QoE Beyond the MOS: An In-Depth Look at QoE via Better Metrics and their Relation to MOS

Tobias Hoßfeld*, Poul E. Heegaard†, Martín Varela‡, Sebastian Möller§
*University of Duisburg-Essen, Modeling of Adaptive Systems, Essen, Germany
Email: tobias.hossfeld@uni-due.de
†Department of Telematics, Norwegian University of Science and Technology (NTNU)
Email: poul.heegaard@item.ntnu.no
‡ VTT Technical Research Centre of Finland — Communication Systems
Email: Martin.Varela@vtt.fi
§ Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Berlin, Germany
Email: Sebastian.Moeller@telekom.de

***Abstract*—While Quality of Experience (QoE) has advanced very significantly as a field in recent years, the methods used for analyzing it have not always kept pace. When QoE is studied, measured or estimated, practically all the literature deals with the so-called Mean Opinion Score (MOS). The MOS provides a simple scalar value for QoE, but it has several limitations, some of which are made clear in its name: for many applications, just having a mean value is not sufficient. For service and content providers in particular, it is more interesting to have an idea of how the scores are distributed, so as to ensure that a certain portion of the user population is experiencing satisfactory levels of quality, thus reducing churn. In this article we put forward the limitations of MOS, present other statistical tools that provide a much more comprehensive view of how quality is perceived by the users, and illustrate it all by analyzing the results of several subjective studies with these tools.**

I. Introduction

Quality of Experience (QoE) is a complex concept, riddled with subtleties regarding several confluent domains, such as systems performance, psychology, physiology, etc., as well as contextual aspects of where, when and how a service is used. For all this complexity, it is most often treated in the most simplistic way in terms of statistical analysis, basically just looking at averages, and maybe standard deviations and confidence intervals.

In this article, we extend our previous work [1], putting forward the idea that it is necessary to go beyond these simple measures of quality when performing subjective assessments, in order to a) get a proper understanding of the QoE being measured, and b) be able to exploit it fully. We present the reasons why it is important to look beyond the Mean Opinion Score (MOS) when thinking about QoE, as well as other measures that can be extracted from subjective assessment data, why they are useful, and how they can be used.

Our main contribution is in highlighting the importance of the insight found in the uncertainty of the opinion scores. This uncertainty is masked by the MOS, and such an insight will enable the service providers to manage QoE in a more effective way. We propose different approaches to quantify the uncertainty; standard deviation, cumulative density functions (CDF), and quantiles, as well as looking into the impact of different types of rating scales on the results. We provide a formal proof that user diversity of a study can be compared by means of the SOS parameter a independent of the used rating scale. We also look at the relationship between quality and acceptance, both implicitly and explicitly. We provide several examples where going beyond simple MOS calculations allows for a better understanding of how the quality is actually perceived by the user population (as opposed to a hypothetical "average user"). A service provider might be interested e.g. for which conditions at least 95% of the users are satisfied with the service quality – which may be quantified in terms of quantiles. In particular, we take a closer look at the link between acceptance and opinion ratings (for a possible classification of QoE measures, cf. Figure 3). Such behavioral metrics like acceptance are important for service providers to plan, dimension and operate their services. Therefore, it is tempting to establish a link between opinion measurements from subjective QoE studies and behavioral measurements which we approach by defining the θ-acceptability. The analysis of acceptance in relation to MOS values is another key contribution in the article. To cover a variety of relevant applications, we consider speech, video, and web QoE.

The remainder of this article is structured as follows. In Section II we discuss why a more careful statistical treatment of subjective QoE assessments is needed. Section III discusses related work. We present our proposed approach and define the QoE metrics in Section IV, while in Section V we look at several subjective assessment datasets, using other metrics besides MOS in our analysis, and also considering the impact of the scales used. We conclude the article in Section VI, discussing the practical implications of our results.

II. Motivation

A. Objective and Subjective QoE Metrics

It is a common and well-established practice to use MOS [2] to quantify perceived quality, both in the research literature, as well as in practical applications such as QoE models. This is simple and useful for some instances of "technical" evaluation

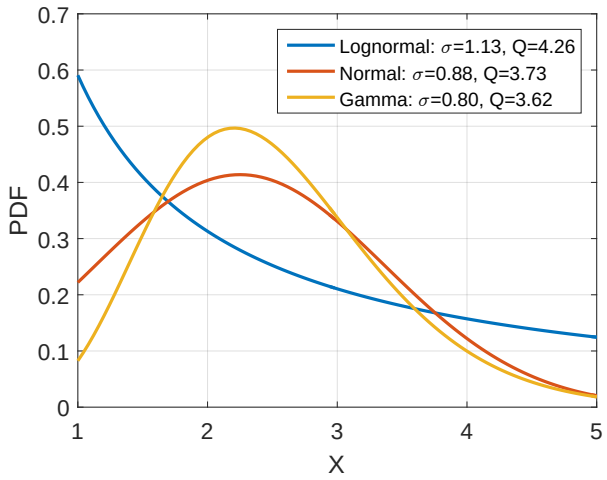


Figure 1: Different continuous distributions with identical mean (2.5) which differ in other measures like standard deviation σ or 90 % quantiles Q.

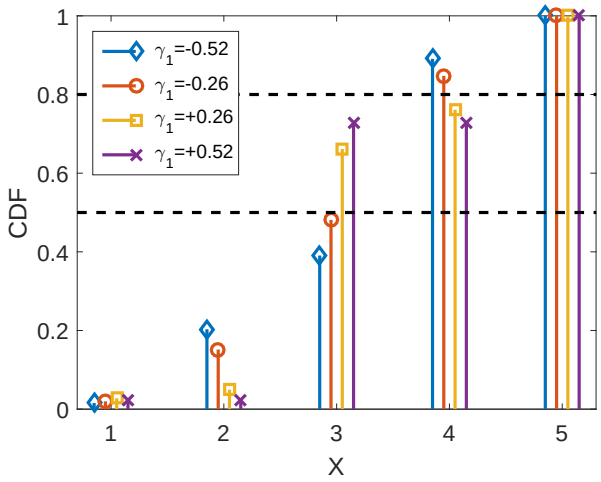


Figure 2: Different discrete distributions with identical mean (3.5) and standard deviation (0.968). It can easily be seen that e.g. the median (and other important quantiles, in fact) are significantly different in each distribution.

of systems and applications such as network dimensioning, performance evaluation of new networking mechansims, assessment of new codecs, etc.

There is a wealth of literature on different objective metrics, subjective methods, models, etc, [3]–[8]. However, none of them consider anything more complex than MOS in terms of analyzing subjective data or producing QoE estimates. In [9], the authors discuss the limitations of MOS and other related issues.

Collapsing the results of subjective assessments into MOS values, however, hides information related to inter-user variation. Simply using the standard deviation to assess this variation might not be sufficient to understand what is really going on, either. Two very different assessment distributions could "hide" behind the same MOS and standard deviation, and in some QoE exploitation scenarios, this could have a significant impact both for the users and the service providers. Figures 1 and 2 show examples of such distributions, continuous and discrete (the latter type being closer to the 5-point scales commonly used for subjective assessment), respectively. As can be seen, while votes following these distributions would present the same MOS (and also standard deviation values in Figure 2), the underlying ground truths would be significantly different in each case. For the discrete case, they differ significantly in skewness and in their quantiles, both of which have practical implications, e.g., for service providers.

When conducting subjective assessments, a researcher may try to answer different types of questions regarding the quality of the service under study. These questions might relate to the overall perception of the quality (probably the most commonly case found in the literature), some more specific perceptual dimensions of quality (e.g., intelligibility, in the case of speech, or blockiness in the case of video), or other aspects such as usability or acceptability of the service. The assessment itself can either explicitly ask opinions from the subjects, or try to infer those opinions through more indirect, behavioral or physiological measurements. Figure 3 presents an overview of approaches to measuring and estimating QoE, both subjectively and objectively.

B. The Need to Go Beyond MOS

Using average values (such as MOS) may be sufficient in some application areas, for instance when comparing the efficiency of different media encoding mechanisms (where quality is not the only consideration, or is a secondary one), or when only a single, simple indicator of quality is sought (e.g., some monitoring dashboard applications). For most other applications — and in particular from a service provider's point of view — however, MOS values are not really sufficient. Averages only consider — well — averages, and do not provide a way to address variations between users. As an extreme example, if the MOS of a given service under a given condition is 3, it is *a priori* impossible to know whether all users perceived quality as acceptable (all scores are 3), or maybe half the users rated the quality 5 while the other half rated it 1, or anything in between, in principle. To some extent, this can be mitigated by quantifying user rating variation via e.g. standard deviations. However, the question often faced by service providers is of the type: "Assuming they observe comparable conditions, are at least 95% of my users satisfied with the service quality they receive?". As we will see, it is a common occurrence that mean quality values indicated as acceptable or better (e.g. MOS 3 or higher) hide a large percentage of users who deem the quality unacceptable. This clearly poses a problem for the service provider (who might get customer complaints despite seeing the estimated quality as "good" in their monitoring systems), and for the users, who might receive poor quality service while the provider is

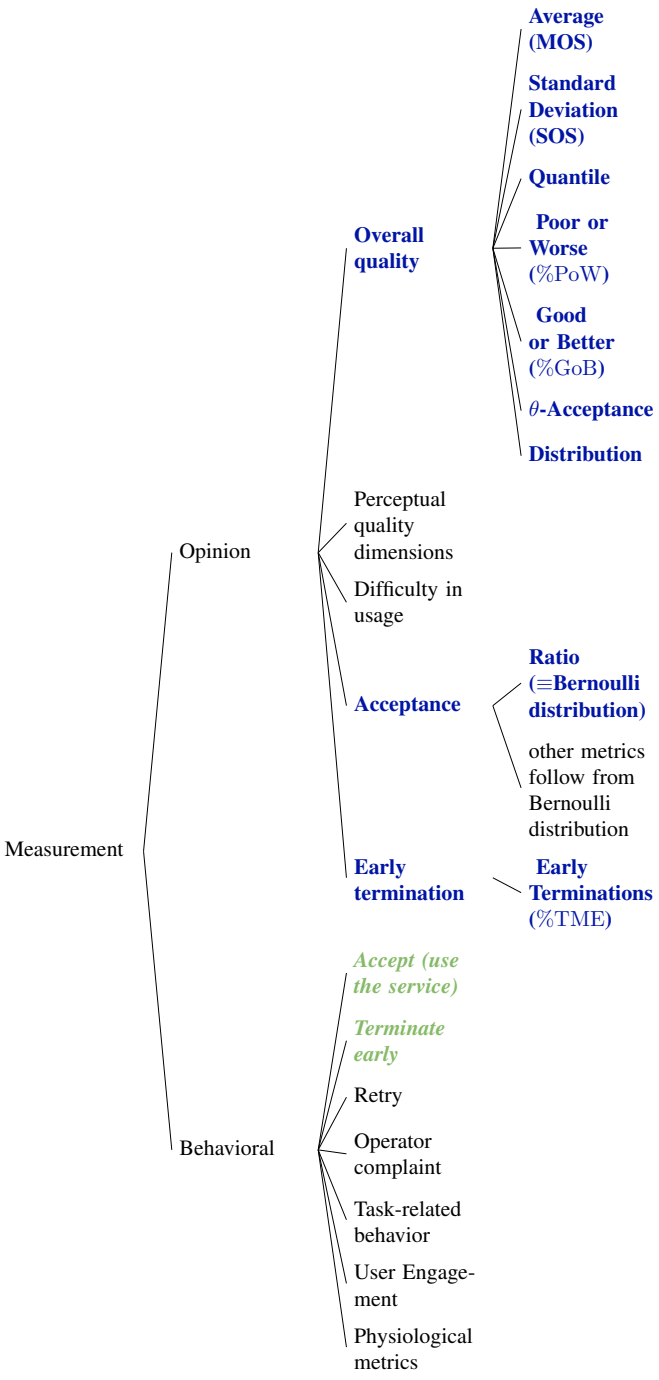


Figure 3: Classification into opinion and behavioral metrics. Perceptual quality dimensions include for example loudness, noisiness, etc. Qualitative opinions are typically 'yes/no' questions like for acceptance. Within the article we address the **bold-faced and blue colored opinion metrics**. Some of the opinion metrics are related to the behavioral metrics ***in italics and colored in green.*** See Section IV for formal definitions of some of the terms above.

unaware of the issue, or worse, believes the problem to be rooted outside of their system.

Likewise, using higher order moments such as skewness and kurtosis can provide insight as to how differently users perceive the quality under a given condition, relative to the mean (e.g. are most users assessing "close" to the mean, and on which side of it).

Very little work has been done on this type of characterization of subjective assessment. One notable exception is [10], where the authors propose a generalized linear model able to estimate a distribution of ratings for different conditions (with an example use case of FTP download times versus link capacity).

III. Background and Related Work

The suitability of the methods used to assess quality has historically been a contentious subject, which in a way reflects the multi-disciplinary nature of QoE research, where media, networking, user experience, psychology and other fields converge.

Qualitative approaches to quality assessment, whereby users describe their experiences with the service in question, have been proposed as tools to identify relevant factors that affect quality [11].

In other contexts (see [12] for a nice example related to subjective validation of objective image quality assessment tools via subjective assessment panels), pair-wise comparisons, or preference rank ordering can be better suited than quantitative assessments.

In practice, most QoE research in the literature typically follows the (quantitative) assessment approaches put forward by the ITU (e.g., ITU-T P.800 [13] for telephony, or ITU-R Rec. BT.500-13 [14] for broadcast video), whereby a panel of users are asked to rate the quality of a set of media samples that have been subjected to different degradations. These approaches have shown to be useful in many contexts, but they are not without limitations.

