DiVA
http://www.diva-portal.org

This is the published version of a paper published in *Journal of Electronic Imaging (JEI)*.

Citation for the original published paper (version of record):

Brunnstrom, K., Barkowsky, M. (2018)
Statistical quality of experience analysis: on planning the sample size and statistical significance testing
*Journal of Electronic Imaging (JEI)*, 27(5): 053013
https://doi.org/10.1117/1.JEI.27.5.053013

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-35233

Journal of Electronic Imaging

JElectronicImaging.org

# Statistical quality of experience analysis: on planning the sample size and statistical significance testing

Kjell Brunnström
Marcus Barkowsky

SPIE. IS&T imaging.org

Kjell Brunnström, Marcus Barkowsky, "Statistical quality of experience analysis: on planning the sample size and statistical significance testing," *J. Electron. Imaging* **27**(5), 053013 (2018), doi: 10.1117/1.JEI.27.5.053013.

# Statistical quality of experience analysis: on planning the sample size and statistical significance testing

**Kjell Brunnström**[a,b,*] **and Marcus Barkowsky**[c]
[a]Visual Media Quality—RISE AB, Acreo, Kista, Sweden
[b]Mid Sweden University, Department of Information Technology and Media (ITM), Sundsvall, Sweden
[c]Deggendorf Institute of Technology (DIT), University of Applied Sciences, Deggendorf, Germany

**Abstract.** This paper analyzes how an experimenter can balance errors in subjective video quality tests between the statistical power of finding an effect if it is there and not claiming that an effect is there if the effect is not there, i.e., balancing Type I and Type II errors. The risk of committing Type I errors increases with the number of comparisons that are performed in statistical tests. We will show that when controlling for this and at the same time keeping the power of the experiment at a reasonably high level, it is unlikely that the number of test subjects that are normally used and recommended by the International Telecommunication Union (ITU), i.e., 15 is sufficient but the number used by the Video Quality Experts Group (VQEG), i.e., 24 is more likely to be sufficient. Examples will also be given for the influence of Type I error on the statistical significance of comparing objective metrics by correlation. We also present a comparison between parametric and nonparametric statistics. The comparison targets the question whether we would reach different conclusions on the statistical difference between the video quality ratings of different video clips in a subjective test, based on the comparison between the student T-test and the Mann–Whitney U-test. We found that there was hardly a difference when few comparisons are compensated for, i.e., then almost the same conclusions are reached. When the number of comparisons is increased, then larger and larger differences between the two methods are revealed. In these cases, the parametric T-test gives clearly more significant cases, than the nonparametric test, which makes it more important to investigate whether the assumptions are met for performing a certain test. *© 2018 SPIE and IS&T* [DOI: 10.1117/1.JEI.27.5.053013]

Keywords: Type-I error; video quality; statistical significance; quality of experience; Student T-test; Bonferroni; Mann–Whitney U-test; parametric versus nonparametric test.

Paper 180214 received Mar. 14, 2018; accepted for publication Aug. 17, 2018; published online Sep. 25, 2018.

## 1 Introduction

Currently, subjective experiments are the best way to investigate the user's Quality of Experience (QoE) for video. Typically, in such experiments, panels of observers rate the quality of video clips that have been degraded in various ways. When analyzing the results, the experimenter often computes the mean over the experimental observations, a.k.a. the Mean Opinion Scores (MOS) and applies statistical hypothesis tests to draw statistical conclusions. A statistical hypothesis test is done by forming a null hypothesis ($H_0$)[1] and an alternative hypothesis ($H_1$) that can be tested against each other. For example, it could be interesting to know whether a new compression algorithm is better than an older one. A way to resolve this question would be to devise a subjective test where two compression algorithms would encode different source video contents at some different bitrates; then, the test subjects could rate the video quality of each video clip, i.e., each combination of source video, algorithm, and bit rate. We will then get for each source content and bitrate two MOS scores that we can compare whether they are statistically different or not. The usual way is to assign the case that the MOS are the same to null hypothesis $H_0$ and the case that they are differ to the alternative hypothesis $H_1$. If we find that we can reject $H_0$, we can then conclude that there is a statistically significant difference between the algorithms at that particular bitrate and source content. Of course, this is just one way this type of test can be used in the analysis of a subjective test.

As in the example above, often, in video quality assessment, the hypothesis test will have the null hypothesis, $H_0$, that the two underlying MOS values are the same and the alternative hypothesis, $H_1$, that they are different. If the result is significant, the experimenter knows with high probability (typically 95%) that $H_1$ is true and thus the MOS values are different. However, there is still a small risk (5% in this case) that this observation is only by chance. If this happens, it is a Type I error—to incorrectly conclude that $H_1$ is true when in reality $H_0$ is true.

When there are more pairs of MOS values to compare, each comparison has the above-mentioned small risk of error. An example is trying to roll the dice and get the number six. If the dice is rolled once, there will be a probability of one-sixth to get the desired number six, and each time the dice is rolled the probability will be the same. However, the overall chance will increase with the number of times the dice is rolled. The same applies to risk of an error, which increases with the number of comparisons and can be estimated by $1-(1-\alpha)^n$, where $\alpha$ is the risk to have an error at a certain confidence level per comparison and n is the number of comparisons.[1] For 100 comparisons at a 95%

*Address all correspondence to: Kjell Brunnström, E-mail: kjell.brunnstrom@ri.se

1017-9909/2018/$25.00 © 2018 SPIE and IS&T

confidence level, this equals >99% risk of at least one Type I error.

The other type of error that can be committed in a statistical inference is to fail to reject the null hypothesis while there is an effect, i.e., not to discover a significant effect. This type of error is referred to as Type II error and usually, has the associated parameter $\beta$, but more common is to talk about power, which is the probability of rejecting $H_0$ when $H_1$ is true and power $= 1 - \beta$.[1] A common value for $\beta$ is 0.2, which is closely connected to the common significance level 0.05. This gives a 4-to-1 relationship between the risk of missing an effect and finding one that is not there; a $\beta$ of 0.2 gives a power of 0.8 that is an 80% probability of finding an effect if it is there. The power will depend on the chosen significance level, the magnitude of the effect of interest, and the sample size. It is most often used for planning the experiments and is not recommended for posthoc analysis, i.e., analysis of the data after the experiments have been done.[2,3]

In subjective video quality assessment based on standardized procedures,[4,5] category scales such as absolute category rating (ACR) is one of the most commonly used scales and especially its five-level version: excellent, good, fair, poor, and bad. The most common approach to analyze the data from an ACR-experiment is to use a parametric approach. That is to translate the categories to numbers using 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. Then, the MOSs are calculated by taking the mean over the test subjects. The corresponding analysis then assumes that the distribution of votes follows a normal distribution. However, this normality assumption in categorical data for subjective assessments has already motivated Thurstone[6] to state "The normal probability curve has been so generally abused in psychological and educational measurement that one has reason to be fearful of criticism from the very start in even mentioning it."[6] There is reason to question this assumption if only the category levels have been presented to the test subjects because then the scale is only an ordinal scale.[7] There is a clear ordering of the categories, but there may be different distances. When calculating the mean, we assume that the scale has equal distance between scale values, i.e., that it has the properties of an interval scale. There are ways to partly assure this by presenting the numbers on a line equally spaced and instructing the test subjects that they should assume equal distances. However, this is not always done and since these labels have an ordinary meaning to people it might not be possible to achieve. Some studies are showing the difficulty in achieving equal distances, especially when also comparing across languages.[8–10] Huynh-Thu et al.[11] made a thorough comparison of different scales but assumed that the ACR scale was an interval scale in their analysis. It is, therefore, of interest to compare a nonparametric approach with parametric analysis, to see how often do we reach different conclusions depending on the analysis performed, which we have done in this paper.

