Experiment Precision Measures and Methods for Experiment Comparisons

Lucjan Janowski*, Jakub Nawała†, Tobias Hoßfeld‡, Michael Seufert‡
*AGH University of Science and Technology, Institute of Telecommunications, Kraków, Poland
lucjan.janowski@agh.edu.pl
†University of Bristol, Department of Electrical & Electronic Engineering, Bristol, UK
jakub.nawala@bristol.ac.uk
‡University of Würzburg, Chair of Communication Networks, Würzburg, Germany
{tobias.hossfeld, michael.seufert}@uni-wuerzburg.de

***Abstract*—The notion of experiment precision quantifies the variance of user ratings in a subjective experiment. Although there exist measures that assess subjective experiment precision, to the best of our knowledge, there is no systematic framework in the Multimedia Quality Assessment (MQA) field for comparing subjective experiments in terms of their precision. Therefore, the main idea of this paper is to propose a framework for comparing subjective experiments in the field of MQA based on appropriate experiment precision measures. We present three experiment precision measures and three related experiment precision comparison methods. We analyze the performance of the measures by using data from real-world Quality of Experience (QoE) subjective experiments. We believe our experiment precision assessment framework will help compare different subjective experiment methodologies. For example, it may help decide which methodology results in more precise user ratings. This may potentially inform future standardization activities.**

I. Introduction

In statistics, the term *precision* quantifies how close measurements are to each other. This is often expressed as the reciprocal of the measurement variance. In this paper, we consider subjective experiments in the field of Multimedia Quality Assessment (MQA). Therefore, the measurements are subjective user ratings. These are simply opinions (expressed using a dedicated scale) of subjective experiment participants about stimuli presented to them (e.g., videos or images). In particular, the quantification of the user-perceived Quality of Experience (QoE) is based on such subjective experiments.

In this context, the term *experiment precision* provides a measure that quantifies the dispersion of the user ratings across different stimuli in a subjective experiment. Typically, experiment precision measures are normalized in the range between 0 (highest possible precision, i.e., minimum variance) and 1 (minimum possible precision, i.e., maximum possible variance). We put forward three novel *experiment precision measures*: 1) g — a Generalized Score Distribution (GSD) [1] based measure. 2) ℓ — a measure based on the subject inconsistency parameter υ of the model presented in [2]. (We later refer to this model as the Li2020 model.) 3) a — based on the so-called "SOS hypothesis" [3], where SOS stands for Standard Deviation of Opinion Scores.

The main idea of this paper is to propose a framework for comparing subjective experiments in the field of MQA based on appropriate experiment precision measures. To this end, we present three *experiment precision comparison methods*. The methods do not depend on the common factor in the compared experiments. The only requirement is to use the same subjective scale. So, we can compare image quality with video quality experiments. The goal of each method is to answer the following question: "Is there a statistically significant difference in experiment precision between a pair of experiments?" (Please note that each experiment precision measure can be treated as a point estimator of experiment precision for a *single* experiment. Each experiment precision method checks for significant differences between a *pair* of point estimators.) Although there are works proposing precision-related measures (e.g., [3], [4]), they do not address the issue of formally comparing a pair of experiments.

To the best of our knowledge, we are the first in the MQA community to propose a systematic framework for comparing subjective experiments in terms of their precision. An important application of such a framework is differentiating between various experiment methodologies. Our framework may help decide which methodology results in higher experiment precision. Such information may guide subjective methodology standardization and help practitioners choose a methodology if high subjective responses' precision is their top priority. In this paper, we make the following claims:

(i) Our experiment precision measures allow one to position a subjective experiment in relation to other experiment types (e.g., speech or video QoE experiments). They also allow one to compare one experiment run with other runs having a similar or modified setup. All of this was done to compare experiment runs in terms of experiment precision.

(ii) The experiment comparison methods we introduce provide a statistical test indicating whether experiment precision measures for a pair of experiments are statistically significantly different.

