Randomised Pair Comparison - An Economic and Robust Method for Audiovisual Quality Assessment

Alexander Eichhorn[1], Pengpeng Ni[1,2], Ragnhild Eg[1]
[1]Simula Research Laboratory, Norway
[2]Department of Informatics, University of Oslo, Norway
{echa, pengpeng, rage}@simula.no

ABSTRACT

Subjective quality perception studies with human observers are essential for multimedia system design. Such studies are known to be expensive and difficult to administer. They require time, a detailed knowledge of experimental designs and a level of control which can often only be achieved in a laboratory setting. Hence, only very few researchers consider running subjective studies at all.

In this paper we present Randomised Pair Comparison (R/PC), an easy-to-use, flexible, economic and robust extension to conventional pair comparison methods. R/PC uses random sampling to select a unique and small subset of pairs for each assessor, thus separating session duration from the experimental design. With R/PC an experimenter can freely define the duration of sessions and balance between costs and accuracy of an experiment.

On a realistic example study we show that R/PC is able to create stable results with an accuracy close to full factorial designs, yet much lower costs. We also provide initial evidence that R/PC can avoid unpleasant fatigue and learning effects which are common in long experiment sessions.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Evaluation

General Terms

Human Factors, Measurement, Experimentation

Keywords

Quality assessment, Pair comparison, Experiment Design

1. INTRODUCTION

Audiovisual quality assessment fundamentally relies on subjective methods to capture the perceived quality experience of human observers. Subjective assessment in general is useful for measuring end-user acceptance, comparing alternative algorithms and finding optimal designs or configurations. Pair comparison is a particularly prominent assessment method because it involves a simple cognitive task, comparing two stimuli in a pair against each other. Results obtained with pair comparison tests are robust and known to closely reflect perceived sensations on a psychological scale [12].

However, the main drawback of pair comparison is that the number of pairs grows exponentially with the number of factors and factor levels under investigation. Audiovisual quality studies are known to contain a large number of factors and levels. Full factorial experiment designs that cover all possible combinations of influential factors at all levels are impractical. For example, the study of a scalable video encoder may require investigation of effects on multiple scaling dimensions at multiple scaling magnitudes on different content types and different display devices. Another example is a comparison of alternative error protection schemes under different loss patterns and loss rates, potentially generating a variety of decoding artifacts and distortions which may have to be considered separately.

Even fractional factorial designs and blocking strategies [7] which systematically reduce the number of pairs by excluding some factor combinations are of limited help. To stay within time and resource limits, an experimenter has to strictly limit the number of factors, also to avoid undesirable fatigue and learning effects. Screen-based tasks are especially susceptible to fatigue effects, even for durations as short as 15 minutes [2]. In video quality assessment, assessors can easily become tired, bored and uncooperative. Their responses will therefore be increasingly unreliable, leading to greater unexplained variance. Moreover, simple cognitive tasks are quickly mastered [8], and discrimination between two visual signals improves over time [13]. It follows that repeated exposure to the same content during an experiment session (although at different quality levels) may lead to undesired training. Assessors tend to focus on salient features of audio and video clips instead of reporting their overall quality impression. This may lead to stricter than necessary interpretations of salient artifacts.

We introduce *Randomised Pair Comparison* (R/PC) as an economic extension to traditional pair comparison designs which become increasingly counterproductive for audiovisual quality studies. The novelty is that, in contrast to full factorial designs, R/PC randomly selects small subsets of pairs and thus creates a unique experiment session for each assessor. An experimenter can control the session duration regardless of the number of factor-level combinations. This allows to make more realistic assumptions about the time

NOSSDAV'10, June 2–4, 2010, Amsterdam, The Netherlands.

assessors have to spend on a study and makes it easier to use the method on assessors with different background and age. Thus we believe R/PC is useful for laboratory experiments, observed field studies and self-controlled web-based studies. R/PC can offer a level of robustness close to traditional experiment designs while effectively avoiding fatigue and learning effects.

