SOS: THE MOS IS NOT ENOUGH!

Tobias Hoßfeld

University of Würzburg
Institute of Computer Science
Würzburg, Germany
tobias.hossfeld@uni-wuerzburg.de

Raimund Schatz, Sebastian Egger

Telecommunications Research
Center Vienna - FTW
Vienna, Austria
{schatz,egger}@ftw.at

ABSTRACT

When it comes to analysis and interpretation of the results of subjective QoE studies, one often witnesses a lack of attention to the diversity in subjective user ratings. In extreme cases, solely Mean Opinion Scores (MOS) are reported, causing the loss of important information on the user rating diversity. In this paper, we emphasize the importance of considering the Standard deviation of Opinion Scores (SOS) and analyze important characteristics of this measure. As a result, we formulate the SOS hypothesis which postulates a square relationship between the MOS and the SOS. We demonstrate the validity and applicability of the SOS hypothesis for a wide range of studies. The main benefit of the SOS hypothesis is that it allows for a compact, yet still comprehensive statistical summary of subjective user tests. Furthermore, it supports checking the reliability of test result data sets as well as their comparability across different QoE studies.

Index Terms— MOS, diversity of user ratings, SOS hypothesis, QoE representation, subjective tests

1. INTRODUCTION

Subjective user studies are the basis for the quantification of user perceived quality of applications and services and therefore have become a fundamental cornerstone of Quality of Experience (QoE) research. Typically, subjective tests are carried out by a test panel of real users who are exposed to defined test conditions and asked to retrospectively rate the perceived quality of each test condition on a category scale. In the QoE domain, the most commonly used scale for quality ratings is a discrete 5-point scale designating the categories 'bad', 'poor', 'fair', 'good', and 'excellent'. The related test method is referred to as Absolute Category Rating (ACR) [1].

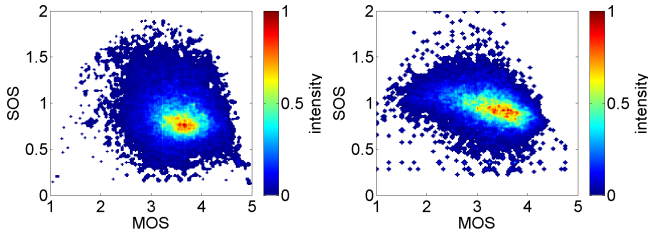
However, in order to obtain accurate results as well as a good understanding of the QoE and its sensitivity to certain parameters, the – sometimes strongly diverging – views of many test subjects on the quality experienced have to be taken into account. In previous research, Karapanos et al. characterized [2] and accounted for [3] the diversity in user ratings. This diversity is caused by several psychological influence factors such as individual expectations regarding quality levels, memory effects due to quality experienced in the past, type of user and sensitivity to impairments, uncertainty how to rate absolutely the quality for a certain test condition, etc. In contrast, the average of these user judgements is supposed to be directly related to non-psychological influence factors on the technical level such as network delivery bandwidth, application level like video codec settings, content level like type of video, or the context level, including location and quality rating scale used in the test.

As a consequence of this inherent diversity in user opinions (and thus rating scores), subjective tests can be both time-consuming and costly, because tests need to be conducted with a considerable number of users to average out this diversity and obtain reliable, statistically significant results. Still, even if large numbers of users are tested, the diversity of opinion remains. In this context, QoE researchers face the challenge that compact, but still comprehensive representations of the numerical results of subjective user studies is required that adequately communicates the underlying opinion diversity. Throughout the remainder of this paper, we refer to such numerical results representations as *QoE measure*.

A commonly used QoE measure is the *Mean Opinion Score (MOS)* [4], which can be determined from subjective ratings by averaging the ratings of subjects with same test conditions. However, the process of averaging scores removes this user diversity and gives only partly insights on the user perceptions [3]. We therefore emphasize the necessity for a proper QoE measure to specify the MOS (quantifying the average rating) as well as additional information that describes the diversity of user ratings.