The following are our contributions that substantiate the claims we make.

C1 We introduce three experiment precision measures (g, a, and ℓ) allowing to assess experiment precision of a single subjective experiment.

C2 We suggest three novel methods to compare experiment precision between a pair of subjective experiments.

C3 We test the three experiment precision comparison methods on real-world subjective data from experiments on VR, speech, image, and video QoE. We also analyzed similar experiments (video quality) to see how the measures behave in such case.

C4 We give guidelines regarding reporting experiment precision and argue why reporting experiment precision is beneficial for the research community.

II. Notion of Experiment Precision and Measures of Its Assessment

We first describe the notion of experiment precision. Then, we present three experiment precision measures. Each of them stems from a model already described in the literature. However, the novelty lies in the fact that we use these methods and their combinations for the first time in the context of experiment precision and validate their usage.

Definition: *The term experiment precision provides a measure that quantifies the dispersion of the user ratings across different stimuli in a subjective experiment.*

A. Notion of Experiment Precision

For real-life subjective responses, the precision of the experiment cannot be directly measured. Instead, a theorized response generating model has to be fitted to the observed responses. Using the parameters of the fitted model, one can then infer the experiment precision. Importantly, the models that are relevant in this context are models that separate bias (i.e., a constant shift in responses, see [5] and [6] for more details) from the variance. Although there are models that partition the variance into per subject, per stimulus, or per distortion condition components [7], [2], it is the total variance that is of our interest. This total variance corresponds to experiment precision. The previous statement also means that we do not treat changes in subject bias as changes in experiment precision. Differently put, having two experiments with the same total variance, but different biases, we treat them as having the same experiment precision. We point out, however, that this is a theoretical assumption, which does not hold in certain corner cases. For one thing, if subjective responses are provided on a discrete scale (which is often the case), then the change in mean response changes the variance. This is the case since all discrete domain probability distributions (and subjective responses can be treated as such [8], [9]) have their mean and variance mutually dependent.

Experiment precision can be used for various reasons. The most obvious one is to report a new database of subjective responses. With experiment precision provided alongside the raw data, a prospective user of the database can quickly learn in what relation to other subjective experiment types this experiment is. For example, one can easily answer the following question: Are these data more or less precise than data coming from a typical video QoE experiment? Furthermore, if an experiment is run in multiple sessions or locations, the notion of experiment precision can help make sure that all experiment runs are similar. Along with other indications, the notion of experiment precision could be used to decide whether the responses gathered in two experiment runs can be merged. Thanks to the information provided by the notion of experiment precision, it could help decide which experiment setup results in more precise measurements as well.

B. GSD Based Measure g

The GSD model and its application to subjective MQA data are explored in [10] and [11]. For the concise description of the model, we refer the reader to [11].

The GSD represents per stimulus response distribution. The distribution is parameterized with two parameters: ψ and ρ. The first one (ψ) defines the central tendency of the data and can be intuitively understood as a drop-in replacement of the MOS measure.[1] The second one (ρ) defines the spread of responses. It acts as a confidence parameter. Thus, the higher the ρ, the higher the confidence of people's opinions and, therefore, the lower opinions' variability.

Since ρ expresses opinions' confidence, it is natural to associate it with experiment precision. As the GSD is fitted per stimulus, there are as many estimated values of ρ, as there are stimuli in a subjective experiment. In other words, if there are K stimuli tested during a subjective experiment, the GSD model is fitted K times, and we end up with K estimates of ρ (each denoted as $\hat{\rho}_j$). Now, to compute the g measure, we simply find the mean of the estimated ρs, that is,

$$g = \frac{1}{K} \sum_{j=1}^{K} \hat{\rho}_j \ , \qquad (1)$$

where K is the number of stimuli tested during the subjective experiment analyzed and $\hat{\rho}_j$ is the estimated value of GSD's ρ for the j-th stimulus.