Randomisation in general is known to yield many benefits for statistical analysis, but the random pair selection in R/PC leads to unbalanced and missing data. Without balance, common statistical tools like ANOVA or GLMs become unstable. That's why the data analysis for R/PC has to either sacrifice some of the quality in the obtained data (e.g. ignore within-subject variability) or use computationally more expensive statistics. We will discuss some alternative ways for data analysis and we will show that it is still possible to find significant main effects.

In the remainder of this paper we first discuss related work in section 2 before we present our R/PC method in depth in section 3. Section 4 provides a first validation of R/PC which is based on an example study we conducted in order to compare a full factorial design against R/PC with the same video material, presentation settings and report scales. Finally, section 5 concludes the paper.

2. RELATED WORK

International recommendation such as ITU BT.500-11 [5], ITU-T P.910 [4], provide instructions on how to perform different types of subjective tests for the assessment of video quality in a controlled laboratory environment. The recommended test methods can be classified as Double Stimulus (DS), Single Stimulus (SS) or Pair Comparison (PC) method. The standard recommendations focus on common aspects of subjective assessment such as viewing conditions, measurement scales and basic statistics for data analysis, while stimuli selection and experiment organisation are left to the experimenter.

In DS methods, assessors are asked to rate the video quality in relation to an explicit reference. In contrast, SS and PC methods do not use explicit references. In SS methods, assessors only see and rate the quality of a single video with an arbitrary length. In the PC method, a pair of clips containing the same content in two different impairment versions is presented and the assessor provides a preference for one version in each pair. The rating procedure of PC method is simpler than that of DS and SS methods and the comparative judgement can be easily verified by examining the transitivity of the ratings.

A comparison of the DS and SS method in [10] shows that the SS method can generate quality estimates comparable to DS methods, but humans consider only the last 9 to 15 seconds of video when forming their quality estimate. While DS and SS methods are mainly used to test the overall quality of a video system, PC methods are well-suited for inspecting the agreement between different users [3].

Pair comparison is widely used in various domains. One example are online voting systems [1] for crowd-sourcing quality assessment tasks. The test design exerts loose control of stimuli presentation only. Assessors are allowed to skip between clips in a pair and they can also decide when to vote. This design is limited to test sequences with constant quality, which restricts its capability of evaluating quality fluctuation within sequences. Our R/PC method is also a variant of the PC test design as defined by ITU-T P.910 [4]. We partially follow standard recommendations and restrict the length of a test sequence to 8 to 10 seconds with the consideration of human memory effects. We also let the experimenter freely select content and quality patterns. One difference to standards is that we do not force assessors to vote in a given time. Instead we measure the timing of responses as well.

The experimental design of R/PC is closely related to designs commonly used in psychological, sociological and biological studies. In particular, completely randomised factorial designs and split-plot factorial designs [7] are closest to R/PC . Such designs are economic in a sense that they require the optimal number of factor combinations and responses to find desired main and interaction effects. They mainly assure that data is balanced so that common statistical assumptions are met. The main difference of R/PC is that our design creates unbalanced data due to random pair selection and early drop-outs and that R/PC allows to choose an arbitrarily small number of pairs per session which is independent of factorial combinations.

3. R/PC METHOD DESIGN

We designed the Randomised Pair Comparison Method with realistic expectations about the time assessors are willing to spend in a study and practical assumptions about the ability of experimenters to control environmental and content-related factors. R/PC is robust and easy to use in laboratory and field studies and is even suitable for web-based self-controlled studies.

Session duration is separated from factorial complexity of an experiment and an experimenter can balance between experiment costs and data accuracy. A session can have an arbitrary duration (down to a single pair) and assessors can quit their session anytime, e.g. when they get distracted by phone calls or have to leave a bus or train.

In contrast to traditional full factorial designs R/PC does not collect a full data sample for all pairs from every assessor. The randomisation procedure in R/PC guarantees that all pairs get eventually voted for, that an experiment session will be unique for every assessor and that all required reference pairs are contained in a session.

