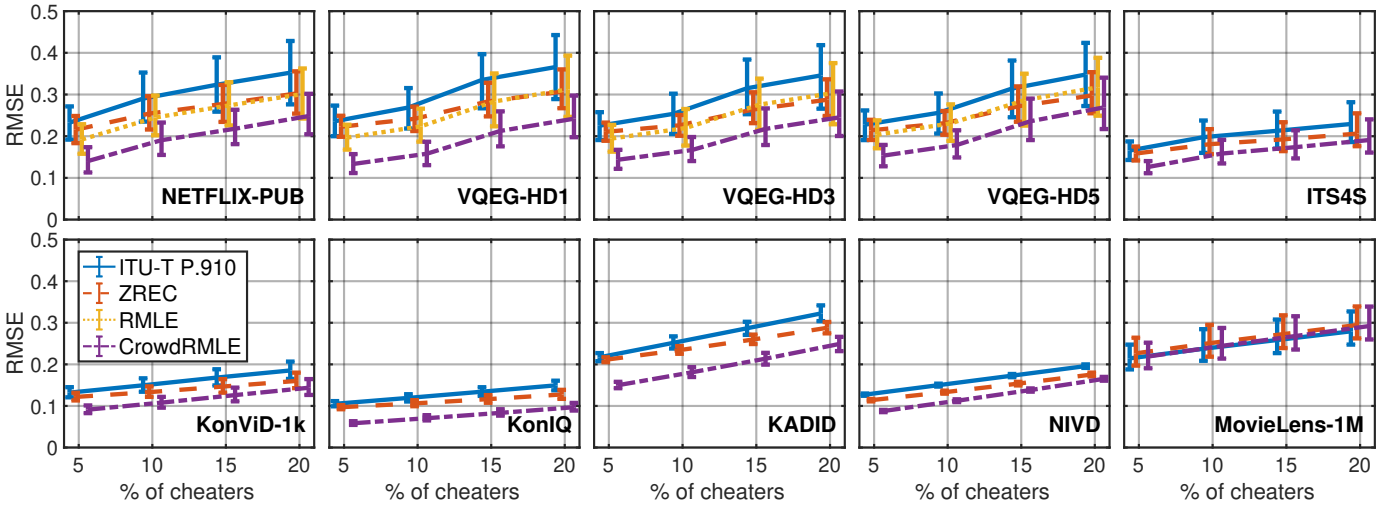

Fig. 1. Robustness of CrowdRMLE under progressively increasing structured noise introduced via an increasing percentage of simulated cheaters, compared with the most recent algorithm recommended in ITU-T P.910 [37] (2023), ZREC [35] (2023), and RMLE [12] (2024). Mean RMSE over 1000 Monte Carlo repetitions, confidence intervals correspond to the 0.025 and 0.975 quantiles. Lower RMSE indicates better robustness. For better visibility of confidence intervals, points are slightly shifted horizontally.

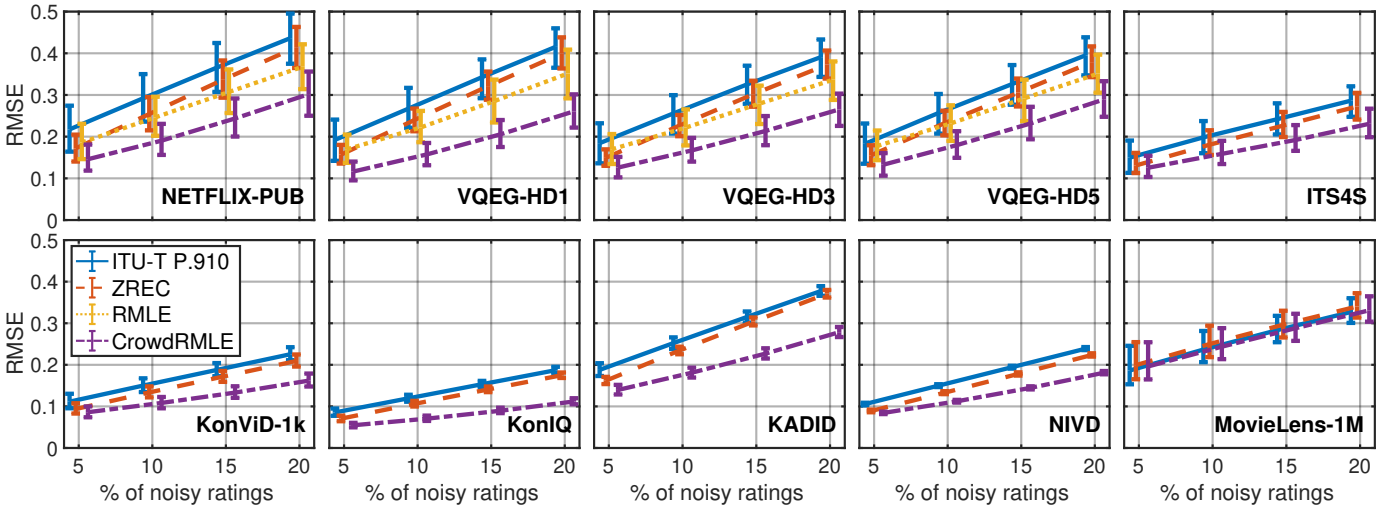

Fig. 2. Robustness of CrowdRMLE under progressively increasing unstructured noise (percentage of randomly corrupted ratings), compared with the most recent algorithm recommended in ITU-T P.910 [37] (2023), ZREC [35] (2023), and RMLE [12] (2024). Mean RMSE over 1000 Monte Carlo repetitions, confidence intervals correspond to the 0.025 and 0.975 quantiles. Lower RMSE indicates better robustness. For better visibility of confidence intervals, points are slightly shifted horizontally.

- **Case 1** (Increasing structured noise): The percentage of peculiar subjects (referred to as cheaters) was increased from 5% to 20%, while the percentage of ratings affected by unstructured noise was fixed at 10%.
- **Case 2** (Increasing unstructured noise): The percentage of peculiar subjects was fixed at 10%, while the proportion of ratings affected by unstructured noise was increased from 5% to 20%.

More precisely, for each noise configuration, a given percentage of subjects was randomly selected from the original dataset and transformed into cheaters by modifying their opinion scores according to predefined behavioral models. Specifically, to simulate a spammer, the selected subject's ratings were replaced with randomly chosen alternative values on the scale; to simulate a central scorer, 95% of the selected subject's ratings were replaced with the central score (3); and to simulate a bimodal scorer, 95% of the ratings of the selected subject were reassigned to scores 2 or 4 depending on whether the mean opinion score of the corresponding stimulus was greater than or less than 3. Given the percentage of cheaters, the scoring behavior of each cheater was selected at random among these three behavioral models. Finally, unstructured noise was introduced separately by randomly replacing a specified percentage of ratings in the dataset with uniformly sampled alternative scale values.

