UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Fundamental advantages of considering quality of experience distributions over mean opinion scores

Michael Seufert

Angaben zur Veröffentlichung / Publication details:

Seufert, Michael. 2019. "Fundamental advantages of considering quality of experience distributions over mean opinion scores." In *Eleventh International Conference on Quality of Multimedia Experience (QoMEX), 5-7 June 2019, Berlin, Germany*, edited by Sebastian Möller and Sebastian Egger-Lampl, 1–6. Piscataway, NJ: IEEE. https://doi.org/10.1109/qomex.2019.8743296.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
Deutsches Urheberrecht
Weitere Informationen finden Sie unter: / For more information see:
https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/

Fundamental Advantages of Considering Quality of Experience Distributions over Mean Opinion Scores

Michael Seufert
University of Würzburg, Würzburg, Germany
michael.seufert@uni-wuerzburg.de

***Abstract*—Due to biased assumptions on the underlying ordinal rating scale in subjective Quality of Experience (QoE) studies, Mean Opinion Score (MOS)-based evaluations provide results, which are hard to interpret and can be little meaningful. This paper proposes to consider the full QoE distribution for evaluating and reporting QoE results instead of only using MOS values. The QoE distribution can be represented in a concise way by using the parameters of a multinomial distribution without losing any information about the underlying QoE ratings, and even keeps backward compatibility with previous, biased MOS-based results. Considering QoE results as a realization of a multinomial distribution allows to rely on a well-established theoretical background, which enables meaningful evaluations also for ordinal rating scales. Exemplary evaluations are described in this work, which demonstrate these fundamental advantages of considering QoE distributions over MOS-based evaluations.**

I. Introduction

The concept of Quality of Experience (QoE) [1] constitutes a major research field, which aims to understand and improve the subjective perception of the quality of a networked service as a whole by the end user. It is widely recognized that the QoE is influenced by different QoE factors, which are characteristics of the user, system, service, application, or context [1]. In order to identify these factors and quantify their influence on the QoE of a service, extensive subjective studies have to be conducted. In these studies, users typically assess their experience with a given stimulus on a rating scale, such as the Absolute Category Rating (ACR) scale [2], which will be considered in the remainder of this work. The ACR scale allows to quantify the user experience as one of five values ranging from 1 (bad) to 5 (excellent). Then, the numerical values of the ratings are typically aggregated by using the arithmetic mean to obtain the Mean Opinion Score (MOS), which has attracted a very high popularity and is widely used as the de facto QoE metric in both industry and academia.

However, the major pitfall of this kind of QoE evaluations is the underlying assumption about the mapping of QoE to the rating scale. When conducting a subjective user study, user ratings are actually collected on a categorical scale, hence the name "Absolute Category Rating", which allows to indicate the subjective QoE as one of five categories, namely, "bad", "poor", "fair", "good", or "excellent". As the different categories can be sorted according to the QoE, i.e., "bad" $<$ "poor" $<$ "fair" $<$ "good" $<$ "excellent", this rating scale also represents an ordinal scale. Although the numerical values associated to the categories might suggest so, however, the rating scale is not an interval scale as the elements of the scale cannot be included into arithmetic operations. The reason is that, while some differences might look numerically equidistant, the corresponding differences between categories might not be actually equal. In particular for QoE ratings, it is unclear and highly questionable if, e.g., the difference in user experience between "bad" (1) and "poor" (2) is the same as between "fair" (3) and "good" (4).

Given that the rating scale of a subjective user study is not an interval scale, averaging ratings by using the arithmetic mean is not an interpretable quantity. As a measure of central tendency, ordinal scales only allow to compute the mode, i.e., the category with the highest number of ratings, as well as the median, which is the 50-percentile of the ratings, i.e., the category, for which 50% of the ratings are lower or equal. If the ratings of a subjective study are nevertheless aggregated in terms of arithmetic mean to a MOS, the implicit assumption is introduced that the differences between numerical values represent the actual differences in QoE. This would imply that all the differences in experience between adjacent QoE rating categories are equal, which is a substantial bias and can lead to systematic errors, e.g., [3].

When quantifying QoE differences or QoE improvements of different stimuli, often differences of MOS values are reported, e.g., the MOS value of stimulus B is by x larger than the MOS value of stimulus A. However, these differences between MOS values face the same issues as differences between the rating categories, and are not a meaningful metric. Other works continue to quantify QoE improvements also in terms of percentages of MOS, e.g., stimulus B has a MOS improvement of $x\%$ over stimulus A. However, such operation would be only interpretable on a ratio scale, which requires an absolute zero, and thereby, allows to compute multiplications and ratios of quantities. Still, an absolute zero for experience is hard to find, and the definition of ratios between categories has strange effects, such that, for example, a MOS increase of 100% is an increase of one category when having "bad" (1) as baseline, but an increase of two categories when considering "poor" (2) as baseline. Consequently, this would allow for highly questionable interpretations that, for example, a "good" (4) experience is two times better than "poor" (2) experience, or four times better than "bad" (1) experience. Therefore, the expression of QoE differences in terms of MOS ratios is also not a meaningful quantity.