The overall costs of R/PC are typically lower than that for comparable full factorial designs. R/PC achieves this by shifting the resource consumption (time and number of assessors, resp. number of responses) to software-based randomisation and a computationally more expensive data analysis. Main effects will be visible with a minimum number of assessors, but with too few responses interaction effects may remain undiscovered. More assessors may increase reliability and the chance to find interaction effects. In total, R/PC may require more individual assessors to achieve significant results, but each assessor has to spend less time.

In the remainder of this section we present the general design of our method and recommendations for applying it to a particular problem. We will also discuss some implications on scales and statistical data analysis.

3.1 Presentation

Similar to conventional Pair Comparison methods [4], audiovisual stimuli are presented as pairs of clips. The duration of each clip should not be longer than 10s, but it can be adjusted to the displayed content and purpose of the

Figure 1: Presentation pattern for a single clip pair.

study. Each clip in a pair is introduced by a 2 second long announcement of the clip name and the letter A or B, displayed as a 50% grey image with black text. This results in the time pattern as shown in figure 1.

After each pair is presented, an assessor is expected to enter a response about the preference on one of the scales defined below. The time to enter the response is recorded and used for later analysis. The session continues with the next pair immediately after the response has been entered.

Because clips in each pair are presented sequentially an assessor's response may be influenced by order which may lead to a systematic bias. We compensate for that by displaying both possible clip orders within a pair (see section 3.3) and randomising the order of pair presentation.

3.2 Factorial Designs

Any two clips in a pair may differ in one or multiple factors, as defined by the experimenter. We call such pairs *contrast pairs*. They are used for actual exploration and hypothesis testing. An experimenter may, for example, base his research hypothesis on assumptions about the visibility and effect size of contrasting factors.

An experimenter should first identify factors which will be controlled in the study and the number of levels for each factor. Factors can be discrete or continuous and the number of levels may differ between factors. Uncontrolled factors such as, for example, an assessor's age and occupation or the time and location of an experiment session should at least be measured if they are regarded as relevant.

Pairs are created for all factor/level combinations. To reduce the overall number of pairs, it is worthwhile to identify factors for blocking. Blocking restricts combinations within each level of the selected factor. The blocking factor can be used in the later analysis to investigate differences between blocks. For example, to explore the effect of video clip content it is not necessary to assess all potential combinations of clips in separate pairs. Instead, only pairs made of the same clip can be assessed.

Blocking by content type also isolates a systematic effect that is introduced by the content itself. Because content is one of the major sources for unexplained variance in audiovisual quality assessment it is desirable to understand and limit its impact. Using a small set of standard sequences would be statistically reasonable, but practically it is undesirable due to the limited relevance findings would have.

3.3 Reference Conditions

Reference conditions are introduced to find unreliable assessors and outliers in the data, but also to understand the perceptual limits of individual assessors.

R/PC uses two types of reference conditions, (1) *equal reference pairs* that contain the same clip at the same quality level twice, and (2) *matched contrast pairs* that just differ in the presentation order of the contained clips, but are regular contrast pairs otherwise. For every contrast pair there should be a matched contrast pair of opposite order. Likewise, for every factor/level combination there should be an equal reference pair, but equal reference pairs don't need a matched counterpart. Reference conditions should be randomly distributed and hidden inbetween other pairs when presented to an assessor to avoid their detection.

Although reference pairs increase the duration of an experiment session, they are necessary to assure data quality and detect systematic problems during data analysis.

3.4 Random Pair Selection and Ordering

In contrast to full factorial designs, where each assessor has to respond to all combinations, the R/PC method creates a *unique random subset of pairs* for each assessor and then randomises the presentation order for pairs.

The procedure ensures that (1) the ratio of contrast to reference pairs is equal in each subset and is also equal to the ratio in a full design, (2) each selected contrast pair is contained in both possible presentation orders (both matched contrast pairs AB and BA are present), (3) equal reference pairs correspond to selected contrast pairs (there is no reference clip version which does not occur in a contrast pair as well), and (4) equal reference pairs are contained only once.