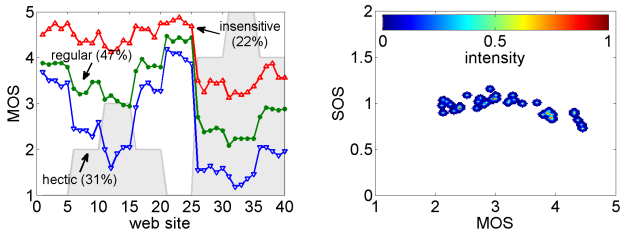
The main contribution of this paper is the so-called *SOS hypothesis* which postulates that the Standard deviation of Opinion Scores (SOS), a measure that reflects the level of rating diversity, can be described as a function of the MOS x. The basic idea behind the SOS hypothesis is that for properly conducted QoE tests, user ratings vary only to a certain extent, since all users experience the same test conditions. We will show that for this reason, user rating diversity within a subjective study can be accurately described by the so called *SOS parameter* a. A comprehensive QoE measure is therefore the MOS in conjunction with the SOS parameter a. We further show that different application categories (Web, Video, Image, etc.) map to different ranges of this SOS parameter a. Consequently, this parameter can be used as an indicator for reliability problems in a subjective test data set, for the type of application being tested and for the comparability of data sets originating from different QoE studies.

The remainder of this paper is structured as follows. Section 2 provides some of our findings regarding the nature of user ratings and the type of user in the context of SOS as a background and briefly reviews related work. Then, the minimum and maximum theoretical bounds of the SOS are derived yielding to the SOS hypothesis in Section 3. The validity of the SOS hypothesis is investigated in Section 4 by means of subjective tests for web traffic. Section 5 shows the application of the SOS hypothesis for a variety of applications and services, including are web surfing, file download, image quality, video streaming, VoIP, and gaming. In this context, we also show how the SOS parameter can be used for comparing and checking subjective tests. Section 6 concludes this paper and provides an outlook on future research.



(a) Internet Movie Database IMDb (b) MovieLens Data Set

Fig. 1: User rating histograms for movies from two different datasets



(a) MOS for user types (b) Nature of ratings

Fig. 2: User ratings as observed in a web traffic experiment

2. BACKGROUND AND RELATED WORK

Before reviewing related literature that address user diversity, we first investigate the nature of non-QoE users ratings as background and motivation for this work. To this end, we statistically analyze end user degree-of-liking ratings of movie files, i.e. quality ratings that are not a function of technical influence factors. As contrast, user QoE ratings generated in the context of a Web QoE test are shown in order to highlight the difference in distributions when technical impairments are present. The results clearly show that different types of users exist that lead to different score distributions. Thus, the MOS is not enough to describe subjective user test results!

2.1. Nature of Non-QoE User Ratings

For investigating the diversity of non-QoE user ratings, we consider now the data from two popular movie databases which are the Internet Movie Database (IMDb) and the MovieLens Data Sets. The MovieLens database contains about 10 million ratings for roughly 10,000 movies and is publicly available at [5]. For retrieving the ratings from IMDb, the website has been parsed and about 2 million user ratings from 40,000 movies have been downloaded. For each movie from both datasets, the average user rating (MOS) and the standard deviation of the ratings (SOS) are computed.

Figure 1 shows a bivariate histogram of the user ratings for the movies from both datasets. While the MOS is depicted on the x-axis, the SOS is given on the y-axis. The ratings are binned into a 200-by-200 grid of equally spaced containers. In order to compare both datasets, each of the resulting histograms is normalized by the bin with the most entries. This is referred to as intensity in Figure 1. Thereby, the intensity for each bin in the grid is color-coded according to the color-bar on the right side. It has to be noted that empty bins into which no (MOS,SOS) tuple falls are plotted in white.

It can be seen that in both cases a large area is covered with measurement points. Thus, a large variety of (MOS,SOS) tuples exist. This indicates that there is no clear relationship between MOS and SOS for non-QoE user ratings, i.e. when users rate their degree-of-liking on the general content level. The coefficients of correlation between MOS and SOS for the IMDb and MovieLens database are -0.18 and -0.29, respectively. Hence, there is no significant statistical dependency between MOS and SOS for these non-QoE user ratings, in contrast to setups in which the QoE is assessed.

2.2. QoE Ratings: Influence of User Type

In order to investigate the diversity of user judgements in the context of QoE tests, we take a closer look at the different types of users in terms of quality rating behavior, using a typical Web QoE study [6] as example. In this study, each test user encountered a series of web pages with defined sequences of page load times (PLT) and submitted a QoE rating after each pageview, respectively.