This paper proposes to consider the full QoE distribution over the ordinal rating categories for evaluating and reporting QoE results instead of using MOS-based metrics. The QoE distribution can be represented in a concise way by using the parameters of a multinomial distribution without losing any information about the underlying QoE ratings, and even keeps backward compatibility with previous, biased MOS-based results. Considering QoE results as a realization of a multinomial distribution allows to rely on a well-established theoretical background, which has various options for more meaningful evaluations. These advantages over MOS-based evaluations are outlined in this work with the help of examples.

Therefore, the remainder of the paper is organized as follows. Section II describes related work on QoE and MOS fundamentals. Section III introduces the theoretical background on multinomial distributions, from which QoE distributions form a small subset. Applications for QoE evaluations based on QoE distributions are described in Section IV, showing their advantages over MOS-based evaluations. Finally, Section V concludes this paper.

II. Related Work

A comprehensive definition of QoE was given in [1] including influence factors of QoE, such as human, system, and context influence factors. However, it was not specified how QoE assessment should be conducted. After a variety of practical implementations in a multitude of studies, cf., e.g., [4]–[6], an overview document was provided in [7], which links to several recommendations for QoE assessment for particular services, such as web browsing [8] or multimedia applications [2]. Here, [7] names MOS as a QoE metric, although it recognizes that test methods can be classified according the applied scaling method and scale level, i.e., nominal, ordinal, interval, and ratio. However, the linked documents might lack this awareness, such as [2], which recommends the usage of the 5-point ACR scale, from which MOS, confidence intervals, and standard deviations shall be computed. However, as the ACR scale is an ordinal scale, but not an interval scale, these metrics are not interpretable without introducing substantial bias. [9] compared the classical assessment of user satisfaction based on MOS with the notion of acceptability of service quality. Evaluation methods are reviewed and differences between both perspectives on QoE assessment are discussed.

Substantial contributions towards improving QoE assessment beyond the MOS were started in [10], which emphasizes that MOS values lose considerable amount of information about the QoE ratings. To overcome this issue, the authors suggested to additionally consider the standard deviation of opinion scores (SOS). However, SOS values face the same substantial bias as MOS, as it is implicitly assumed that the rating scale of user experience is an interval scale. The work in [10] was extended in [11], in which quantiles, entropy, and probability distribution were added to a recommended set of QoE descriptors. In contrast to MOS and SOS, the newly added descriptors do not face the issues that were previously discussed. Additionally, [11] postulated the idea that individual ratings for a single test condition can be described as realizations of a binomial distribution. [12] continued the previous works and elaborated more on the value of quantiles and acceptance thresholds, such as percentage of Poor-or-Worse (%PoW) and Good-or-Better (%GoB). [13] modeled an individual user rating with a truncated normal distribution. Most recently, the concept of QoE was extended to QoE fairness [14], i.e., the notion that users in a shared system should experience a fair QoE distribution. The proposed fairness metric is based on the standard deviation of individual QoE ratings, which is again the SOS. Thus, the fairness metric also inherits the problems of SOS, which were described above.

This work will avoid this pitfall of QoE assessment by considering that all ratings of a test condition follow a multinomial distribution on the ordinal rating categories, which also takes a more holistic perspective of the subjective user study. Consequently, it also allows to obtain all the previously proposed metrics, which is shortly discussed. Additionally, this theoretical framework provides several advantages over MOS-based QoE evaluations, such as simple testing and quantification of QoE differences between two QoE distributions.

III. Theoretical Background on QoE Distributions

This section introduces QoE distributions as a subset of multinomial distributions and shortly recaps the theoretical background. Afterwards, it is outlined how previously used MOS-based evaluations could be obtained from QoE distributions. However, except for some backward compatibility, this would not be recommended due to the inherent bias when applied to QoE ratings on ordinal scales.