For each noise configuration, the simulation was repeated 1000 times to ensure statistical stability. For each method, we computed the RMSE between the outputs obtained from the original dataset and those obtained from the noisy versions of the same dataset. The mean RMSE values and 95% confidence intervals were estimated. The confidence interval was obtained using a quantile-based approach, i.e., computing the 0.025 and 0.975 quantiles from the sample of RMSE obtained from the 1000 repetitions.

Figures 1 and 2 show the evolution of the RMSE as a function of increasing structured and unstructured noise, respectively. Across all image and video quality datasets, CrowdRMLE consistently yields the lowest RMSE values under both experimental scenarios.

The 95% confidence intervals obtained from 1000 repetitions remain well separated from those of the competing approaches in most configurations. In particular, the mean RMSE curve of CrowdRMLE consistently lies outside the confidence intervals of the other approaches across a wide range of noise levels.

D. Accuracy of CrowdRMLE

The objective of this experiment is to evaluate the estimation accuracy of CrowdRMLE in a controlled setting where the true quality of each stimulus is known. Such an analysis cannot be performed on real subjective datasets since only noisy opinion scores are available and the underlying perceptual qualities are unknown. Therefore, following a common practice in the literature [11], [32], [36], we resort to simulation in order to create a dataset with known ground-truth qualities. It should be noted that the purpose of this experiment is not to reproduce all the complexities of a real subjective study, but rather to provide a controlled benchmark that enables a direct evaluation of estimation accuracy.

No artificial image or video stimuli are generated in this experiment. This is because subjective quality recovery algorithms, including CrowdRMLE and the competing approaches considered in this paper, operate exclusively on opinion scores rather than on the underlying media content. Therefore, for the purpose of evaluating estimation accuracy, it is sufficient to generate a synthetic rating matrix with known ground-truth qualities and simulated subject ratings. Specifically, we consider 100 hypothetical stimuli and 25 reliable subjects evaluating them on a five-point Absolute Category Rating (ACR) scale. The ground-truth quality values are defined as 100 equally spaced points in the interval $[1, 5]$. Denoting by q_i the ground-truth quality of stimulus i, the opinion score assigned by each subject is generated as a realization of a Gaussian random variable with mean q_i and standard deviation

$$\sigma_i = 0.2(-q_i^2 + 6q_i - 5), \tag{22}$$

in accordance with the Standard deviation of Opinion Scores (SOS) hypothesis [50], which describes the relationship between perceived quality and score variability in subjective quality assessment. The generated values are subsequently rounded and clipped to the discrete scale $\{1, \ldots, 5\}$, yielding a 100×25 rating matrix. From now on, we will refer to this rating matrix as the "Simulated Dataset".

Note that CrowdRMLE does not assume that opinion scores follow a Gaussian distribution around the underlying quality. Therefore, unlike some benchmarking approaches whose assumptions closely match the simulation model, CrowdRMLE is not inherently favored by the adopted simulation setting.

We then progressively add cheaters and unstructured noise to the Simulated Dataset following the same procedure described in Section IV-C for real datasets. For each configuration, we compute the RMSE between the estimated quality scores and the ground-truth qualities q_i.

The results in Figure 3 show that CrowdRMLE consistently achieves the lowest RMSE across all noise levels. The mean RMSE of CrowdRMLE lies outside the corresponding confidence intervals of competing approaches in most configurations, indicating that CrowdRMLE estimates the ground-truth quality with higher accuracy.

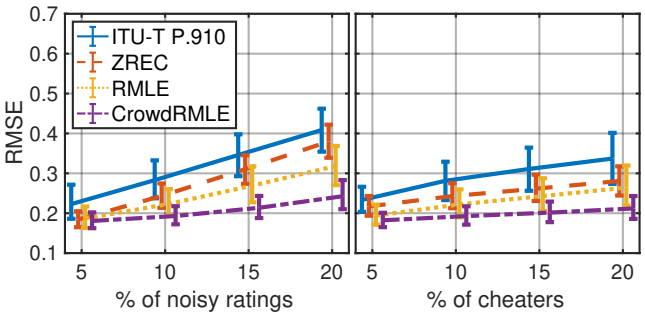

Fig. 3. Accuracy of CrowdRMLE and other benchmarking approaches on a synthetic dataset with known ground truth. Ratings are progressively corrupted by the insertion of unstructured noise (left) and cheaters (right). Mean RMSE over 1000 repetitions, confidence intervals correspond to the 0.025 and 0.975 quantiles. Lower RMSE indicates higher accuracy. For better visibility of confidence intervals, points are slightly shifted horizontally.

E. Statistical Significance Analysis

The confidence intervals shown in Figs. 1, 2, and 3 suggest that CrowdRMLE generally achieves lower RMSE values than the competing approaches. However, visual inspection of confidence intervals obtained through Monte Carlo simulation does not constitute a formal statistical significance test. We therefore performed a dedicated statistical analysis to assess whether the observed robustness and accuracy improvements are statistically significant.

For each dataset, noise level, and competing method, we considered the paired differences

$$d_i = \mathrm{RMSE}(\mathrm{CrowdRMLE}, i) - \mathrm{RMSE}(\mathrm{competing}, i), \tag{23}$$

where $i = 1, 2, \ldots, 1000$ indexes the random seed used in the Monte Carlo simulation. Preliminary normality tests revealed that the distributions of the paired differences d_i were generally not Gaussian. We therefore employed a non-parametric test, i.e., the one-sided Wilcoxon signed-rank test. Since negative values of d_i correspond to CrowdRMLE producing lower RMSE values than the competing method, the null and alternative hypotheses are: H_0: CrowdRMLE does not achieve lower RMSE values than the competing method; H_1: CrowdRMLE achieves lower RMSE values than the competing method. The test was performed independently for every dataset, noise level, and competing method. To provide a conservative summary, Table III reports, for each dataset and competing method, the largest (i.e., least significant) p-value obtained across all considered noise levels.

The results confirm the conclusions suggested by the confidence intervals. CrowdRMLE significantly outperforms the

TABLE III
WILCOXON SIGNED-RANK TEST (H_1: CROWDRMLE ACHIEVES LOWER RMSE THAN THE COMPETING METHOD). THE TEST P-VALUES CORRESPONDING TO THE 22 DATASET-CASE COMBINATIONS IN FIG. 1, 2, AND 3 ARE REPORTED. CASE 1 (C1) AND CASE 2 (C2) (INCREASING STRUCTURED/UNSTRUCTURED NOISE) ARE DESCRIBED IN SECTION IV-C.