C. Li2020 Based Measure ℓ

In [2] Li *et al.* introduce another model that represents the generation process of subjective responses. To make the discussion more comprehensible, let us refer to the model introduced in [2] as the Li2020 model. The Li2020 model has three parameters. The three parameters correspond to: (i) true quality ψ (conceptually similar to GSD's ψ), (ii) subject bias Δ (representing a systematic shift in the responses of a single subject, relative to the opinion of all the other subjects) and (iii) subject inconsistency υ (representing a random error). Although on the surface, the GSD and Li2020 models look

[1] Another intuitive description of the ψ parameter is that it represents the mean opinion of the complete population of observers. In other words, it represents the MOS, as would be observed, if we asked about the opinion, all people, whose opinion we are interested in.

similar, their internal structures differ significantly. For one thing, the Li2020 model uses an underlying continuous normal distribution (that is mapped to a discrete domain to reflect actually observed responses), whereas the GSD does not.[2]

As mentioned, one of the parameters in the Li2020 model relates to subject inconsistency (the υ parameter). Intuitively, subject inconsistency must be related to experiment precision. Thus, we use this parameter to assess experiment precision. Since subject inconsistency is estimated on the per subject basis, we get as many estimated subject inconsistencies $\hat{\upsilon}$, as there are subjects taking part in a subjective experiment. Now, to arrive at the measure ℓ, we compute the average estimated υ. Assuming that N subjects take part in the experiment, we can find the value of the measure ℓ as follows. With $\hat{\upsilon}_i$ being the estimated subject inconsistency υ for the i-th subject,

$$\ell = \frac{1}{N} \sum_{i=1}^{N} \hat{\upsilon}_i \ . \tag{2}$$

D. SOS Hypothesis Based Measure a

The SOS hypothesis based experiment precision measure a uses the SOS parameter a of the experiment of interest. In turn, the SOS parameter a is based on the so-called "SOS hypothesis" [3]. The SOS hypothesis states that in a typical QoE experiment, there is a simple quadratic relationship between mean opinion scores (MOS) m_j and the variance of opinion scores (SOS) v_j. The function has the following form for a 5-point rating scale:

$$v = f_a(m) = a \cdot (5 - m) \cdot (m - 1) \ , \quad 1 \leq m \leq 5 \ . \tag{3}$$

For calculating the SOS parameter a of an experiment, we take the MOS values m_j and rating variances v_j of all stimuli of that experiment and fit the corresponding MOS–SOS curve to obtain the parameter a. The SOS parameter can be directly computed, see [12], via ordinary least-squares (OLS) regression:

$$a = \frac{\sum_{j=1}^{K} (5 - m_j) \cdot (m_j - 1) \cdot v_j}{\sum_{j=1}^{K} (5 - m_j)^2 \cdot (m_j - 1)^2} \ . \tag{4}$$

We directly use the a parameter computed this way as an indication of experiment precision (referring to the approach as the experiment precision measure a).

III. Experiment Precision Comparison Methods

In this section, we describe three experiment precision comparison methods. Importantly, the methods are based on the three experiment precision measures.

Definition: *An experiment precision comparison method allows assessing whether there is a statistically significant difference in terms of experiment precision between a pair of subjective experiments.*

A. Comparison Method Based on the g Measure

As we already mentioned in Sec. II-B, the GSD model is estimated on the per stimulus basis. Thus, there are as many estimated values of ρ, as there are stimuli in a subjective experiment. Let us denote such a vector of ρ estimates as $\hat{\boldsymbol{\rho}}$. Having two such vectors from a pair of experiments we wish to compare ($\hat{\boldsymbol{\rho}}_1$ and $\hat{\boldsymbol{\rho}}_2$), we apply a two-sample independent t-test on them, assuming unequal variances in the two samples. The null hypothesis is that the two vectors have the same average value. In other words, the null hypothesis states that the two experiments have the same value of the g measure. Finally, we use t-test's p-value as an indication of whether the two experiments differ significantly in terms of precision.