A. Multinomial Distributions

Multinomial distributions describe probabilities in an experiment where n balls are drawn with replacement from a bag with balls of k different colors. The probability that a ball of color i is drawn is p_i with $\sum_{i=1}^{k} p_i = 1$. The random variables X_i count how often a ball of color i is drawn. Then, the probability mass function of the multinomial distribution is given as:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \\ = \begin{cases} \frac{n!}{x_1! \cdot x_2! \cdot \ldots \cdot x_k!} \cdot p_1^{x_1} \cdot p_2^{x_2} \cdot \ldots \cdot p_k^{x_k}, \text{ when } \sum_{i=1}^{k} x_i = n, \\ 0, \text{ otherwise.} \end{cases} \tag{1}$$

Thus, Equation 1 describes the joint probability for all $i = 1, \ldots, k$ that in an experiment, in which n balls are drawn with replacement, $X_i = x_i$ balls are drawn with color i.

B. QoE Distributions

This experiment, which constitutes multinomial distributions, can be easily mapped to QoE studies, in which n participants rate the QoE of a stimulus. There are k categories on the rating scale, and the numbers X_i count the participants, which

rate category i. The parameters p_i describe the underlying and hidden probability that the presented stimulus gives an experience in category i. In case of the 5-point ACR scale, which is considered in the remainder of this work, $k = 5$ and i represents the numerical value assigned to the rating categories, i.e., "bad" ($i = 1$), "poor" ($i = 2$), "fair" ($i = 3$), "good" ($i = 4$), and "excellent" ($i = 5$). Thus, the result of a QoE study $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)$ is a realization of a QoE distribution, which comprise a subset of multinomial distributions.

The vector notation $\boldsymbol{x}$ is a very compact and concise way to report the results of a QoE study, and allows to fully make use of the advantages of considering QoE distributions. From this representation, also the underlying parameters of the QoE distribution p_i can be estimated using a maximum likelihood approach. Thereby, the estimated parameters $\hat{p}_i$ can be obtained as:

$$\hat{p}_i = \frac{x_i}{n} = \frac{x_i}{\sum_{j=1}^{5} x_j}, i = 1, \ldots, 5. \tag{2}$$

Following Equation 2, the outcome of a QoE study can also be reported with another compact representation $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5, n)$, from which one of the $\hat{p}_i$ could be omitted as $\sum_{i=1}^{5} \hat{p}_i = 1$. Obviously both representations $\boldsymbol{x}$ and $\hat{\boldsymbol{p}}$ can be easily converted into the other representation.

The compact representations allow to compute quantiles easily, which are a meaningful metric for ordinal scales. Let $\hat{\boldsymbol{c}} = (\hat{c}_1, \hat{c}_2, \hat{c}_3, \hat{c}_4, \hat{c}_5, n)$ be the vector containing cumulative probabilities computed from $\hat{\boldsymbol{p}}$, i.e., $\hat{c}_i = \sum_{j=1}^{i} \hat{p}_j$. Note that $\hat{\boldsymbol{c}}$ is also a representation equivalent to $\boldsymbol{x}$ and $\hat{\boldsymbol{p}}$. Then, the q-quantile Q_q is the category i given by:

$$Q_q = \min\{i | \hat{c}_i \geq q\}. \tag{3}$$

Moreover, it is possible to directly compute a more intuitive percentage of Poor-or-Worse (%PoW) and Good-or-Better (%GoB) , which is different from the previous definition based on the E-model [12]. This means, it is possible to literally obtain the %PoW as the percentage of users who rated the category "poor" (2) or worse, i.e., "bad" (1), and also the %GoB as the percentage of users who rated the category "good" (4) or better, i.e., "excellent" (5):

$$\%PoW = \hat{c}_2 \cdot 100\%, \ \%GoB = (1 - \hat{c}_3) \cdot 100\%. \tag{4}$$

C. Backward Compatibility towards MOS-based Evaluations

Although MOS-based evaluations face the issues described above, for the sake of backward compatibility, MOS-based QoE metrics can be computed from QoE results expressed as QoE distributions. In the following, these computations are outlined briefly.

First, the sample mean of ratings, or MOS value, can be obtained from $\boldsymbol{x}$ or $\hat{\boldsymbol{p}}$ as follows:

$$MOS = \frac{\sum_{i=1}^{5} i \cdot x_i}{\sum_{i=1}^{5} x_i} = \frac{\sum_{i=1}^{5} i \cdot x_i}{n} = \sum_{i=1}^{5} i \cdot \hat{p}_i. \tag{5}$$

TABLE I: Exemplary QoE distributions from conducted study.