In this paper, we will investigate and discuss statistical methods used for QoE assessment, especially targeting the planning of the sample size for subjective experiments, i.e., the number of test subjects, before the actual experiment is executed. We will show that with applying traditional statistical methods, strong inferences can be obtained of the planned sample size. The purpose of the paper is not to improve the statistical analysis per se but to demonstrate its usage and typical consequences when conducting subjective experiments for QoE analysis. Furthermore, we believe that this type of paper is lacking within the scientific domain of QoE. Many articles and reports address the problem partly but in most cases, they use the statistical methods as a tool to solve their primary goal of the paper, for example performance evaluation of objective metrics or compression algorithms. It has been part of Video Quality Experts Group (VQEG) reports, e.g., Refs. 12–15 and International Telecommunication Union (ITU) recommendations, e.g., Refs. 5, 16, and 17 but also part of scientific papers, e.g., Refs. 18 and 19. Although, Bayesian statistical analysis has been brought forward as a more powerful alternative to traditional hypothesis testing,[20] we have concentrated on the former in this work as this is still by far the dominating way of statistical analysis in QoE. It is partly based on previous material,[21,22] but this paper goes deeper into the analysis and is substantially extended.

## 2 Method

There are various statistical methods to safeguard against Type I errors. Here, it is important to distinguish between planned comparison and posthoc testing. If some multiple comparisons are planned before the data are collected, then this number is what is used to safeguard against Type I errors. Then, of course, only these multiple comparisons should be performed when the data are collected.[1] Otherwise, all possible comparisons should be considered. An intuitive argument for that is that when observing the actual MOS values and then decides on what comparisons to perform, implicitly all of the comparisons have already been made when picking out the cases to compare.

A common way to compare a set of means is to perform an analysis of variance (ANOVA) followed by a posthoc test. This is a two-step approach where first ANOVA indicates whether there is an overall effect, then a more refined test [such as Tukey honestly significant difference (HSD)] analyzes whether there are any significant pairwise differences. However, it is quite difficult to estimate the influence of a particular number of comparisons on the efficiency of the statistical test. Fortunately, there is also a rather straightforward method, suggested by Bonferroni,[1] where the considered significance level ($\alpha$) is divided by the number of comparisons ($n$) so that the significance level for each comparison will be $\alpha/n$. The advantage here is that it can be combined with simple tests, such as the Student's T-test. The disadvantage is that it can be overly conservative. For example, if there are 10 comparisons and the overall $\alpha = 0.05$, then each comparison should have a significance level of $0.05/10 = 0.005$.

We will study the statistical significance level α at three different cases.

C1. 95% confidence level, i.e., $\alpha = 0.05$.

C2. Compensating for Type I error based on the number of the videos involved in the test, using Bonferroni method $\alpha/n$ ($n$ = number of comparisons).

C3. Compensating for Type I error, using Bonferroni, when the full posthoc pairwise comparison is performed per experiment.

These three cases are considered for the following reasons: C1 represents the no-control case—i.e., the occurrence of Type I errors is accepted to be on a comparison-by-comparison basis and conclusions based on multiple comparisons would exceed the 5% error margin. This is what most experimenters within the QoE community have used in the past. C2 is the typical case for technical studies on performance gains, such as old versus new video coding algorithms. For the same set of source videos (SRC), any two hypothetical reference circuits (HRC) get compared, i.e., a specific bitrate or coding quality setting. In this case, the error control needs to extend over each HRC separately, but it is assumed that the experimenter is not interested in comparing two distinct individual SRC. C3 provides this full control, i.e., any combination of SRC and HRC can be compared with any other combination. This is typically the case for training, verifying, and validating objective measures in which each processed video sequences (PVS) is checked separately and put in relation to any other PVS by pooling measures, such as Pearson or Spearman rank order correlation or root mean square error. Furthermore, C3 represents the case when no preplanning of comparisons has been made, but the data are explored for possible interesting differences and then all possible comparisons need to be compensated for.

For the test design, there are two important cases to distinguish, which in turn affects the statistical analysis. The two cases are whether it is a between-group design or within-subject design. The first means that the same test subject has just been used once or giving their ratings once, but in the other case, the same test subject has provided answers more than once.

The within-subject design is very common for video quality experiments. Usually, different degraded versions of video clips are presented to the same observer that is asked to give a quality score for each of them. For the analysis, as there are dependencies between the scores, we need to use the dependent T-test for paired samples.[23]

The pure between-group case is not that common because it would usually require quite a few test subjects but could occur, for instance, when experiments have been repeated by different labs or repeated by different panels of observers in the same lab. For instance, when comparing two experiments using the same distorted videos. The experimenter might want to test whether there are differences in MOS between the two panels for the same video clips. For the analysis, the different scores are independent, and we can use the independent two-sample T-test.

The cases C1-3 can be applied to the analysis regardless whether the study is within-subject design or between-group design.

In video quality experiments, there are different options for the experimental methods that can be used. Some of them are standardized by the ITU (ITU-R Rec. BT.500-13; ITU-T Rec. P.910, ITU-T Rec. P.913).[4,5,24] The method could be single stimulus as in the absolute category rating (ACR) method or double stimulus as in the double stimulus continuous quality scale (DSCQS). Central to the methods are the rating scales that could be discrete in, e.g., five levels as in the ACR method or continuous as in the DSCQS methods. Here, we will assume a quality scale that can be mapped to the range of 1 to 5, where the discrete levels correspond to poor, bad, fair, good, and excellent. Furthermore, we will assume that it has been statistically confirmed that parametric statistics can be applied and the underlying distribution is essentially normal. These two assumptions can be questioned in the sense that the ACR scale is a discrete ordinal scale and therefore should be analyzed with nonparametric methods. However, the parametric analysis is still very commonly applied and what is recommended by the ITU, although strictly speaking this is not statistically correct.

In this study, we look at the interesting cases of MOS differences of 0.5 and 1.0 on a five-level scale. A MOS difference of 0.5 and 1 was chosen in the following as they represent typical targets. They may be motivated in two distinct ways. First, due to the quantization of the ACR scale, the observers are forced to vote for an integer value even if their opinion is in between two attributes. A single observer who decides one way or the other changes the MOS score by $1/m$ (m being the number of observers). To obtain a MOS difference of 0.5, half the observers need to change their opinion and all need to change their opinion to get a MOS difference of 1. The smaller the MOS difference, the more likely it is that the result is due to quantization noise, a MOS difference $>1$ is unlikely to satisfy the accuracy goals of a study. The second way to motivate the choice is based on the distribution of the observers as a Gaussian distribution with a given sigma value, typically 0.8. If we could achieve a significant MOS difference close to 0, the Gaussian distribution for the two stimuli would overlap and the preference result for a single observer would be random. The larger the difference, the more of the population agrees on the ordering; according to the cumulative Gaussian distribution function with a sigma of 0.8, 73% agree at a MOS difference of 0.5, and 89% at 1.[6]

A different view is from a macroscopic perspective. When voting, the test subjects are forced to select a level even if he/she consider the quality to be between those values. Then, MOS values between these levels mean that some test subjects have selected a higher level and some have gone for a lower level. When we get MOS difference of 0.5, it is the half way between these levels and the MOS difference of 1.0 is a whole level, e.g., in one lab a video is rated as "good," but at another, it is just rated "fair." These five quality labels are used in other constructions of quality scales, where test subjects are able to give values between the labels, e.g., DSCQS[4] or the nine-point ACR[5] scale, but will still anchor the voting. Other differences can of course be of interest depending on the particular focus of a study, but we believe that the chosen cases have practical usage in themselves and are good cases to discuss around.

We then consider the influence of multiple comparisons on the number of test subjects required and on the differences between MOS that are statistically significant. The cases that are analyzed are single comparison, a fixed number of preplanned number of comparison, and all possible pairwise comparisons.

An area where statistical significance testing is often neglected is when the performance of objective metrics is analyzed. This is often done by comparing the output of the objective metrics with MOS from one or more subjective experiments using Pearson correlation (PCC). The Recommendation ITU-T Rec. P.1401[17] gives guidelines how this analysis can be done in a statistically better way, but

it does not consider multiple comparison and we have, therefore, analyzed that in this article. Nuutinen et al.[25] proposed an interesting performance evaluation method of objective metrics for video and image quality assessments, by developing the subjective root-mean-square error (SRMSE), which ties the performance metrics to the variance of subjective data and gives its output in the number of average observers and can predict whether an algorithm is likely to be able to replace a subjective experiment. However, they did not consider multiple comparisons in their analysis.