B. Comparison Method Based on the ℓ Measure

As Li2020's subject inconsistency υ is estimated on the per subject basis, we get as many estimated subject inconsistencies $\hat{\upsilon}$, as there are subjects taking part in a subjective experiment. Let us denote this vector of estimated subject inconsistencies as $\hat{\boldsymbol{\upsilon}}$. Having two such vectors from a pair of experiments we wish to compare ($\hat{\boldsymbol{\upsilon}}_1$ and $\hat{\boldsymbol{\upsilon}}_2$), we apply a two-sample independent t-test on them, assuming unequal variances in the two samples. The null hypothesis is that the two vectors have the same average value. Differently put, the null hypothesis is that the two experiments have the same value of the ℓ measure. We use t-test's p-value to assess whether the two experiments differ significantly in terms of experiment precision.

C. Comparison Method Based on the a Measure

We estimate the parameters a_1 and a_2 from the MOS-SOS tuples of both subjective experiments using Eq. 4 and obtain the variances of the parameter estimates from the quadratic OLS regression. We can now apply the independent two-sample t-test, assuming unequal sample variances. The null hypothesis is that the two a parameters are the same, and we reject it based on the p-value. Please note that since we assume that the estimated SOS parameter a is equivalent in value to the experiment precision measure a, this procedure effectively compares a measures between a pair of experiments.

IV. Practical Use Case

Here, we present how the experiment precision measures perform in practice. This practical example is an indication of what can be achieved using the experiment precision measures. We show that our experiment precision measures can show differences between different types of subjective experiments. Importantly, measures' indications are in line with our expectations regarding experiment type's precision. We hope that this section will convince the reader of the practical value our notion of experiment precision brings.

Our analysis is based on subjective experiments of three types: i) virtual reality (VR) QoE, ii) speech QoE, and iii) video QoE. Specifically, we use one VR QoE experiment, one speech QoE experiment, and six video QoE experiments. The VR experiment comes from an international multilaboratory QoE study [13]. The one experiment we used from this

[2] Readers interested in learning more about the differences between the GSD and Li2020 models are encouraged to take a look at [11].

TABLE I
Experiment Precision Measures For QoE Subjective Experiments of 3 Types: VR(VR), Speech (S), and Video (V-n).

Exp.	$\ell\downarrow$	SE(ℓ)	$g\uparrow$	SE(g)	$a\downarrow$	SE(a)
V-6	0.574	0.014	0.908	0.0050	0.137	0.0020
V-1	0.583	0.011	0.891	0.0068	0.149	0.0022
V-4	0.610	0.020	0.826	0.0056	0.224	0.0021
V-3	0.613	0.016	0.863	0.0066	0.188	0.0021
V-5	0.627	0.019	0.871	0.0059	0.190	0.0021
V-2	0.627	0.022	0.867	0.0070	0.191	0.0021
S	0.953	0.028	0.744	0.0083	0.281	0.0015
VR	1.059	0.037	0.692	0.0093	0.335	0.0040

Arrows point in the direction of high precision. SE($\cdot$) stands for standard error of a particular precision measure.

study occurred in Wuhan, used the ACR methodology, and made use of video sequences being 10 s long. (From now on, we refer to this experiment as VR.) The speech QoE experiment comes from the "ITU-T Coded-Speech Database" (constituting Supplement 23 to the P series of ITU-T Recommendations) [14]. We use only the responses collected by Nortel during the first of the three experiments reported in this database. (We later refer to this experiment as S.) Finally, the six video QoE experiments we use come from the international multilaboratory VQEG HDTV Phase I study [15]. (We refer to the experiments in this study as V-n, with $n \in \{1, 2, \ldots, 6\}$.)