Dataset	Case	ITU-T P.910	ZREC	RMLE
NETFLIX-PUB	C1	$p < .001$	$p < .001$	$p < .001$
	C2	$p < .001$	$p < .001$	$p < .001$
VQEG-HD1	C1	$p < .001$	$p < .001$	$p < .001$
	C2	$p < .001$	$p < .001$	$p < .001$
VQEG-HD3	C1	$p < .001$	$p < .001$	$p < .001$
	C2	$p < .001$	$p < .001$	$p < .001$
VQEG-HD5	C1	$p < .001$	$p < .001$	$p < .001$
	C2	$p < .001$	$p < .001$	$p < .001$
ITS4S	C1	$p < .001$	$p < .001$	–
	C2	$p < .001$	$p < .001$	–
KonIQ	C1	$p < .001$	$p < .001$	–
	C2	$p < .001$	$p < .001$	–
KonViD-1k	C1	$p < .001$	$p < .001$	–
	C2	$p < .001$	$p < .001$	–
KADID	C1	$p < .001$	$p < .001$	–
	C2	$p < .001$	$p < .001$	–
NIVD	C1	$p < .001$	$p < .001$	–
	C2	$p < .001$	$p < .001$	–
MovieLens-1M	C1	$p = 1.00$	$p < .001$	–
	C2	$p = 1.00$	$p < .001$	–
Simulated Dataset	C1	$p < .001$	$p < .001$	$p < .001$
	C2	$p < .001$	$p < .001$	$p < .001$

TABLE IV
AVERAGE SQI AND INCONSISTENCY (INC) VALUES FOR ALL DATASETS. FOR CROWDSOURCING EXPERIMENTS, P-VALUES OF A TWO-SAMPLE WELCH'S T-TEST ASSESS WHETHER THE MEAN DIFFERS SIGNIFICANTLY FROM THE CORRESPONDING MEAN COMPUTED OVER ALL LAB EXPERIMENTS.

Datasets	**Avg SQI**	**SQI p-values**	**Avg Inc**	**Inc p-values**
NETFLIX-PUB	0.26	–	0.62	–
VQEG-HD1	0.25	–	0.58	–
VQEG-HD3	0.31	–	0.62	–
VQEG-HD5	0.31	–	0.63	–
ITS4S	0.30	–	0.68	–
KonViD-1k	0.44	$p < .001$	0.68	$p < .001$
KonIQ	0.37	$p < .001$	0.56	$p = 1.00$
KADID	0.44	$p < .001$	0.69	$p < .001$
NIVD	0.50	$p < .001$	0.70	$p < .001$
MovieLens-1M	0.37	$p < .001$	0.90	$p < .001$

competing methods in almost all cases, with p-values below 0.001 for nearly all dataset and method pairs. The only exception is the comparison with ITU-T P.910 on the MovieLens-1M dataset, where no statistically significant difference is observed. Overall, the statistical analysis provides strong evidence that the robustness and accuracy improvements achieved by CrowdRMLE are unlikely to be due to random variation.

V. ANALYSIS OF CROWDSOURCING DATASETS WITH CROWDRMLE

A. Statistical Comparison of SQI and Subject Inconsistency

Although it is widely acknowledged that data collected in crowdsourcing settings may exhibit lower reliability than those gathered in controlled environments, identifying statistical indicators that quantify this difference remains an open problem in image and video quality assessment. Table IV reports the average SQI and subject inconsistency (Inc) values for all datasets. For SQI, the separation between laboratory and crowdsourcing datasets is systematic: the smallest average SQI observed in crowdsourcing (0.37) exceeds the largest average observed in laboratory datasets (0.31).

To assess the statistical significance of this systematic difference, we compared the mean of subject-level SQI values of each crowdsourcing dataset against the mean of the pooled laboratory SQI values. All five laboratory datasets were merged to avoid bias due to specific experimental settings. In order to compare the mean of SQI values in the two environments, the two-sample Welch's t-test was used since equal variances could not be assumed between laboratory and crowdsourcing environments. The Welch's t-test is applicable in our case because established statistical literature [51]–[53] has proven its accuracy when dealing with large sample sizes (from 125 to 12,812 in our case) and data that are moderately skewed (from 0.63 to 1.15 for our samples) with moderate kurtosis (from 4.25 to 6.63 in our samples). All comparisons yielded p-values smaller than 0.001, confirming that scale quantization is significantly more pronounced in crowdsourcing settings.

In contrast, subject inconsistency does not exhibit the same systematic separation. Although inconsistency values are generally higher in crowdsourcing datasets, two crowdsourcing datasets out of five present average inconsistency values that are smaller or equal to those observed in laboratory experiments, and for one crowdsourcing experiment (KonIQ) the average inconsistency is even smaller than the average inconsistency of controlled environment with statistical significance. Therefore, while inconsistency is a valuable measure of subject inaccuracy, it does not systematically distinguish crowdsourcing from controlled environments in image and video quality assessment.

To go beyond the analysis of average SQI values, we investigated whether SQI provides a better characterization than subject inconsistency regarding the differences in scoring behavior between laboratory-controlled and crowdsourcing environments in image and video quality assessment. To this end, we performed a Receiver Operating Characteristic (ROC) analysis using both SQI and inconsistency as classification scores to distinguish between subjects participating in laboratory-controlled and crowdsourcing experiments. The results shown in Figure 4 indicate that, across all four image and video quality datasets (KonViD-1k, KonIQ, KADID, and NIVD), SQI consistently achieves higher Area Under the Curve (AUC) values than inconsistency. This suggests that SQI provides a superior characterization of the peculiar scoring behaviors that differentiate laboratory-controlled and crowdsourcing environments.

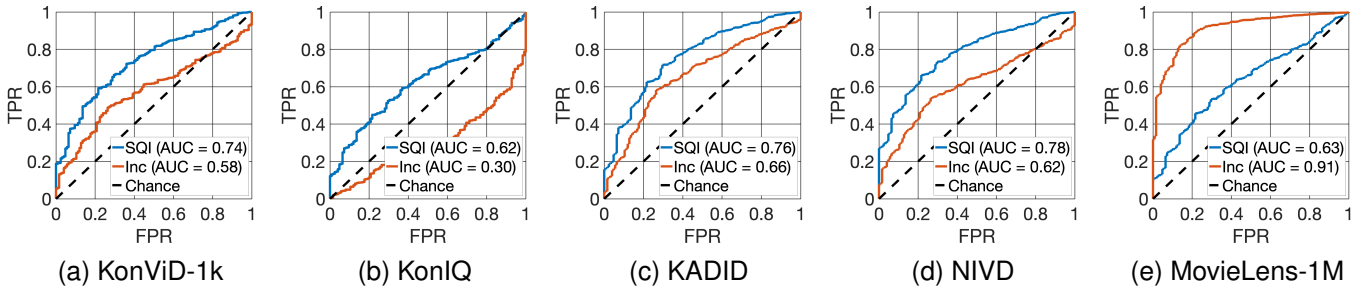

Fig. 4. ROC curves comparing the ability of SQI and subject inconsistency values to distinguish between controlled and crowdsourcing environments. The five laboratory datasets were merged to obtain a consistent sample of values in controlled environments and to mitigate the effects of dataset-specific peculiarities on the analysis. AUC values indicate discriminative performance.

Figure 5 provides additional insight into the nature of the deviations captured by the SQI and the inconsistency. Using the maximum value of SQI and maximum value of inconsistency in controlled laboratory environments as reference thresholds, subjects were categorized according to whether they exceed laboratory levels in SQI, inconsistency, or both. For all image and video quality datasets, the proportion of subjects exhibiting scale quantization beyond laboratory standards (Q3 + Q4) exceeds the proportion exhibiting inconsistent behavior (Q2 + Q3). This indicates that unbalanced use of the quality scale, as measured by SQI, is in general more prevalent than pure inconsistency-induced deviations in crowdsourcing experiments for image and video quality assessment.