### 2.1 *Within-Subject Design*

The Student T-test for a within-subject design is a dependent T-test for paired samples. The equation is

$$t_{\text{obs}} = \frac{\mu_D - \mu_o}{\sigma_D}\sqrt{n}, \tag{1}$$

for calculating the observed $t$-value for the paired samples. Where $\mu_D$ is the difference between the paired samples or ratings from the same test subject and $\sigma_D$ is the standard deviation of the paired samples. $n$ is the number of paired samples. $\mu_0$ is used if the comparison is done against another value than zero. We will assume in our analysis this value to be zero. The degrees of freedom are $(n-1)$. For any given values of the difference mean $\mu_D$ between two means ($\mu_1$ to $\mu_2$), the number of data points ($n$), and the standard deviations ($\sigma_D$), we can calculate the probability of significance, $p$.

For the power analysis we have used the pwr-package,[26] in $R$[27] and for the within-subject design case, we specified the "paired" keyword for the "type" parameter in the function "pwr.t.test."[28]

### 2.2 *Between-Group Design*

To analyze an effect in the between-group design case, we assume the Student's T-test with equal standard deviations and the same number of data points in the two mean values, based on independent data samples. This gives the simplified equation:

$$t_{\text{obs}} = \frac{\mu_1 - \mu_2}{\sqrt{2}\sigma}\sqrt{n}. \tag{2}$$

The degree of freedom is in this case $(2n-2)$. We can, in the same way as above, analyze the requirements for getting statistical significance by calculating the probability $p$ for different input values.

For the power analysis, we have also here used the pwr-package[26] in $R$[27] and for the between-group design case we specified the "two.sample" keyword for the "type" parameter in the function "pwr.t.test."[28]

### 2.3 *Pearson Correlation*

The PCC is usually calculated between human subjects and predicted scores from objective measures. For estimating the probability significance for the PCC, we follow ITU-T Rec. P.1401.[17] The PCC is defined as follows:

$$\text{PCC} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})\cdot(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2}\cdot\sqrt{\sum(Y_i - \overline{Y})^2}}, \tag{3}$$

where $n$ is the total number MOS scores that are compared to the same number of predicted MOS scores. $X_i$ is the subjective MOS scores and $\overline{X}$ is their mean. $Y_i$ is the predicted MOS scores and $\overline{Y}$ is their mean.

The PCC is not normally distributed, but if the Fisher $z$ transformation is applied we can get a normally distributed variable:

$$z = 0.5\cdot\ln\left(\frac{1+\text{PCC}}{1-\text{PCC}}\right); \qquad \sigma_z = \sqrt{\frac{1}{n-3}}. \tag{4}$$

We can see that the standard deviation only depends on the number of points used in the correlation, i.e., the number of subjective and predicted MOS scores that are compared.

We can then form a test statistic to evaluate against for a two-tailed Student's T-distribution:

$$z_n = \frac{z_1 - z_2}{\sqrt{2}}\cdot(n-3), \tag{5}$$

with the degrees of freedom of: $2n-2$ if we are comparing PCC with the same number of involved subjective and predicted MOS scores.

### 2.4 *Parametric versus Nonparametric*

For parametric statistics when comparing two mean values, the Student's T-test can be used. For between-group design the different scores are independent, and we can use the independent two-sample T-test (ITT), and in the other case there is a dependency between the scores, and we need to use the dependent T-test (DTT) for paired samples.[1]

For the nonparametric statistics, we have chosen to use the Mann–Whitney U-test, which is usually put forward as the nonparametric counterpart of the T-test.[1] It has an independent (IMW) and dependent variant as well (DMW).

We have compared the outcome of the statistical testing parametric and nonparametric from two large video quality investigations. One was the HDTV phase I test[14] by VQEG and an adaptive streaming investigation based on three different subjective experiments. Here all the videos were the same each time, but the experimental conditions were different in the three cases.[29]

#### 2.4.1 *Normality analysis*

When deciding on whether to use a parametric or a nonparametric analysis, it is important to investigate whether the data are normally distributed. We have, therefore, applied a few statistical tests for normality to the subjective video quality datasets presented below and the result is presented in Sec. 3.4.1.

#### 2.4.2 *VQEG HDTV phase I-test*

In VQEG HDTV phase I investigation,[14] there were six subjective experiments performed in different labs in the various countries. It contained a common set, consisting of of 24 PVS in total, formed from four SRCs crossed with six treatments or HRC. These PVSs were the same in all the subjective tests allowing to compare the ratings given across experiments. This is particularly interesting as here the test subjects have had their category scale levels given in different languages, which based on the earlier investigations[8–10] would give different distances in quality between the scale values and, therefore, it could be argued that a nonparametric

analysis would be preferable. Each video clip across experiments and labs have been scored by different panels of observers and can, therefore, be analyzed with an independent hypothesis test. Each PVS can then be tested pairwise between all experiments (case C2), i.e., $6 \times 5/2 = 15$ times, which gives in total $15 \times 24 = 360$ hypothesis test comparisons. The $\alpha$ levels used here are 0.05 (C1) and $0.05/360 = 0.00014$ (C2).

Then, each experiment also contained PVSs unique for each experiment. There were in total 168 PVS including the common set. These have been scored with a within-subject design, so when comparing these with each other, we must use hypothesis test for dependent samples. If we perform all pairwise comparisons within each experiment, we will get $168 \times 167/2 = 14{,}028$ pairwise comparisons and in total $6 \times 168 \times 167/2 = 84{,}168$ hypothesis test comparisons. The $\alpha$ levels used here are: 0.05 (C1), $0.05/168 = 0.002$ (C2), and $0.05/14{,}028 = 3.6 \cdot 10^{-6}$ (C3). There were 24 test subjects used in each of the six experiments.

#### 2.4.3 *Adaptive streaming investigation*

The adaptive streaming investigation consisted of three experiments, where the same PVSs were presented in a different way and rated with test subjects partly from various countries and then different mother tongue. It could, therefore, be possible that scales have not been experienced in the same way and that the scale distances have not been the same. As in the HDTV test above, we can analyze the results in two ways: across the three experiments and then compare each PVS with each other using independent tests and within each experiment with dependent tests. There were 132 PVSs used, and they were rated by 21 test subjects in two experiments and 20 in one.

For the independent hypothesis tests comparison, this gives $3 \times 132 = 396$ hypothesis test comparisons $(\alpha = 0.05/396 = 0.00013)$(C2), and for the dependent test comparison, there were $3 \times 132 \times 131/2 = 25{,}938$ hypothesis test comparisons $(\alpha = 0.05/25{,}938 = 1.9 \cdot 10^{-6})$(C3).

## 3 Results

### 3.1 *Within-Subject Design*

Figure 1 shows curves for simulated MOS differences, for experiments using within-subject design, ranging from 0.2 to 1.4 along the $x$-axis. The standard deviation used was motivated by actual experiments: VQEG HDTV test,[14] where the average standard deviation was about 0.8, which included six different subjective tests. We observed similar or slightly lower average standard deviations in our previous adaptive streaming quality experiment.[29] Along the $y$-axis are the $p$-values. The plotted curves are for 20 (black curve), 30 (medium gray curve), and 40 (light gray curve) test subjects. Different $\alpha$ levels have been indicated with horizontal lines. Dotted line shows $\alpha = 0.05$ (C1), short dashed line shows $\alpha = 0.05/100 = 0.0005$ (C2), corresponding to 100 comparisons and dashed line all pairwise comparisons among 100 cases, i.e., 4950 comparisons $(\alpha = 0.05/4950 = 0.00001)$ (C3). The different curves must be below the $\alpha$ threshold for the Student's T-test to detect a difference in MOS at the 95% confidence level.