For identifying different types of users, we utilized different cluster analysis methods. In particular, we used the Expectation Maximization (EM) clustering algorithm as implemented by the machine learning software Weka [7]. As variables we used all QoE ratings a test user provided for the web pages throughout the test sequence. As a result of this cluster analysis, we identified three different user groups: (1) *Hectic users* (32 %) often change their user rating, even if the test conditions do not change, and tend to to be more critical. (2) *Regular users* constitute the largest user group (47 %) and show less variance in the user ratings and higher MOS compared to the hectic users. (3) *Insensitive users* (22 %) are least sensitive to the underlying network conditions and are more or less always satisfied with the actual service quality. This clustering is in line with the observations made during the test.

Figure 2(a) shows the MOS for the 40 web page sequence of the subjective test. However, the MOS for each web page is aggregated according to each user group. For the sake of readability, we depicted the PLT adjusted for the different web pages as light gray area plot in the background. Evidently, the curves for the different user groups significantly deviate from each other, as the average MOS over all web pages is 2.50, 3.25, and 4.13 for the hectic, regular, and insensitive users, respectively.

In contrast to the non-QoE movie ratings, the users in this test experienced technical impairments which also manifests in different user groups. Thus, the diversity of users is caused by different quality perception as seen in the rating behavior. This may be expressed adequately by the SOS across the different user groups and is shown as bivariate histogram in Figure 2(b). When comparing the different histograms, clearly visible differences between the distributions of movie and Web QoE ratings become evident. While the distributions for the movie ratings form elliptic or circular clusters with very little correlation between SOS and MOS, the web experiment generates a cluster that is rather banana-shaped. Indeed, we observe a coefficient of correlation of SOS and MOS about -0.63 for the web experiment. Thus, there seems to exist a significant dependency between SOS and MOS in such technical QoE study setting which will be expressed by the SOS hypothesis later.

2.3. Related Work

For comparing multimedia content or communication systems regarding their quality from a user perspective, subjective quality assessment methodologies have been developed [4, 8, 9]. The main objective of these methodologies is to obtain opinion ratings that are comparable across different studies. Typically, such methodologies recommend usage of the MOS, standard deviation and confidence intervals of the obtained scores for comparisons [4, 8]. Also statisti-

cal literature [10,11] does emphasize the importance of these figures for result comparison. However, many publications in the HCI and QoE domain do not really follow these recommendations [12].

In a comparison of different quality rating scales, [13] found evidence that the aforementioned statistical measures do not only depend on the underlying technical conditions of the system under test but are also affected by the rating scales used. In particular, the rating scale chosen directly influences the distribution of the SOS due to discretization. In addition, [3] analyzes how the practice of score averaging eliminates important information, arguing for result presentations that include information on the user opinion diversity.

From these observations, we conclude that the current status quo of subjective test results reporting tends to omit important features of the underlying data. Therefore, we analyze properties of different user studies' rating score distributions for a variety of applications and discuss characteristic relationships between them in this paper.

3. THEORETICAL VIEW ON THE SOS

Subjective tests are conducted by a finite test panel of N users which rate the quality on a K-point scale. In this section, we investigate the theoretical upper and lower bounds of the SOS. To this end, we derive in Subsection 3.1 the minimum SOS (which depends on N), and in Subsection 3.2 the maximum SOS (which depends on K). These theoretically derived bounds of the SOS guide us to the SOS hypothesis as formulated in Subsection 3.3.

3.1. Minimum SOS

In the following, we derive the minimum SOS_* for a given MOS_*. There are N test users in total which rate the MOS on a 5-point scale. The minimum SOS_* for a given $\mathrm{MOS}_* \in [i; i+1[$ is obtained when y users rate i on the quality scale, while the other $N-y$ users rate $i+1$.

$$\mathrm{MOS}_*(y) \in [i; i+1[\text{ with } i \in \{1, 2, \ldots, K\} \quad (1)$$

$$\mathrm{MOS}_*(y) = i + 1 - \frac{y}{N} \quad (2)$$

$$\mathrm{SOS}_*(y) = \frac{1}{N}\sqrt{(N-y)y} \quad (3)$$

For an infinite number of users we arrive at the following equations by defining the ratio κ of users rating i. Thus, we have a square relationship between the minimum variance of opinion scores SOS_*^2 and the corresponding MOS, i.e. $\mathrm{SOS}_*^2 \propto \mathrm{MOS}_*^2$.