In particular, different scales, labels, and rating mechanisms have been proposed (e.g. [15]), as well as other mechanisms for assessing quality in more indirect ways, for example, by seeing how it affects the way users perform certain tasks [16]–[19]. These approaches provide, in some contexts, a more useful notion of quality, by considering its effects on the users, rather than considering user ratings. Their applicability, however, is limited to services and use cases where a clear task with measurable performance can be identified. This is limiting in many common scenarios, such as entertainment services. Moreover, the use of averages is still pervasive in them, posing the same type of limitations that the use of MOS values has. Other indirect measures of quality and how it affects users can be found in willingness to pay studies, which aim at understanding how quality affects the spending behavior of users [20], [21].

Other approaches of quality assessment focus on (or at least explicitly include) the notion of acceptability [22]–[25]. Acceptability is a critical concept in certain quality assessment

contexts[1] and application domains, both from the business point of view ("will customers find this level of quality acceptable, given the price they pay?") and on more technical aspects, for instance for telemedicine applications, where applications often have a certain quality threshold below which they are not longer acceptable to use safely. Later in the article we discuss the relation between quality and acceptability (by looking at measures such as "Good or Better", "Poor or Worse", and introducing a more generic one, θ-acceptability) in more detail.

A. QoE and Influence Factors on User Ratings

From the definition of quality first introduced by [27], it follows that quality is the result of an individual's perception and judgment process, see also [28]. Both processes lead to a certain degree of delight or annoyance of the judging individual when s/he is using an application or service, i.e. the Quality of Experience (QoE). The processes are subject to a number of influence factors (IFs) which are grouped in [28] into human, system and context influence factors. Human IFs are static or dynamic user characteristics such as the demographic and socio-economic background, the physical or mental constitution, or the user's mental state. They may influence the quality building processes at a lower, sensory level, or at a higher, cognitive level. System IFs subsume all technical content, media, network and device related characteristics of the system which impact quality. Context IFs "embrace any situational property to describe the user's environment in terms of physical, temporal, social, economic, task, and technical characteristics" [28], [29] which impact the quality judgment. Whereas the impact of System IFs is a common object of analysis when new services are to be implemented, with few exceptions little is known about the impact of User and Context IFs on the quality judgment.

Two well-known examples of actually including context factors into quality models are the so-called "advantage of access" factor in the E-model [30], and the type of conversation and its impact on the quality judgment with respect to delay in telephony scenarios [31], [32]. Some of these contextual factors, such as the aforementioned "advantage of access" incorporated in the E-model might even vary with time, as different usage contexts become more or less common.

B. Influence Factors in Subjective Experiments

In order to cope with the high number of IFs, subjective experiments which aim at quantifying QoE are usually carried out under controlled conditions in a laboratory environment, following standardized methodologies [2], [14], [33] in order to obtain quality ratings for different types of media and applications. These methodologies have been designed with consistency and reproducibility in mind, which allow results to be comparable across studies done in similar conditions. For the most part, these methodologies result in MOS ratings, along with standard deviation and confidence intervals, whereas even early application guidelines (such as the ones given in the ITU-T Handbook on Telephonometry [34]) already state that the consideration of distributions of subjective ratings would be more appropriate, given the characteristics of the obtained ratings.

Regarding the Context IFs, the idea of laboratory experiments is to keep the usage context as far as possible constant between the participants of an experiment. This is commonly achieved by designing a common test task, e.g. perceiving prerecorded stimuli and providing a quality judgment task, with or without a parallel (e.g. content-transcription) task, or providing scenarios for conversational tasks [35]. A context effect within the test results from presenting different test conditions (e.g. test stimuli) is a sequence, so that the previous perception process sets a new reference for the following process. This effect can partially be ruled out by factorial designs, distributing test conditions across participants in a mostly balanced way, or (approximately) by simple randomization of test sequences. Another context effect results from the rating scales which are used to quantify the subjective responses.

System IFs also carry an influence on the test outcome, in terms of the selection of test conditions chosen for a particular test (session). It is commonly known that a medium-quality stimulus will obtain a relatively bad judgment in a test where all the other stimuli are of better quality; in turn, the same stimulus will get a relatively positive judgment if it is nested in a test with only low-quality stimuli. This impact of the test conditions was ruled out in the past by applying the same stimuli with known "reference degradations" in different tests. In speech quality evaluation, for example, the Modulated Noise Reference Unit (MNRU) was used for this purpose [36].

C. Service Provider's Interest in QoE Metrics

In order to stay in business in a free market, ISPs and other service providers need to maintain a large portion of their users satisfied, lest they stop using the service or change providers — the dreaded "churn" problem. For any given service level the provider can furnish, there will be a certain proportion of users who might find it unacceptable, and the perceived quality of the service is one of the key factors determining user churn [37]. Moreover, a large majority ($\sim 90\%$) of users will simply defect a service provider without even complaining to them about service quality, and report their bad experience within their social circles [38], resulting in a possibly even larger business impact in terms of e.g., brand reputation. With only a mean value as an indicator for QoE, such as the MOS, the service provider cannot know what this number of unsatisfied users might be, as user variation is lost in the averaging process.

For many applications, however, it is desirable to gauge the portion of users that is satisfied given a set of conditions (e.g., under peak-time traffic, for an IPTV service). For example, a service provider might want to ensure that at least, say, 95% of its users find the service acceptable or better. In order to ascertain this, some knowledge of how the user ratings are

[1] Arguably, and going by the ITU-T definition of QoE, it is at the core of QoE: "QoE is the overall acceptability of an application or service, as perceived subjectively by the end user" [26].

distributed for any given condition is needed. In particular, calculating the 95% quantile (keeping in line with the example above) would be sufficient for the provider.

In the past, service providers have also based their planning on (estimated) percentages of users judging a service as "poor or worse" (%PoW), "good or better" (%GoB), or the percentage of users abandoning a service (Terminate Early, %TME). These percentages have been calculated from MOS distributions on the basis of large collections of subjective test data, or of customer surveys. Whereas the original source data is proprietary in most cases, the resulting distributions and transformation laws have been published in some instances. One of the first service providers to do this was Bellcore [39], who provided transformation laws between an intermediate variable, called the Transmission Rating R, and %PoW, %GoB and %TME. These transformation were further extended to other customer behavior predictions, like retrial (to use the service again) and complaints (to the service provider). The Transmission Rating could further be linked to MOS predictions, and in this way a link between MOS, %PoW and %GoB could be established. The E-model, a parametric model for planning speech telephony networks, took up this idea and slightly modified the Transmission Rating calculation and the transformation rules between R and MOS, see [40]. The resulting links can be seen in Fig. 4. Such links can be used for estimating the percentage of dissatisfied users from the ratings of a subjective laboratory test; there is, however, no guarantee that similar numbers would be observed with the real service in the field. In addition, the subjective data the links are based on mostly stem from the 1970-1980s; establishing such links anew, and for new types of services, is thus highly desirable.

In an attempt to go beyond user satisfaction and into user acquisition, many service providers have turned to the Net Promoter Score (NPS)[2], which purports to classify users into "promoters" (enthusiastic users likely to will keep buying the service and "promoting growth", "passives" (users that are apathetic towards the service and might churn if a better offer from a competitor comes along) and "detractors" (vocal, dissatisfied users who can damage the service's reputation). While popular with business people, the research literature on the NPS is critical of the reliability of such subjective test assessments (e.g. [41], [42]). The NPS is based on a single-item questionnaire whereby a user is asked how likely they are to recommend the service or product to a friend or colleague, which might explain its shortcomings.

IV. Definition of QoE Metrics

The key QoE metrics are defined in this section: the mean of the opinion scores (MOS); the standard deviation of opinion scores (SOS) reflecting the user diversity and its relation to MOS; the newly introduced θ-acceptability as well as acceptance; the ratio of (dis-)satisfied users rating good or better %GoB and poor or worse %PoW, respectively. The detailed formal definitions of the QoE metrics are added in the technical report [43].

[2] http://www.netpromoter.com/why-net-promoter/know

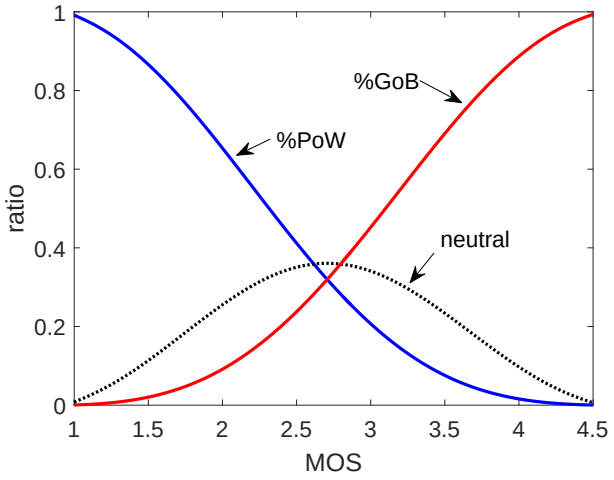


Figure 4: Relationship between MOS, %PoW and %GoB as used in the E-model [40]. The ratio of users not rating poor or worse as well as good or better is referred to as 'neutral' and is computed by $1 - \%\text{GoB} - \%\text{PoW}$.

A. Preamble

In this article we consider studies where users are asked their opinion on the overall quality (QoE) of a specific service. The *subjects* (the participants in a study that represent users), rate the quality as a *quality rating* on a *quality rating scale*. As a result, we obtain an *opinion score* by interpreting the results on the rating scale numerically. An example is a discrete 5-point scale with the categories 1≙'bad', 2≙'poor', 3≙'fair', 4≙'good', and 5≙'excellent', referred to as an Absolute Category Rating (ACR) scale [30].

B. Expected value and its estimate: MOS

Let U be a random variable (RV) that represents the quality ratings, $U \in \Omega$, where Ω is the rating scale, which is also the state space of the random variable U. The RV U can be either discrete, with probability mass function f_s, or continuous, with probability density function $f(s)$ for rating score s. The estimated probability of opinion score s from the R user ratings U_i is

$$\hat{f}_s = \frac{1}{R} \sum_{i=1}^{R} \delta_{U_i,s} \tag{1}$$

with the Kronecker delta $\delta_{U_i,s} = 1$ if user i is rating the quality with score s, i.e. $U_i = s$, and 0 otherwise.

The Mean Opinion Score (MOS) is an estimate of $E[U]$.

$$u = \hat{U} = \frac{1}{R} \sum_{i=1}^{R} U_i \tag{2}$$

C. SOS as function of MOS

In [44], the minimum, $S^-(u)$, and the maximum SOS, $S^+(u)$ were obtained, as a function of the MOS u. The minimum SOS is $S^-(u) = 0$ on a continuous scale, $[U^-; U^+]$, and

$$S^-(u) = \sqrt{u(2\lfloor u \rfloor + 1) - \lfloor u \rfloor(\lfloor u \rfloor + 1) - u^2} \quad (3)$$

on a discrete scale, $\{U^-, \ldots, U^+\}$.

The maximum SOS is, on both continuous and discrete scales (the scales as above).

$$S^+(u) = \sqrt{-u^2 + (U^- + U^+)u - U^- \cdot U^+} \quad (4)$$

The SOS hypothesis [44], formulates a generic relationship between MOS and SOS values independent of the type of service or application under consideration.

$$S(u) = \sqrt{a} \cdot S^+(u) \quad (5)$$

It has to be noted that the SOS parameter a is scale invariant when linearly transforming the user ratings and computing MOS and SOS values for the transformed ratings. The SOS parameter allows to compare user ratings across various rating scales. Thus, any linear transformation of the user ratings does not affect the SOS parameter a which is formally proven in the Appendix B. However, it has to be clearly noted that if the participants are exposed to different scales, then different SOS parameters may be observed. This will be shown in Section V-D e.g. for the results on speech QoE in Figure 12a. The parameter a, depends on the application or service, and the test conditions. The parameter is derived from subjective tests, and in the Section V-D a few examples are included.

D. θ-Acceptability

For service providers, acceptance is an important metric to plan, dimension and operate their services. Therefore, we would like to establish a link between opinion measurements from subjective QoE studies and behavioral measurements. In particular, it would be very useful to derive the "accept" behavioral measure from opinion measurements of existing QoE studies. This would allow to reinterpret existing QoE studies from a business oriented perspective. Therefore, we introduce the notion of θ-acceptability which is based on opinion scores.

The θ-acceptability, $\mathbb{A}_\theta$, is defined as the probability that the opinion score is above a certain threshold θ, $P(U \geq \theta)$, and can be estimated by $\hat{f}_s$ from Eq. (1) or by counting all user ratings $U_i \geq \theta$ out of the R ratings.

$$\mathbb{A}_\theta = \int_{s=\theta}^{U^+} \hat{f}_s ds = \frac{1}{R} \left|\{U_i \geq \theta : i = 1, \ldots, R\}\right| \quad (6)$$

E. Acceptance

When a subject is asked to rate the quality as either *acceptable* or *not acceptable*, this means that U is *Bernoulli*-distributed. The quality ratings are then samples of $U_i \in \{0, 1\}$, where 1≙'accepted' and 0≙'not accepted'. The probability of acceptance is then $f_u = P(U = u)$, $U \in \{0, 1\}$, and can be estimated by Eq. (1) with $u = 1$:

$$\hat{f}_1 = \frac{1}{R} \sum_{i=1}^{R} \delta_{U_i,1} \quad (7)$$

(this is equal to $\mathbb{A}_1$ in Eq. (6) with $U^- = 0$ and $U^+ = 1$ on a discrete scale).

F. %GoB and %PoW

Section III-C describes the use of the percentage of *Poor-or-Worse* (%PoW) and *Good-or-Better* (%GoB). These are quantile levels in the distribution of the quality rating U, or in the empirical distribution of $\mathcal{U} = \{U_i\}$.

The two terms are used in the E-model [40] where the RV of the quality rating, $U \in [0; 100]$ is referring to Transmission Rating R that represents objective (estimated) rating of the *voice quality*. The E-model assumes that $U \sim N(0, 1)$, which is the standard normal distribution.