The scatter plots in Figure 5 also show that SQI and inconsistency are not correlated: subjects with high SQI do not necessarily exhibit high inconsistency, and vice versa. This weak correlation between the two measures indicates that they capture complementary aspects of scoring behavior.

We notice that a different behavior is observed for the MovieLens-1M dataset, which addresses overall QoE or preference assessment rather than image and video quality assessment. In this case, inconsistency achieves higher discriminative power. This result highlights a limitation of the current SQI formulation, which assumes the presence of a structured latent perceptual distribution. In tasks characterized by inherently higher subjectivity and potentially flatter opinion score distributions, scale quantization may be less dominant. Extending the approach to such scenarios represents a future direction of improvement.

Overall, the results in this section show that the mean of SQI values is significantly higher in crowdsourcing settings than in controlled environments, whereas the same conclusion is not always observed for a conventional measure of reliability such as the subject inconsistency. Furthermore, the ROC analysis shows that the SQI value of an individual subject is more informative than the inconsistency value for determining whether the subject participated in a controlled or crowdsourcing experiment. Taken together, these results indicate that the SQI better captures peculiar scoring behaviors that are particularly prevalent in crowdsourcing-based subjective media quality assessment than the subject inconsistency. This provides evidence that minimizing SQI should be considered an important design objective for future crowdsourcing methodologies aiming to approach the reliability of controlled environments.

B. Analysis of the SQWs in Highly Inconsistent Subjects

Given the significant prevalence of scale quantization in crowdsourcing, as discussed in the previous section, a natural question arises: do highly inconsistent subjects in crowdsourcing tests for image quality assessment score entirely at random, thus not exhibiting any preference for specific quality levels on the scale, or do they exhibit a systematic pattern of scale quantization? To address this question, we ordered the subjects in each crowdsourcing dataset by decreasing inconsistency and analyzed their scale quantization weights.

Figure 6 shows the scale quantization weights of the 100 most inconsistent subjects in each crowdsourcing dataset. A consistent pattern emerges: highly inconsistent subjects in image and video quality assessment tend to avoid using the central opinion scores. Specifically, the center of the heatmaps is rather blue, while the extremes are red.

To numerically verify this visually observed trend, we conducted statistical tests to evaluate whether highly inconsistent subjects use the extremes of the quality scale more than the central part. For each crowdsourcing dataset, we selected all subjects whose inconsistency exceeded the maximum inconsistency observed in controlled-environment datasets. From the scale quantization weights of these subjects, we computed two proportions:

1) $P_{central}$: The proportion of highly inconsistent subjects for whom at least two of the three scale quantization weights associated with ”2,” ”3,” and ”4” are positive.
2) $P_{extremes}$: The proportion of highly inconsistent subjects for whom at least one of the two scale quantization weights associated with ”1” or ”5” is positive, while the condition defining $P_{central}$ does not hold.

$P_{central}$ represents the proportion of highly inconsistent subjects who overuse at least two central opinion scores. Conversely, $P_{extremes}$ represents those who avoid the central part of the scale, underusing at least two of the three central opinion scores, but overuse at least one of the two extreme opinion scores.

To assess whether $P_{extremes}$ is significantly larger than $P_{central}$ among highly inconsistent subjects, we performed the McNemar’s exact test for paired nominal data. For each subject, two binary indicators were defined corresponding to the conditions underlying $P_{central}$ and $P_{extremes}$, and the test

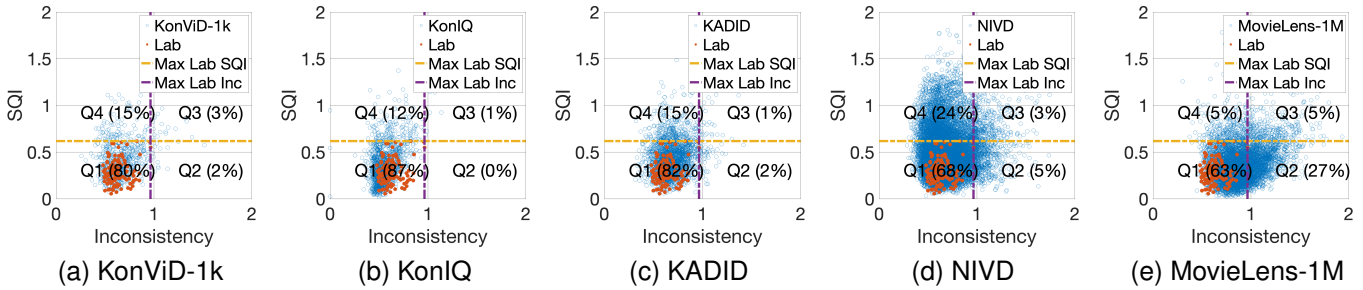

Fig. 5. Scatter plots of subject inconsistency versus SQI in controlled and crowdsourcing environments. Each point represents one subject. Dashed lines indicate the maximum SQI and inconsistency observed in the five laboratory datasets, which were merged to obtain a consistent sample of values in controlled environments and to mitigate the effects of dataset-specific peculiarities on the analysis.

was applied to the resulting 2×2 contingency table. The exact version of the test was adopted to obtain accurate p-values also for datasets with a small number of observations. The results in Table V show that $P_{extremes}$ is significantly larger than $P_{central}$ across all crowdsourcing datasets. This confirms the hypothesis suggested by the patterns in Figure 6. Therefore, highly inconsistent scorers in crowdsourcing tests for image and video quality assessment tend to underuse the central part of the scale. While this behavioral pattern could be expected in uncontrolled subjective experiments as a strategy to complete the test as soon as possible, the present analysis is, to the best of our knowledge, the first to reveal it within the context of image and video quality assessment.

It is important to note, however, that not all sources of inconsistency are linked to scale quantization. For instance, a subject who scores at random or inverts the quality scale would not exhibit preferences for specific opinion scores. Nevertheless, our analysis suggests that the pattern in Figure 6 is prevalent among highly inconsistent subjects. Indeed, the average value of $P_{extremes}$ across the five datasets in Table V indicates that over 70% of highly inconsistent subjects underuse at least two of the three central opinion scores.

Based on the results in Figure 6 and Table V, it can be inferred that using approaches such as pair comparison or reducing the number of options (for instance, from 5 to 2 or 3) might help mitigate high inconsistency values in crowdsourcing, thus improving the overall quality of the gathered data. This conclusion is in line with those in [54], [55], where the authors, using a completely different approach from CrowdRMLE concluded, among other things, that the use of pair comparison would increase the reliability of the collected data. Our analysis, therefore, serves as further motivation to revisit the number of options offered on the quality scale in future methodologies for collecting opinion scores in crowdsourcing.

TABLE V
COMPARISON OF THE FRACTIONS P_{CENTRAL} AND P_{EXTREME}. THE P-VALUES REFER TO THE EXACT MCNEMAR'S TEST CONDUCTED TO ASSESS WHETHER THE TWO FRACTIONS ARE DIFFERENT WITH STATISTICAL SIGNIFICANCE.

