https://doi.org/10.2352/ISSN.2470-1173.2017.14.HVEI-122

Balancing Type I Errors and Statistical Power in Video Quality Assessment

Kjell. Brunnström[a,b], and Marcus Barkowsky[c]
[a]Acreo Swedish ICT AB, Kista, Sweden
[b]Mid Sweden University, Sundsvall, Sweden
[c]University of Nantes, Nantes, France

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 it 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 will require more test subjects than are normally used and recommended by international standardization bodies like the ITU. Examples will also be given for the influence of Type I error on the statistical significance of comparing objective metrics by correlation.

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) (Maxwell & Delaney, 2003) [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 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 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 α 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% confidence level, this equals more than a 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 β, but more common is to talk about power, which is the probability of rejecting H_0 when H_1 is true and *power* = 1 - β[1]. A common value for β 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 β 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 post-hoc analysis i.e. analysis of the data after the experiments have been done [2]. At least this should be done with great care.

In this paper, we will investigate how to balance the Type I errors and the power of video quality assessment, to find the effects that could be reasonable to look for. It tries to predict the number of test subjects that would be required, for finding different effects in video quality tests.

The paper is an extension of the short paper about Type I error from QoMEX 2015[3], which only covered the between subject case and not investigated the balance with the Type II errors as in this paper. We were motivated to investigate these statistical properties by our recent study (Tavakoli, Brunnström, & Garcia, 2015)[4], where we despite following common practice we did find any statistically significant difference, although we observed large absolute differences between the MOS values,

There are also important discussions when to use parametric or non-parametric statistical methods and if normal distribution assumptions are valid or not in video quality assessment, but those are outside the scope of this paper. We will assume for this discussion the normal distribution assumption is valid and that parametric statistical tests can be used. The motivation for this is that a parametric test will in most cases have greater power than the non-parametric test and would, therefore, act as the limiting case i.e. at least these number of test subjects would be required.

Method

There are various statistical methods to safeguard against Type I errors. Here, it is important to distinguish between planned comparison and post-hoc testing. If some multiple comparisons are planned before the data is 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 is collected. (Maxwell & Delaney, 2003)[1]. Otherwise, all possible comparisons should be taken into account. An intuitive argument for that is that when observing the actual MOS values and then decide on what comparisons to perform, implicitly all 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 post-hoc 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 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 (α) is divided by the number of comparisons (n) so that the significance level for each comparison will be α/n. The advantage here is that it can be combined with simple tests like the Student's T-test. The disadvantage is that it can be overly conservative. For example, if there are ten comparisons and the overall α = 0.05, then each comparison should have a significance level of 0.05/10 = 0.005.

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 [1]. In the first case, the different scores are independent, and we can use the independent two-sample T-test, and in the other case there is a dependency between the scores, and we need to use the dependent T-test for paired samples [5].

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. 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.

In video quality experiments there are different options for the experimental methods that can be used. Some of them are standardized by the International Telecommunication Union (ITU) (ITU-R Rec. BT.500-13; ITU-T Rec. P.910, ITU-T Rec. P.913)[6-8]. 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 non-parametric methods. However, the parametric analysis is still very commonly applied and also 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 5-level scale. In the second case the MOS difference is so large the evaluation level has changed one step from e.g. in one lab a video is rated "good", but at another, it is just rated "fair". 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. We also consider the performance evaluation of objective metrics, based on ITU-T Rec. P.1401[9]. To this end, we analyze Pearson correlation for multiple comparisons.

Within subject design

The Student T-test for a within subject design is a dependent T-test for paired samples. The formula is: $t_{obs} = \frac{\mu_D - \mu_o}{\sigma_D}\sqrt{n}$. Where μ_D is the difference between the paired samples or ratings from the same test subject and σ_D the standard deviation of the paired samples. n is the number of paired samples. μ_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 μ_D between the means (μ_1-μ_2), the number of data points (n) and the standard deviations (σ_D), we can calculate the probability of significance, *p*.

For the power analysis we have used the pwr-package[10] in R [11], and for the within subject design case we specified the "paired" keyword for the "type" parameter in the function "pwr.t.test"[12].