As an example, test case, we assume that we are planning the number of subjects for a study in which we use only 1

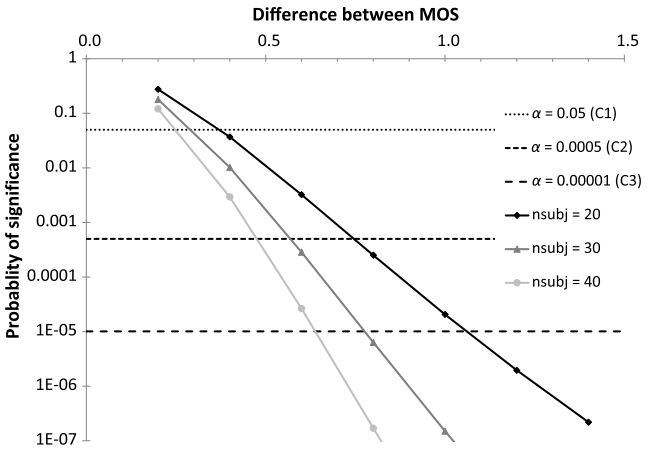


**Fig. 1** Probability of significance as a function of the difference between compared simulated MOS values for subjective experiments using within-subject design. The different curves show the probability for significance for 20 (black curve), 30 (medium gray curve), and 40 (light gray curve) test subjects and with an assumed standard deviation of 0.8 estimated for the VQEG HDTV test.

comparison (C1), a set of comparison such as a codec comparison with 100 preplanned comparisons (C2), and the full set of all 4950 possible comparison as for the validation of objective measures (C3). The corresponding $\alpha$ values are 0.05 (C1), 0.0005 (C2), and 0.00001 (C3), respectively. We plan for a significance at two MOS differences, 0.5 and 1.0 as explained earlier. We have calculated the probability of significance for different number of test subjects which is in the range of what is typically used, as shown in Fig. 1: 20 subjects (black curve), 30 (medium gray), and 40 (light gray). It can be noted from the curves that 20 subjects will not be completely sufficient to reliably discover a statistical difference of 1.0 MOS when all pairwise comparisons are considered but 30 and 40 test subjects will. For 100 comparisons, all the calculated numbers of test subjects will be able to show a difference of 1.0, but for a difference of 0.5 we need to use about 40 test subjects or more, as shown in Fig. 1, neither 20 or 30 test subjects will be sufficient. We have tabulated the minimum number of subjects for each test condition in Table 1. Please keep in mind that the ITU recommends a minimum of 15, and VQEG used 24 subjects.

In Fig. 2, we have plotted the curves for the probability of significance for simulated MOS differences of 1.0 (medium gray curve) and 0.5 (black curve) as a function of the number of test subjects. We have also indicated with vertical lines the minimum number of test subjects recommended by ITU, i.e., 15 (long dashed line)[4] and what has been used by VQEG, i.e., 24 (dot dashed line) see, e.g., (VQEG, 2008, 2010).[14] For a simulated MOS difference of 1.0, we can see that 15 test subjects would not be sufficient to conclude significance with an overall significance level of 95% with all pairwise comparisons compensated for, but for preplanned 100 comparisons or just one comparison it would work just fine. Twenty-four test subjects would be good in all the three analyzed cases. For a simulated MOS difference of 0.5, only one comparison will be significant for both 15 and 24 test subjects, but the other cases will not.

Furthermore, Fig. 2 shows a view that may be more practical for preplanning: first, the number of preplanned

**Table 1** The number of required test subjects (sample size) for obtaining a power of 0.8 and for different significance levels $\alpha$ and effect sizes (simulated MOS differences). Two samples sizes are shown in the two rightmost columns, based on two different estimated standard deviations in the experiment, 0.8 and 1.0.

| Design type | # comparisons | $\alpha$ | Simulated MOS difference | Sample size Std dev 0.8 | Sample size Std dev 1.0 |
|---|---|---|---|---|---|
| Within | 1 | 0.05 | 0.5 | 23 | 34 |
| | | | 1.0 | 8 | 10 |
| | 100 | 0.0005 | 0.5 | 54 | 81 |
| | | | 1.0 | 18 | 25 |
| | 4950 | 0.00001 | 0.5 | 81 | 121 |
| | | | 1.0 | 27 | 37 |
| Between | 1 | 0.05 | 0.5 | 42 | 64 |
| | | | 1.0 | 12 | 17 |
| | 100 | 0.0005 | 0.5 | 99 | 153 |
| | | | 1.0 | 27 | 41 |
| | 4950 | 0.00001 | 0.5 | 147 | 227 |
| | | | 1.0 | 41 | 61 |

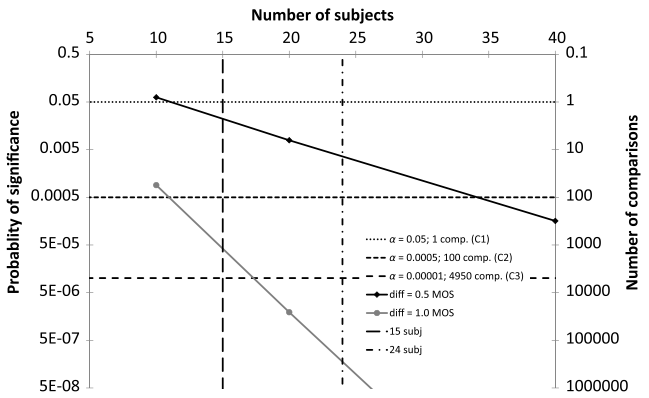


**Fig. 2** Probability of significance as a function of the number of test subjects for subjective experiments using within-subject design. The different curves show the probability for significance for a simulated MOS difference of 1.0 (medium gray curve), and an MOS difference of 0.5 (black curve) and with an assumed standard deviation of 0.8 estimated for the VQEG HDTV test. The vertical lines indicate 15 (long dashed line) and 24 (dot dashed line) test subjects. The secondary *y*-axis to the right shows the corresponding number of comparison, when using Bonferroni to safeguard the Type I error to $\alpha = 0.05$.

comparisons according to the specific setup (C1, C2, and C3) is used to calculate the $\alpha$ value that forms a horizontal line parallel to the *x*-axis. Then, the interception point of the MOS difference curves with that line is searched for and the minimum number of subjects on the *x*-axis can be determined graphically.

In Fig. 3, we have drawn the sample size, i.e., the number of the subjects as a function of effect size, i.e., the difference in the simulated MOS that would be planned to be resolved for a power of 0.8. The different graphs in Fig. 3 are drawn for different significance levels $\alpha$: [$\alpha = 0.05$, solid curves (C1)], 100 comparisons [$\alpha = 0.0005$, dashed curves (C2)], and 4950 comparisons [$\alpha = 0.00001$, dotted curves (C3)]. We have marked the specific cases of simulated MOS difference of 1.0 [Fig. 3(a)] and 0.5 [Fig. 3(b)]. The calculated numbers are summarized in Table 1.

For the preplanned case of 100 comparisons, we would need 18 test subjects, which is lower than what VQEG is normally using, i.e., 24, but more than what is recommended by ITU-R BT.500-13,[4] which is 15. It is only without compensating for multiple comparisons we can get by with less than what is recommended in ITU-R BT.500-13,[4] and here we get 8. For a simulated MOS difference of 0.5, we need at least 23 test subjects for just one comparison and then even higher numbers for the other cases, see Table 1.

### 3.2 *Between-Group Design*

In a similar way as before, Fig. 4 shows curves for simulated MOS differences, for experiments using between-group design, ranging from 0.2 to 1.4 along the *x*-axis. The standard deviation used here was 0.8, which is the same as before. Along the *y*-axis are the *p*-values. The plotted curves are for 20 (black curve), 30 (medium gray curve), and 40 (light gray curve) test subjects. Different $\alpha$ levels have been indicated with horizontal lines. The dotted line shows $\alpha = 0.05$ (C1), short dashed line $\alpha = 0.0005$ (C2), corresponding to 100 comparisons and finally, the dashed line shows all pairwise comparisons among 100 cases, i.e., 4950 comparisons ($\alpha = 0.00001$) (C3). The different curves must be below the $\alpha$ threshold for the Student's T-test to detect a difference in MOS at the 95% confidence level. We can see that 30 test subjects will be just about sufficient to reliably discover a statistical difference of 1.0 MOS when all pairwise comparisons are considered. For 100 comparisons, 20 test subjects will be needed for a MOS difference of 1.0 to be significant. However, for a MOS difference of 0.5 we only get significance if they do not compensate for multiple comparisons as in C1, with the considered number of test subjects. Furthermore, for 20 test subjects we only have borderline significance.