QoE Distribution	MOS	SOS	$CI_{MOS}^{0.95}$	F
$S1 = (48, 20, 4, 3, 0) =$ $(0.64, 0.27, 0.05, 0.04, 0.00, 75)$	1.49	0.78	$[1.32; 1.67]$	0.61
$S_2 = (11, 25, 18, 7, 1) =$ $(0.18, 0.40, 0.29, 0.11, 0.02, 62)$	2.39	0.96	$[2.15; 2.63]$	0.52
$S_3 = (13, 15, 16, 21, 3) =$ $(0.19, 0.22, 0.24, 0.31, 0.04, 68)$	2.79	1.20	$[2.51; 3.08]$	0.40

The sample standard deviation of ratings, or SOS value [10], is given by:

$$SOS = \sqrt{\frac{x_i \cdot (i - MOS)^2}{(\sum_{i=1}^{5} x_i) - 1}} = \sqrt{\frac{n}{n-1} \cdot \hat{p}_i \cdot (i - MOS)^2}. \tag{6}$$

The confidence interval (CI) of the MOS for a confidence level of $1 - \alpha$ can be computed for large enough n (cf. central limit theorem) using the $(1 - \frac{\alpha}{2})$-quantile of the standard normal distribution $z_{(1-\frac{\alpha}{2})}$:

$$CI_{MOS}^{1-\alpha} = \left[MOS - z_{(1-\frac{\alpha}{2})} \frac{SOS}{\sqrt{n}}; MOS + z_{(1-\frac{\alpha}{2})} \frac{SOS}{\sqrt{n}} \right]. \tag{7}$$

Note that for small sample sizes, the standard normal distribution should be replaced by Student's t distribution. By substituting a desired CI width d in the error margin $\frac{d}{2} = z_{(1-\frac{\alpha}{2})} \frac{SOS}{\sqrt{n}}$ of Equation 7 and solving for n, also required sample sizes n can be easily obtained. Finally, also the QoE fairness index F [14] can be obtained as:

$$F = 1 - \frac{SOS}{2}. \tag{8}$$

Given the inherent bias of these MOS-based evaluations, in the following, improved QoE evaluations will be presented, which leverage the advantages of QoE distributions.

IV. Applications of QoE Distributions

This section presents applications of QoE distributions, which give more meaningful QoE evaluations based on the ordinal rating scales of QoE studies. To demonstrate the improved evaluations, the ratings for three stimuli S_1, S_2, and S_3 are considered, which have been collected in a past crowdsourcing QoE study and have been filtered to exclude unreliable ratings [15]. These exemplary QoE distributions are described in Table I. S_1 has a significantly lower MOS than the other stimuli, but the highest fairness score. S_3 has a higher MOS than S_2, but the 95% CIs overlap, and the fairness score is lower for S_2. The QoE distributions of S_1 (black), S_2 (dark brown) and S_3 (light brown) are also visualized in Figure 1 as PDFs ($\hat{\boldsymbol{p}}$, bars) and CDFs ($\hat{\boldsymbol{c}}$, dashed lines).

A. Confidence Intervals and Sample Size

Equation 2 described the maximum likelihood estimation of each of the parameters p_i of the QoE distribution. To obtain confidence intervals, a binomial confidence interval can be

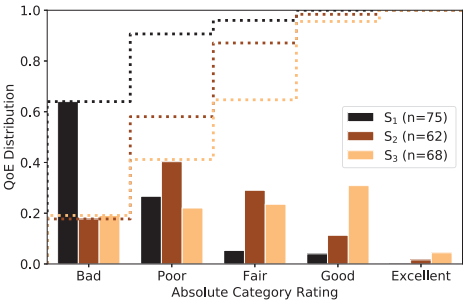

Fig. 1: Exemplary QoE distributions from conducted study.

computed for each parameter $\hat{p}_i$ individually for large enough n (cf. central limit theorem):

$$CI_{\hat{p}_i}^{1-\alpha} = \left[\hat{p}_i - z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}; \hat{p}_i + z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}\right]. \quad (9)$$

This results in five confidence intervals for each parameter p_i of the QoE distribution, such as $CI_{S_1}^{0.95} = ([0.53; 0.75], [0.17; 0.37], [0.00; 0.10], [0.00; 0.10], [0.00; 0.00])$, $CI_{S_2}^{0.95} = ([0.08; 0.27], [0.28; 0.52], [0.18; 0.40], [0.03; 0.19], [0.00; 0.05])$ and $CI_{S_3}^{0.95} = ([0.10; 0.28], [0.12; 0.32], [0.13; 0.34], [0.20; 0.42], [0.00; 0.09])$. Some CIs for the same parameter do not overlap, which indicates that there is significant difference for this parameter on a significance level of 5%, e.g., for p_1 between S_1 and the other two QoE distributions, or for p_4 between S_2 and S_3. Again, Equation 9 also allows to compute sample sizes n_{S_i} for a desired width d_i of $CI_{\hat{p}_i}$, i.e., $CI_{\hat{p}_i} = \left[\hat{p}_i - \frac{d_i}{2}; \hat{p}_i + \frac{d_i}{2}\right]$:

$$n_{S_i} = \frac{4 \cdot z_{(1-\frac{\alpha}{2})}^2 \cdot \hat{p}_i(1-\hat{p}_i)}{d_i^2}. \quad (10)$$