$$\lim_{n\to\infty} \frac{y}{N} = \kappa \text{ with } 0 \le \kappa \le 1 \quad (4)$$

$$\mathrm{MOS}_* = i + 1 - \kappa \quad (5)$$

$$\mathrm{SOS}_* = \sqrt{\kappa - \kappa^2} \quad (6)$$

3.2. Maximum SOS

The maximum standard deviation SOS^* can be obtained when a ratio α of the test users rate 1 and $1-\alpha$ rate 5 on a 5-point rating scale. Then, the following equations hold.

$$\mathrm{MOS}^*(\alpha) = \alpha \cdot 1 + (1-\alpha)\cdot 5 = 5 - 4\alpha \quad (7)$$

$$\mathrm{SOS}^*(\alpha)^2 = \alpha \cdot 1^2 + (1-\alpha)\cdot 5^2 - (\mathrm{MOS}^*(\alpha))^2 = 16(\alpha - \alpha^2) \quad (8)$$

Thus, for a given MOS^* $x \in [1;5]$, the ratio α follows as $\alpha = \frac{5-x}{4}$. Then, we arrive at the following equation.

$$\mathrm{SOS}^*(\alpha) = \mathrm{SOS}^*(\tfrac{5-x}{4}) = \sqrt{-x^2 + 6x - 5} = \mathrm{SOS}^5(x) \quad (9)$$

If a different K-scale than the 5-point rating scale was used for rating, then the range of the scale goes from v_1 to v_K. Typically, we have a) $v_1 = 1$, $v_K = K$ or b) $v_1 = 0$, $v_K = K-1$. Again, a square interdependency between the maximum SOS^2 and the MOS x is observed.

$$\mathrm{SOS}^K(x)^2 = -x^2 + (v_1 + v_K)x - v_1 v_K \quad (10)$$

According to Eq.(10), the maximum coefficient of variation of the opinion scores can be computed in dependence of the MOS x by $\mathrm{COS}^K(x) = {}^{\mathrm{SOS}^K(x)}/_{x}$. Using a 5-point scale, the function COS^K has a maximum at $\left(\frac{5}{3}; \frac{\sqrt{20}}{5}\right)$. This means that the maximum coefficient of variation that can be observed is around 0.89; in that case, the MOS is around 1.33.

3.3. SOS Hypothesis: Square Relationship

We consider now the user diversity for subjective user ratings on a 5-point quality scale. The maximum user rating diversity is observed at MOS $x = 3$, which can be proven by computing the maximum of Eq.(9), $\frac{d}{dx}\mathrm{SOS}^5(x)^2 = -2x + 6 = 0$. The maximum diversity at a MOS of 3 has been also found in other subjective studies [13, 14]. At the extremes, SOS obviously has to be zero, since a MOS of 1 and 5 can only be reached if all users rate 1 and 5, respectively.

Grounded on above analysis of the theoretical bounds of the SOS, we assume a square relationship between $\mathrm{SOS}(x)^2$ and MOS x. Together with the observation of no diversity at the edges and maximal diversity at $x = 3$, we arrive at the following equation which is parametrized by the SOS parameter a. For the remainder of this paper, we will denote Eq.(11) as *SOS hypothesis*.

$$\mathrm{SOS}(x)^2 = -ax^2 + 6ax - 5a \quad \text{(SOS Hypothesis)}$$

$$= a(-x^2 + 6x - 5) \quad (11)$$

4. EXAMPLE: THE SOS HYPOTHESIS IN THE CONTEXT OF WEB QOE

The purpose of this section is to validate the SOS hypothesis based on subjective assessment studies for web traffic we conducted. The measurement setup has been already sketched in Section 2.2 and described detailed in [6]. Three Web QoE experiments have been conducted which are referred to as 'local test', 'online #1', and 'online #2'. The local and the online test differ by realizing impairments of web traffic in terms of delay in a local testbed and across the Internet.

The important aspect for our analysis of SOS is that within each of the three experiments, all participants experienced the same test conditions. This allows to investigate the diversity of user rating behavior. For each set of test conditions i we therefore obtain one particular MOS value MOS_i and one SOS value SOS_i. For the sake of clarity we first discuss the MOS ratings from one of the web traffic experiment, before we consider the SOS and validate the SOS hypothesis.