The vertical lines in Fig. 5 indicate 15 (long dashed line) and 24 (dot dashed line) test subjects. For 15 test subjects, it is only possible to show significance for one comparison and with a MOS difference of 1.0 (intersection of the medium gray curve and long dashed line). It can be observed, on the other hand, that for 24 subjects and one comparison, we get significance for both simulated MOS differences of 0.5 and 1.0 (the intersection of both curves and the dot dashed line). With 100 comparisons, only a simulated MOS difference of 1.0 is significant (intersection of the medium curve and dot dashed line). With all 4950 pairwise comparisons, 24 test subjects cannot detect a simulated MOS difference of 1.0.

In the same way as for Fig. 2, Fig. 5 could also be used for preplanning as described above.

In Fig. 6 we have drawn the sample size, i.e., the number of the subjects as a function of effect size, i.e., the difference in the simulated MOS that would be planned to be resolved for a power of 0.8. The different graphs in Figs. 6(a) and 6(b) are drawn for different significance levels $\alpha$: 0.05 (C1), 0.0005 (C2), and 0.00001 (C3). We have marked the specific cases of simulated MOS difference of 1.0 [Fig. 6(a)] and 0.5 [Fig. 6(b)], respectively. The calculated numbers are summarized in Table 1 for these cases as well. We can then see that if we want to make the trade-off

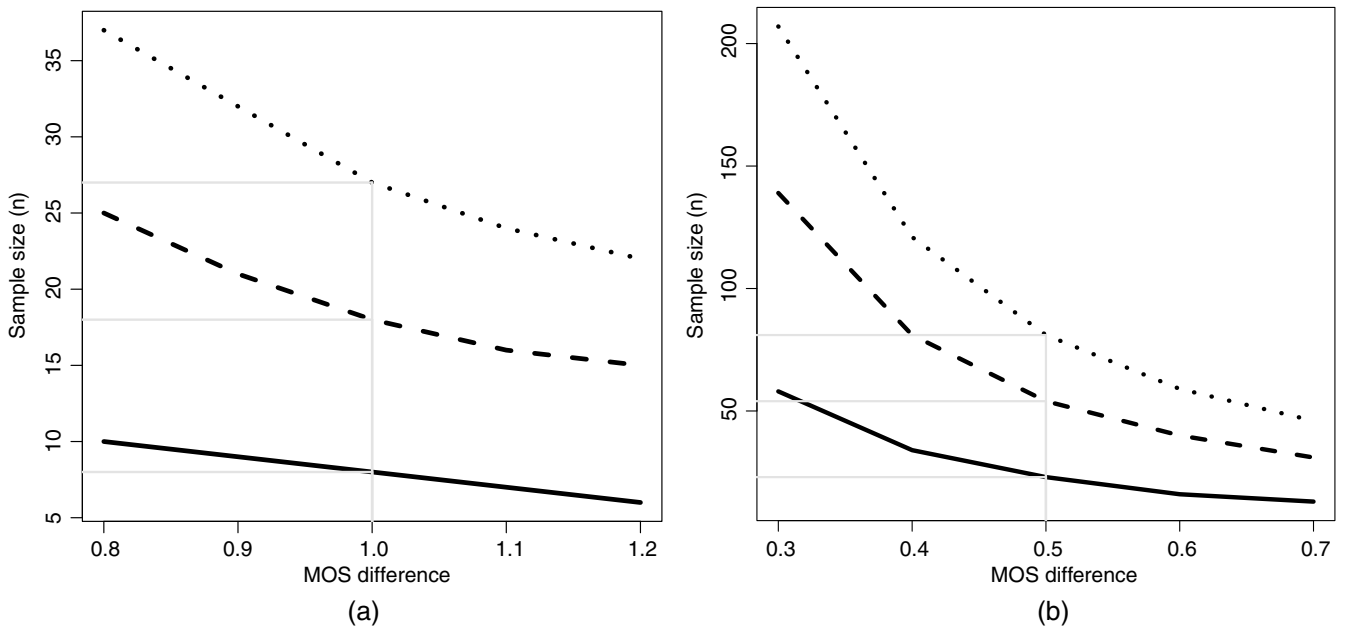


**Fig. 3** The sample size, i.e., the number of test subjects required for a within-subject designed video quality experiment with a power of 0.8 as a function MOS difference for three different significance levels $\alpha$. (C1) $\alpha = 0.05$ (solid curve). (C2) $\alpha = 0.0005$ (dashed curve) (C3) $\alpha = 0.00001$ (dotted curve). (a) The sample sizes required (gray lines) to be able to find an MOS difference of 1.0 and (b) the same for a MOS difference of 0.5.

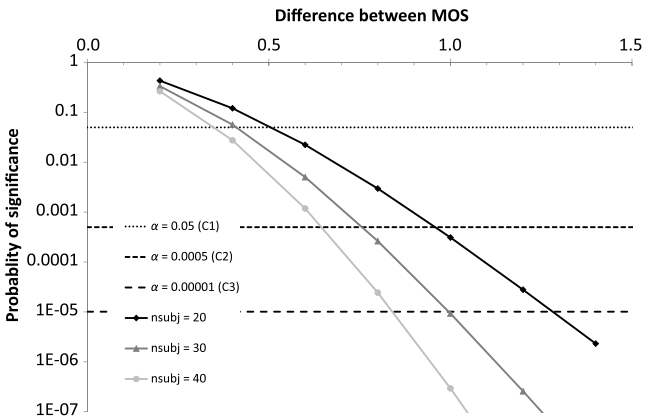


**Fig. 4** Probability of significance as a function of the difference between compared simulated MOS values for subjective experiments using between-group design. The different curves show the probability for significance for 20 (black curve), 30 (medium gray curve), and 40 (light gray curve) test subjects and with an assumed standard deviation of 0.8 estimated for the VQEG HDTV test.

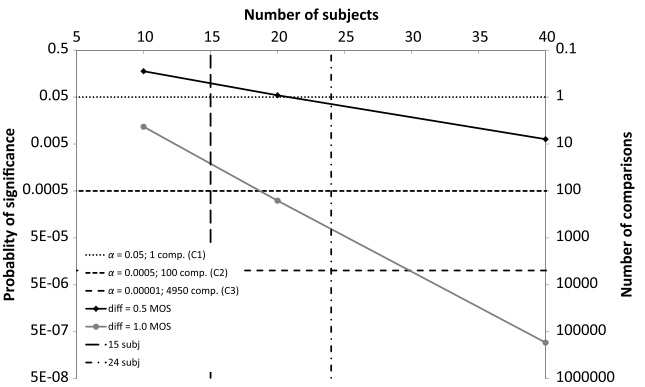


**Fig. 5** Probability of significance as a function of the number of test subjects for subjective experiments using between-group design. The different curves show the probability for significance for an MOS difference of 1.0 (medium gray curve), and a MOS difference of 0.5 (black gray curve) and with an assumed standard deviation of 0.8 estimated for the VQEG HDTV test. The vertical lines indicate 15 (long dashed line) and 24 (dot dashed line) test subjects. The secondary *y*-axis to the right shows the corresponding number of comparison, when using Bonferroni to safeguard the Type I error to $\alpha = 0.05$.

and reach a power of 0.8 and at the same time compensate for all possible comparisons of 100 PVSs, we would need 41 test subjects for finding a simulated MOS difference of 1.0. For the preplanned case of 100 comparisons, we would need 27 test subjects. It is only without compensating for multiple comparisons we can get by with about the same as what is recommended in ITU-R BT.500-13,[4] which is 15, and here we get 12. For a simulated MOS difference of 0.5, we need at least 42 test subjects for just one comparison and then even higher numbers for the other cases.