Under this assumption, the measures have been defined as[3]

$$\text{GoB}(u) = F_U\left(\frac{u - 60}{16}\right) = P_U\left(U \geq 60\right) \quad (8)$$

$$\text{PoW}(u) = F_U\left(\frac{45 - u}{16}\right) = P_U\left(U \leq 45\right) \quad (9)$$

The E-model also defines a transformation of the U onto a continuous scale of MOS$\in [1; 4.5]$, by the following relation:

$$MOS(u) = 7{\cdot}(u{-}60){\cdot}(100{-}u){\cdot}u{\cdot}10^{-6}{+}0.035{\cdot}u{+}1 \quad (10)$$

The plot of (continuous) MOS ($\in [1; 4.5]$) in Figure 4 is an example where this transformation has been applied to map the MOS to %GoB and %PoW. Observe that the sum of %GoB + %PoW does not add up to 100%, because the probability (denoted "neutral" in the figure), $P(45 < U < 60)$, is not included in neither %PoW nor %GoB. The quantiles used (i.e. 45 and 60) for the two measures, and the assumed standard normal distribution, are chosen as a result of a large number of subjective *audio quality tests* conducted while developing the E-model [40]. Table I includes the MOS and the Transmission Rating R, with their corresponding values[4] of the %PoW and %GoB.

The measures are estimated based on the ordered set of quality ratings, $\mathcal{U} = \{U^{(i)}\}$, by using the θ-Acceptability estimator from Eq. (6). First, discretise the quality rating scale $\mathcal{U} \in \{0, 100\}$. Then, using the Eq. (6), the following applies

$$\hat{\%\text{GoB}} = \mathbb{A}_{\theta_{gb}} \quad (11)$$

$$\hat{\%\text{PoW}} = 1 - \mathbb{A}_{\theta_{pw}} \quad (12)$$

For example, in the E-model the $\theta_{gb} = 60$ and $\theta_{pw} = 45$ for $\mathcal{U} \in \{0, 100\}$, and $\theta_{gb} = 3.1$ and $\theta_{pw} = 2.3$ on a $\mathcal{U} \in \{1, 5\}$ scale (when using Eq. 10).

[3]When $U \sim N(0, 1)$ then $F_U(u) = 1 - F_U(-u)$, which is applied for the GoB definition

[4]Observe: all values of MOS on the ACR scale are included, even for MOS=5 where the Transmission Rating R is not defined.

Table I: E-model: MOS and Transmission Rating R with the quantile measures for speech quality.

MOS	R	%PoW	%GoB
1.00	6.52	99.192	0.041
1.50	27.27	86.611	2.039
2.00	38.68	65.349	9.139
2.32	45.00	50.000	17.425
2.50	48.57	41.176	23.747
3.00	58.08	20.685	45.221
3.10	60.00	17.425	50.000
3.50	67.96	7.563	69.062
4.00	79.37	1.585	88.699
4.50	100.00	0.029	99.379
5.00	undefined	0.000	100.000

The purpose of the example above is to demonstrate GoB and PoW using an ACR scale (1-5). This is a theoretical exercise (valid for the E-model) where we apply the transformation from R to "MOS" (term used when E-model was introduced) as given in Eq. (10), and transform Eq. (8)-(9) into Eq. (11)-(12), using the notation introduced in Section IV-D. Samples from Eq. (10) are given in Table I. The $\%\text{GoB} = P(R \geq 60)$ corresponds to $\%\text{GoB} = P(\text{MOS} \geq 3.1)$ which on an integer scale is $\%\text{GoB} = P(\text{MOS} \geq 4)$. Correspondingly, for $\%\text{PoW} = P(R \leq 45) = P(\text{MOS} \leq 2.32) = P(\text{MOS} \leq 2)$.

It is important to note that the quantiles in the examples are valid for *speech quality tests* under the assumptions given in the E-model. The mapping of the MOS to the %PoW and %GoB metrics in Table I are specific for this E-model, but the %PoW and %GoB metrics are general and can be obtained from any quality study, provided that the thresholds θ_{gb} and θ_{pw} are determined.

In the following we demonstrate the use of %PoW and %GoB metrics also for other quality tests.

V. Application to Real Data Sets: Some Examples

A. Overview on Selected Applications and Subjective Studies

The presented QoE measures are applied to real data sets available in the literature[5], comparing MOS values to other quantities. To cover a variety of relevant applications, we consider speech, video, and web QoE. The example studies highlight which conclusions can be drawn from other measures beyond the MOS, such as SOS, quantiles, or θ-acceptability. The limitations of MOS become clear from the results. These additional insights are valuable e.g., to service providers to properly plan or manage their systems.

Section V-B focuses on the link between acceptance and opinion ratings. The study considers web QoE, however, users have to complete a certain task when browsing. Test subjects are asked to rate the overall quality as well as answering an acceptance question. This allows to investigate the relation between MOS, acceptance, θ-acceptability, %GoB, and %PoW based on the subjects' opinions. The relation between acceptance as a behavioral measure and overall quality as opinion measure is particularly interesting. To wit, it would be very useful to be able to derive the "accept" behavioral measure from QoE studies and subjects' opinions. This would provide a powerful tool to re-interpret existing QoE studies from a different, more business-oriented perspective.

Section V-C investigates the ratio of (dis-)satisfied users. The study on speech quality demonstrates the impact of rating scales and compares %PoW and %GoB related to MOS when subjects are rating on a discrete and a continuous scale. The results are also checked against the E-model to analyze its validity when linking overall quality (MOS) to those quantities. Additional results for web QoE can be found in the Appendix A-C in Figure 13. In this subjective study on web QoE, page load times are varied while subjects are viewing a simple web page. The web QoE results confirm the gap between the %GoB and %PoW estimates (as defined e.g. for speech QoE by the E-model), and the measured %GoB and %PoW.

Section V-D relates the diversity in user ratings in terms of SOS to MOS. Results from subjective studies on web, speech, and video QoE are analyzed. As a result of the web QoE study, we find that the opinion scores for this study can be very well approximated with a binomial distribution – which allows us to fully specify the voting distribution using only the SOS parameter a. For the video QoE study, a continuous rating scale was used and we find that the opinion scores follow a truncated normal distribution. Again, the SOS parameter a derived for this video QoE study fully describes then the distribution of opinion scores for any given MOS value. Thus, the SOS parameter allows to model the entire distribution and then to derive measures such as quantiles. We highlight the discrepancy between quantiles and MOS, which is of major interest for service providers.

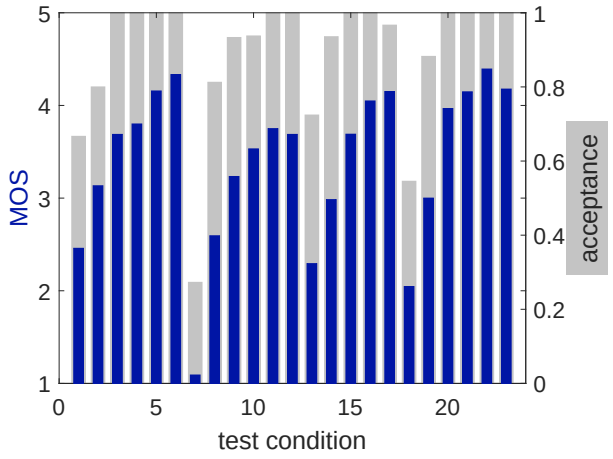
Section V-E provides a brief comparison of the studies presented in the article. It serves mainly as an overview on interesting QoE measures beyond MOS and a guideline how to properly describe subjective studies and their results.

For the sake of completeness, the reader finds a detailed summary of the experimental description in the Appendix A.

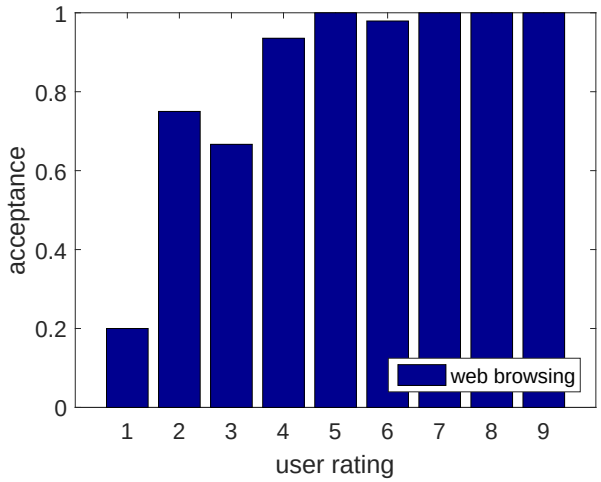
B. θ-Acceptability Derived from User Ratings

The experiments in [45] investigated task-related web QoE in conformance with ITU-T Rec. P.1501 [46]. In the campaign conducted, subjects were asked to carry out a certain task, e.g. 'Browse to search for three recipes you would like to cook in the given section.' on a certain cooking web page (cf. Table III). The network conditions were changed and the impact of page load times during the web session was investigated. Besides assessing the overall quality of the web browsing session, subjects additionally answered an acceptance question. In particular, after each condition, subjects were asked to rate their overall experienced quality on a 9-point ACR scale, see Figure 11, as well as a binary acceptance question. The experiment was carried out in a laboratory environment, with 32 subjects.

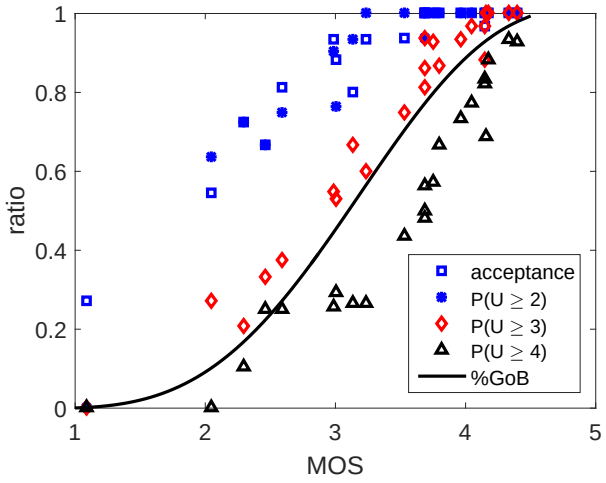
[5]We ask the reader to take notice that we did not conduct new subjective studies, but rather used the opinion scores from the existing studies to apply the QoE measures and interpret the results in a novel way, obtaining a deeper understanding of them.



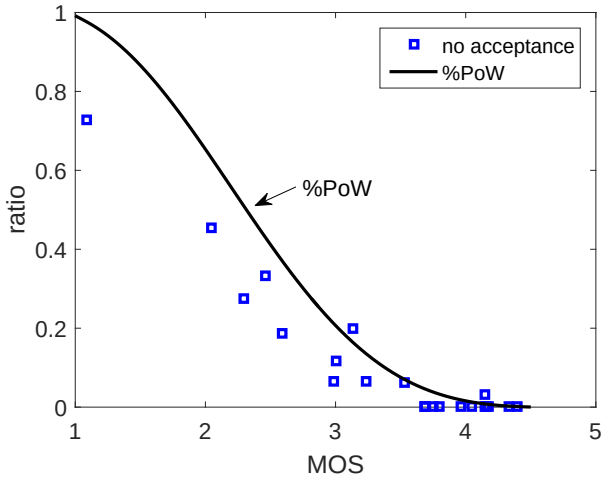
(a) MOS & Acceptance per Condition. The blue bars in the foreground depict the MOS values per test condition on the left y-axis. The grey bars in the background depict the acceptance values for that test condition on the right axis. While the acceptance values reach the upper bound of 100 %, the maximum MOS observed is 4.39. The minimum MOS over all test conditions is 1.09, while the minimum acceptance ratio is 27.27 %.



(b) Acceptance per Rating Category. The users are rating the overall quality on a 9-point ACR scale and additionally answer an acceptance question. All users who rate an arbitrary test condition with x are considered and the acceptance ratio y is computed. The plot shows how many users accept a condition and rate QoE with x. For each rating category $1, \cdots, 9$, there are at least 20 ratings. Still, 20 % of the users accept the service, although the overall quality is bad.



(c) %GoB-MOS Plot. The markers depict θ-acceptability $P(U \geq \theta)$ depending on the MOS for $\theta = 3$ '◊' and $\theta = 4$ '△' i.e. %GoB. The %GoB (solid line) overestimates the true ratio of users rating good or better ($\theta = 4$). This can be adjusted by considering users rating fair or better $P(U >= 3)$ which is close to the %GoB estimation. In addition, the acceptance ratio '□' is plotted depending on the MOS. However, the θ-acceptability curves as well as the %GoB estimates do not match the acceptance curve. In particular, for the minimum MOS of 1.09, the θ-acceptability is 0 %, while the acceptance ratio is 27.27 %.



(d) %PoW-MOS Plot. The markers depict the the ratio of users not accepting a test condition '□' depending on the MOS for all 23 test conditions. The results are compared with %PoW estimation, but again the characteristics are not matched. Especially, 27.27 % of users are still accepting the service, although the MOS value is 1.09. The %PoW is close to 0 %. Nevertheless, this indicates that overall quality can be mapped roughly to other dimensions like 'no acceptance'.

Figure 5: Task-Related Web QoE and Acceptance. Results of the task-related web QoE and acceptance study [45] in Section V-B. The data is based on a subjective lab experiment in which participants had to browse four different websites at different network speeds resulting in different levels of experienced responsiveness. The network speeds determined the page load times while browsing and executing a certain task. Defined tasks for each technical condition should stimulate the interaction between the web site and the subject for each test condition, see Table III. In total, there are 23 different test conditions in the data set. The overall quality for each test condition was evaluated by 10–30 subjects on a discrete 9-point scale which was subsequently mapped into a 5-point ACR scale. Furthermore, subjects gave their opinion on the acceptance (yes/no) of that test condition.

Figure 5 quantifies the acceptance and QoE results from the subjective study in [45]. This study also considered web QoE; however, users must complete a certain task when browsing. The test subjects were asked to rate the overall quality as well as answering an acceptance question. This allowed to investigate the relation between MOS, acceptance, θ-acceptability, %GoB, and %PoW based on the subjects' opinions.