After the sample sizes n_{S_i} have been computed with a desired width d_i for all parameters $\hat{p}_i$, the maximum sample size $n_S = \max_i n_{S_i}$ should be used as the sample size of the entire QoE study. For the considered stimuli, a desired width of $d = 0.1$ for all CIs would result in $n_S = 355$ for S_1, $n_S = 370$ for S_2, and $n_S = 328$ for S_3. Note that simultaneous CIs can be computed following the method presented in [16].

B. Comparison of QoE Results

For comparing different QoE distributions, the concept of stochastic dominance [17] from decision theory can be utilized and transferred. Stochastic dominance describes a partial ordering between random variables. It can indicate if a gamble, i.e., a probability distribution over possible outcomes, is dominant and should be preferred. For QoE distributions, this means, that, if ratings (outcomes) are obtained from a superior QoE distribution, the corresponding stimulus (gamble) should be preferred. Different orders of dominance exist, but as it is a partial ordering, there might not always be a dominant distribution in comparisons of QoE results.

A QoE distribution B with cumulative representation $\hat{\boldsymbol{c}}^B$ has a first-order stochastic dominance (FSD) over a QoE distribution A with $\hat{\boldsymbol{c}}^A$, if:

$$\hat{c}_i^B \leq \hat{c}_i^A, \quad \forall i = 1, \ldots, 5. \quad (11)$$

Intuitively, this FSD of B indicates that the probability of having a rating of at least category i, i.e., $1 - \hat{c}_{i-1}^B$ is higher than the corresponding probability for A, i.e., $1 - \hat{c}_{i-1}^A$, for all categories. A weaker form of dominance is second-order stochastic dominance. QoE distribution B has a second-order stochastic dominance (SSD) over a QoE distribution A, if:

$$\sum_{i=1}^{j} \hat{c}_i^B \leq \sum_{i=1}^{j} \hat{c}_i^A, \quad \forall j = 1, \ldots, 5. \quad (12)$$

The intuitive explanation of SSD is that overall differences in probability mass between B and A are shifted more towards categories with higher QoE, i.e., $\sum_{i=1}^{j} \hat{c}_i^A - \hat{c}_i^B \geq 0$ for all j. Obviously, FSD implies SSD. Note that the definition of SSD in this work avoids the typical definition via integrals, cf. [17], as integrals are not meaningful for ordinal scales. For the exemplary QoE distributions, S_2 and S_3 show FSD over S_1, while for S_2 and S_3, neither FSD nor SSD can be observed in any direction.

C. Testing for Significant QoE Differences

After two QoE distributions have been compared, it has to be tested if there is a significant difference between them. The null hypothesis is that two realizations were drawn from the same QoE distributions. The p-value is the probability of facing the observed or a more extreme realization assuming that the null hypothesis was true. If the p-value is below the significance level α, which is selected by the researcher, the null hypothesis is rejected, and thus, the two QoE distributions are considered as being significantly different.

While many non-parametric statistical tests exist, which compare probability distributions, the Mann-Whitney U test should be considered for ordinal data [18]. It computes the U statistics for both QoE distributions A and B from the ranks of the ratings, considering the number of tied ranks $t_i = x_i^A + x_i^B$:

$$U = \sum_{i=1}^{5} \left(x_i \cdot (1 + \sum_{j=1}^{i-1} t_j + \frac{t_i - 1}{2}) \right) - \frac{n(n+1)}{2}. \quad (13)$$

The smaller of both U values is considered and its significance can be looked up in dedicated tables. For large samples, the standardized value $z_U = \frac{U - \mu_U}{\sigma_U}$ with mean $\mu_U = \frac{n^A \cdot n^B}{2}$ and tie-corrected standard deviation $\sigma_U = \sqrt{\frac{n^A n^B}{12}\left((n^A + n^B + 1) - \sum_{i=1}^{5} \frac{t_i^3 - t_i}{(n^A+n^B)(n^A+n^B-1)}\right)}$ approximately follows a standard normal distribution, and thus, can be compared to the critical values $\pm z_{(1-\frac{\alpha}{2})}$. In the considered QoE study, the p-value for the Mann-Whitney U test between S_2 and S_3 is 0.04 (two-tailed), i.e., the null hypothesis that both QoE distributions are equal has to be rejected on a significance level of $\alpha = 5\%$. The p-values between S_1 and S_2 and between S_1 and S_3 are much smaller ($< 10^{-7}$), thereby, also indicating significant differences. Note that the Kruskal-Wallis test, which is the one-way analysis of variance (ANOVA) on ranks, extends the Mann-Whitney U test to compare multiple QoE distributions.