### 3.3 *Pearson Correlation*

Let us now consider the impact of multiple comparisons when evaluating objective metrics with PCC.[3] Figure 7 shows the probability of significance for two correlation coefficients PCC1 and PCC2 when the difference between the correlation coefficients is PCC1 − PCC2 = 0.05 (for example, a difference between correlations of PCC1 = 0.90 and PCC2 = 0.85). The *y*-axis shows PCC2. The different curves represent different numbers of data points (10, 100, and 1000). 100 data points (i.e., video sequences)

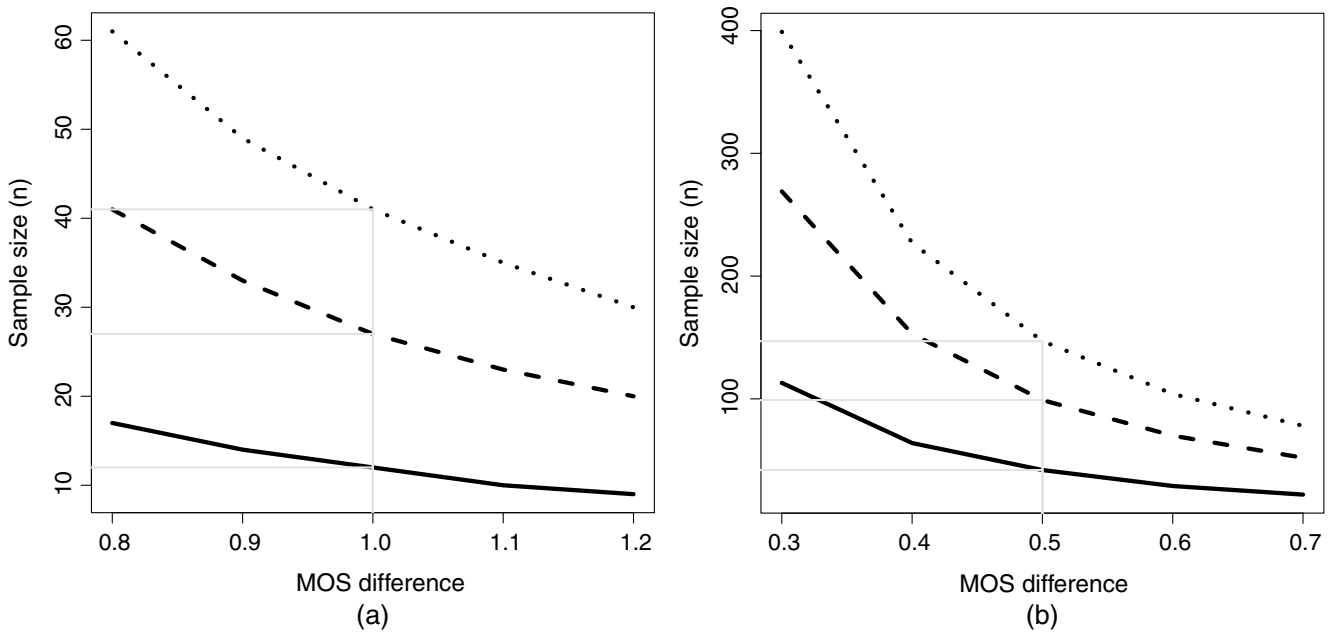


**Fig. 6** The sample size, i.e., the number of test subjects required for a between-group designed video quality experiment with a power of 0.8 as function MOS difference for three different significance levels $\alpha$. (C1) $\alpha = 0.05$ (solid curve). (C2) $\alpha = 0.0005$ (dashed curve) (C3) $\alpha = 0.00001$ (dotted curve). (a) The sample sizes required (gray lines) to be able to find an MOS difference of 1.0 and (b) the same for a MOS difference of 0.5.

are a common number in a single video quality experiment. We assume that we like to compare in total the prediction performance of 10 different objective measures, we indicate the significance level of 1 comparison ($\alpha = 0.05$) with a dotted horizontal line (one measure to one other measure), 9 comparisons ($\alpha = 0.0056$) with a short dashed line (one measure to all others), and 45 comparisons ($\alpha = 0.0011$) with a dashed line (all measures to all measures, the case most often claimed). Looking at the intersection of the medium gray curve with the dotted line, we see that the significant differences can be expected first when the correlation is about PCC2 = 0.92 (PCC1 = 0.97) and then only when we are doing just one comparison. When doing multiple comparisons, no significance can be detected from 100 data points, even if we get perfect correlation of 1.0 for one measure. With more data points the situation improves, so for 1000 data points, which is rare to have in a subjective test, we can expect significance for difference of 0.05 from 0.8 correlation and for all comparisons among 10 different models.

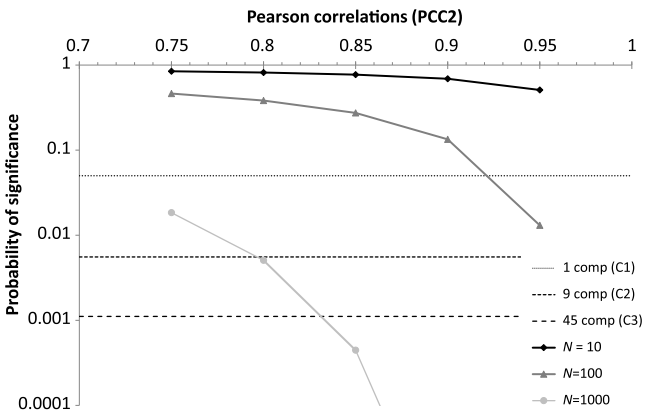


**Fig. 7** Probability of significance for PCCs with a difference of 0.05, where *N* is the number of data points. The *y*-axis shows PCC2, which is compared with PCC1, which is then PCC1 = PCC2 + 0.05.

### 3.4 *Parametric versus Nonparametric*

#### 3.4.1 *Normality analysis*

As mentioned in Sec. 1, parametric tests require the underlying data to follow a normal distribution. Therefore, 10 different statistical tests for normality with their default parameters have been calculated using the statistics package R on the votes of each PVS.[30,31] The statistical tests differ in sensitivity—while the Hegazy–Green goodness-of-fit test using order statistics alerts on 2% (27 out of 1406 PVS), Cramer–von-Mises test for the composite hypothesis of normality did not find a single case. However, it can be noted from Table 2, which shows the confusion matrix that the statistical tests refuse the same sets of votes.

To illustrate the difficulty, a border case for the Hegazy–Green normality test is shown in Fig. 8 as a quartile–quartile plot. It may be interpreted either as a normal distribution with a standard deviation that is larger than the one-to-five scale or as a uniform distribution with some random accumulation region in the center. As this is a typical border case and <2% of the data are concerned, parametric tests may be justified. On the other hand, the example shows that with only five choices for only 24 subjects, typical statistical analysis methods such as the quartile–quartile plot reach their limits and nonparametric tests may stabilize the analysis.

#### 3.4.2 *VQEG HDTV phase I test*

For the HDTV experiment, we got statistical significance for C1 148 times using ITT and 142 times using IMW out of 360, which is a difference of about 1.7%. For C2 we got

**Table 2** Confusion matrix of the number of PVS (out of a total of 1406 PVS) that are identified as being nonnormally distributed for 10 frequently used normality tests implemented in the statistics package R. The confusion matrix shows that the same PVS get identified, but the sensitivity is different. The methods used are: M1, Hegazy–Green2; M2, Weisberg–Bingham; M3, Shapiro–Francia; M4, Shapiro–Wilk; M5, Lilliefors; M6, Hegazy–Green1; M7, Anderson–Darling; M8, Frosini; M9, Cramer–von-Mises; and M10, Pearson chi-squared test for the composite hypothesis of normality.

| | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 |
|---|---|---|---|---|---|---|---|---|---|---|
| M1 | 27 | 13 | 12 | 8 | 2 | 2 | 1 | 1 | 0 | 0 |
| M2 | 13 | 13 | 12 | 8 | 2 | 2 | 1 | 1 | 0 | 0 |
| M3 | 12 | 12 | 12 | 8 | 2 | 2 | 1 | 1 | 0 | 0 |
| M4 | 8 | 8 | 8 | 8 | 2 | 2 | 1 | 1 | 0 | 0 |
| M5 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| M6 | 2 | 2 | 2 | 2 | 0 | 2 | 1 | 1 | 0 | 0 |
| M7 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 |
| M8 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 |
| M9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

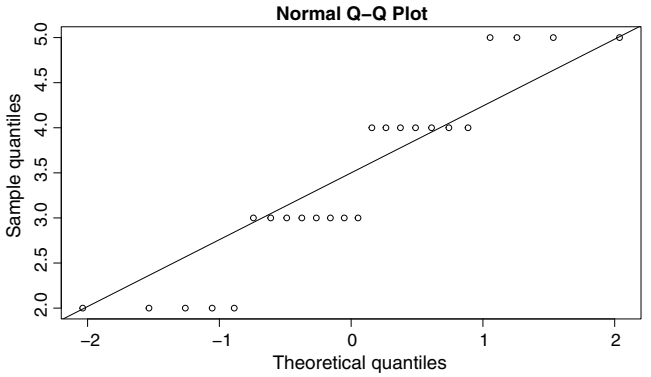


**Fig. 8** Quartile–Quartile plot of the vote distribution for a borderline case of normality: due to the coarseness of the ACR scale and the few observers, the distribution may be interpreted either as a normal distribution or as uniform distribution (VQEG HD4, PVS 72).