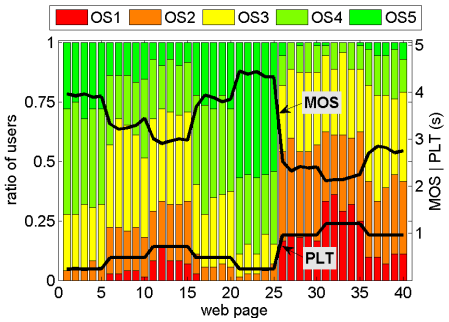


Fig. 3: CDF of user ratings for web traffic 'online #2'

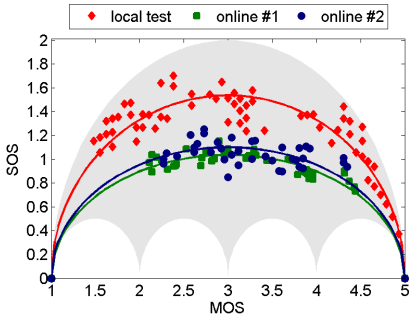


Fig. 4: SOS hypothesis for web traffic experiments

4.1. MOS for Web Traffic: Online Test 1

Figure 3 illustrates the influence of the network transmission on QoE showing the results of online test #1. The sequence of 40 web pages downloaded by the test user is plotted along the x-Axis. On the left y-axis, the share of users rating the QoE in each category OS1 (bad) to OS5 (excellent) of the 5-point ACR scale is illustrated as stacked bar plot. Thus, the bars represent the relative distribution of participants opinion scores for each page of the test sequence. In addition, Figure 3 displays the MOS per web page, which is the average over all user ratings for this web page, as well as the instrumented PLT in seconds (as scaled on the right y-axis). Overlaying both curves reveals the inverse relationship between the PLT and Web QoE: the higher the PLT set for a webpage, the lower the resulting MOS score.

4.2. Validation of the SOS Hypothesis

Next, we investigate the applicability of the SOS hypothesis in detail. Figure 4 shows the SOS in dependence of the MOS for the three web traffic experiments. Each measurement point depicts a certain test condition, which in our case is the download of a web page with a defined, preset page load time. The solid lines represent the fitted square functions f_a according to Eq.(11), whereby the SOS parameter a is obtained by minimizing the least squared errors between the measurement data and the fitting function. The gray area plot in the background illustrates the upper and lower bound of the SOS as derived in Section 3.

Our first observation is that all measurement points for the three studies $(\mathrm{MOS}_i, \mathrm{SOS}_i)$ are located close to the theoretical square SOS function f_a, in contrast to the movie user rating $(\mathrm{MOS}, \mathrm{SOS})$ tuples which are widely spread throughout the entire parameter space (see Section 2.2). Here, the range around f_a represents the mean squared error (MSE) between the measurements and the fitting. The MSE is influenced by the actual number N of test users. If N is too low, the test panel is not representative in that sense that not all users groups are taken into account properly. Thus, the diversity of users, i.e. the SOS, is not representative, too. Other influence factors of MSE are psychological factors like the uncertainty of users how to rate, e.g. due to missing training phases or unclear briefing. This explains the larger MSE for the local test which has a value of 0.09 in contrast to the online tests with 0.01 and 0.03, respectively. Nevertheless, the SOS hypothesis cannot be rejected for these experiments.

The second observation concerns the SOS parameter a. For both online tests, a is in the same order of magnitude with 0.269 and 0.303, which indicates strong similarity of both tests. The SOS parameter of the local test, however, is about 0.590. This large difference indicates that the local and the online tests must be different in some sense and must be compared with care. In fact, the tests are different in that we used a 3-point scale for the local test while we relied on a 5-point scale for the online tests. The results from the local test were then mapped to a 5-point scale. The rating scales used may influence test users and imply a difference in the used quality ranges [15] which is shown by the different instantiations of a.

5. APPLICATION OF THE SOS HYPOTHESIS

In this section, we apply the SOS hypothesis to a number of QoE studies covering a wide range of application classes. As application classes we consider web surfing, file download, video streaming, Voice-over-IP, viewing images, as well as cloud gaming. The latter class refers to a new kind of service, which combines the successful concepts of Cloud Computing and Online Gaming and provides the entire game experience to the users remotely from a data center [16]. Thus, the game is rendered within the cloud and the video is streamed to the end user device. The data sets for video streaming, VoIP, and image quality are publicly available and can be downloaded at the locations given in the bibliography. The web surfing studies in Section 4 and in this section were conducted at the University of Würzburg and at FTW, respectively, while the cloud gaming experiments have been conducted at the University of Würzburg.