Following our intuition and SOS–MOS curves for various types of subjective experiments (cf. Fig. 5 in [3]), we expect video and speech QoE experiments to be *more* precise than VR QoE experiments. If we were to take the SOS–MOS curves presented in [3] for granted, we should also expect video QoE experiments to be *more* precise than speech QoE experiments. (Speech QoE experiments correspond to VoIP experiments in Fig. 5 from [3].) All in all, our expected ordering of experiment types in terms of precision (starting from the least precise experiment type) is as follows: VR, speech, and video.

Table I presents experiment precision results for eight subjective experiments of interest. The table is sorted in ascending order, according to the measure ℓ. The arrows next to the headings identifying three experiment precision measures indicate the direction of higher precision. For example, the higher the g, the more precise the experiment is. Our experiment precision measures order three experiment types in line with the prior expectations. That is, the one VR experiment is assessed to be the least precise, with the speech experiment following, and the six video experiments assessed to be the most precise. Importantly, this ordering is reflected in the readings from all three experiment precision measures. This suggests that all measures are capable of estimating subjective experiment precision in line with the intuition of experts.

In practice, we may need to compare a pair of subjective experiments in terms of their precision. Thus, it is interesting to check whether our experiment precision measures indicate statistically significant differences between each pair of experiments from our pool of eight experiments. We use the 5% significance level, which means that the null hypothesis (i.e., experiment precision is the same) is rejected only if the p-value is less than or equal to 0.05. We generally expect to see statistically significant differences between experiments of different types (e.g., VR QoE vs speech QoE). We do not expect statistically significant differences between experiments of the same type (e.g., V-1 experiment vs V-2 experiment).

TABLE II
p-Values Resulting From Comparing Experiment Precision Measures Between All Video QoE Experiments. We Expect No Statistically Significant Differences.

Exp. 1	Exp. 2	ℓ p-Value	g p-Value	a p-Value
V-6	V-1	6.07E-01	4.76E-02	4.44E-05
V-6	V-4	1.45E-01	5.76E-24	2.60E-97
V-6	V-3	7.70E-02	1.21E-07	2.11E-50
V-6	V-5	2.97E-02	3.10E-06	9.07E-53
V-6	V-2	5.10E-02	3.72E-06	5.10E-53
V-1	V-4	2.36E-01	1.23E-12	1.21E-78
V-1	V-3	1.30E-01	3.12E-03	4.60E-32
V-1	V-5	4.89E-02	2.70E-02	2.30E-34
V-1	V-2	8.29E-02	1.57E-02	1.35E-34
V-4	V-3	9.12E-01	3.21E-05	2.06E-28
V-4	V-5	5.42E-01	7.03E-08	1.03E-25
V-4	V-2	5.73E-01	5.33E-06	1.71E-25
V-3	V-5	5.77E-01	3.54E-01	4.97E-01
V-3	V-2	6.11E-01	6.27E-01	4.59E-01
V-5	V-2	9.96E-01	6.95E-01	9.51E-01

p-Values smaller than 0.05 are marked with a purple background.

TABLE III
p-Values Resulting From Cross-type Comparisons of Experiment Precision Measures. We Expect Statistically Significant Differences.

Exp. 1	Exp. 2	ℓ p-Value	g p-Value	a p-Value
V-6	S	6.02E-14	3.29E-45	1.99E-167
V-1	S	3.09E-13	1.84E-34	9.65E-148
V-4	S	1.21E-12	3.82E-15	7.02E-66
V-3	S	1.07E-12	5.82E-25	1.34E-112
V-5	S	4.28E-12	2.06E-29	3.14E-109
V-2	S	1.06E-11	7.78E-26	5.03E-109
V-6	VR	1.10E-14	1.49E-37	3.78E-66
V-1	VR	3.95E-14	9.45E-36	3.49E-65
V-4	VR	8.09E-14	1.46E-22	3.33E-44
V-3	VR	1.09E-13	5.90E-30	8.57E-55
V-5	VR	2.62E-13	1.20E-31	2.03E-54
V-2	VR	3.41E-13	7.54E-31	2.28E-54
S	VR	2.54E-02	5.42E-05	2.52E-21