18 for ITT and 14 IMW, which is a difference of 1%, also out of 360 comparisons.

For dependent case C1 (see Fig. 9 gray bars), we got 14,389 for DTT and 14,267 for DMW out of 84,168, which is a difference of 0.1%. For C2, we had 11,556 for DTT and 10,189 for DMW, which is a difference of about 2%. And finally, for C3 we got 9592 for DTT and 20 for DMW, which is a difference of about 11%. We can observe an increased difference of significant cases when the number of comparisons increases. The reason is that for increased differences between the compared MOS values in the comparisons, the p-values based on the T-test will continue to decrease, but the $p$-values based on values of Mann–Whitney U-test will level out and stop decreasing at some

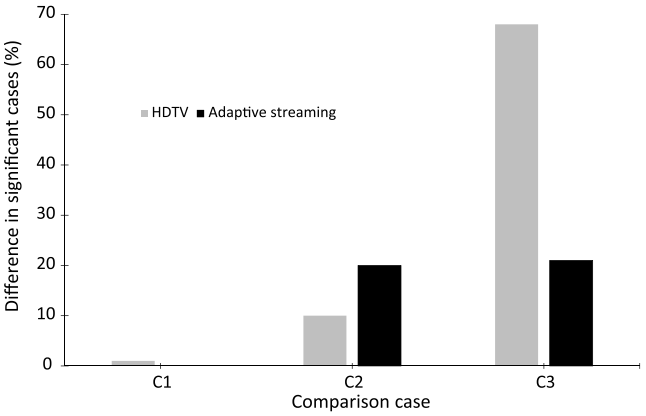


**Fig. 9** The relative increase in statistically significant cases for the parametric T-test as compared with the nonparametric Mann–Whitney U-test.

point. Minimum $p$-value in the HDTV test for the T-test was as low as $10^{-31}$, whereas the lowest $p$-value for Mann–Whitney U-test found here was about $10^{-6}$.

### 3.4.3 *Adaptive streaming investigation*

For the adaptive streaming investigation, we got for C1 68 ITT and 64 IMW out of 396, which is difference of about 1%. For C2 we got 4 for ITT and 4 IMW, which is no difference.

For dependent case C1 (see Fig. 9 black bars), we got 17,543 for DTT and 17,951 for DMW out of 25,938, which is a difference of 2%. For C2 we had 5188 for DTT and 3704 for DMW, which is a difference of 21%. And finally, for C3 we got 5192 for DTT and 0 for DMW, which is a difference of almost 18%. In this case, we do not see the same increase in difference between C2 and C3, which is due to the distribution of qualities in this test that are more compressed than in the HDTV test and very large differences between the MOS do not occur in the same way as for the HDTV test. Min $p$-value for the T-test was $10^{-16}$ and for the Mann–Whitney U-test was $10^{-5}$.

## 4 Discussion

In this paper, we have been using the Bonferroni[1] method for the simplicity to illustrate the different cases described in the paper. This model is simple to compute, but very conservative. To broaden the discussion, we have performed a comparison to some other methods. Closely related to Bonferroni method is Holm–Bonferroni.[32] It was named in the original paper "sequentially rejective Bonferroni test", and this is a good description of the procedure. The p-values are sorted from the lowest to the largest. Then the smallest $p$-value is compared against the Bonferroni $\alpha$-level, i.e., $\alpha/n$. If this is smaller it is considered significant. The next $p$-value in the list is compared with an updated $\alpha$-level, i.e., $\alpha/(n-1)$ and so on until there is a $p$-value that is larger than the recomputed $\alpha$-level. The method has as good protection against Type I errors as the Bonferroni method, but is more efficient. We have analyzed what this could mean in our video quality tests, by comparing the number of significant cases on our datasets, see Table 3. To compare we have also computed Tukey HSD.[1] Tukey HSD was computed using Statistica 64 10. The significant values for the case C2 were picked out of the full posthoc comparison

**Table 3** The ratio of significant cases to the total number of comparisons for the VQEG HDTV[14] test and the Adaptive streaming investigation[29] as defined in Sec. 2. The case C2 is the preplanned number of comparisons corresponding to the number of PVS involved in the experiment and the case C3 is the full posthoc pairwise comparison, see also Sec. 2. The $\alpha$-level used is 0.05 and the false discovery rate 0.1. The number in parentheses is the actual number and is given for some close cases.

| Method/experiment | No control (%) | Bonferroni (%) | Holm (%) | Tukey HSD (%) | Benjamini–Hochberg (%) | Benjamini–Yekutieli (%) |
|---|---|---|---|---|---|---|
| VQEG HDTV C2 | 41 | 5 (18) | 5 (19) | 6 (21) | 31 | 11 |
| VQEG HDTV C3 | 82 | 50 | 52 | 60 | 80 | 70 |
| Adaptive streaming C2 | 17 | 1 (4) | 1 (4) | 0 | 2 | 0 (1) |
| Adaptive streaming C3 | 69 | 20 | 21 | 29 | 66 | 50 |

table, which means the compensation for Type I was higher than necessary for C2 and gives Tukey HSD a slightly unfair advantage in this comparison. However, this is a common implementation of Tukey HSD found in, e.g., both Statistica and R.

It is possible to take a different view, which has recently become popular especially in data mining, machine, and deep learning. There will potentially become a huge number of comparison and controlling the Type I errors in the traditional way, will have a very low power. It was the introduced by Benjamini and Hochberg,[33] that instead of trying to control the risk for Type I errors, the false discovery rate (FDR) should be controlled. Meaning that we know that some statistically significant cases may be wrong, but we control the amount of it to some prescribed level, e.g., 10%. This has led to much more powerful methods, which we also have compared with Benjamini and Hochberg[33] and Benjamini and Yekutieli.[34] One view that has been put forward is that one experiment is not performed in isolation, but rather in a series and then the significant cases can be candidates for further studies. An example in QoE could be that one experiment is performed by ACR and a second is followed using pair comparison to separate the close but significant cases from the first study.

What we can see from the comparison in Table 3 is that the traditional posthoc methods give almost the same number of significant cases, whereas the FDR-based ones give substantially more significant cases.

The expected standard deviation also has a substantial impact on the sample size, as shown in the rightmost column of Table 3. It shows the impact on the sample size if the expected standard deviation becomes 1.0. For instance, the number of test subjects needed for C2 for the within case and MOS difference of 1.0 will increase from 18 to 25.

Not all scales allow for a parametric evaluation and should be analyzed with nonparametric methods. However, the parametric test will in most cases have greater power than the nonparametric tests and would, therefore, act as the limiting case, i.e., at least these number of test subjects would be required. On the other hand, we have used the Bonferroni model for compensating, which is perhaps a bit too safe. In the case where a parametric model can be used the current simulation may give too conservative numbers, but, for the nonparametric method, they may be a better match to what is required. This needs to be further investigated, though, but has been out of scope in the current investigation.

Our investigation shows that in most cases, the number of test subjects should increase in comparison with what is traditionally recommended. That does not mean the experiments cannot be performed using this lower number of test subjects. If a statistically significant effect is found in a particular study, it can be reported as existing within the local context of this study with the safe guards against Type I errors used, regardless whether the effect can be globally observed or reproduced. However, there is an obvious risk that significant effects will be missed if the number of test subjects are not preplanned to find effects of a certain size.