Dataset	$\mathbf{P}_{\text{central}}$	$\mathbf{P}_{\text{extreme}}$	p-value
KonViD-1k	0.2058	0.7941	$p < .001$
KonIQ	0.1428	0.8571	$p = .015$
KADID	0.2361	0.7222	$p < .001$
NIVD	0.1872	0.7800	$p < .001$
MovieLens-1M	0.3624	0.6309	$p < .001$

C. SQI and SQWs vs. Subject Bias

As previously mentioned in Section II, modern, standardized approaches for analyzing subjectively annotated datasets in image and video quality assessment rely on measures of subject bias and inconsistency. In the previous section, we studied the link between scale quantization and subject inconsistency. In this section, we show that the scale quantization weights, ν_k^j, calculated by CrowdRMLE, generalize the traditional concept of bias in image and video quality assessment.

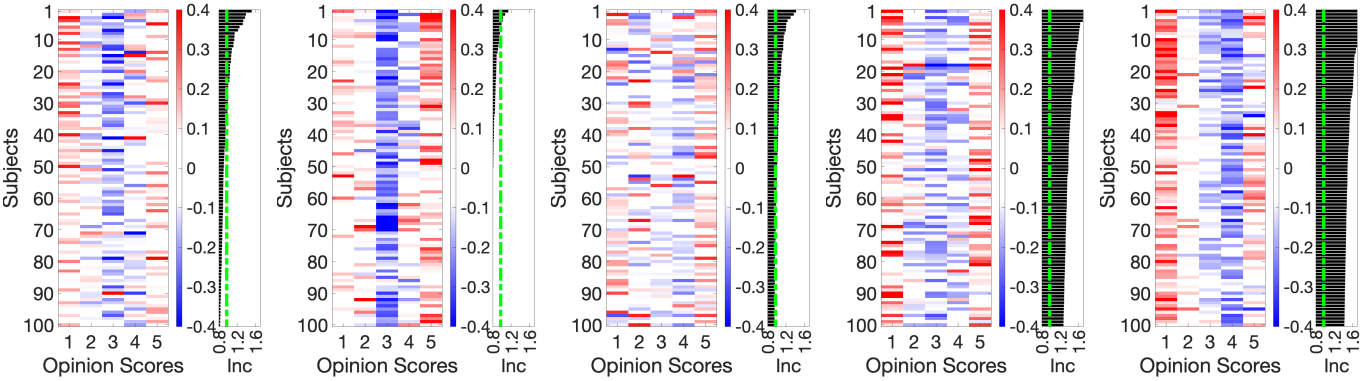

Fig. 6. Scale quantization weights of the top 100 most inconsistent subjects in each crowdsourcing dataset. The subjects are sorted in decreasing order of inconsistency values. The dashed green line ($\sigma = 0.972$) represents the maximum value of inconsistency in experiments conducted in controlled environments.

TABLE VI
SPEARMAN RANK CORRELATION BETWEEN THE SUBJECT BIAS CALCULATED FROM THE FIVE-SCALE QUANTIZATION WEIGHTS AND THE BIAS COMPUTED ACCORDING TO THE ITU-T P.910 AND ZREC.

Dataset	ITU-T P.910	ZREC
NETFLIX-PUB	0.9989	0.9959
VQEG-HD1	0.9996	0.9893
VQEG-HD3	0.9996	0.9869
VQEG-HD5	0.9997	0.9983
ITS4S	1.0000	0.9988
KonViD-1k	0.9987	0.9970
KonIQ	0.9983	0.9964
KADID	0.9973	0.9896
NIVD	0.9996	0.9990
MovieLens-1M	0.9923	0.9895

Bias is defined in the media quality assessment literature as the systematic tendency of a subject to overuse the left part (negative bias) or the right part (positive bias) of the quality scale. In the introduction, we defined scale quantization as the systematic tendency of a subject to overuse any subset of opinion scores. From both definitions, it is evident that bias is simply a specific type of scale quantization. In fact, we have conducted experiments to show that the bias of a subject can be computed from the scale quantization weights produced by CrowdRMLE.

In particular, given the five scale quantization weights $\nu_k^j \;\; k = 1, 2, \ldots, 5$, of subject j as computed by CrowdRMLE, we estimate the bias b_j of subject j as follows:

$$b_j = \sum_{k=1}^{5} \nu_k^j \cdot k. \tag{24}$$

Thus, bias is estimated from the quantization weights by weighting each opinion score k on the quality scale with the subject's tendency to overuse or underuse that opinion score.

We then compared this bias, estimated from the quantization weights, to the bias estimated according to the latest ITU recommendations, i.e., ITU-T P.910 [37], and a recent algorithm, ZREC [35]. The results, shown in Table VI, indicate that the bias computed from the scale quantization weights is strongly correlated with the bias obtained using state-of-the-art approaches.

Thus, the quantization weights of a subject contain all the information needed to compute the bias of that subject. However, the reverse is not true, i.e., the information contained in the quantization weights cannot be recovered if only the bias is known. In fact, it is evident from Eq. (24) that, for a given bias b_j, there are infinitely many ways to choose the values of the scale quantization weights $\nu_k^j, \;\; k = 1, 2, \ldots, 5$. Thus, the bias does not uniquely determine the scale quantization weights.

The previous discussion suggests that scale quantization weights provide new information that bias alone does not capture. For instance, beyond identifying which subjects are biased, dataset creators might also want to understand the portion of the left part of the quality scale disregarded by positively biased subjects in their experiments. Clearly, the larger this portion, the more problematic the presence of biased subjects becomes.

Figure 7 shows the scale quantization weights of the 100 subjects with the largest positive bias in each crowdsourcing dataset. For instance, in the NIVD and MovieLens-1M datasets, positive bias mainly arose from the overuse of "5" rather than a consistent preference for the right part of the quality scale, as observed in the KADID dataset. In the KonViD-1k and KonIQ datasets, positively biased subjects used the scores "1" and "2" more or less as expected but underused the score "3", replacing it with "4" or "5". This behavior resulted in a slight overuse of "4" and "5".

The analysis in Figure 7 therefore allows dataset creators to infer the origin of the bias in order to determine: i) if it might derive from the test settings, as in the case of the KonViD-1k and KonIQ datasets, where the use of a scale with fewer opinion scores could have helped; ii) if it derives from the intention of some "lazy" subjects to score almost every stimulus whose quality is not strongly impaired as '5', as observed in the NIVD dataset; iii) if the bias can be effectively corrected, as in the case of the KADID dataset, where it is a slight and consistent shift of the opinion scores toward the right.