We have applied the SOS hypothesis Eq.(11) to the data set from different user studies that represent aforementioned application classes. As result, the SOS parameters a and the mean squared errors (MSE) of the optimal fittings are shown for each study in Table 1. We have grouped the user ratings belonging similar test conditions in order to derive the MOS and SOS across the different users. However, we differentiate the user studies into several subcategories according to a specific criteria, e.g. the type of web site for web browsing, content class or the type of game for cloud gaming (see second left column).

MOS and SOS parameter as comprehensive QoE measure. First, we notice that for each individual user study the SOS parameters of the different categories lie very close together. For example, for [17], the influence of delay on the QoE for video streaming is investigated for different types of videos. In that case, the a is in the range $0.13 \leq a \leq 17$. The difference for web surfing [6] was already explained in the previous section because of differently applied quality scales in the local and online tests. Thus, the user ratings in a properly executed QoE test vary only to a certain extent, since all users experience the same test conditions. Hence, the SOS hypothesis cannot be rejected in the cases analyzed. Furthermore, in these cases the user rating diversity of the subjective studies can be described using a without critical information loss. A comprehensive QoE measure can therefore be composed from the MOS

depending on the variety of test conditions in conjunction with the SOS parameter a.

Comparison between user studies. Secondly, we see that even for the different subjective user studies, the SOS parameter a lies in the same range. For example, the SOS parameter of the web surfing studies is about $0.23 < a < 0.30$. It has to be noted that the MSE for the web surfing study [18] is large in comparison to the other cases, because we split the user tests according to the type of website, leading to a small number of users who experienced the same test conditions. Nevertheless, the SOS parameter a lies in the range observed for the web surfing study [6] (except for the ORF website). This finding suggests that the SOS parameter could be used as criterion for deciding whether results data from similar QoE experiments can be pooled – if the SOS parameters are sufficiently close to each other – in order to increase the statistical reliability of results without having to perform costly large-scale user studies. In the case of the the web surfing study [18], we obtain $a = 0.2751$ with an MSE of 0.0219 when pooling the results from all websites.

This observation that a has the same order of magnitude for the same application can be explained by the psychological influence factors causing user opinion diversity, such as the uncertainty how to rate absolutely the quality for a certain test condition. Large values of a indicate that it is more difficult to judge the user experienced quality. This is typically easier for video experiments (with artifacts) than for web traffic or cloud gaming where the QoE impact of technical impairments is not always salient. Furthermore, it has to be noted that for the two image quality studies we analyzed the SOS parameters differ significantly, because different rating scales were used: while [19] used a continuous quality scale, [20] relied on a 5-point scale. Thus, a can be utilized i) to check whether different subjective studies can be directly compared (or even pooled), ii) to support checking for reliability problems of subjective test and its test environment, as well as iii) to help determining the underlying type of application.

QoE similarity wrt. user diversity. Figure 5 shows SOS^2 depending on the MOS for the various applications based on the subjective tests, see Table 1. The color represents the value of the SOS parameter a which may range from 0.1 as observed for video streaming to 0.35 as observed for cloud gaming. We used the average SOS parameter a_j over the different categories and user studies for the same application j. This figure depicts that quality ratings for gaming results into the largest user diversity, followed by web traffic, VoIP, video streaming, image quality. It has to be noted that we do not show clear thresholds for the SOS parameter for the various applications on purpose, since the ranges may be overlapping and several applications may fall into the same QoE application class with respect to user diversity, like video streaming and image quality. Nevertheless, the intention of this plot is to support the comparison and checking the reliability of different user studies. For unknown applications, the SOS parameter may imply the underlying type of application reflecting user uncertainty regarding how to rate quality for this kind of application.

6. CONCLUSIONS AND OUTLOOK

In this paper, we showed that the diversity of user ratings in subjective studies is an important issue in the context of analyzing and interpreting QoE results. While Mean Opinion Scores (MOS) are widely used for reporting measurement studies, the process of averaging subjective ratings eliminates the diversity of user assessments. We therefore emphasized the necessity to report also information on user opinion diversity using the different user types found in a typical Web QoE study.