Figure 5a shows the MOS and the acceptance ratio for each test condition. The blue bars in the foreground depict the MOS values on the left y-axis. The grey bars in the background depict the acceptance values on the right y-axis. While the acceptance values reach the upper bound of 100 %, the maximum MOS observed is 4.39. The minimum MOS over all test conditions is 1.09, while the minimum acceptance ratio is 27.27 %. These results indicate that users may tolerate significant quality degradation for web services, provided they are able to successfully execute their task. This result contrasts with e.g., speech services, where very low speech quality makes it almost impossible to have a phone call, and hence results in non-acceptance of the service. Accordingly, the %PoW estimator defined in the E-model is almost 100 % for low MOS values.

Figure 5b makes this even more clear. The plot shows how many users accept a condition and rate QoE with x for $x = 1, \ldots, 9$. All users who rate an arbitrary test condition with x are considered and the acceptance ratio y is computed over those users. For each rating category $1, \cdots, 9$, there are at least 20 ratings. Even when the quality is perceived as bad ('1'), 20 % of the users accept the service. For category '2' between 'poor' and 'bad' (see Figure 11), up to 75 % accept the service at an overall quality which is at most 'poor'.

Figure 5c takes a closer look at the relation between MOS and acceptance, θ-acceptability, as well as the %GoB estimation as defined in Section IV-F. The markers depict θ-acceptability $P(U \geq \theta)$ depending on the MOS for $\theta = 3$ '◇' and $\theta = 4$ '△' i.e. %GoB. The %GoB estimator (solid line) overestimates the true ratio of users rating good or better ($\theta = 4$). This can be adjusted by considering users rating fair or better $P(U >= 3)$ which is close to the %GoB estimator. In addition, the acceptance ratio '□' is plotted depending on the MOS. However, the θ-acceptability curves as well as the %GoB do not match the acceptance curve. In particular, for the minimum MOS of 1.09, the θ-acceptability is 0 %, while the acceptance ratio is 27.27 %.

The discrepancy between acceptance and the %GoB estimator is also rather large, see Figure 5c. The estimator in the E-model maps a MOS value of 1 to a %GoB of 0 %, as a speech service is not possible any more if the QoE is too bad. In contrast, in the context of web QoE, a very bad QoE can still result in a usable service which is accepted by the end user. Thus, the user can still complete for example the task to find a wikipedia article, although the page load time is rather high. This may explain why 20 % of the users accept the service even though they rate the QoE with bad quality (1).

We conclude that it is not generally possible to map opinion ratings on the overall quality to acceptance[6]. The conceptual difference between acceptance and the concept of θ-acceptability is the following. In a subjective experiment, each user defines his own threshold determining when the overall quality is good enough to accept the service. Additional contextual factors like task or prices influence strongly acceptance [47]. In contrast, θ-acceptability considers a globally defined threshold (e.g. defined by the ISP) which is the same for all users. Results that are only based on user ratings do not reflect user acceptance, although the correlation is quite high (Pearson's correlation coefficient of 0.93).

Figure 5d compares acceptance and %PoW. The markers depict the ratio of users not accepting a test condition '□' depending on the MOS for all 23 test conditions. The %PoW is a conservative estimator of the 'no acceptance' characteristics. Especially, 27.27 % of users are still accepting the service, although the MOS value is 1.09. The %PoW is close to 0 %. This indicates that overall quality can only be roughly mapped to other dimensions like 'no acceptance'.

C. %GoB and %PoW: Ratio of (Dis-)Satisfied Users

The opinion ratings of the subjects on speech quality are taken from [48]. The listening-only experiments were conducted by 20 subjects in an environment fulfilling the requirements in ITU-T Rec. P.800 [2] using the source speech material in [49]. The subjects assessed the same test stimuli on two different scales: the ACR scale (Figure 6) and the extended continuous scale (Figure 7). To be more precise, each subject was using both scales during the experiment. The labels were internally assigned to numbers of the interval [0,6] in such a manner that the attributes corresponding to ITU-T Rec. P.800 were exactly assigned to the numbers $1, \cdots, 5$.

excellent	good	fair	poor	bad
5	4	3	2	1

Figure 6: Five point discrete quality scale as used for the speech QoE experiments [48].

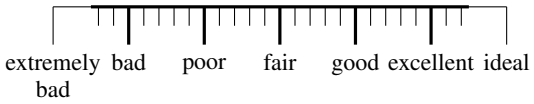
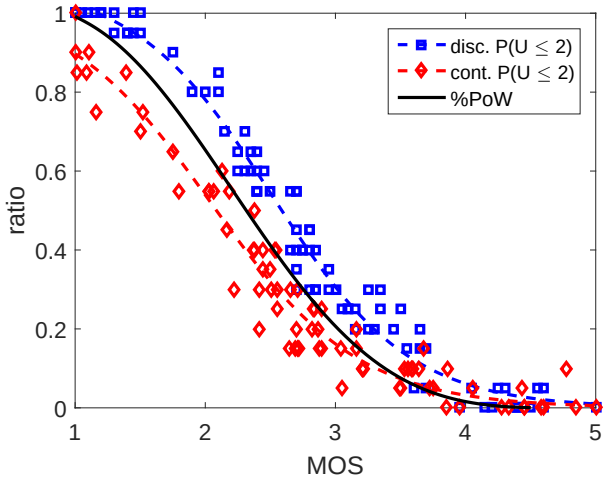


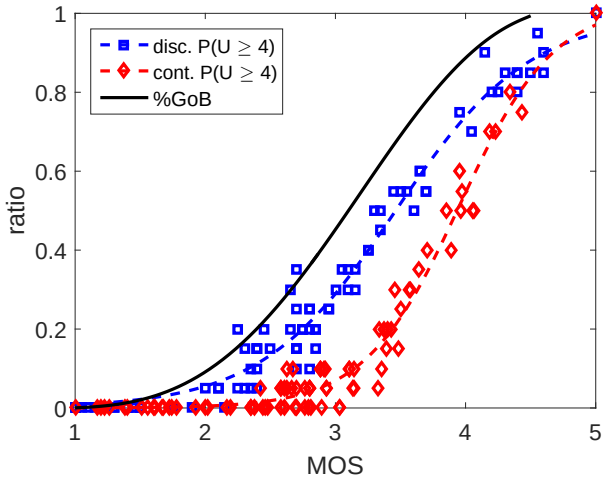
Figure 7: Five point continuous quality scale as used for the speech QoE experiments [48].

Figure 8a investigates the impact of the rating scale on the ratio of dissatisfied users. For 86 test conditions, the MOS, %PoW, and %GoB values were computed over the opinions from the 20 subjects on the discrete rating scale and the continuous rating scale. The results for the discrete scale are

[6]Note that θ-acceptability is defined on the user quality ratings on a certain rating scale and a global threshold θ. In contrast, acceptance is the subject's rating on a binary scale whether the quality is either acceptable or not acceptable.



(a) %POW-MOS Plot. The markers depict the MOS and the ratio $P(U \leq 2)$ from the subjective study on the discrete and the continuous scale. The solid black line shows the %PoW ratio depending on MOS for the E-model. The E-model underestimates the measured %PoW on the discrete scale which is larger than the %PoW on the continuous scale.



(b) %GoB-MOS Plot. The markers depict the MOS and the ratio $P(U \geq 4)$ from the subjective study on the discrete and the continuous scale. The solid black line shows the %GoB ratio depending on MOS for the E-model. The E-model overestimates the ratio of satisfied users on the discrete scale which is smaller than the %GoB on the continuous scale.

Figure 8: Speech QoE. Results of the speech QoE study [48]. For the 86 test conditions, the MOS and %PoW, %GoB values were computed over the 20 subjects for the discrete 5-point ACR scale (Figure 6) and the extended continuous scale (Figure 7). The results for the discrete scale are marked with '□', while the QoE measures for the continuous scale are marked with '◊'. The dashed lines represent logistic fitting functions of the subjective data.

Although the MOS is larger than 3, about 30 % and 20 % of the users are not satisfied rating poor or worse on the discrete and the continuous scale, respectively. The results are also checked against the E-model to analyze its validity when linking overall quality (MOS) to %PoW. We consider the ratio $P(U \leq 2)$ of users rating a test condition poor or worse. For that test condition, the MOS value is computed and each marker in Figure 8a represents the measurement tuple (MOS, $P(U \leq 2)$ for a certain test condition. In addition, a logistic fitting is applied to the measurement values depicted as dashed line. It can be seen that the ratio %PoW of the subjects on the discrete rating scale is always above the E-model (solid curve). The maximum difference between the logistic fitting function and the E-model is 13.78 % at MOS 2.29. Thus, the E-model underestimates the measured %PoW for the discrete scale.

For the continuous rating scale, the ratio $P(U \leq 2)$ is below the E-model. However, we can determine the parameter θ in such a way that the mean squarred error (MSE) between the %PoW of the E-model and the subjective data $P(U \leq \theta)$ is minimized. In the appendix, Figure 12b shows the MSE for different realizations of θ. The value $\theta = 2.32 > 2$ leads to a minimum MSE regarding %PoW. The E-model overestimates the measure %PoW, i.e. $P(U \leq 2)$, for the continuous scale. However, $P(U \leq \theta)$ leads to a very good match with the E-model.

In a similar way, Figure 8b investigates the θ-acceptability and compares the results with %GoB of the E-model. Even when the MOS is around 4, the subjective results show that the ratio of users rating good or better is only 80 % and 70 % on the discrete and the continuous scale, respectively. The E-model overestimates the ratio $P(U \geq 4)$ of satisfied users rating good or better on the discrete scale. The maximum difference between the logistic fitting function and the %GoB of the E-model is 17.49 % at MOS 3.34. For the continuous rating scale, the E-model further overestimates the ratio of satisfied users, with the maximum difference being 46.20 % at MOS 3.49. The value $\theta = 3.01$ leads to a minimum MSE between the E-model and $P(U \geq \theta)$ on the continuous scale, as numerically derived from Figure 12b. Thus, for the speech QoE study, the %GoB of the E-model corresponds to the ratio of users rating fair or better.

In summary, the E-model does not match the results from the speech QoE study for PoW, i.e. $P(U \leq 2)$, and GoB, i.e. $P(U \geq 4)$, on both rating scales. The results on the discrete rating scale lead to a higher ratio of dissatisfied users rating poor or worse than a) the %PoW of the E-model and b) the %PoW for the continuous scale. The %GoB of the E-model overestimates the %GoB on the discrete and the continuous scale.[7] Thus, in order to understand the ratio of satisfied and dissatisfied users it is necessary to compute those QoE metrics

[7] Similar results can also be found for the web QoE experiments with users rating QoE for varying page load times on a discrete rating scale, see Figure 13 in the appendix.

for each subjective experiments since the E-model does not match for all subjective experiments. Due to the non-linear relationship between MOS and θ-acceptability, the additional insights get evident. For service providers, the θ-acceptability allows to go beyond the 'average' user in terms of MOS and to derive the ratio of satisfied users with ratings larger than θ.

D. SOS Hypothesis and Modeling of Complete Distributions

We relate the SOS values to MOS values and show that the entire distribution of user ratings for a certain test condition can be modeled by means of the SOS hypothesis. A discrete and continuous rating scale will lead to a discrete and continuous distribution respectively.

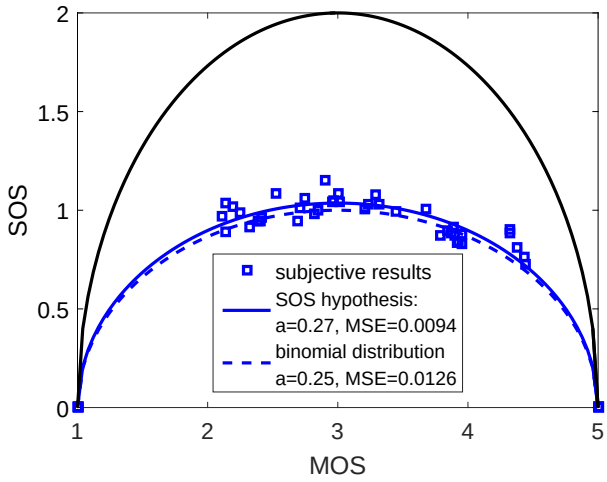
1) Results for Web QoE on a Discrete Rating Scale: Figure 9 shows the results of the web QoE study [50]. In the study, the page load time was influenced for each test condition and 72 subjects rated the overall quality on a discrete 5-point ACR scale. Each user viewed 40 web pages with different images on the page and page load times (PLTs) from 0.24 s to 1.2 s resulting into 40 test conditions per user.[8] For each test condition, MOS and SOS are computed over the opinions of the 72 subjects. As users conducted the test remotely, excessively high page load time might have caused them to cancel or restart the test. In order to avoid this, only a maximum PLT of 1.2 s was chosen. As a result, the minimum MOS value observed is 2.11 for the maximum PLT.

Figure 9a shows the relationship between SOS and MOS and reveals the diversity in user ratings. The markers '□' depict the tuple (MOS,SOS) for each of the 40 test conditions. For a given MOS the individual user rating is relatively unpredictable due to the user rating diversity (in terms of standard deviation).