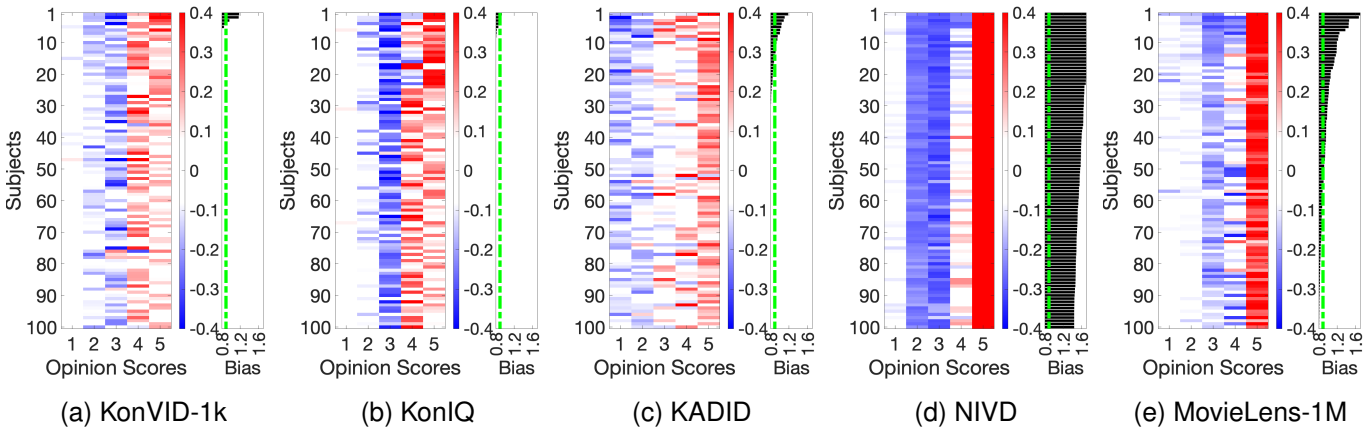

Fig. 7. Scale quantization weights of the top 100 positively biased subjects in each crowdsourcing dataset. Subjects are sorted in decreasing bias order. The dashed green line ($b = 0.972$) represents the maximum value of bias in experiments conducted in controlled environments. Bias was computed as per the ITU-T P.910 recommendation.

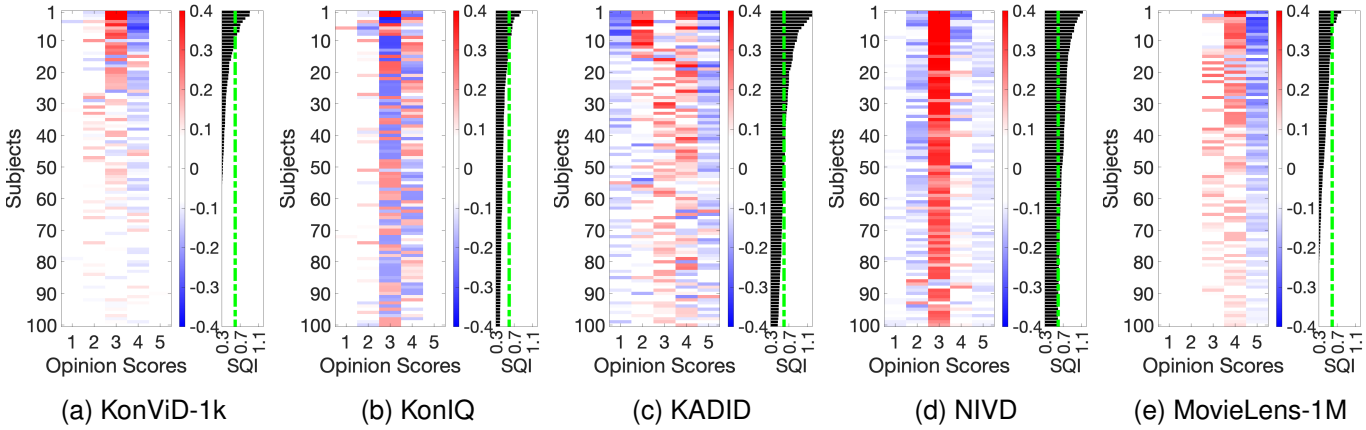

Fig. 8. Scale quantization weights of unbiased and not inconsistent subjects sorted in decreasing order of the subjects' SQI. The dashed green line represents the threshold τ of the SQI defined in Eq. (25).

D. Advancing Peculiar Scoring Patterns Detection with the SQI

In general, in state-of-the-art approaches for peculiar behavior detection in subjective media quality assessment, the greater the subject's inconsistency and the absolute value of their bias, the more peculiar the subject is considered. In this section, we aim to identify a new type of peculiar subjects, i.e., those who are neither inconsistent nor biased, but still peculiar due to their improper use of the quality scale.

To identify subjects who are neither biased nor inconsistent, we computed the bias and inconsistency of each subject according to ITU-T P.910 [37]. We then selected subjects who could be considered unbiased when using a five-point quality scale (absolute value of bias < 0.25 [11]) and with low inconsistency, i.e., smaller than the average inconsistency value computed from the five datasets gathered in controlled environments.

The selected subjects in each crowdsourcing dataset were then ordered in decreasing order of their Scale Quantization Index (SQI) value, and heatmaps of the scale quantization weights of the 100 subjects with the largest SQI values are shown in Figure 8.

By examining each heatmap in Figure 8 from top to bottom, it can be observed that unbiased and consistent subjects who strongly quantize the quality scale are generally "central" scorers. In almost all cases, the preferred opinion scores by the subjects are "2", "3", or "4". For instance, in the NIVD dataset, the analysis revealed subjects who are peculiar due to a high tendency to score the quality as "3". This behavior typically arises from insufficient attention to the task or a desire to complete the test quickly without providing ratings that can be trivially judged as wrong. In the MovieLens-1M dataset, the analysis highlighted subjects who exhibit a peculiar scoring behavior characterized by a consistent misuse of the highest opinion score. Specifically, while they appropriately utilize the first three opinion scores (with scale quantization weights close to 0), they fail to distinguish between "4" and "5", using only "4" even when "5" would be more appropriate.

Test designers are therefore strongly encouraged to replicate the analysis described here on their own datasets. Specifically, they should plot the heatmap of the scale quantization weights of unbiased and consistent subjects in decreasing SQI order to check for the presence of central scorers. They may wish to exclude or correct their ratings depending on the application for which the dataset is intended. For instance, if the dataset is meant for training an objective measure, the presence of central scorers is highly problematic, as they skew the distribution of subjective scores (typically considered as the ground truth) toward the center of the scale, making the distribution of labels imbalanced. In particular, note that machine learning algorithms are known to not generalize well when trained on datasets in which the labels are not uniformly distributed [25], [56].

We also investigated whether the algorithm proposed in the ITU-R BT.500 recommendation for peculiar subject exclusion is capable of recognizing and excluding unbiased and consistent subjects who highly quantize the quality scale during the experiment. In order to compare CrowdRMLE to BT.500, we needed a threshold for the SQI above which CrowdRMLE would recommend excluding an unbiased and consistent subject. In particular, denoting by $\mathbf{SQI_{lab}}$ the sample of all SQI values for the 125 subjects who participated in the five experiments conducted in controlled environments (see Table I), we defined this threshold using the Tukey outlier detection formula:

$$\tau = q_{0.75} + 1.5 \cdot (q_{0.75} - q_{0.25}), \tag{25}$$

where $q_{0.25}$ and $q_{0.75}$ denote the first and third quartiles of $\mathbf{SQI_{lab}}$ respectively. Thus, from each crowdsourcing dataset we recommend excluding any unbiased and consistent subjects whose SQI is an outlier of the distribution of SQI values observed in controlled environments.