Table 1: SOS parameter a for various applications derived from own and existing subjective measurement studies

appl.	subjective user study	SOS a	MSE
web surfing	type of website [18]: Spiegel, Wiki, ORF, Expedia, Chefkoch, Google, Photo	0.2342, 0.2774, 0.3497, 0.2428, 0.2806, 0.2617, 0.2816	0.0498, 0.1553, 0.2204, 0.0562, 0.0957, 0.0379, 0.0456
	type of delay pattern [6]: local, online #1, online #2	0.5902, 0.2685, 0.3027	0.0909, 0.0094, 0.0315
file download	type of file [18]: mp3, zip	0.2777, 0.2417	0.0359, 0.1034
video streaming	effect of delay for different videos [17]: football, foreman, news	0.1593, 0.1668, 0.1334	0.0074, 0.0047, 0.0077
	type of resolution for IRCCyN/IVC: SD RoI [21], 1080i [22]	0.2116, 0.2061	0.0405, 0.0312
	type of codec for IT-IST Lisbon [23]: H.264, MPEG-2	0.1078, 0.1137	0.0122, 0.0212
VoIP	host laboratory for ITU-T P.Sup23 [24]: CNET, CSELT, NTT, BNR	0.2487, 0.1808, 0.1702, 0.1991	0.0105, 0.0135 0.0257, 0.0217
images	LIVE image quality assessment database [19]: JPEG, JPEG2000	0.0400, 0.0377	0.0172, 0.0289
	IRCCyN/IVC Scores on Toyama [20]	0.1715	0.0262
cloud gaming	fast-, medium-, slow-paced gameplay [16]	0.2718, 0.3287, 0.3466	0.0358, 0.0863, 0.2140

As a result, we formulated the SOS hypothesis which postulates a square relationship between the MOS and the SOS^2 which depends only on a single parameter, the SOS parameter a. We demonstrated the validity and applicability of the SOS hypothesis for a number of existing QoE studies on different application categories. Thus, the SOS hypothesis enables the development of a compact, but comprehensive representation of the numerical results of subjective user studies. In particular, a proper QoE measure is therefore the MOS (depending on the variety of test conditions) in conjunction with this SOS parameter.

We have also shown that the SOS parameter a can be used i) to check whether different subjective studies can be directly compared (or even pooled), ii) to test the reliability of the subjective tests conducted, as well as iii) to help determining the underlying type of

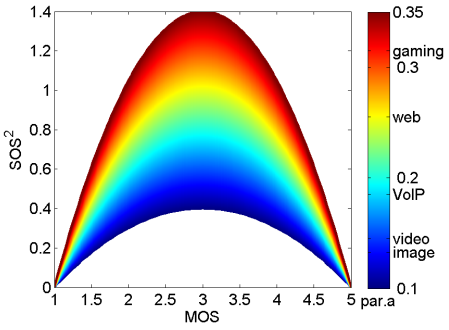


Fig. 5: SOS hypothesis parameter a for various applications

application.

As concerns future work, we plan to analyze datasets from additional subjective QoE assessment studies in order to test the generalizability of the SOS hypothesis across different application categories and experimental setups. This way, also the mapping between application types and SOS parameter ranges as well as the applicability of the SOS hypothesis to study results pooling can be investigated more throughly. Furthermore, we will analyze in depth the mathematical properties of the SOS parameter a in to which extent it can be used for reconstructing the score distribution for a given experiment.

We believe this work is an important step towards better understanding and describing fundamental mathematical properties of QoE user ratings. To this end, we cordially invite the QoE research community to contribute test data or SOS statistics from studies conducted in order to jointly advance our knowledge on this subject.

7. REFERENCES

[1] Sebastian Moeller, *Assessment and Prediction of Speech Quality in Telecommunications*, Springer, 1st edition, August 2000.

[2] Evangelos Karapanos and Jean-Bernard Martens, "Characterizing the diversity in users' perceptions," in *Proc. of 11th IFIP TC 13 international conference on Human-computer interaction*, Berlin, Heidelberg, 2007, INTERACT'07, pp. 515–518, Springer-Verlag.

[3] Evangelos Karapanos, Jean-Bernard Martens, and Marc Hassenzahl, "Accounting for diversity in subjective judgments," in *Proc of 27th International Conference on Human factors in Computing Systems*, New York, NY, USA, 2009, CHI '09, pp. 639–648, ACM.

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

[5] GroupLens Research, "Movielens data sets," `www.grouplens.org`, 2011.