Tables II and III show p-values resulting from comparing experiment precision measures between all pairs of the eight experiments of interest. The headings identify which column corresponds to which experiment precision measure. The first two columns indicate which two experiments are compared. We highlight in purple comparisons that resulted in significant differences (assuming a 5% significance level). Note that Tab. II presents same-type pairs (i.e., the two experiments in the pair are of the same type), whereas Tab. III presents cross-type pairs (i.e., the two experiments in the pair are of two different types). All precision measures flag cross-type pairs as corresponding to significant differences in experiment precision. This is desirable. However, out of 15 same-type pairs (cf. Tab. II), the measures g and a, mark 12 as indicating significant differences in terms of precision. This is counter-intuitive and suggests that there may be a problem with these two measures. The behavior of the ℓ measure is generally

TABLE IV
Raw Experiment Precision Measures Results For Two Image QoE Experiments—VIME1 (I-V) and CCRIQ2 (I-C).

Exp.	$\ell\downarrow$	SE(ℓ)	$g\uparrow$	SE(g)	$a\downarrow$	SE(a)
I-V	1.053	0.0330	0.717	0.0085	0.314	0.0025
I-C	1.100	0.0316	0.683	0.0103	0.347	0.0030

in line with our prior expectations. It flags as significantly different only two out of 15 same-type pairs. It is worth keeping in mind that since we assume the 5% significance level, flagging roughly one comparison as significant may happen purely due to randomness. Thus, the measure flagging as significant only two out of 15 same-type pairs is very close to our prior expectation of no true differences.

A. *Detecting Imprecise Experiments*

It is interesting to check whether the precision measures would be able to detect problems with experiments that are known to be flawed. One such example of experiments with insufficient precision are experiments VIME1 and CCRIQ2 described in [16]. (We later refer to these experiments as V&C.) The two experiments investigated the quality of a set of images taken with consumer capture devices (e.g., smartphones or tablets). Due to a few unusual experiment design choices, the subjective responses gathered during the two experiments were identified in [16] as less precise than would be typical for a standard image QoE subjective experiment.

To check whether the experiment precision measures can detect the low precision of V&C, we apply the measures to raw subjective responses. In other words, we do *not* apply any data cleansing procedure before running the measures. The only preprocessing step that we take is to remove the responses of two subjects—one with ID 259 from the VIME1 experiment and one with ID 270 from the CCRIQ2 experiment. We do so since the two subjects did not assess the quality of all stimuli.

Table IV presents the results of applying our measures to the data originating from V&C experiments. The first thing to notice is how the results compare to the results presented in Tab. I. According to the precision measures, experiments V&C have the precision similar to the VR QoE experiment. Both VIME1 and CCRIQ2 are also statistically significantly *less* precise than the speech QoE experiment, and this is true for all precision measures (assuming a 5% significance level). This is unusual. According to [3], image QoE experiments are generally *more* precise than speech QoE experiments (cf. Fig. 5 of [3]). Precision smaller than that of the speech QoE experiment and similar to the VR QoE experiment indicates that there is a problem with the precision of V&C experiments.

As image QoE experiments, V&C should have the precision similar to that of other image or video QoE experiments [3]. Table 1 in [3] shows that typical image and video experiments correspond to a between 0.0377 and 0.2116. However, VIME1 and CCRIQ2 correspond to a of 0.314 and 0.347, respectively. Such high readings of a make the two experiments resemble cloud gaming QoE experiments, which are classified as one of the *least* precise QoE experiments in [3]. Taking this and previous observations into account, it is clear that our experiment precision measures correctly detected the low precision of V&C experiments.

V. Discussion

The results in Sec. IV suggest that the ℓ experiment precision measure is the most reliable one. Our intuition here is that this measure's dominance over the two other measures stems from its ability to ignore subject bias. We can observe how ignoring bias influences the results in Tab. II. From Tab. 1 in [5], we know that the HDTV4 experiment (denoted V-4 in Tab. II) recruited people with subject biases significantly exceeding (in terms of its range of values) subject biases present in all other HDTV experiments. Yet, the ℓ measure does not flag this experiment as significantly different from other HDTV experiments. This is a strong suggestion that this measure follows our theoretical assumptions regarding the notion of experiment precision (cf. Sec. II-A) and truly compares experiments without considering potential differences in response biases.