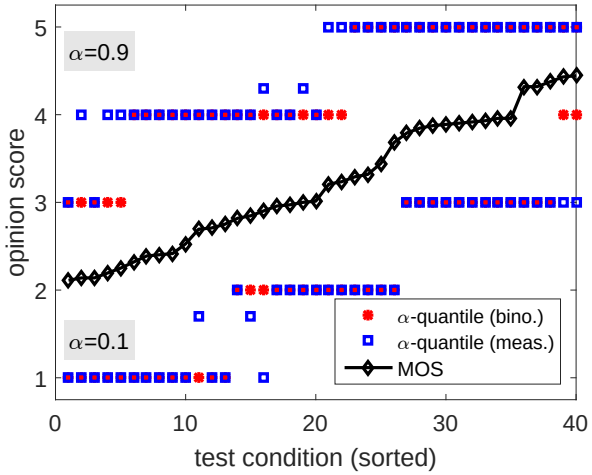
The results in Figure 9a confirm the SOS hypothesis and the SOS parameter is obtained by minimizing the least squared error between the subjective data and Eq. 5. As a result, a SOS parameter of $\tilde{a} = 0.27$ is obtained. The mean squarred error between the subjective data and the SOS hypothesis (solid curve) is close to zero (MSE 0.01), indicating a very good match. In addition, the MOS-SOS relationship for the binomial distribution ($a_B = 0.25$) is plotted as dashed line. To be more precise, if user ratings U follow a binomial distribution for each test condition, the SOS parameter is $a_B = 0.25$ on a 5-point scale. The parameters of the binomial distribution per test condition are given by the fixed number $N = 4$ of rating scale items and the MOS value μ which determines $p = (\mu - 1)N$. Since the binomial distribution is defined for values $x = 0, \cdots, N$, the distribution is shifted by one to have user ratings on a discrete 5-point scale from 1 to 5. Thus, for a test condition, the user ratings U follow the shifted binomial distribution with $N = 4$ and $p = (\mu - 1)N$ for a MOS value μ, i.e. $U \sim B(N, (\mu - 1)N) + 1$ and $P(U = i) = \binom{N}{i-1} p^{i-1}(1-p)^{n-i+1}$ for $i = 1, \cdots, N + 1$ and $\mu \in [1; 5]$.

We observe that the measurements can be well approximated by a binomial distribution with $a_B = 0.25$ (MSE=0.01)

[8] More details on the experimental setup can be found in the Appendix A-C.



(a) SOS-MOS Plot. The markers '□' depict the tuple (MOS,SOS) for each of the 40 test conditions. The solid blue line line shows the SOS fitting function with the SOS parameter $a = 0.27$. The resulting MSE is 0.01. We observe that the measurements can be well approximated by a binomial distribution with $a = 0.25$ (MSE=0.01) plotted as dashed curve. The solid black curve depicts the maximum SOS.



(b) Quantile-MOS Plot. The 10 %- and 90 %- quantiles '□' for the web browsing study as well as the MOS '♦' are given for the different test conditions (increasingly sorted by MOS). There are strong differences between the MOS and the quantiles. The maximum difference between the 90 %-quantile and MOS is $4 - 2.14 = 1.86$. The quantiles for the shifted binomial distribution •' are also given which match the empirically derived quantiles.

Figure 9: Web QoE for PLT only. Results of the web QoE study [50]. The page load time was influenced for each test condition and 72 subjects rated the overall quality on a discrete 5-point ACR scale. Each user viewed 40 web pages with different images on the page and PLTs from 0.24 s to 1.2 s resulting into 40 test conditions per user. For each test condition, the MOS, SOS, as well as 10 %- and 90 %-quantiles are computed over the opinions of the 72 subjects.

plotted as dashed curve. The SOS parameter of the measurement data is only $\sqrt{a/a_B} = 1.04$ higher than the SOS for the binomial distribution. The SOS parameter a is a powerful approach to select appropriate distributions of the user opinions. In the study here, we observe roughly $a = 0.25$ on a discrete 5-point scale which means that the distribution follows the aforementioned shifted binomial distribution. Thus, for any MOS value, the entire distribution (and deducible QoE metrics like quantiles) can be derived.

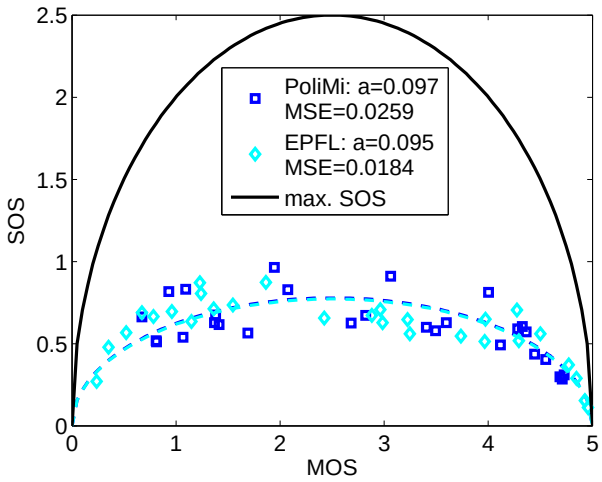
Figure 9b shows the measured α-quantiles '□' as well as the quantiles from the binomial distribution •' compared to the MOS values '♦'. The quantiles for the shifted binomial distribution •' match the empirically derived quantiles very well. The 10 %- and 90 %-quantiles quantify the opinion score of the 10 % of the most critical and the most satisfied users, respectively. There are strong differences between the MOS and the quantiles. The maximum difference between the 90 %-quantile and MOS is $4 - 2.14 = 1.86$. For the 10 %-quantile, we observe a similarly strong discrepancy, $2.90 - 1 = 1.90$.

This information, while very significant to service providers, is masked out by averaging used to calculate MOS values. As a conclusion from the study, we recommend to report different quantities beyond the MOS to fully understand the meaning of the subjective results. While the SOS values reflect the user diversity, the quantiles help to understand the fraction of users with very bad (e.g. 10 % quantile) or very good quality perception (e.g. 90 % quantile).

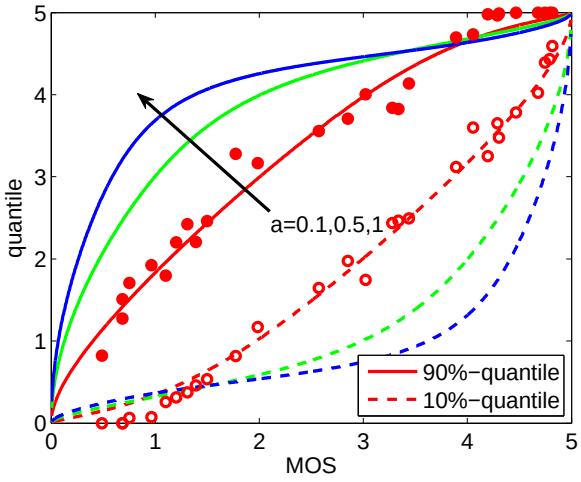
2) Results for Video QoE on a Continuous Rating Scale: Figure 10 shows the results of the video QoE study [51]. A continous rating scale from 0 to 5 (cf. Figure 14) was used. The two labs where the study was carried out are denoted as "EPFL" and "PoLiMi" in the result figures. The packet loss in the video transmission was varied in $p_L \in \{0; 0.1; 0.4; 1; 3; 5; 10\}$ (in %) for four different videos. In total, 40 subjects assessed 28 test conditions. The MOS, SOS, as well as the 10 %- and 90 %-quantile were computed for each test condition over all 40 subjects from both labs. More details on the setup can be found in the Appendix A-D.

Fiugre 10a provides a SOS-MOS plot. The markers depict the tuple (MOS,SOS) for each of the 28 test conditions (PoliMi '□' and EPFL '◊'). The dashed lines shows the SOS fitting function with the corresponding SOS parameters for the two labs which are almost identical. When merging the results from both labs, we arrive at the SOS parameter $a = 0.10$. Due to the user diversity, we observe of course positive SOS values for any test condition (the theoretical minimum SOS is zero for the continuous scale), but the diversity is lower than for web QoE. Subjects are presumably more confident on (or familiar with) how to rate an impaired video, while the impact of temporal stimuli i.e. PLT for web QoE is more difficult to evaluate.

For each test condition, we observe a MOS value and the corresponding SOS value according to the SOS parameter. We fit the user ratings per packet loss ratio with a truncated normal distribution in $[0; 5]$ with the measured mean μ (MOS) and standard deviation σ (SOS). Thus, the user ratings U follow the truncated normal distribution, i.e. $U \sim N(\mu; \sigma; 0; 5)$



(a) SOS-MOS Plot. The markers depict the tuple (MOS,SOS) for each of the 28 test conditions (PoliMi '□' and EPFL '◊'). The dashed lines shows the SOS fitting function with the corresponding SOS parameters for the two labs which are almost identical. When merging the results from both labs, we arrive at the SOS parameter $a = 0.10$. But the diversity is lower than for web QoE. Subjects are more sure how to rate an impaired video, while the impact of temporal stimuli i.e. PLT for web QoE is more difficult to evaluate for subjects. The solid black curve depicts the maximum SOS.



(b) Quantile-MOS Plot. The markers depict the empirically derived 90 %-quantiles '•' and 10 %-quantiles '○', respectively. Furthermore, we plot the quantiles depending on MOS for user ratings following a truncated normal distribution and SOS parameter $a = 0.1, 0.5, 1$. The SOS hypothesis returns for each MOS value μ the related SOS value σ which allows to compute the quantiles of the truncated normal distribution, i.e. $U \sim N(\mu; \sigma; 0; 5)$. The solid and dashed lines depict the 90 %- and 10 %-quantiles, respectively.

Figure 10: Video QoE. Results of the video QoE study [51]. A continous rating scales from 0 to 5, cf. Figure 14, was used in the experiments for subjects evaluating the quality of videos transmitted over a noisy channel [51]. The study was repeated in two different labs denoted as 'EPFL' and 'PoLiMi' in the result figures. The packet loss in the video transmission was varied in $p_L \in \{0; 0.1; 0.4; 1; 3; 5; 10\}$ (in %) for four different videos. In total, 40 subjects evaluated 28 test conditions.

with $U \in [0;5]$. We observe a very good match between the empirical CDF and the truncated normal distribution, see Figure 15b in the appendix. This is not obivous and no trivial result, although the first two moments of both distributions are identical, the underlying distributions could be very different, see Section II. Thus, together with the SOS parameter a, the user voting distribution is completely specified for any MOS value μ on the rating scale i.e. $\mu \in [0;5]$.

Figure 10b shows the quantiles as a function of MOS. The filled '•' and non-filled markers '○' depict the empirically derived 90 %- and 10 %-quantiles for the 28 test conditions, respectively. Furthermore, we plot the quantiles depending on MOS for user ratings U following a truncated normal distribution and SOS parameter $a = 0.1, 0.5, 1$. Note that we measure $a = 0.10$ in the experiments on video QoE. The SOS parameter 0.5 leads to $\sqrt{0.5/0.1} = 2.24$ higher SOS values for an observed MOS. The SOS parameter 1 leads to the maximum possible SOS which is 3.16 times higher than in the subjective data. Due to the SOS hypothesis and a given SOS parameter a, we obtain for each MOS value μ the related SOS value $\sigma(\mu; a)$, see Eq.(5). Thereby, a MOS value represents the outcome of a concrete test condition. The parameters μ and σ are input parameters of the truncated normal distribution which allows us to compute the α-quantile of the truncated normal distribution, i.e. $U \sim N(\mu; \sigma; 0; 5)$. The solid and dashed lines depict the 90 %- and 10 %-quantiles, respectively. We observe that the truncated normal distribution corresponding to the SOS parameter $a = 0.1$ fit very well the empirical quantiles. With the information of the SOS parameter, the quantiles, etc., can be completely derived for any MOS value. Similarly to the discrete rating scale results from the web QoE study, we observe strong differences between the MOS and the quantiles when using a continuous rating scale. The maximum difference between the 90 %-quantile and MOS is $3.62 - 2.42 = 1.20$. Also on the continuous scale, the MOS masks out such meaningful information for providers.

3) Results for Speech QoE – Comparison between Continuous and Discrete Rating Scale: When comparing the SOS values from the web and video study, we observe that the discrete rating scale leads to higher SOS values than the continous scale. However, the higher user diversity may be caused by the application [44]. Therefore, we briefly discuss the speech QoE study (as already discussed in Section V-C and described in the Appendix A-B). Subjects rate the QoE for certain test conditions on a discrete and a continuous scale which allows a comparison.

As a result (cf. Figure 12a), the SOS parameter $a_d = 0.23$ and $a_c = 0.12$ are obtained for the discrete and the continuous scales, respectively. For the discrete scale, we observe larger SOS values than for the continous scale, which can also be seen by the larger SOS parameter $a_d > a_c$. In particular, on the discrete scale, the SOS values are larger by a factor of $\sqrt{a_d/a_c} \approx 1.38$. This observation seems to be reasonable, as the continuous scale has more discriminatory power than the discrete scale. Subjects can assess the quality more fine granular on the continuous scale by choosing a value $x \in [i; i+1]$, while the subject has to decide between i and $i+1$ on a discrete scale. The minimum SOS for a given MOS value is zero for a continuous scale, while the minimum SOS is larger than zero and depends on the actual MOS value, cf. Eq.(3).

Although the results seem to be valid from a statistical point of view, the literature shows conflicting results. In [52], subjective studies on the image aesthetic appeal were conducted using a discrete 5-point ACR scale as well as a continuous scale. However, similar SOS parameters were obtained for both rating scales. [53] compared two different subjective quality assessment methodologies for video QoE: *absolute category rating (ACR)* using a 5-point discrete rating scale and *subjective assessment methodology for video quality (SAMVIQ)* using a continuous rating scale. As a key finding, SAMVIQ is more precise (in terms of confidence interval width of a MOS value) than ACR for the same number of subjects. However, SAMVIQ uses multiple stimuli assessment, i.e. multiple viewing of a sequence. There are further works [54]–[57] comparing different (discrete and continuous) rating scales as well as assessment methodologies like SAMVIQ in terms of reliability and consistency of the user ratings. We note, however, that they do not address the issues of using averages to characterize the results of those assessments. A detailed analysis of the comparison of continuous and discrete rating scales and their impact on QoE metrics is left for future work.

E. Comparison of Results

All experiments and some key quantities are summarized in Table II, which may serve as a guideline to properly describe subjective studies and their results in order to extract as much insight from them as possible. For comparing the key measures across the experiments with different rating scales, the user ratings in all experiments are mapped on a scale from 1 (bad quality) to 5 (excellent quality).

The user rating diversity seems to be lower when using a continuous rating scale than a discrete one. This can be observed from the SOS parameter a, but also the maximum SOS at a certain MOS. It should be noted, however, that in more interactive services such as web QoE, there might be an inherently higher variation of user ratings, due e.g., to uncertainty on how to rate the overall quality.