D. Quantification of QoE Differences

To quantify differences between two QoE distributions, there exist a plethora of statistical distances. Simple examples include the total variation distance $\delta(A,B) = \max_i |\hat{p}_i^A - \hat{p}_i^B|, i = 1,\ldots,5$, which is the largest difference between the probabilities that both distributions assign to the same category, or the Kolmogorov-Smirnov test statistic $D_{KS}(A,B) = \max_i |\hat{c}_i^A - \hat{c}_i^B|, i = 1,\ldots,5$, which is the maximum vertical distance between the corresponding cumulative probability distributions. The widely used Kullback-Leibler divergence D_{KL}, however, is not recommended as it is not a metric. Moreover, if one of the categories was never rated by any users, i.e., its probability is zero, D_{KL} and its derived symmetric versions become ∞, e.g., in S_1 for "excellent" (5).

A more robust and intuitive distance metric is given by the Wasserstein metric, which is also called earth mover's distance D_{EM} [19]. It indicates the minimal amount of probability mass that has to be moved to change the shape and make one probability distribution look exactly the same as the other probability distribution. Obviously, the more different the distributions are, the more probability mass has to be moved, hence, D_{EM} will be larger. A simple formula exists to compute D_{EM} between QoE distributions A and B:

$$D_{EM}(A,B) = \sum_{j=1}^{4} |\hat{c}_j^A - \hat{c}_j^B|. \tag{14}$$

Note that D_{EM} indicates the absolute value of probability mass, which has to be shifted. However, the probability mass is counted for each of the intermediate categories, if it flows between categories that are not adjacent. Thus, it can only be interpreted as the shifted probability mass weighted by the number of categories that it has to be shifted. For example, considering $A = (0,0,0.1,0,0.9)$, $B = (0,0,0,0.2,0.8)$, and $I_5 = (0,0,0,0,1)$, both $D_{EM}(A,I_5) = D_{EM}(B,I_5) = 0.2$. However, in the case of A, it means that a probability mass of 0.1 has to be shifted by two categories, while, in case of B, a probability mass of 0.2 has to be shifted by one category. Note once again that it has to be carefully avoided to interpret these numbers in terms of numerical differences or ratios between QoE rating categories, which is not possible for ordinal rating scales and would again introduce the inherent bias discussed above. This means, for example, that although the above discussed shifts from A to I_5 (0.1 for two categories) and from B to I_5 (0.2 for one category) are numerically equal, they cannot be considered equal in terms of QoE improvement, which is also indicated by the fact that $D_{EM}(A,B) = 0.2 \neq 0$.

For two arbitrary QoE distributions A and B, $\max_{A,B} D_{EM}(A,B) = 4$, which is reached for the distance between $I_1 = (1,0,0,0,0)$ and $I_5 = (0,0,0,0,1)$, i.e., a probability mass of 1 has to be shifted by four categories. Thus, it is possible to normalize the D_{EM} to the unit interval $[0,1]$ by $D_{EM,norm}(A,B) = \frac{1}{4} D_{EM}(A,B)$. For the considered QoE study, it can again be seen from $D_{EM,norm}(S_1,S_2) = 0.22$ and $D_{EM,norm}(S_1,S_3) = 0.33$ that S_1 is not very close to S_2 and S_3. In contrast, $D_{EM,norm}(S_2,S_3) = 0.11$, which confirms that S_2 and S_3 are rather similar. $D_{EM,norm}$ also allows to construct a QoE deficit index QDI of a QoE distribution A. For this, QDI is defined as the normalized distance to the ideal QoE distribution $I_5 = (0,0,0,0,1)$, for which all participants rated an "excellent" (5) experience:

$$QDI(A) = D_{EM,norm}(A,I_5) = \frac{1}{4}\sum_{j=1}^{4} \hat{c}_j^A. \tag{15}$$