Between subject design

To analyze an effect in the between subject 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 formula $t_{obs} = \frac{\mu_1 - \mu_2}{\sqrt{2}\sigma}\sqrt{n}$. 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 used the pwr-package in R and for the between subject design case we specified the "two.sample" keyword for the "type" parameter in the function "pwr.t.test"[12].

Pearson correlation

The Pearson correlation is usually calculated between human subjects and predicted scores from objective measures. For estimating the probability significance for the Pearson correlation (PCC), we follow ITU-T Rec. P.1401[9], which is defined as follows

$$PCC = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \cdot \sqrt{\sum(Y_i - \bar{Y})^2}}$$

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 $\bar{X}$ their mean. $Y_i$is the predicted MOS scores and $\bar{Y}$ their mean.

The PCC is not normally distributed, but if the Fisher z transformation is applied we can get a normally distributed variable: $z = 0.5 \cdot \ln\left(\frac{1+PCC}{1-PCC}\right)$; $\sigma_z = \sqrt{\frac{1}{n-3}}$. 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: $z_n = \frac{z_1 - z_2}{\sqrt{2}} \cdot (n-3)$, with the degrees of freedom of : $2n-2$ if we are comparing PCC with the same number of involved subjective and predicted MOS scores.

Results

Within subject design

Figure 1 shows curves for MOS differences ranging from 0.2 to 1.4 along the x-axis. The standard deviation used was motivated by actual experiments: VQEG HDTV test (VQEG, 2010)[13], where the average standard deviation was about 0.8, which included six different subjective tests. We observed similar or slightly higher average standard deviations in our previous adaptive streaming quality experiment[4]. Along the y-axis are the p-values. The plotted curves are for 20 (black curve), 30 (blue curve) and 40 (green curve) test subjects. Different alpha levels have been indicated with horizontal lines. Red line shows alpha = 0.05, yellow line alpha (0.0005) corresponding to 100 comparisons and blue line all pairwise comparisons among 100 cases i.e. 4950 comparisons (alpha = 0.00001). 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. 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 subject or more, as shown in Figure 1, neither 20 or 30 test subjects will be sufficient.

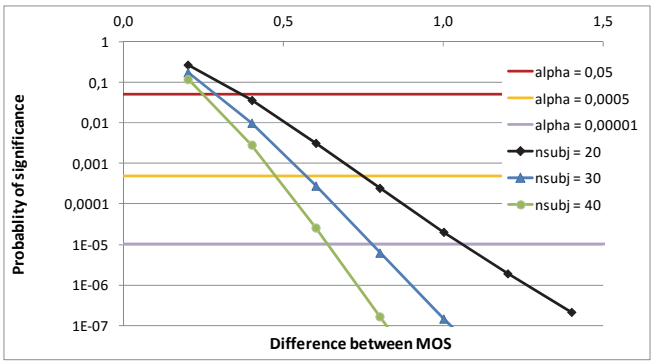

*Figure 1: Probability of significance as a function of the difference between compared MOS values for subjective experiments using **within-subject design**. The different curves show the probability for significance for 20 (black curve), 30 (blue curve) and 40 (green curve) test subjects and with an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

In Figure 2 we have plotted the curves for the probability of significance for MOS differences of 1.0 (black curve) and 0.5 (green 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 (blue line)[6] and what has been used by VQEG i.e. 24 (green line) see e.g. (VQEG, 2008, 2010)[13]. For a MOS differences 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 pair-wise comparisons compensated for, but for pre-planned 100 comparisons or just one comparison it would work just fine. 24 test subjects would be good in all the three analyzed cases. For a MOS differences of 0.5, only one comparison will be significant for both 15 and 24 test subjects, but the other cases will not.