The MSE-optimal parameter θ is determined by minimizing the MSE between the θ-acceptability of the measurement data and the %GoB-MOS. The discrete rating scale can only find a discrete value θ and therefore stronger deviations between the %GoB estimator and the θ-acceptability arise. We see that for the task-related web QoE, the MSE-optimal parameter is $\theta = 3$. This means that the ratio of users rating fair or better match the %GoB curve. For the continuous rating scales, optimal continuous thresholds can be derived. For the speech QoE and the video QoE on continuous scales, a value of θ around 3 matches the %GoB curve.

The limitations of MOS are made evident by the minimum %GoB ratio $P(U \geq 4)$ for all test conditions which lead to a MOS value equal or larger than 4. The ratio shows how

Table II: Description of the subjective studies conducted for analyzing QoE for different applications. Key QoE metrics for the subjective results are depicted. For comparing the metrics across the experiments with different rating scales, the user ratings in all experiments are mapped on a scale from 1 (bad quality) to 5 (excellent quality). The MSE-optimal parameter θ is determined by minimizing the MSE between the θ-acceptability of the measurement data and the $\%\mathrm{GoB}$, cf. Eq. (8).

Experiment	Speech QoE Disc.	Speech QoE Cont.	PLT Web QoE	Video QoE	Task-related Web QoE
Rating Scale Type	Discrete	Continuous	Discrete	Continuous	Discrete
Ω – Rating Scale	$\{1, \cdots, 5\}$	$[0; 6]$	$\{1, \cdots, 5\}$	$[0; 5]$	$\{1, \cdots, 5\}$
R – #subjects	20	20	72	40	10–30 per test condition
J – #conditions	86	86	40	28	23
a – SOS parameter	0.230	0.123	0.268	0.099	0.266
Max. SOS (at MOS)	1.342 (2.700)	0.954 (3.953)	1.153 (2.903)	0.709 (2.420)	1.335 (3.000)
MSE-optimal θ	4	3.014	4	3.065	3
Min. $\%\mathrm{GoB}$ $P(U \geq 4)$	0.700	0.500	0.806	0.950	0.688
Max. difference between 90 %-quantile and MOS	5-2.700=2.300	4.840-3.426=1.414	4-2.139=1.861	3.623-2.420=1.203	4.900-3.000=1.900
Max. difference between MOS and 10 %-quantile	3.300-1.5=1.800	3.953-2.423=1.530	2.903-1=1.903	3.416-2.397=1.020	3-1.100=1.900

many users accept (or do not accept) the condition, although the MOS exceeds the threshold.

Another limitation of the MOS is highlighted by the quantiles. In particular, the maximum difference between the 90 %-quantile and the MOS values is shown to reach up to 2 points on the 5-point scale. This highlights the importance of considering QoE beyond the MOS.

VI. Conclusions

In this article, we argued for going beyond MOS when performing subjective quality tests. While MOS is a practical way to convey QoE measures and a simple to interpret scalar value, it hides important details about the results. These details often have a significant impact in terms of the service technical performance, and on the business aspects of the service.

Our contributions are many-fold. Firstly, while there are many works in the literature dealing with subjective and objective quality assessment, they are mostly limited to MOS, while ignoring higher order statistics and the relation between quality and acceptance. Our first contribution is thus that there are other tools available for understanding QoE besides the MOS, their importance, and how they are used. A second contribution is a survey of the available QoE measures, their definition and interpretation. Using these tools brings more insight into QoE analysis. Our third contribution is a showcase, by means of analyzing several concrete use cases, of how these analysis tools are used, highlighting the extra insight they bring beyond that of the MOS. We analyze e.g., the impact of using continuous vs. discrete scales on the accuracy of the assessment, the relation between quality and acceptance.

Concerning acceptability ratings, we note the following difference between acceptability (as an explicit question to the users) and the concept of θ-acceptability. In a subjective experiment, each user defines their own threshold reflecting the point where QoE is good enough to accept the service. This is the result of a complex cognitive process. In contrast, θ-acceptability considers a globally defined threshold (e.g. defined by the ISP, or whoever designed the subjective test scale used) which is the same for all users. This leads to a discrepancy with the subjective results, which can vary significantly with the application considered. For instance, in the case of Web QoE with a task, the discrepancy is rather large. In the case of speech, the E-Model-inspired $\%\mathrm{GoB}$ estimator in Eq. (8) maps a MOS value of 1 to a $\%\mathrm{GoB}$ of 0 %, as a speech service is not possible any more if the quality is too degraded, and hence it is unacceptable. In contrast, in the Web QoE case, a very bad QoE can still result in a usable service which is accepted by the end user. Thus, the user can still complete for example the task of finding a wikipedia article, although the page load times are very high. This may explain why 20 % of the users accept the service although they rate the QoE with bad quality (1). From this, we can recommend that acceptability be included explicitly as part of subjective assessment, as it cannot be directly inferred from user ratings on the quality of a service, e.g., on a 5-point MOS scale.

These differences in the way that users accept (or not) the service quality, and how this relates to MOS values can provide key insights to providers when assessing the QoE delivered to their users, and how it may relate to issues such as churn. Asking explicitly about acceptability seems like a necessary step to consider in certain use cases (where business considerations are important, for example). Likewise, thinking in terms of distributions, or at least quantiles, provides more actionable information to service and content providers, as it allows them to better grasp how their users actually perceive the quality of the service, and how many of those users may be happy or unhappy (or, in following with the QoE definition, delighted or annoyed) with it. This implies that existing quality models that provide MOS estimates should be complemented (or eventually replaced) by new models that estimate rating distributions, or key quantiles. These results are directly relevant to several aspects of service provisioning, from the more technical ones, such as network management, to marketing and pricing strategies, to customer support.

In summary, we have made the case for going beyond the MOS, and delving deeper into the analysis of QoE assessment results, with practical applications (e.g., business and engineering considerations on the service providers' part) in mind.

VII. Acknowledgements

The authors would like to thank the anonymous reviewers for their very constructive comments which helped to improve the contributions of this article.

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grants HO 4770/1-2 and TR257/31-2 and in the framework of the COST ACROSS Action. Martín Varela's work was partially funded by Tekes, the Finnish agency for research innovation, in the context of the CELTIC+ project NOTTS. The authors alone are responsible for the content.

VIII. Supplementary Material

Matlab scripts for computing the QoE metrics for given data sets are available as supplementary material to this publication as well as in GitHub [58]. The formal definition of the QoE metrics is available as supplementary material as well as technical report [43].

References

[1] T. Hoßfeld, P. E. Heegaard, and M. Varela, "QoE beyond the MOS: Added Value Using Quantiles and Distributions", in *7th International Workshop on Quality of Multimedia Experience (QoMEX 2015)*, Costa Navarino, Greece, May 2015.

[2] ITU-T Recommendation P.800.1, *Mean Opinion Score (MOS) Terminology*, International Telecommunication Union, Mar. 2003.

[3] U. Engelke and H.-J. Zepernick, "Perceptual-based Quality Metrics for Image and Video Services: A Survey", in *3rd EuroNGI Conference on Next Generation Internet Networks*, Trondheim, Norway, May 2007.

[4] A. Van Moorsel, "Metrics for the Internet Age: Quality of Experience and Quality of Business", in *5th International Workshop on Performability Modeling of Computer and Communication Systems (PMCCS 5)*, Erlangen, Germany, Sep. 2001.

[5] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison", *IEEE Transactions on Broadcasting*, vol. 57, no. 2, 2011.

[6] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and Objective Quality Assessment of Image: A Survey", *arXiv preprint arXiv:1406.7799*, 2014.

[7] J. Korhonen, N. Burini, J. You, and E. Nadernejad, "How to Evaluate Objective Video Quality Metrics Reliably", in *4th International Workshop on Quality of Multimedia Experience (QoMEX 2012)*, Melbourne, Australia, Jul. 2012.

[8] M. Mu, A. Mauthe, G. Tyson, and E. Cerqueira, "Statistical Analysis of Ordinal User Opinion Scores", in *IEEE Consumer Communications and Networking Conference (CCNC 2012)*, Las Vegas, USA, Jan. 2012.

[9] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives", *Multimedia Systems*, vol. 22, no. 2, Mar. 2014.

[10] L. Janowski and Z. Papir, "Modeling Subjective Tests of Quality of Experience with a Generalized Linear Model", in *1st IEEE International Workshop on Quality of Multimedia Experience (QoMEX 2009)*, San Diego, USA, Jul. 2009.

[11] A. Bouch, M. A. Sasse, and H. DeMeer, "Of Packets and People: A User-centered Approach to Quality of Service", in *8th International Workshop on Quality of Service (IWQOS '00)*, Pittsburgh, USA, Jun. 2000.

[12] H. Nachlieli and D. Shaked, "Measuring the Quality of Quality Measures", *IEEE Transactions on Image Processing*, vol. 20, no. 1, 2011.

[13] ITU-T Recommendation P.800, *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, Aug. 1996.

[14] ITU-R Recommendation BT.500-13, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, International Telecommunication Union, Jan. 2012.

[15] A. Watson and M. Sasse, "Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications", in *ACM Multimedia '98*, Bristol, UK, Sep. 1998.

[16] H. Knoche, H. G. De Meer, and D. Kirsh, "Utility Curves: Mean Opinion Scores Considered Biased", in *7th International Workshop on Quality of Service (IWQoS '99)*, London, UK, Jun. 1999.

[17] L. Gros, N. Chateau, and A. Macé, "Assessing Speech Quality: A New Approach", in *4th European Congress on Acoustics (Forum Acusticum)*, Budapest, Hungary, Aug. 2005.

[18] L. Gros, N. Chateau, and V. Durin, "Speech Quality: Beyond the MOS Score", in *Measurement of Speech and Audio Quality in Networks Workshop (MESAQIN'06)*, Prague, Czech Republic, Jun. 2006.

[19] V. Durin and L. Gros, "Measuring Speech Quality Impact on Tasks Performance", in *9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, Brisbane, Australia, Sep. 2008.

[20] A. Sackl, P. Zwickl, and P. Reichl, "The Trouble with Choice: An Empirical Study to Investigate the Influence of Charging Strategies and Content Selection on QoE", in *9th International Conference on Network and Service Management (CNSM 2013)*, Zürich, Switzerland, Oct. 2013.

[21] T. Mäki, P. Zwickl, and M. Varela, "Network Quality Differentiation: Regional Effects, Market Entrance, and Empirical Testability", in *IFIP Networking 2016*, Vienna, Austria, May 2016.

[22] M. H. Pinson, S. Wolf, and R. B. Stafford, "Video Performance Requirements for Tactical Video Applica-

tions", in *IEEE Conference on Technologies for Homeland Security (HST)*, Woburn, MA, USA, May 2007.

[23] M. A. Sasse and H. Knoche, "Quality in Context – An Ecological Approach to Assessing QoS for Mobile TV", in *2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems (PQS 2006)*, Berlin, Germany, Sep. 2006.

[24] P. Spachos, W. Li, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Acceptability and Quality of Experience in Over The Top Video", in *IEEE ICC 2015 - Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI)*, London, UK, Jun. 2015.

[25] T. D. Pessemier, K. D. Moor, A. J. Verdejo, D. V. Deursen, W. Joseph, L. D. Marez, L. Martens, and R. V. de Walle, "Exploring the Acceptability of the Audiovisual Quality for a Mobile Video Session based on Objectively Measured Parameters", in *3rd International Workshop on Quality of Multimedia Experience (QoMEX 2011)*, Mechelen, Belgium, Sep. 2011.

[26] ITU-T Recommendation P.10/G.100, Amendment 2, *Vocabulary and Effects of Transmission Parameters on Customer Opinion of Transmission Quality*, International Telecommunication Union, 2006.

[27] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. Springer Science & Business Media, 2005, ISBN: 9783540288602.

[28] P. Le Callet, S. Möller, and A. Perkis (eds.), "Qualinet White Paper on Definitions of Quality of Experience", *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, Mar. 2013.

[29] S. Jumisko-Pyykkö and T. Vainio, "Framing the Context of Use for Mobile HCI", *International Journal of Mobile Human Computer Interaction*, vol. 2, no. 4, 2010.

[30] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. Springer US, Aug. 2000, ISBN: 0792378946.

[31] S. Egger, P. Reichl, and K. Schoenenberg, "Quality of Experience and Interactivity", in *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, Eds. Springer International Publishing, 2014, ISBN: 978-3-319-02681-7.

[32] ITU-T Recommendation G.107, *The E-Model, a Computational Model for Use in Transmission Planning*, International Telecommunication Union, Apr. 2011.

[33] ITU-T Recommendation P.910, *Subjective Video Quality Assessment Methods for Multimedia Applications*, International Telecommunication Union, Apr. 2008.

[34] ITU-T Handbook on Telephonometry, International Telecommunication Union, 1992.

[35] ITU–T Recommendation P.805, *Subjective Evaluation of Conversational Quality*, International Telecommunication Union, Apr. 2007.

[36] ITU-T Recommendation P.810, *Modulated Noise Reference Unit (MNRU)*, International Telecommunication Union, Feb. 1996.

[37] H.-S. Kim and C.-H. Yoon, "Determinants of Subscriber Churn and Customer Loyalty in the Korean Mobile Telephony Market", *Telecommunications Policy*, vol. 28, no. 9-10, 2004, ISSN: 0308-5961.

[38] D. Soldani, M. Li, and R. Cuny, *QoS and QoE Management in UMTS Cellular Systems*. Wiley, 2006, ISBN: 9780470016398.

[39] ITU-T P.Sup3 – Suppl. 3 to ITU-T Series P Recommendations, *Models for Predicting Transmission Quality from Objective Measurements*, International Telecommunication Union, Mar. 1993.

[40] ETSI Technical Report ETR 250, *Transmission and Multiplexing (TM); Speech Communication Quality from Mouth to Ear for 3,1 kHz Handset Telephony across Networks*, European Telecommunications Standards Institute, Jul. 1996.