We stated in Sec. IV that both the g and a experiment precision measures labelled 12 out of 15 same-type experiment pairs as significantly different. The same is not true for the ℓ measure, which detected significant differences in only two pairs. Although these statements point to ℓ measure's superiority, there is one caveat that we must mention. The g and a measures internally use per stimulus estimated parameters. For the case of HDTV experiments, this means that the two measures operate on 168 stimuli (i.e., on a sample with 168 observations). On the other hand, the ℓ measure internally uses per subject estimated parameters. Since there were 24 participants in each HDTV experiment, the ℓ measure operates on a sample of 24 observations. Now, in general, the more observations in a sample, the more precise the estimation process. It is thus natural that g and a measures are more sensitive to changes in experiment precision than the ℓ measure. Hence, there is a chance that the increased sensitivity of g and a measures is responsible for the higher number of significant differences detected by these measures. In other words, it is possible that ℓ measure's seeming superiority stems from its lesser sensitivity to experiment precision changes, rather than from its true accuracy.

Notwithstanding the caveat mentioned above, the ℓ measure seems to be the best in estimating experiment precision (at least within the boundaries of our definition of the concept given in Sec. II-A). Still, our recommendation is to compute all three measures (ℓ, g, and a) for each subjective experiment conducted. A measure's indications should be reported, along with their standard errors and the number of observations they are based on. For example, if we were to report the g measure for the sixth experiment of the VQEG HDTV Phase I study (cf. the first row of Tab. I), we would give measure's indication (0.908), its standard error (0.0050) and the number of observations it is based on (168, since that many video stimuli were presented to experiment participants).

Although we present a set of measures assessing experiment precision, we would like to stress that these measures are *not* sufficient to compare the precision of a pair of experiments. In other words, our measures should not be used as the *only* mean used to compare a pair of experiments in terms of their precision. Experiment precision is a multifaceted concept. We thus recommend approaching the topic comprehensively. When comparing a pair of subjective experiments in terms of their precision, we suggest considering, among others, the following factors: i) inter-rater reliability, ii) the number of subjects discarded due to the post-experimental screening of subjects using Pearson linear correlation (cf. clause 11.4 of [17]), iii) width of MOS confidence intervals (cf. clause A1-2.2 of [18]), and iv) to what extent an experiment conforms to the ITU and the research community guidelines.

VI. Conclusions and Further Work

In this work, we propose a notion of experiment precision. We also define and test three experiment precision measures and related experiment precision comparison methods. We do so by using real data from subjective experiments on VR, speech, image, and video QoE. We provide guidelines regarding reporting experiment precision measures as well. We hope that these guidelines will be followed by practitioners creating new subjective data sets. At last, our results suggest that the Li2020 model [2] based ℓ measure performs best in assessing experiment precision.

We demonstrate that all three measures order (in terms of experiment precision) the three types of real-world subjective experiments (VR, speech, and video QoE) identically. Still, the experiment precision ordering of individual experiments within one experiment type differs depending on which measure we use. Nonetheless, the indications of all three measures are in line with expert intuition and domain knowledge regarding the three experiment types investigated.

We believe that with this work we provide sufficient evidence to support the claims we put forward in Sec. I. Specifically, our experiment precision measures turn out to be able to position an experiment in terms of its precision in relation to various experiment types (e.g., video or speech QoE experiments). Our experiment precision comparison methods make it possible to statistically compare (in terms of precision) multiple experiment runs.