In articles about comparisons of performances between different objective video quality measurement methods, correlations coefficients are often reported with four decimal digits. The analysis in this paper shows that we could expect at most two decimal digits to be significant. Furthermore, comparisons are also reported without supporting statistical significance tests, and current analysis indicates that many reported differences in performance have been nonsignificant unless the number of fitted data points has been large. If PCC is used as the performance criteria, then this analysis gives indications of the number of sample videos that are needed to find reasonably significant differences between the objective metrics. Similar type of analysis should also be performed on other performance metrics, e.g., the root mean squared error and the outlier ratio, which we intend to do in future work, which then could be partly based on the work by Nuutinen et al.[25]

There was hardly any difference when the significance level was about 0.05, so if just a few pairs are compared then almost the same conclusions are reached. However, if the number of comparisons is increased and the significance level is lowered based on, e.g., Bonferroni to compensate for Type I errors, then larger and larger differences between the two methods are revealed. Then, the parametric T-test gives clearly more significant cases, than nonparametric test.

The consequence of this is that when claiming significance based on the T-test based on very low p-values, where the nonparametric test does not give a significant value, it becomes much more important to have checked that the assumptions for a parametric analysis have been met.

## 5 Conclusion

In this paper, we investigated how to balance the trade-off between compensating for multiple comparisons and still have large power, i.e., probability of finding an effect if it is there, in subjective video quality experiments. The conclusion is that we need to use, in most cases, a larger number of test subjects, than current recommendations. For studies

using within-subject design, when preplanning the number of comparisons to perform and the standard deviation can be kept sufficiently low, it comes down to the number of test subjects usually used by VQEG, i.e., 24, or even below.

For objective metric comparisons using correlation coefficients, it is difficult to find any significance with few data points and correlations below 0.9. In this case, multiple comparisons have a large impact on the final conclusions that can be drawn.

We have also analyzed the difference between parametric and nonparametric analysis when it comes to taking the decisions whether there is a statistically significant difference in video quality between two video clips.

We found that there was hardly no difference when few comparisons are compensated for, i.e., then almost the same conclusions are reached. When the number of comparisons is increased, then larger and larger differences between the two methods are revealed. In these cases, the parametric T-test gives clearly more significant cases, than the nonparametric test, which makes it more important to investigate whether the assumptions are met for performing a certain test.

To provide practical guidance, we have proposed a simple method to estimate the number of required observers in function of the planned comparisons and the targeted significant MOS difference for typical values of subjective evaluations in video quality.

## Acknowledgments

The support from Knowledge Foundation (Grant no. 20160194, Funder ID https://doi.org/10.13039/100003077) and VINNOVA (Sweden's innovation agency) is hereby gratefully acknowledged.

## References

1. S. E. Maxwell and H. D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 2nd ed., Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey (2003).
2. L. Thomas, "Retrospective power analysis," *Conserv. Biol.* **11**(1), 276–280 (1997).
3. S. J. Walters, "Consultants' forum: should post hoc sample size calculations be done?" *Pharm. Stat.* **8**(2), 163–169 (2009).
4. ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R Rec. BT.500-13, International Telecommunication Union, Radiocommunication Sector (2012).
5. ITU-T, "Subjective video quality assessment methods for multimedia applications," ITU-T Rec. P.910, International Telecommunication Union, Telecommunication standardization sector (1999).
6. L. L. Thurstone, "Psychophysical analysis," *Am. J. Psychol.* **38**(3), 368–389 (1927).
7. J. C. Baird, *Fundamentals of Scaling and Psychophysics / John C. Baird, Elliot Noma*, Wiley series in behavior, E. J. Noma, Ed., Wiley, New York (1978).
8. R. Hamberg and H. de Ridder, "Continuous assessment of time-varying image quality," *Proc. SPIE* **3016**, 248–259 (1997).
9. K. Teunissen, "The validity of CCIR quality indicators along a graphical scale," *SMPTE J.* **105**(3), 144–149 (1996).
10. N. Narita, "Graphic scaling and validity of Japanese descriptive terms used in subjective -evaluation tests," *SMPTE J.* **102**(7), 616–622 (1993).
11. Q. Huynh-Thu et al., "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Trans. Broadcast.* **57**(1), 1–14 (2011).
12. VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," VQEG Final Report of FR-TV Phase II Validation Test, Video Quality Experts Group (VQEG) (2003).
13. VQEG, "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I," VQEG Final Report of MM Phase I Validation Test, Video Quality Experts Group (VQEG) (2008).
14. VQEG, "Report on the validation of video quality models for high definition video content," Video Quality Experts Group (VQEG), www.vqeg.org (2010).
15. VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Video Quality Experts Group (VQEG), ITU (2000).
16. ITU-R, "Subjective assessement for image quality in high-definition television," Rec. ITU-R BT.710-3, International Telecommunication Union, Radiocommunication Sector (1997).
17. ITU-T, "Statistical analysis, evaluation and reporting guidelines of quality measurements," ITU-T P.1401, International Telecommunication Union, Telecommunication standardization sector, Geneva, Switzerland (2012).
18. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.* **15**(11), 3440–3451 (2006).
19. J. Breebaart, "Evaluation of statistical inference tests applied to subjective audio quality data with small sample size," *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(5), 887–897 (2015).
20. J. K. Kruschke and T. M. Liddell, "The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective," *Psychonomic Bull. Rev.* **25**(1), 178–206 (2018).
21. K. Brunnström, S. Tavakoli, and J. Søgaard, "Compensating for Type-I errors in video quality assessment," in *7th Int. Workshop on Quality of Multimedia Experience (QoMEX 2015)*, Messina, Greece, IEEE Xplore (2015).
22. K. Brunnström and M. Barkowsky, "Balancing Type I errors and statistical power in video quality assessment," in *IS&T Int. Symp. on Electronic Imaging, Human Vision and Electronic Imaging*, p. HVEI-122, Society for Imaging Science and Technology, Burlingame, California (2017).
23. J. H. McDonald, *Handbook of Biological Statistics*, 3rd ed., Sparky House Publishing, Baltimore, Maryland (2014).
24. 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," ITU-T Rec. P.913, International Telecommunication Union, Telecommunication standardization sector (2014).
25. M. Nuutinen, T. Virtanen, and J. P. Häkkinen, "Performance measure of image and video quality assessment algorithms: subjective root-mean-square error," *J. Electron. Imaging* **25**(2), 023012 (2016).
26. S. Champely, "pwr: basic functions for power analysis," http://CRAN.R-project.org/package=pwr (11 July 2018).
27. R. C. Team, "R: a language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ (2015).
28. R. I. Kabacoff, *R in Action—Data Analysis and Graphics in R*, Manning Publications Co., Shelter Island, New York (2011).
29. S. Tavakoli et al., "Quality of experience of adaptive video streaming: investigation in service parameters and subjective quality assessment methodology," *Signal Process. Image Commun.* **39**(B), 432–443 (2015).
30. I. Gavrilov and R. Pusev, "normtest: Tests for Normality—tests for the composite hypothesis of normality," https://CRAN.R-project.org/package=normtest (11 July 2018).
31. J. Gross and U. Ligges, "normtest: Tests for Normality—five omnibus tests for testing the composite hypothesis of normality," https://CRAN.R-project.org/package=nortest (11 July 2018).
32. S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.* **6**(2), 65–70 (1979).
33. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Series B (Methodol.)* **57**(1), 289–300 (1995).
34. Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Stat.* **29**(4), 1165–1188 (2001).

**Kjell Brunnström** is a senior scientist at RISE AB, leading visual media quality and an adjunct professor at Mid Sweden University. He is cochair of the Video Quality Experts Group (VQEG). His research interests are in Quality of Experience (QoE) for video and display quality assessment (2D/3D, VR/AR, immersive). He is an associate editor of the *Journal of Advances in Multimedia* and has written more than 100 articles in international peer-reviewed scientific journals and conferences.

**Marcus Barkowsky** received his Dr.-Ing. degree from the University of Erlangen-Nuremberg in 2009. He joined the University of Nantes and was promoted to an associate professor in 2010. In 2018, he obtained the professorship on interactive systems and internet of things at the Deggendorf Institute of Technology, University of Applied Sciences. His activities range from designing 3-D interaction and measuring visual discomfort using psychometric measurements to computationally modeling spatial and temporal effects of the human perception.