[41] T. L. Keiningham, B. Cooil, T. W. Andreassen, and L. Aksoy, "A Longitudinal Examination of Net Promoter and Firm Revenue Growth", *Journal of Marketing*, vol. 71, no. 3, Jul. 2007.

[42] E. de Haan, P. C. Verhoef, and T. Wiesel, "The Predictive Ability of Different Customer Feedback Metrics for Retention", *International Journal of Research in Marketing*, vol. 32, no. 2, 2015, ISSN: 0167-8116.

[43] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, "Formal Definition of QoE Metrics", *Arxiv cs.MM*, Jul. 2016.

[44] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!", in *3rd International Workshop on Quality of Multimedia Experience (QoMEX 2011)*, Mechelen, Belgium, Sep. 2011.

[45] R. Schatz and S. Egger, "An Annotated Dataset for Web Browsing QoE", in *6th International Workshop on Quality of Multimedia Experience (QoMEX 2014)*, Singapore, Sep. 2014.

[46] ITU-T Recommendation P.1501, *Subjective Testing Methodology for Web Browsing*, International Telecommunication Union, Apr. 2013.

[47] P. Reichl, S. Egger, S. Möller, K. Kilkki, M. Fiedler, T. Hoßfeld, C. Tsiaras, and A. Asrese, "Towards a Comprehensive Framework for QoE and User Behavior Modelling", in *7th International Workshop on Quality of Multimedia Experience (QoMEX 2015)*, Costa Navarino, Greece, May 2015.

[48] F. Köster, D. Guse, M. Wältermann, and S. Möller, "Comparison between the Discrete ACR Scale and an Extended Continuous Scale for the Quality Assessment of Transmitted Speech", in *Fortschritte der Akustik - DAGA 2015: Plenarvortr. u. Fachbeitr. d. 41. Dtsch. Jahrestg. f. Akust.*, Nürnberg, Germany, Mar. 2015.

[49] D. Gibbon, "EUROM. 1 German speech database", *ESPRIT project 2589 Report (SAM, Multi-Lingual Speech*

Input/Output Assessment, Methodology and Standardization), 1992.

[50] T. Hoßfeld, R. Schatz, S. Biedermann, A. Platzer, S. Egger, and M. Fiedler, "The Memory Effect and Its Implications on Web QoE Modeling", in *23rd International Teletraffic Congress (ITC 23)*, San Francisco, USA, Sep. 2011.

[51] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective Assessment of H. 264/AVC Video Sequences Transmitted over a Noisy Channel", in *1st International Workshop on Quality of Multimedia Experience (QoMEX 2009)*, San Diego, US, 2009.

[52] E. Siahaan, J. A. Redi, and A. Hanjalic, "Beauty is in the Scale of the Beholder: Comparison of Methodologies for the Subjective Assessment of Image Aesthetic Appeal", in *6th International Workshop on Quality of Multimedia Experience (QoMEX 2014)*, Singapore, 2014.

[53] S. Péchard, R. Pépion, and P. Le Callet, "Suitable Methodology in Subjective Video Quality Assessment: A Resolution Dependent Paradigm", in *3rd International Workshop on Image Media Quality and its Applications (IMQA2008)*, Kyoto, Japan, Sep. 2008.

[54] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video", in *2nd International Workshop on Quality of Multimedia Experience (QoMEX 2010)*, Trondheim, Norway, 2010.

[55] M. H. Pinson and S. Wolf, "Comparing Subjective Video Quality Testing Methodologies", in *Visual Communications and Image Processing 2003*, International Society for Optics and Photonics, 2003.

[56] M. D. Brotherton, Q. Huynh-Thu, S David, and K. Brunnstrom, "Subjective Multimedia Quality Assessment", *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 89, no. 11, 2006.

[57] Q. Huynh-Thu and M. Ghanbari, "A Comparison of Subjective Video Quality Assessment Methods for Low-bit Rate and Low-resolution Video", in *Signal and Image Processing (SIP 2005)*, Honolulu, Hawaii, USA, Aug. 2005.

[58] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, *Scripts for the Computation of QoE Metrics beyond the MOS*, Jul. 2016. [Online]. Available: https://github.com/hossfeld/QoE-Metrics.git.

Appendix A
Experimental Setup & Additional Results

A. Experimental Setup for Task-Related Web QoE

The experiments in [45] investigated web QoE. In the campaign conducted, subjects were asked to carry out a certain task, and the impact of page load times (PLT) during the web session was investigated. Besides assessing the overall quality of the web browsing session, subjects additionally answered an acceptance question. The experiment was carried out in a laboratory environment, with 32 subjects. In contrast to the PLT experiments on web QoE in Section A-C, subjects carried out a task and evaluated the overall quality of the complete web session related to the task in Section V-B – as opposed to giving their opinions based on the PLT of a single page.

Four different websites (encyclopedia, cooking community, news portal, travel portal) were used in the test, with strongly varying page complexity in terms of number of visual elements and modalities (textual, visual, audio-visual). For each of the websites, subject were asked to perform a certain task (cf. Table III), while network conditions were changed. In particular, six downlink bandwidth conditions were tested, leading to different page load times for the presented web pages during each session. In total, 23 different test conditions (i.e. website and bandwidth condition) were tested. However, each participant rated only a subset of those conditions, resulting in 418 opinion ratings and acceptance answers with 10–30 opinions per test condition.

After each condition, subjects were asked to rate their overall experienced quality on a 9-point ACR scale, see Figure 11, as well as acceptability. Note that this test methodology conforms with ITU-T Rec. P.1501 [46].

Each session lasted for approximately two hours, including a briefing, training conditions, debriefing interviews and a break of roughly 10 min halfway through the test. For the web browsing tasks, the test operator set different maximum network downlink bandwidth conditions to be experienced, remotely started a browser session with the corresponding website, asked the user to perform a certain browsing scenario on a notebook Windows PC and triggered electronic rating prompts after each condition. The session duration for each condition, from starting the browser session until the display of the electronic rating prompt was approximately 180 s.

9	excellent
8	
7	good
6	
5	fair
4	
3	poor
2	
1	bad

Figure 11: 9-point quality scale as used for the task-related web QoE experiments [45] in Section V-B.

Table III: Subjects conducted a certain task in the web browsing experiments in Section V-B.

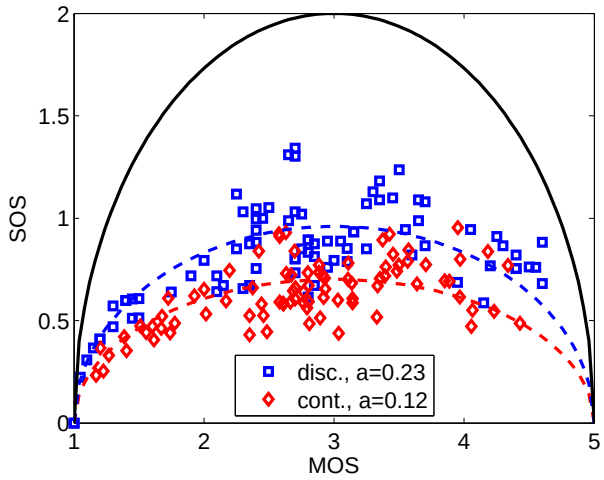
Encyclopedia	Find the article [*GIVEN TERM* e.g New York City]. Try to answer the following five questions .... by browsing through the article by clicking on links
Cooking	Browse to search for three recipes you would like to cook in the given section.
News	Try to get an overview of the current news in the given section.
Travel	Browse through the hotels in ..... and select five you would like to stay in.

B. Speech Quality on Discrete and Continuous Scale

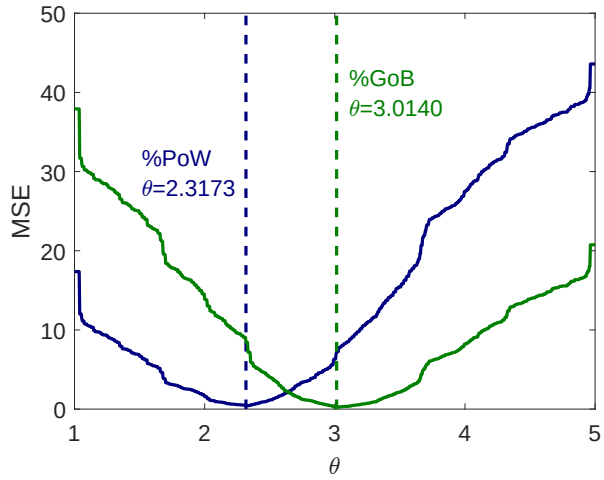
The opinion ratings of the subjects on speech quality are taken from [48]. We briefly describe the experimental setup and focus only on the details relevant for our analysis. The listening-only experiments were conducted by 20 subjects in an environment fulfilling the requirements in ITU-T Rec. P.800 [2]. As source speech material, recorded sentences by one male and one female German speakers from the EUROM database [49] were used and sampled at 8 kHz (narrowband) and 16 kHz (wideband), respectively. The narrowband test consisted of 18 conditions including different loudness levels, noise types (babble and hoth), bandpasses, codecs, codec tandems, MNRUs, and packet losses. The wideband tests consisted of 25 conditions including clean speech and different loudness levels, noise types (babble and hoth), bandpasses, wideband codecs, codec tandems, wideband MNRUs, and packet losses. The test conditions were rated for the male and the female speaker content resulting into 86 different test conditions in total.

The subjects assessed the different test conditions on two different scales: the ACR scale (Figure 6) and the extended continuous scale (Figure 7). To be more precises, each subject was using both scales during the experiment. The scales were incorporated in a software program that led the participants through the experiments. The ACR scale was realized with software buttons pre-annotated with the numbers 5 to 1 and labelled according to ITU-T Rec. P.800 [2]. The extended continuous scale was depicted as a bitmap, together with a software slider. The labels were internally assigned to numbers of the interval [0,6] in such a manner that the attributes corresponding to ITU-T Rec. P.800 were exactly assigned to the numbers $1, \cdots, 5$.

The test was split into four sessions, i.e. narrowband and wideband test for the discrete and the continuous scale. The samples of the different test conditions were randomized per session, so as to avoid learning effects. The first two sessions always consisted of the narrowband context, whereas the last two sessions consisted of the wideband context. For both contexts, the scales were presented in random order. The order of narrowband and wideband context was fixed to consider internal quality expectation of the participants when migrating from traditional narrowband to modern wideband speech. Dedicated training samples were asked to be rated in prior to each session in order to foster the sense for the range of quality to be expected.



(a) SOS-MOS Plot. The markers depict the tuples (MOS,SOS) for any test condition of the subjective study on the discrete and the continuous rating scale. The dashed lines show the fitting function according to the SOS hypothesis with the SOS parameter a specified in the legend. The solid black line shows the maximum SOS for a given MOS value. For the discrete scale, we observe larger SOS values than for the continous scale.



(b) Parameter θ fitting. For the continuous rating scale, the parameter θ is determined which minimizes the mean squarred error (MSE) between the E-model and the subjective data. For the ratio $\%\mathrm{PoW}$ of dissatisfied users rating poor or worse in the E-model, the value $\theta = 2.32$ leads to a minimum MSE. For the ratio $\%\mathrm{GoB}$ of satisfied users rating good or better in the E-model, the value $\theta = 3.0140$ leads to a minimum MSE. Obviously, the MSE optimal value is larger than 2 for $\%\mathrm{PoW}$ and smaller than 4 for $\%\mathrm{GoB}$, respectively. The E-model overestimates $\%\mathrm{GoB}$ for given MOS values.

Figure 12: Speech QoE. Results of the speech QoE study [48]. For the 86 test conditions, QoE metrics were computed over the 20 subjects for the discrete 5-point ACR scale (Figure 6) and the extended continuous scale (Figure 7). The results for the discrete scale are marked with '□', while the QoE measures for the continuous scale are marked with '◊'.

C. Web QoE and Discrete Rating Scale

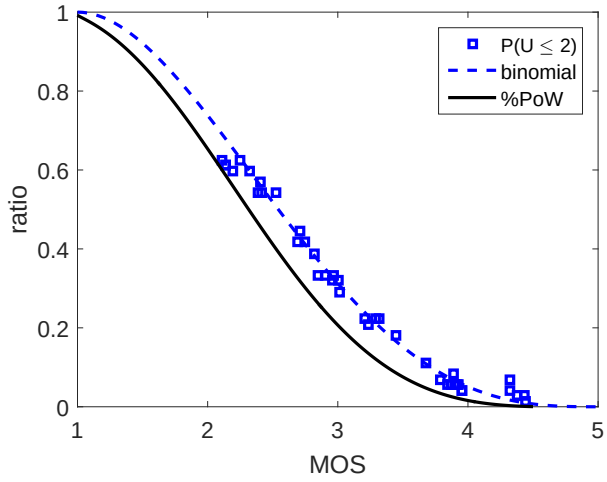
The opinion ratings from the subjective user study on web QoE is based on the experiments in [50]. Subjects sequentially browsed a set of web pages while the page loading times (PLT) were varied in order to quantify the impact of PLT on web QoE. In total, 72 subjects completed the online test in their preferred environment. The test was implemented by means of a Java applet to ensure that all participants experienced the same pre-defined sequences of PLTs – regardless of their Internet access' performance. The participants interacted with the Java applet that already contained the contents of the websites. The applet simulated the download of various web pages with predefined page load times.

The content consisted of a simple photo web page displaying a single image in order to avoid any content specific influences on user quality perception and rating behavior. During the tests, a user downloaded and viewed sequentially 40 different web pages with predefined page load times. The maximum PLT was 1.2 s. The minimum and the mean PLT were 0.24 s and 0.66 s, respectively.