We hope that our notion of experiment precision will help MQA practitioners differentiate between subjective experiments. In particular, we envision that experiment precision measures may be part of a set of tools aimed at detecting differences between subjective experiments performed following distinct methodologies. For example, the notion of experiment precision may help decide which experiment methodology results in generally more precise responses. The two methodologies compared may, for example, use two different response scales (a five-point scale vs. a seven-point scale). Thanks to our experiment precision measures, it may be easier to decide which experiment methodology should be followed when response precision is key.

To facilitate adoption of our experiment precision measures, we make available a source code implementing the experiment precision measures and experiment comparison methods presented in this paper at https://github.com/Qub3k/qoe-experiment-precision.

Acknowledgments

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grant SE 3163/3-1, project number: 500105691. The research leading to this results and done by Lucjan Janowski and Jakub Nawała has received funding from the Norwegian Financial Mechanism 2014-2021 under project 2019/34/H/ST6/00599. The authors alone are responsible for the content.

References

[1] L. Janowski *et al.*, “Generalized score distribution,” *arXiv*, 2019. [Online]. Available: http://arxiv.org/abs/1909.04369

[2] Z. Li *et al.*, “A Simple Model for Subject Behavior in Subjective Experiments,” in *Human Vision and Electronic Imaging (HVEI) 2020*, 2020. [Online]. Available: http://arxiv.org/abs/2004.02067

[3] T. Hoßfeld, R. Schatz, and S. Egger, “SOS: The MOS is not enough!” *Quality of Multimedia Experience (QoMEX)*, 2011.

[4] M. H. Pinson, “Confidence intervals for subjective tests and objective metrics that assess image, video, speech, or audiovisual quality,” NTIA Report 21-550, 2020.

[5] L. Janowski and M. Pinson, “Subject bias: Introducing a theoretical user model,” in *Quality of Multimedia Experience (QoMEX)*, 2014.

[6] ——, “The Accuracy of Subjects in a Quality Experiment: A Theoretical Subject Model,” in *IEEE Transactions on Multimedia*, vol. 17, no. 12, 2015, pp. 2210–2224.

[7] Z. Li and C. G. Bampis, “Recover Subjective Quality Scores from Noisy Measurements,” *Data Compression Conference*, 2017.

[8] M. Seufert, “Fundamental advantages of considering quality of experience distributions over mean opinion scores,” *QoMEX*, 2019.

[9] ——, “Statistical methods and models based on quality of experience distributions,” *Quality and User Experience*, vol. 6, no. 1, pp. 1–27, 2021.

[10] J. Nawała *et al.*, “Describing Subjective Experiment Consistency by p-Value P-P Plot,” in *ACM Multimedia (MM)*, 2020.

[11] J. Nawała *et al.*, “Generalised score distribution: A two-parameter discrete distribution accurately describing responses from quality of experience subjective experiments,” 2022, accepted for publication in IEEE Transactions on Multimedia. [Online]. Available: https://arxiv.org/abs/2202.02177

[12] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, “Formal definition of qoe metrics,” *arXiv*, 2016. [Online]. Available: http://arxiv.org/abs/1607.00321

[13] J. Gutierrez *et al.*, “Subjective evaluation of visual quality and simulator sickness of short 360° videos: Itu-t rec. p. 919,” *IEEE Transactions on Multimedia*, vol. 24, pp. 3087–3100, 2021.

[14] ITU-T Study Group 12, “ITU-T Coded-Speech Database,” 1998. [Online]. Available: http://handle.itu.int/11.1002/1000/4415

[15] M. Pinson *et al.*, “Report on the validation of video quality models for high definition video content,” *Video Quality Experts Group*, 2010. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/hdtv/

[16] J. Nawała *et al.*, “Study of subjective data integrity for image quality data sets with consumer camera content,” *Journal of Imaging*, vol. 6, pp. 1–21, 2020.

[17] ITU-T, “Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment,” Rec. ITU-T P.913, 2021.

[18] ITU-R, “Methodologies for the subjective assessment of the quality of television images,” Rec. ITU-R BT.500-14, 2019.