QDI is in the unit interval, i.e., a QoE deficit index of 0 indicates an ideal QoE distribution ($A = I_5$), and a QDI of 1 means that A has the worst possible QoE distribution $I_1 = (1,0,0,0,0)$. Also, a corresponding QoE level index QLI of a QoE distribution A can be derived as $QLI(A) = D_{EM,norm}(A,I_1) = 1 - QDI(A)$. As QDI and QLI are based on D_{EM}, the same limitations apply in terms of interpretation. Here again, consider the example discussed for D_{EM} above, which equally applies to QDI. Note that there is also a mathematical relation to MOS via $MOS(A) = 5 - D_{EM}(A,I_5) = 5 - 4 \cdot QDI(A) = 1 + 4 \cdot QLI(A)$. It allows to define MOS based on a distance metric between QoE distributions over ordinal categories, rather than relying on a biased cast of ordinal rating data to an interval scale. Thus, it allows for an unbiased interpretation of MOS in terms of QoE probability masses, which are shifted and weighted by the number of shifted rating categories. Consequently, the ranking of the stimuli S_1, S_2, and S_3 in terms of QLI with $QLI(S_1) = 0.12 < QLI(S_2) = 0.35 < QLI(S_3) = 0.45$ is equivalent to the ranking based on MOS. The ranking and the QLI scores indicate that the highest QoE deficit is in S_1, in terms of the number of ratings and/or number of categories that would have to be shifted to reach an ideal QoE.

Next, the net flow of probability mass $NF_i(A \to B)_i$ from each category i of A towards category $i+1$ of B can be obtained from the terms of the sum in Equation 14:

$$NF_i(A \to B) = \hat{c}_i^A - \hat{c}_i^B,\ i = 1,\ldots,4. \tag{16}$$

Here, a positive $NF_i(A \to B)$ means that probability mass of A flows from category i towards $i+1$ in B, i.e., towards higher QoE. In contrast, if $NF_i(A \to B)$ is negative, A's probability mass flows from category $i+1$ to i in B, i.e., towards lower QoE. Note that, in contrast to D_{EM}, NF_i is signed and directed, such that $NF_i(A \to B) = -NF_i(B \to A)$. This concept also allows to count the number of categories with a positive or negative net flow from A to B and vice versa. At the same time, $NF_i(A \to B)$ also quantifies the net probability mass, which flows between the categories. Confer with Equation 11, which indicates FSD when all $NF_i(A \to B)$ are positive. When all signed net flows are added, the resulting number indicates the net balance, i.e., the overall directed net probability flow from A to B:

$$NB(A \to B) = \sum_{i=1}^{4} NF_i(A \to B). \tag{17}$$

Note the relation to SSD in Equation 12, which follows if all partial sums of $NB(A \rightarrow B)$ are positive. Generally speaking, $NB(A \rightarrow B)$ is a signed number that for positive values indicates a shift of probability mass towards higher QoE categories, such as in the considered example, in which $NB(S_1 \rightarrow S_2) = 0.89$ and $NB(S_2 \rightarrow S_3) = 0.41 > 0$. Again, it is weighted by the number of categories and, as differently signed shifts of probability mass have been canceled out, it should not be interpreted in terms of quantitative differences or ratios between QoE rating categories, which cannot be obtained from ordinal scales.

E. Metric for QoE Fairness

Finally, also the QoE fairness of a QoE distribution can be assessed. For any given QoE distribution A, the closest, perfectly fair QoE distribution I_{m_A} is the monolithic distribution, for which all participants have rated the modal QoE category of A, i.e., the category of A with the highest number of participants. The fair QoE distribution I_m, which has category $m \in \{1, \ldots, 5\}$ as mode, can be described by $p_m = 1$ and $p_i = 0,\ \forall i \neq m$. Consequently, a simple QoE fairness metric F_a can be described by the level of agreement on the modal category normalized to the unit interval:

$$F_a(A) = \frac{5}{4} \cdot (\hat{p}_{m_A} - \frac{1}{5}) = \frac{5}{4} \cdot (\max_i \hat{p}_i - \frac{1}{5}). \qquad (18)$$

The normalization takes into account that, due to the five rating categories, the minimum mode of any QoE distribution is $\frac{1}{5}$. A fairness score of 1 indicates that all participants have rated the same category, while a fairness score of 0 indicates a uniform rating distribution. In the considered example, the QoE distributions reach the following fairness scores: $F_a(S_1) = 0.55$, $F_a(S_2) = 0.25$, and $F_a(S_3) = 0.14$.