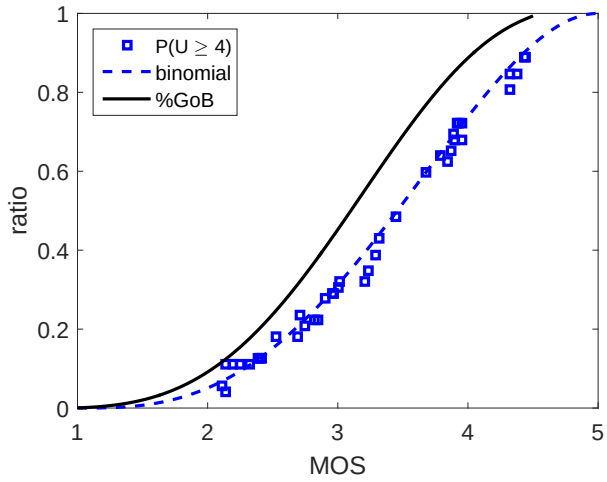
After the download of each web page, the user was prompted for his or her opinion about the overall QoE on a given rating scale. The web page contained rating buttons from 1 to 5 (similar to Figure 6), which were used by the test subjects to give his/her personal opinion score on the overall quality during the browsing session. In particular, subjects were asked to answer the question *"Are you satisfied with this download speed?"*.

Figure 13a depicts the %PoW in relation to MOS. The markers depict the MOS and the ratio $P(U \leq 2)$ for each of the test conditions. The dashed blue line represents the corresponding ratio for the shifted binomial distribution. The solid black line shows the %PoW ratio depending on MOS using the definitions in Section IV-F. The empirical results highlight the averaging effect of the MOS, clearly showing that it is not a sufficient measure to fully understand the results of a subjective study. It can be seen that even for fair or good overall MOS values, a significant number of users perceives the quality as poor or bad. Simply using the %PoW estimator underestimates the ratio of dissatisfied users $P(U \leq 2)$ with a maximum difference 11.01 % at MOS 2.731.

Figure 13b illustrates the results for ratio %GoB of users rating good or better i.e. $P(U \geq 4)$. The markers depict the MOS and the %GoB for each test condition. The dashed blue line represents the corresponding ratio for the binomial distribution. The solid black line shows the %GoB ratio depending on MOS. Even for good MOS values larger than 4, up to 20 % of the users are rating fair or worse. The %GoB again overestimates the ratio of satisfied users $P(U \geq 4)$ with a maximum difference 17.24 % at MOS 3.572. In summary, the web QoE results confirm the observations for the speech QoE results on the discrete scale.



(a) %PoW-MOS Plot. The markers depict the MOS and the ratio $P(U \leq 2)$ for each of the test conditions. The dashed blue line represents the corresponding ratio for the shifted binomial distribution. The solid black line shows the %PoW ratio estimation depending on MOS as defined in Eqs. (9). For the minimum MOS 2.111 observed, 37.50 % if the users are rating fair or better. The %PoW estimator underestimates the ratio of dissatisfied users $P(U \leq 2)$ with a maximum difference 11.01 % at MOS 2.731.



(b) %GoB-MOS Plot. The markers depict the MOS and the %GoB i.e. $P(U \geq 4)$ for each test condition of the web QoE study. The dashed blue line represents the corresponding ratio for the shifted binomial distribution. The solid black line shows the %GoB ratio estimation depending on MOS, see Eqs. (8). Although the MOS is larger than 4, up to 20 % of the users are rating fair or worse. The %GoB estimator overestimates the ratio of satisfied users $P(U \geq 4)$ with a maximum difference 17.24 % at MOS 3.572.

Figure 13: Web QoE for PLT only. Results of the web QoE study [50]. The page load time was influenced for each test condition and 72 subjects rated the overall quality on a discrete 5-point ACR scale. Each user viewed 40 web pages with different images on the page and PLTs from 0.24 s to 1.2 s resulting into 40 test conditions per user. For each test condition, the MOS, SOS, entropy, %GoB, %PoW, as well as 10 %- and 90 %-quantiles are computed over the opinions of the 72 subjects.

D. Video QoE and Continuous Rating Scale

The subjective results on the experiments on video QoE are publicly available and taken from [51]. The study aimed at investigating the impact of transmitting video sequences over a noisy channel on the video quality experienced by the end users. Subjects were in a room with controlled lighting and color temperature, and seated directly in line with the center of the video display at a fixed viewing distance. The test was conducted in two different laboratories with identical test conditions which resulted into 40 subjects in total.

In the analysis, we consider four different video sequences of 10 s available at CIF spatial resolution (352×288 pixels) at a frame rate of 30 fps. Additionally two other sequences were used for training the subjects (and subsequently not used in the actual test session). The video sequences were encoded with H.264/AVC. Details on the encoding parameters and other experimental parameters can be found in [51]. For each of the original H.264/AVC bitstreams, corrupted test sequences were generated by dropping IP packets according to a two-state Gilbert's model to generate burst loss pattern due to the noisy channel. Six different packet loss ratios were applied $[0.1\,\%, 0.4\,\%, 1\,\%, 3\,\%, 5\,\%, 10\,\%]$. This results into a reference sequence and the six degraded ones for each of the video contents. In total, 28 different test video sequences were considered.

Each test session involved only one subject per display assessing the test material. The subject was asked to rate the quality of the presented test sequence using the 5-point ITU continuous scale in the range $[0; 5]$ as described in ITU-R Rec. BT.500-13 [14]. The five point continuous rating scale is depicted in Figure 14. The presentation order of test sequences for each subject was randomized, taking care that consecutive conditions did not use the same content. A training session was performed during which the meaning of the labels were explained by the test moderator. After the training, the actual test session was carried out.

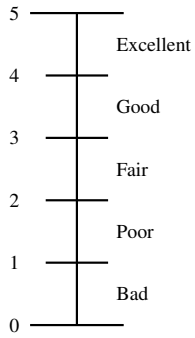
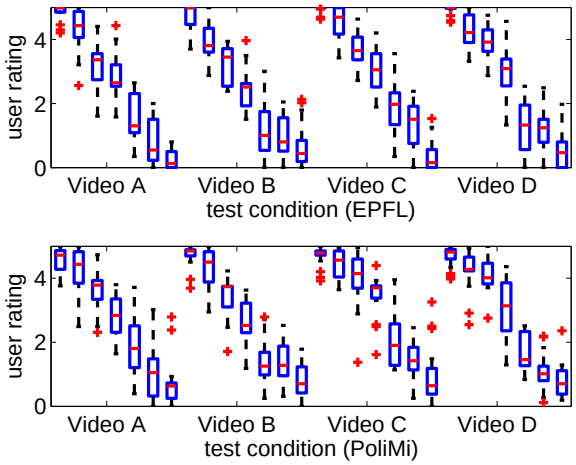
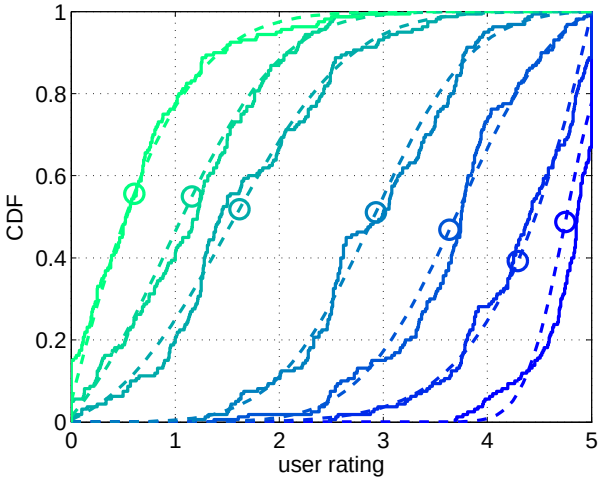


Figure 14: Five point continuous quality scale as used for the video QoE experiments [51]. It has to be noted that the numerical values $(0, \cdots, 5)$ attached to the scale were used only for data analysis and were not shown to the subjects during the test.



(a) Box Plot. A graphical illustration of the subjective ratings on a continous scale is a box plot. For each test condition, the user ratings represent a continuous random variable. The box '□' quantifies the lower quartile and the upper quartile values as box. The median is provided as line '-' in the box. Whiskers (dashed lines) extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the ends of the box. Outliers '+' are data with values beyond the ends of the whiskers. This plot shows also the test settings (four videos, seven packet loss settings) for the two labs (upper and lower subplot).



(b) CDF Plot. The user ratings represent a continous random variable which can be visualized by a cumulative distribution funtion (CDF). For each of the seven packet loss ratios, we consider the user ratings for the four videos used in the test and plot the empirical CDF as solid line. In addition, we fit the user ratings per packet loss ratio with a truncated normal distribution in $[0; 5]$ with the measured mean (MOS) and standard deviation (SOS). The marker '○' indicates the MOS value for that packet loss condition.

Figure 15: Video QoE. Results of the video QoE study [51]. A continous rating scales from 0 to 5, cf. Figure 14, was used in the experiments for subjects evaluating the quality of videos transmitted over a noisy channel [51]. The study was repeated in two different labs denoted as 'EPFL' and 'PoLiMi' in the result figures. The packet loss in the video transmission was varied in $p_L \in \{0; 0.1; 0.4; 1; 3; 5; 10\}$ (in %) for four different videos. In total, 40 subjects evaluated 28 test conditions.

Figure 15a shows a box plot of the results — an appropriate graphical illustration of the subjective ratings on a continous scale. For each test condition, the user ratings represent a continuous random variable. The box '□' quantifies the lower quartile and the upper quartile values. The median is depicted as a line '-' in the box. Whiskers (dashed lines) extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the ends of the box. Outliers '+' are data with values beyond the ends of the whiskers. This box plot also visualizes the test settings (four videos, seven packet loss settings) for the two labs (upper and lower subplot).

Figure 15b shows the cumulative distribution function (CDF) of the user ratings for the packet loss ratios tested in the video QoE study. The user ratings represent a continous random variable which can be visualized by a CDF. For each of the seven packet loss ratios, we consider the user ratings for the four videos used in the test and plot the empirical CDF as solid line. In addition, we fit the user ratings per packet loss ratio with a truncated normal distribution in $[0;5]$ with the measured mean μ (MOS) and standard deviation σ (SOS). Thus, the user ratings U follow the truncated normal distribution, i.e. $U \sim N(\mu;\sigma;0;5)$ with $U \in [0;5]$. The marker '○' indicates the MOS value for that packet loss condition. We observe a very good match between the empirical CDF and the truncated normal distribution. This is not obivous and no trivial result, althought the first two moments of both distributions are identical, the underlying distributions could be very different, as pointed out in Figure 1 in Section II.

Appendix B
Invariance of SOS Parameter for Linearly Transformed Ratings

In a subjective experiment, we observe the random variable U_c which represents the quality ratings of the subjects for a certain test condition c. In the experiment, a continuous rating scale is used with lower bound L_1 and higher bound H_1, i.e. $U_c \in [L_1;H_1]$. We observe the SOS parameter a.

Now, the user ratings are linearly transformed to another rating scale $[L_2;H_2]$ by the transformation function

$$\tau(u) = \frac{u - L_1}{H_1 - L_1}(H_2 - L_2) + L_2. \tag{13}$$

Then, the transformed user ratings $\tau(U_c)$ for any test condition c will lead to the same SOS parameter a.

We consider a certain test condition U_c. Then, the expected value is $E[U_c] = x$ and $Var[U_c] = V_1(x)$ according to the SOS hypothesis with

$$V_1(x) = a_1(-x^2 + (L_1 + H_1)x - L_1 \cdot H_1). \tag{14}$$

The variance of the transformed user ratings is

$$\begin{aligned} Var[\tau(U_c)] &= Var\left[\frac{U_c - L_1}{H_1 - L_1}(H_2 - L_2) + L_2\right] && (15)\\ &= Var\left[\frac{H_2 - L_2}{H_1 - L_1}U_c\right] && (16)\\ &= \left(\frac{H_2 - L_2}{H_1 - L_1}\right)^2 \cdot Var[U_c]. && (17) \end{aligned}$$

However, the latter term is equivalent to the variance according to the SOS hypothesis with SOS parameter a on the transformed rating scale, i.e.

$$V_2(\tau(x)) = a_2(-\tau(x)^2 + (L_2 + H_2)\tau(x) - L_2 \cdot H_2)\,. \tag{18}$$

For the user ratings transformed on the second rating scale, it holds

$$V_2(\tau(x)) = Var[\tau(U_c)] \tag{19}$$

which leads to

$$a_2 = a_1 = a\,. \tag{20}$$

As a result, the SOS hypothesis holds with the same SOS parameter a. The SOS parameter a is scale invariant when linearly transforming the user ratings in a mathematical way. However, it has to be clearly noted that subjective studies using different rating scales may lead to different SOS parameters. This has been observed e.g. for the results for speech QoE in Figure 12a.

As an implication, the numerical derivation (by solving the optimization problem) of the SOS parameter a for given MOS and SOS values can be done with linearly transformed user ratings, see Figure 16. Thus, the SOS parameter reflects the user rating diversity independent of the rating scale.

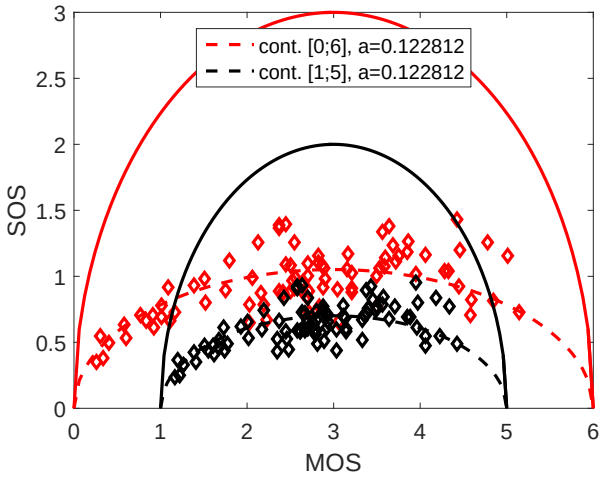


Figure 16: Transformation of user ratings from speech QoE results from rating scale $[0;6]$ to rating scale $[1;5]$ leads to the same SOS parameter a. However, the values of MOS and SOS (tuple depicted as diamond marker) as well as the maximum SOS for a given MOS (solid lines) are changing of course.