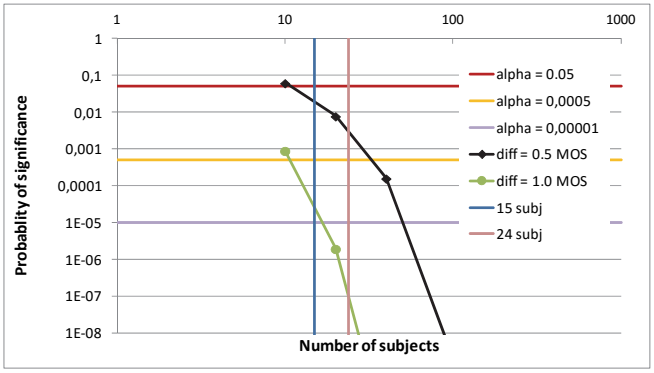

*Figure 2: Probability of significance as a function of the number of test subjects for subjective experiments **using within-subject design**. The different curves show the probability for significance for a MOS difference of 1.0 (black curve), and a MOS difference of 0.5 (green curve) and with an assumed standard deviation of 0.8 estimated for the VQEG HDTV test. The vertical lines indicate 15 (blue line) and 24 (brown line) test subjects.*

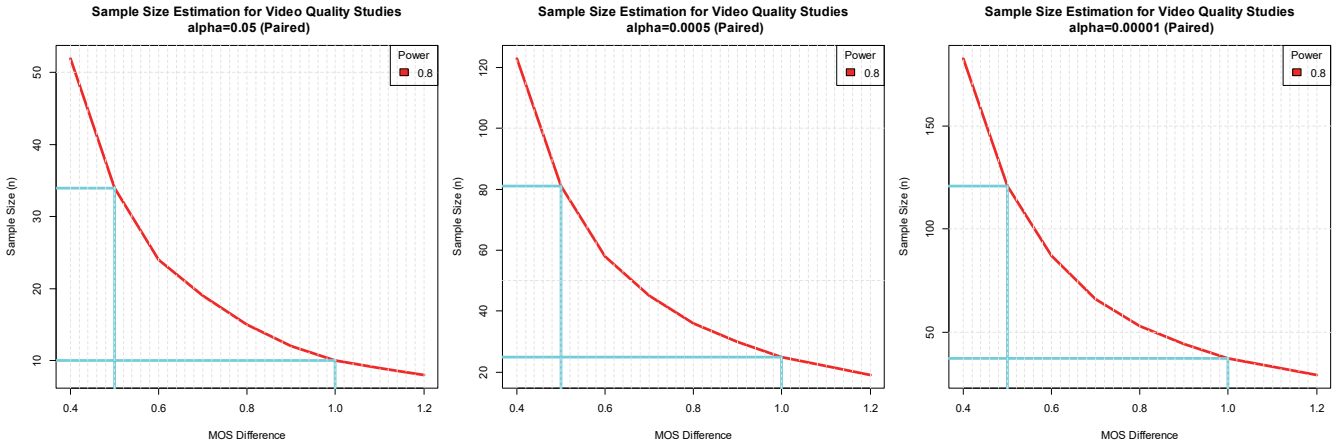

*Figure 3: The sample size i.e. number of test subjects required for a **within subject** designed video quality experiment with a power of 0.8 as function of effect size (MOS difference) for three different significance levels alpha. a) alpha = 0.05. b) alpha = 0.0005 c) alpha = 0.00001*

In Figure 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 MOS that would be planned to be resolved for a power of 0.8. The different graphs in Figure 3 a) to c) are drawn for different significance levels alpha: 0.05, 0.0005 and 0.00001. We have marked the specific cases of MOS difference of 0.5 and 1.0. The calculated numbers are summarized in Table 1 for these cases as well. We can then see if we want to make the trade-off and reach a power of 0.8 and at the same time compensate for all possible comparisons of 100 PVSs (4950 comparisons) we would need 37 test subjects for finding a MOS difference of 1.0. For the pre-planned case of 100 comparisons, we would need 25 test subjects, which is very close to what VQEG is normally using i.e. 24. It is only without compensating for multiple comparisons we can get by with less than what is recommended in ITU-R BT.500-13[6], which is 15, and here we get 10. For a MOS difference of 0.5 we need at least 34 test subjects for just one comparison and then even higher numbers for the other cases, see Table 1

Between subject design

The vertical lines in Figure 4, indicates 15 (blue line) and 24 (pink line) test subjects. The horizontal lines show the p-value indicated by the Bonferroni formula when making one comparison (alpha = 0.05), 100 comparisons (alpha = 0.0005), and 4950 comparisons (alpha = 0.00001).