[6] Tobias Hoßfeld, Raimund Schatz, Alexander Platzer, Sebastian Egger, Sebastian Biedermann, and Markus Fiedler, "The Memory Effect and Its Implications on Web QoE Modeling," in *currently under review*, 2011.

[7] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[8] International Telecommunication Union, "Subjective video quality assessment methods for multimedia applications," *ITU-T Recommendation P.910*, April 2008.

[9] International Telecommunication Union, "Methodology for the subjective assessment of the quality of television pictures," *ITU-R Recommendation BT.500-12*, Mar. 2009.

[10] Jürgen Bortz and Nicola Döring, *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*, Springer-Lehrbuch. Springer-Verlag Berlin and Heidelberg GmbH & Co. KG, Heidelberg, 4., überarb. Aufl. edition, 2006.

[11] Warren S. Torgerson, *Theory and Methods of Scaling*, John Wiley & Sons,Inc., 1963.

[12] Paul Cairns, "HCI... not as it should be: inferential statistics in HCI research," in *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*, 2007, p. 195201.

[13] Stefan Winkler, "On the properties of subjective ratings in video quality experiments," in *Proc. of International Workshop on Quality of Multimedia Experience (QoMEX*, 2009.

[14] Marcus Barkowsky, Björn Eskofier, Jens Bialkowski, and Andre Kaup, "Influence of the presentation time on subjective votings of coded still images," in *Proc. of IEEE International Conference on Image Processing*, 2006.

[15] Philip Corriveau, Christina Gojmerac, Bronwen Hughes, and Lew Stelmach, "All subjective scales are not created equal: the effects of context on different scales," *Signal Process.*, vol. 77, pp. 1–9, August 1999.

[16] Michael Jarschel, Daniel Schlosser, Sven Scheuring, and Tobias Hoßfeld, "An Evaluation of QoE in Cloud Gaming Based on Subjective Tests," in *Proc. of 1st International Workshop on Future Internet and Next Generation Networks (FINGNet-2011) in conjunction with IMIS 2011*, June 2011.

[17] Vasanthi Dwaraka Bhamidipati and Swetha Kilari, "Effect of Delay/ Delay Variable on QoE in Video Streaming," M.S. thesis, School of Computing at Blekinge Institute of Technology, Karlskrona, Sweden, May 2010.

[18] Peter Reichl, Bruno Tuffin, and Raimund Schatz, "Logarithmic laws in service quality perception: Where microeconomics meets psychophysics and quality of experience.," *Accepted for: Telecommunication Systems Journal, Springer*, 2011, in print.

[19] Hamid Rahim Sheikh, Muhammad Farooq Sabir, and Alan C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, Nov. 2006, Data available at `http://live.ece.utexas.edu/research/quality/subjective.htm`.

[20] Sylvain Tourancheau, Florent Autrusseau, Parvez Sazzad, and Yuukou Horita, "Impact of the subjective dataset on the performance of image quality metrics," in *Proc. of IEEE International Conference on Image Processing (ICIP 2008)*, San Diego, US, 10 2008, Data available at `ftp://ftp.ivc.polytech.univ-nantes.fr/IRCCyN_IVC_Scores_On_Toyama_Images/`.

[21] Fadi Boulos, Wei Chen, Benoit Parrein, and Patrick Le Callet, "Region-of-Interest Intra Prediction for H.264/AVC Error Resilience," in *Proc. of IEEE International Conference on Image Processing*, Cairo, Egypt, 2009, Data available at `ftp://ftp.ivc.polytech.univ-nantes.fr/IRCCyN_IVC_SD_RoI_Database/`.

[22] Stephane Pechard, Romuald Pepion, and Patrick Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in *Proc. of International Workshop on Image Media Quality and its Applications (IMQA2008)*, Kyoto, Japan, 2008, Data available at `ftp://ftp.ivc.polytech.univ-nantes.fr/IRCCyN_IVC_1080i_Database/`.

[23] Tomas Brandao, Luis Roque, and Maria Paula Queluz, "Quality assessment of H.264/AVC encoded video," in *Proc of Confernce on Telecommunications - ConfTele*, Sta. Maria da Feira, Portugal, 2009, Data available at `http://amalia.img.lx.it.pt/~tgsb/H264_test/`.

[24] ITU, *ITU-T coded-speech database (ITU-T Recommendation P.Sup23)*, International Telecommunications Union, Feb. 1998.