Table VII shows the number of unbiased and consistent subjects that CrowdRMLE and BT.500 suggest to exclude from each of the five crowdsourcing datasets used in our analysis. It can be observed that BT.500 struggles to recognize peculiarity not induced by bias and inconsistency. In fact, it only excluded a total of 6 subjects, of which only one belongs to the set of 178 subjects excluded by CrowdRMLE.

Thus, the peculiar subjects identified by CrowdRMLE in this section are not only overlooked by state-of-the-art ap-

TABLE VII
NUMBER OF SUBJECTS IDENTIFIED AS EXHIBITING PECULIAR SCORING BEHAVIOR BY THE ITU-R BT.500 RECOMMENDED ALGORITHM AND CROWDRMLE AMONG THE UNBIASED AND CONSISTENT SUBJECTS.

Datasets	ITU-R BT.500	CrowdRMLE	Common
KonViD-1k	0	8	0
KonIQ	0	12	0
KADID	5	56	1
NIVD	0	91	0
MovieLens-1M	1	11	0
Total	6	178	1

proaches relying on bias and inconsistency (since, by definition, they are unbiased and consistent), but also by a well-known and widely used algorithm such as BT.500 for peculiar scorer detection in media quality assessment.

VI. CONCLUSION

This paper proposes CrowdRMLE, an approach to analyze subject scoring behavior in crowdsourcing subjective tests for image and video quality assessment. We formulated the optimization problem underlying CrowdRMLE and derived an analytical expression of its optimal solution. This makes CrowdRMLE a very efficient approach, as this analytical expression allows the computation of the CrowdRMLE parameters without the need of a computationally complex search for the optimal solution. Experiments conducted on 10 different datasets show that CrowdRMLE: i) is more accurate and robust to noise than other state-of-the-art approaches; ii) provides important insights into the sources of subject bias and inconsistency as well as guidance on how to improve future crowdsourcing test methodologies in image and video quality assessment; iii) can detect peculiar scorers that are overlooked by previous approaches.

REFERENCES

[1] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, *Factors Influencing Quality of Experience*. Cham: Springer International Publishing, 2014, pp. 55–72.

[2] X. Min, H. Duan, W. Sun, Y. Zhu, and G. Zhai, "Perceptual video quality assessment: a survey," *Science China Information Sciences*, vol. 67, no. 11, p. 211301, Oct. 2024.

[3] ITU-T PSTR-CROWDS, "Subjective evaluation of media quality using a crowdsourcing approach," May 2018.

[4] H. Lin, G. Chen, M. Jenadeleh, V. Hosu, U.-D. Reips, R. Hamzaoui, and D. Saupe, "Large-scale crowdsourced subjective assessment of picturewise just noticeable difference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5859–5873, 2022.

[5] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Online HodgeRank on random graphs for crowdsourceable QoE evaluation," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 373–386, Feb. 2014.

[6] Q. Xu, M. Yan, C. Huang, J. Xiong, Q. Huang, and Y. Yao, "Exploring Outliers in Crowdsourced Ranking for QoE," in *Proc. of 25th Intl. Conf. on Multimedia*. Mountain View, CA, USA: ACM, Oct. 2017, pp. 1540–1548.

[7] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014.

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

[9] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Prague, Czech Republic: IEEE, 2011, pp. 2416–2419.

[10] S.-H. Kim, H. Yun, and J. S. Yi, "How to filter out random clickers in a crowdsourcing-based study?" in *BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*. Seattle, Washington, USA: ACM, Oct. 2012, pp. 1–7.

[11] Z. Li, C. G. Bampis, L. Janowski, and I. Katsavounidis, "A simple model for subject behavior in subjective experiments," *Electronic Imaging*, vol. 32, no. 11, pp. 131–1–131–14, 2020.

[12] L. Fotio Tiotsop, A. Servetti, M. Barkowsky, and E. Masala, "Modeling subject scoring behaviors in subjective experiments based on a discrete quality scale," *IEEE Trans. on Multimedia*, vol. 26, pp. 8742–8757, Mar. 2024.

[13] L. Fotio Tiotsop, T. Mizdos, E. Masala, M. Barkowsky, and P. Pocta, "How to train no reference video quality measures for new coding standards using existing annotated datasets?" in *23rd Intl. Workshop on Multimedia Signal Processing (MMSP)*. Tampere, Finland: IEEE, 2021, pp. 1–6.

[14] L. Fotio Tiotsop, F. Agboma, G. Van Wallendael, A. Aldahdooh, S. Bosse, L. Janowski, M. Barkowsky, and E. Masala, "On the link between subjective score prediction and disagreement of video quality metrics," *IEEE Access*, vol. 9, pp. 152 923–152 937, 2021.

[15] M. Cheon and J.-S. Lee, "Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 7, pp. 1467–1480, 2018.

[16] L. Fotio Tiotsop, A. Servetti, and E. Masala, "Full reference video quality measures improvement using neural networks," in *Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 2737–2741.

[17] L. Po, M. Liu, W. Y. F. Yuen, Y. Li, X. Xu, C. Zhou, P. H. W. Wong, K. W. Lau, and H. Luk, "A novel patch variance biased convolutional neural network for no-reference image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1223–1229, 2019.

[18] L. Fotio Tiotsop, E. Masala, A. Aldahdooh, G. V. Wallendael, and M. Barkowsky, "Computing quality-of-experience ranges for video quality estimation," in *Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, Jun. 2019, pp. 1–3.

[19] M. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," in *SPIE Video Communications and Image Processing Conference*, 2003, pp. 8–11.

[20] L. Krasula, Y. Baveye, and P. Le Callet, "Training objective image and video quality estimators using multiple databases," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 961–969, 2020.

[21] L. Fotio Tiotsop, T. Mizdos, M. Uhrina, P. Pocta, M. Barkowsky, and E. Masala, "Predicting single observer's votes from objective measures using neural networks," *Electronic Imaging*, vol. 32, no. 11, pp. 130–1–130–8, Jan. 2020.

[22] L. Fotio Tiotsop, T. Mizdos, M. Barkowsky, P. Pocta, A. Servetti, and E. Masala, "Mimicking individual media quality perception with neural network based artificial observers," *ACM Trans. on Multimedia Computing, Communications, and Applications*, vol. 18, no. 1, p. 1–25, 2022.

[23] L. Fotio Tiotsop, A. Servetti, M. Barkowsky, P. Pocta, T. Mizdos, G. Van Wallendael, and E. Masala, "Predicting individual quality ratings of compressed images through deep CNNs-based artificial observers," *Signal Processing: Image Communication*, vol. 112, p. 116917, Mar. 2023.