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 green curve and blue line). It can be observed, on the other hand, that for 24 subjects and one comparison, we get significance for both MOS differences of 0.5 and 1.0 (the intersection of both curves and the green line). With 100 comparisons, only a MOS difference of 1.0 is significant (intersection of the blue curve and purple line). With 4950 comparisons, 24 test subjects cannot detect a MOS difference of 1.0. This is illustrated differently in Figure 5, where we have drawn the probability of significance for the cases of 20, 30 and 40 test subjects as a function of MOS difference. When all pairwise comparisons are considered, as is typical, 30 test subjects are needed for the Student's T-test to reach the low probability p that compensates for all comparisons done, so that 1.0 MOS difference can be safely concluded as significant.

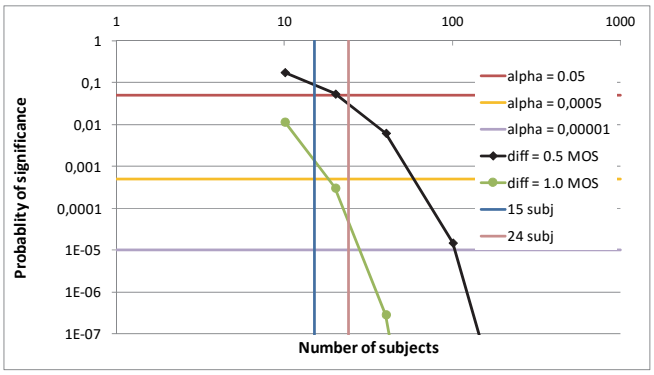

Figure 4: Probability of significance for subjective experiments. 'alpha' and 'diff' denote the confidence level per comparison and MOS difference in order.

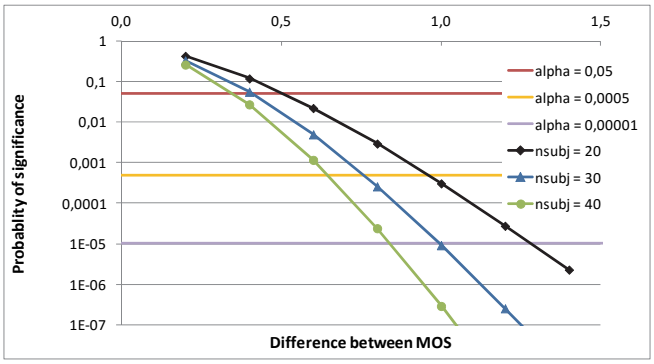

*Figure 5: Probability of significance as a function of the difference between compared MOS values for subjective experiments using **between-subject design**. The different curves show the probability for significance for 20 (black curve), 30 (blue curve) and 40 (green curve) test subjects and with an assumed standard deviation of 0.8 estimated for the VQEG HDTV test.*

In Figure 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 MOS that would be planned to be resolved for a power of 0.8. The different graphs in Figure 6 a) to c) are drawn for different significance levels alpha: 0.05, 0.0005 and 0.00001. We have marked the specific cases of MOS difference of 0.5 and 1.0. 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 and reach a power of 0.8 and at the same time compensate for all possible comparisons of 100 PVSs, we would need 61 test subjects for finding a MOS difference of 1.0. For the pre-planned case of 100 comparisons, we would need 41 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[6], which is 15, and here we get 17. For a MOS difference of 0.5 we need at least 64 test subjects for just one comparison and then even higher numbers for the other cases, see Table 1.

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 (MOS differences).