This concept of fairness towards a monolithic distribution also allows to define a more advanced QoE fairness score F_d, which is based on the D_{EM} distance between A and its corresponding I_{m_A}. Using $\max_A D_{EM}(A, I_{m_A}) = \frac{7}{3}$, which is the maximum distance between any QoE distribution A and its closest, perfectly fair QoE distribution I_{m_A}, the QoE fairness score can be normalized to the unit interval:

$$F_d(A) = 1 - \frac{D_{EM}(A, I_{m_A})}{\frac{7}{3}} = 1 - \frac{3 \cdot D_{EM}(A, I_{m_A})}{7}. \qquad (19)$$

Here again, a fairness score of 1 indicates perfect fairness of the QoE ratings, i.e., all participants have rated the same category, which is the mode of A. In contrast, a fairness score of 0 indicates the highest unfairness in the QoE ratings in terms of D_{EM}. This is achieved, e.g., for $A = (\frac{1}{3}, \frac{1}{3} - \varepsilon, 0, 0, \frac{1}{3} + \varepsilon)$ with a small $\varepsilon > 0$, which has mode $m = 5$. The distance to the corresponding $I_5 = (0, 0, 0, 0, 1)$ is $D_{EM}(A, I_5) = \frac{7}{3} - 3\varepsilon$, which approaches the maximum value. In the considered QoE study, $F_d(S_1) = 0.79$, $F_d(S_2) = 0.68$, and $F_d(S_3) = 0.45$, i.e., the fairness decreases from S_1 to S_3, with S_1 being closest to a monolithic QoE distribution.

V. Conclusion

This work described the inherent bias in many MOS-based evaluations of QoE studies, which is caused by too simplistic assumptions about the mapping of QoE to the rating scale. Typically QoE studies use only ordinal rating scales, such as the 5-point ACR scale, for which means, differences, and ratios between categorical values are not meaningful. To overcome this issue, this work considered QoE distributions, which can be based on the well-established theoretical framework of multinomial distributions. Exemplary evaluations based on QoE distributions were described, which give meaningful results also for ordinal rating scales, and thus show fundamental advantages over biased MOS-based evaluations. In future works, the concept of QoE distributions has to be extended towards even more applications, e.g., how QoE models can be formulated based on QoE distributions.

References

[1] P. Le Callet, S. Möller, and A. Perkis (eds), "Qualinet White Paper on Definitions of Quality of Experience," Lausanne, Switzerland, Tech. Rep., 2013, version 1.2.

[2] International Telecommunication Union, "ITU-T Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications," 2008.

[3] T. M. Liddell and J. K. Kruschke, "Analyzing Ordinal Data with Metric Models: What Could Possibly Go Wrong?" *Journal of Experimental Social Psychology*, vol. 79, 2018.

[4] M. Alreshoodi and J. Woods, "Survey on QoE\QoS Correlation Models for Multimedia Services," *International Journal of Distributed and Parallel Systems*, vol. 4, no. 3, 2013.

[5] M.-N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnström, and A. Raake, "Quality of Experience and HTTP Adaptive Streaming: A Review of Subjective Studies," in *QoMEX*, 2014.

[6] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, 2015.

[7] International Telecommunication Union, "ITU-T Recommendation G.1011: Reference Guide to Quality of Experience Assessment Methodologies," 2015.

[8] ——, "ITU-T Recommendation P.1501: Subjective Testing Methodology for Web Browsing," 2013.

[9] R. Schatz, S. Egger, and A. Platzer, "Poor, Good enough or Even Better? Bridging the Gap between Acceptability and QoE of Mobile Broadband Data Services," in *ICC*, 2011.

[10] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *QoMEX*, 2011.

[11] T. Hoßfeld, P. E. Heegaard, and M. Varela, "QoE beyond the MOS: Added Value using Quantiles and Distributions," in *QoMEX*, 2015.

[12] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, "QoE beyond the MOS: An In-depth Look at QoE via Better Metrics and their Relation to MOS," *Quality and User Experience*, vol. 1, no. 1, 2016.

[13] T. Hoßfeld, P. E. Heegaard, L. Skorin-Kapov, and M. Varela, "No Silver Bullet: QoE Metrics, QoE Fairness, and User Diversity in the Context of QoE Management," in *QoMEX*, 2017.

[14] T. Hoßfeld, L. Skorin-Kapov, P. Heegaard, and M. Varela, "Definition of QoE Fairness in Shared Systems," *IEEE Communications Letters*, vol. 21, no. 1, 2017.

[15] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, "Quantification of YouTube QoE via Crowdsourcing," in *MQoE*, 2011.

[16] C. P. Sison and J. Glaz, "Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions," *Journal of the American Statistical Association*, vol. 90, no. 429, 1995.

[17] J. Hadar and W. R. Russell, "Rules for Ordering Uncertain Prospects," *The American Economic Review*, vol. 59, no. 1, 1969.

[18] S. Siegel, *Nonparametric Statistics for the Behavioral Sciences*, 1956.

[19] Y. Rubner, C. Tomasi, and L. J. Guibas, "A Metric for Distributions with Applications to Image Databases," in *ICCV*, 1998.