[24] L. Fotio Tiotsop, A. Servetti, and E. Masala, "Opinion score distribution prediction via AI-based observers in media quality assessment," in *IEEE 18th Intl. Conf. on Application of Information and Communication Technologies (AICT)*. IEEE, Sep. 2024, pp. 1–6.

[25] L. Fotio Tiotsop, T. Mizdos, M. Uhrina, M. Barkowsky, P. Pocta, and E. Masala, "Modeling and estimating the subjects' diversity of opinions in video quality assessment: A neural network based approach," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3469–3487, Jan. 2021.

[26] R. R. R. Rao, S. Göring, and A. Raake, "Towards high resolution video quality assessment in the crowd," in *13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, Jun. 2021, pp. 1–6.

[27] A. Laghari, H. He, A. Khan, R. Laghari, S. Yin, and J. Wang, "Crowdsourcing platform for QoE evaluation for cloud multimedia services," *Computer Science and Information Systems*, vol. 19, no. 3, p. 1305, 2022.

[28] A. Goswami, A. Ak, W. Hauser, P. Le Callet, and F. Dufaux, "Reliability of crowdsourcing for subjective quality evaluation of tone mapping operators," in *23rd Intl. Workshop on Multimedia Signal Processing (MMSP)*. IEEE, Oct. 2021, pp. 1–6.

[29] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Intl. Symp. on Multimedia*. Dana Point, CA, USA: IEEE, Dec. 2011, pp. 494–499.

[30] C.-C. Wu, K.-T. Chen, Y.-C. Chang, and C.-L. Lei, "Crowdsourcing multimedia QoE evaluation: A trusted framework," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1121–1137, Aug. 2013.

[31] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proceedings of the 17th ACM International Conference on Multimedia*. Beijing, China: ACM, 2009, pp. 491–500.

[32] A. Altieri, L. Fotio Tiotsop, and G. Valenzise, "Subjective media quality recovery from noisy raw opinion scores: A non-parametric perspective," *IEEE Transactions on Multimedia*, vol. 26, pp. 9342–9357, Apr. 2024.

[33] J. Li, S. Ling, J. Wang, and P. Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *28th Intl. Conf. on Multimedia*. Seattle, WA, USA: ACM, Oct. 2020, pp. 3339–3347.

[34] L. Fotio Tiotsop, A. Servetti, and E. Masala, "A scoring model considering the variability of subjects' characteristics in subjective experiments," in *15th International Conference on Quality of Multimedia Experience (QoMEX)*. Ghent, Belgium: IEEE, Jun. 2023, pp. 43–48.

[35] J. Zhu, A. Ak, P. Le Callet, S. Sethuraman, and K. Rahul, "ZREC: Robust recovery of mean and percentile opinion scores," in *International Conference on Image Processing (ICIP)*. Kuala Lumpur, Malaysia: IEEE, Oct. 2023, pp. 2630–2634.

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

[37] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," Oct. 2023.

[38] T. Hoßfeld, M. Seufert, and B. Naderi, "On inter-rater reliability for crowdsourced QoE," in *13th Intl. Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2021, pp. 37–42.

[39] J. Li and P. Le Callet, "Improving the discriminability of standard subjective quality assessment methods: A case study," in *10th Intl. Conf. on Quality of Multimedia Experience (QoMEX)*, May 2018, pp. 1–3.

[40] L. Fotio Tiotsop, A. Servetti, P. Pocta, G. Van Wallendael, M. Barkowsky, and E. Masala, "Multiple image DNN modeling individual subject quality assessment," *ACM Trans. on Multimedia Computing, Communications, and Applications*, vol. 20, no. 8, pp. 1–27, Jun. 2024.

[41] D. Saupe and S. H. Del Pin, "National differences in image quality assessment: An investigation on three large-scale IQA datasets," in *16th International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2024, pp. 214–220.

[42] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Data Compression Conference (DCC)*. Snowbird, UT, USA: IEEE, Apr. 2017, pp. 52–61.

[43] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[44] VQEG, "Report on the validation of video quality models for high definition video content (v. 2.0)," https://bit.ly/2Z7GWDI, Jun. 2010, (retrieved Jan 15th, 2025).

[45] L. Fotio Tiotsop, A. Servetti, and E. Masala, "Investigating prediction accuracy of full reference objective video quality measures through the ITS4S dataset," *Electronic Imaging*, vol. 32, pp. 93–1–93–1, 2020.

[46] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, Germany, May 2017, pp. 1–6.

[47] C. G. Bampis, L. Krasula, Z. Li, and O. Akhtar, "Measuring and predicting perceptions of video quality across screen sizes with crowdsourcing," in *15th Intl. Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2023, pp. 13–18.

[48] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 4, pp. 1–19, Dec. 2015.

[49] L. Fotio Tiotsop, A. Servetti, M. Barkowsky, and E. Masala, "RMLE software," 2024, accessed: 2025-01-15. [Online]. Available: https://github.com/polito-img/RMLE-quality

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

[51] T. Lumley, P. Diehr, S. Emerson, and L. Chen, "The importance of the normality assumption in large public health data sets," *Annual review of public health*, vol. 23, no. 1, pp. 151–169, 2002.

[52] M. W. Fagerland, "t-tests, non-parametric tests, and large studies—a paradox of statistical practice?" *BMC medical research methodology*, vol. 12, no. 1, p. 78, 2012.

[53] E. Skovlund and G. U. Fenstad, "Should we always choose a nonparametric test when comparing two apparently nonnormal distributions?" *Journal of clinical epidemiology*, vol. 54, no. 1, pp. 86–92, 2001.

[54] H. Narimanzadeh, A. Badie-Modiri, I. G. Smirnova, and T. H. Y. Chen, "Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method," *Proc. of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, Oct. 2023.

[55] 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.

[56] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

Lohic Fotio Tiotsop received his M.Sc. degree in Mathematical Engineering in 2017 and his Ph.D. in Control and Computer Engineering in 2021, both from the Politecnico di Torino, Italy. He is currently an Assistant Professor in the Department of Control and Computer Engineering at the same institution. His research focuses on the development and application of statistical approaches, machine learning, and deep learning models to predict and evaluate media quality as perceived not only on average by groups of users, but also at the individual level, accounting for variability across different user experiences. His work therefore aims to optimize user Quality of Experience (QoE) across a variety of scenarios, including on-demand content streaming and video conferencing.

Antonio Servetti has been an Assistant Professor with the Department of Control and Computer Engineering at Politecnico di Torino, Italy, since 2007. His research interests lie in the processing and transmission of multimedia content in real-time Internet communication scenarios. His studies focus on the optimization of quality of experience and end-to-end latency, particularly in the context of Web-based video conferencing and adaptive streaming applications. His recent work explores the integration of artificial intelligence techniques to improve media transmission and to ensure high visual and auditory quality under varying conditions.

Enrico Masala (M'01 SM'14) received his Ph.D. degree in Computer Engineering in 2004 at the Politecnico di Torino, Italy, where he is currently associate professor. His main research interests include video coding, perceptual quality optimization, and network-aware multimedia streaming. More recently his work focused on investigating techniques to analyze the perceived media quality by means of neural network approaches and the use of large scale datasets. 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).