Design type	Alpha	MOS difference	Sample size
Within	0.05	0.5	34
		1.0	10
	0.0005	0.5	81
		1.0	25
	0.00001	0.5	121
		1.0	37
Between	0.05	0.5	64
		1.0	17
	0.0005	0.5	153
		1.0	41
	0.00001	0.5	227
		1.0	61

Pearson correlation

Let us now consider the impact of multiple comparisons when evaluating objective metrics with Pearson correlation [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 different curves represent different numbers of data points (10, 100 and 1000). 100 data points (i.e. video sequences) is a common number in a single video quality experiment. We assume we like to compare in total the prediction performance of 10 different objective measures, we indicate the significance level of 1 comparison (0.05) with a red horizontal line (one measure to one other measure), 9 comparisons (0.0056) with a yellow line (one measure to all others) and 45 comparisons with a grey line (all measures to all measures, the case most often claimed). Looking at the intersection of the blue curve with the red 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.

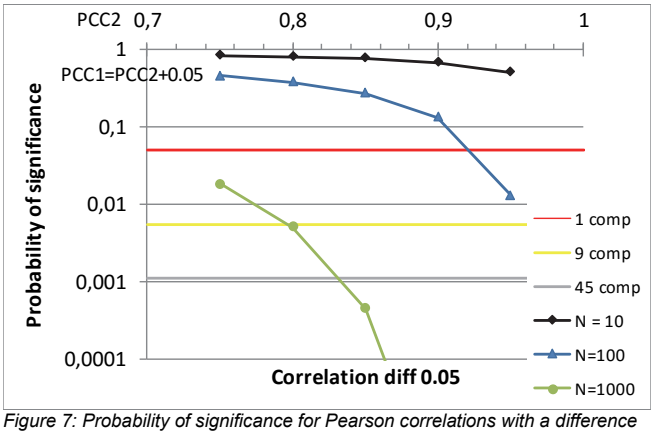

Figure 7: Probability of significance for Pearson correlations with a difference of 0.05, where N is the number of data points.

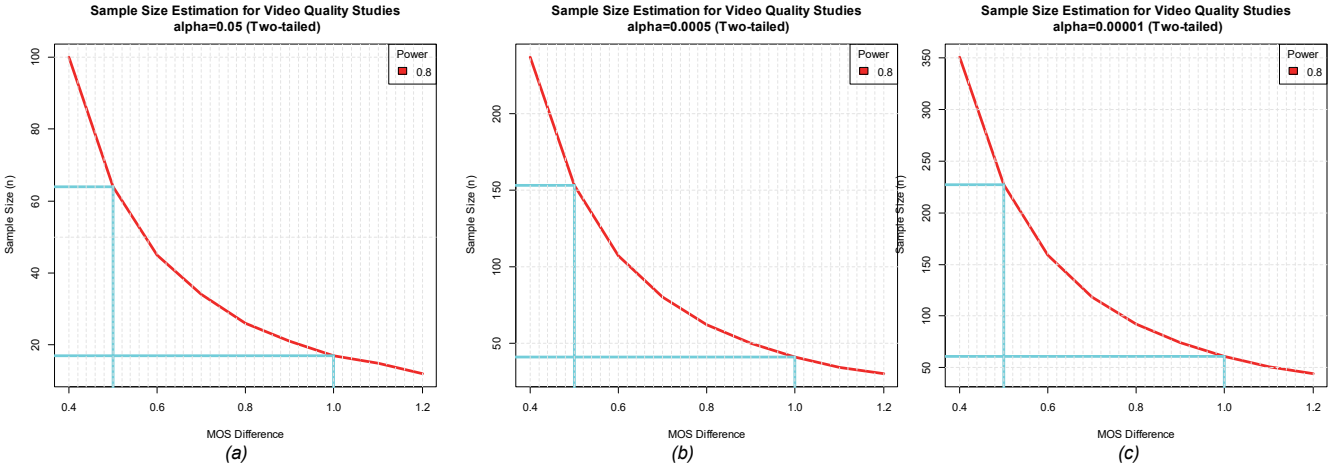

*Figure 6: The sample size i.e. number of test subjects required for a **between subject** designed video quality experiment with a power of 0.8 as function effect size (MOS difference) for three different significance levels alpha. a) alpha = 0.05. b) alpha = 0.0005 c) alpha = 0.00001*

Discussion

In this study, we have used a fairly simple model to compensate for multiple comparisons i.e. Bonferroni [1]. This model may be overly safe, so more efficient models can be used e.g. Tukey HSD [1], or it is possible to raise the overall statistical safeguard against Type I error e.g. setting the overall alpha to 0.1 instead of 0.05. At least, in this case, a conscious choice has been made, and the experimenter is aware of the tradeoff made. However, other methods can be problematic too, for instance [1] recommends Tukey HSD for pairwise comparisons for between-subject designs, but Bonferroni for within subject designs, since Tukey does not always maintain its overall alpha-level.

Not all scales allow for a parametric evaluation and should be analyzed with non-parametric methods. However, the parametric test will in most cases have greater power than the non-parametric 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 non-parametric 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 pre-planned 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 non-significant unless the number of fitted data points has been large (PCC > 0.9 and n > 100). 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.

Conclusions

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 and can pre-plan the number of tests to perform it comes down to the number of test subjects usually used by VQEG.

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.

References

[1]. Maxwell, S.E. and H.D. Delaney, *Designing experiments and analyzing data : a model comparison perspective.* 2nd ed. 2003, Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, Inc.

[2]. Thomas, L., *Retrospective power analysis.* Conservation Biology, 1997. **11**(1): p. 276-280.

[3]. Brunnström, K., S. Tavakoli, and J. Søgaard. *Compensating for Type-I Errors in Video Quality Assessment*. in *7th International Workshop on Quality of Multimedia Experience (QoMEX 2015),*. 2015. Messina, Greece, 26-29 May, 2015: IEEE Xplore.

[4]. Tavakoli, S., K. Brunnström, J. Gutiérrez, and N. Garcia, *Quality of Experience of Adaptive Video Streaming: Investigation in Service Parameters and Subjective Quality Assessment Methodology.* Signal Processing: Image Communication, 2015(doi:10.1016/j.image.2015.05.001).

[5]. McDonald, J.H., *Handbook of Biological Statistics, 3rd ed.* 2014, Baltimore, Maryland, USA: Sparky House Publishing.

[6]. ITU-R. (2012). *Methodology for the subjective assessment of the quality of television pictures* (ITU-R Rec. BT.500-13). International Telecommunication Union, Radiocommunication Sector.

[7]. ITU-T. (1999). *Subjective video quality assessment methods for multimedia applications* (ITU-T Rec. P.910). International Telecommunication Union, Telecommunication standardization sector.

[8]. ITU-T. (2014). *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.

[9]. ITU-T. (2012). *Statistical analysis, evaluation and reporting guidelines of quality measurements* (ITU-T P.1401). International Telecommunication Union, Telecommunication standardization sector: Geneva, Switzerland.

[10]. Champely, S. *pwr: Basic Functions for Power Analysis*. 2015; Available from: http://CRAN.R-project.org/package=pwr.

[11]. Team, R.C. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. 2015 [cited 2015; Available from: http://www.R-project.org/.

[12]. Kabacoff, R.I., *R in Action - Data analysis and graphics in R*. 2011, Shelter Island, NY, USA: Manning Publications Co.

[13]. VQEG. (2010). *Report on the Validation of Video Quality Models for High Definition Video Content*. Video Quality Experts Group (VQEG), www.vqeg.org.

Author Biography

Kjell Brunnström, Ph.D., is a Senior Scientist at Acreo Swedish ICT AB and Adjunct Professor at Mid Sweden University. He is an expert in image processing, computer vision, image and video quality assessment having worked in the area for more than 25 years. Currently, he is leading standardization activities for video quality measurements as Co-chair of the Video Quality Experts Group (VQEG). His current research interests are in Quality of Experience for visual media in particular video quality assessment both for 2D and 3D, as well as display quality related to the TCO requirements.

Marcus Barkowsky received the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 2009. He joined the Image and Video Communications Group at IRCCyN at the University of Nantes in 2008, and was promoted to associate professor in 2010. His activities range from modeling effects of the human visual system, in particular the influence of coding, transmission, and display artifacts in 2D and 3D to measuring and quantifying visual discomfort and visual fatigue on 3D displays using psychometric and medical measurements. He currently co-chairs the VQEG "Joint Effort Group Hybrid" activities.