First, an experimenter must pre-determine the subset size. Assuming all pairs are of equal duration, the size s is calculated as $s = d_s/d_p$, where d_s is the total session duration as defined by the experimentor and d_p is the duration of a pair including an estimated time for voting. The subset size should be equal for all assessors in an experiment.

Then contrast pairs are randomly chosen and their matched contrast pair counterparts are added. Assuming there are in total p contrast pairs (in AB order), the same amount of matched contrast pairs (in BA order), and e equal reference pairs, then $s(p/(2p+e))$ (matched) contrast pairs and $s(e/(2p+e))$ equal reference pairs are selected. This ensures the same ratio between contrast pairs and equal reference pairs as in a full factorial design. Note that equal reference pairs have to match the selection of contrast pairs so that no reference pair exists which does not occur otherwise.

Randomisation in general has many benefits for statistical analysis. The randomised pair selection in R/PC , however, leads to an unbalanced data matrix, where (i) the total number of responses per item may be unbalanced, (ii) the number of responses per item can be zero, (iii) each assessor votes for a small percentage of pairs only and thus many empty within-subject cells exist, (iv) the number of responses per assessor may be unbalanced when the assessor quits a session before it ends. Statistical tools for data analysis have to be robust against these uncommon features or a pre-processing step will be required to create the desired features for statistical tests an experimenter would like to employ.

Because R/PC presents a small amount of pairs per session only, it is expected that the number of assessors may be higher than for full factorial designs to achieve stable statistical results. Overall, each assessor spends less time in a session and less responses are collected which may lead to confounding of estimated main effects with factor interactions. Due to the complete randomisation in R/PC the confounding effects are limited. This is because each pair that would be used in a full factorial design will eventually contribute in R/PC as well. For confounding effects to become negligible and results to become stable a sufficient number of assessors is required. A minimal number may dif-

fer between studies and we are investigating the influencing factors further.

3.5 Assessment Task and Reporting Scales

The main task for assessors is to *compare the overall quality* of the presented pairs and to report their preference using a self-report scale. At the beginning of a session a brief introduction about the purpose of the study and the scale which is used may be given. Assessors should be reminded to pay close attention, but an experimenter should avoid specific restrictions or hints which might guide an assessor's focus. A training session is not required.

For voting we suggest not to impose time limitation which would force a decision. Instead we propose to measure the time it takes an assessor to respond and use this in the data analysis. Assessors should report their preference on one of the following comparison scales:

- a *binary preference scale* which allows to express either a preference for clip A or a preference for clip B (forced preference selection results in random votes for equal reference pairs and close to just noticeable differences, JND)
- a *3-point Likert scale* which contains a neutral element in addition to a preference for A or B (promotes indecisiveness, but allows to detect JND thresholds)
- a *4-point Likert scale* which provides items for weak and strong preference for either clip, but lacks a neutral element (has a higher resolution than the binary scale and forces preference selection)
- a *5-point Likert scale* which contains a neutral element as well as items to express weak and strong preference (high resolution, may promote indecisiveness, but allows JND detection)

From a statistical perspective the data obtained with such scales is binomial or ordinal at most. Although some psychometrics researchers argue that data on a 5-point Likert scale can be considered as interval-type because this scale measures psychological units (perceptual differences in our case), we advise to apply non-parametric statistics.

3.6 Data Analysis

Proper data analysis for R/PC is currently a work in progress. In this paper we briefly discuss some implications of our method design and options on how to deal with unbalanced data. For an in-depth discussion on non-parametric procedures see [11].

The nature of the binomial and Likert scales suggests non-parametric statistics for data analysis. Whether the reason for a study is hypothesis testing or exploratory analysis, care should be exercised when responses from assessors are extremely skewed or inconsistent. Even though non-parametric statistics are robust against outlier values, because they rely on medians instead of means, unreliable assessors should be completely removed. Assessor validity can be verified based on reference pairs, in particular, by comparing matched contrast pairs for inconsistent responses.

Useful non-parametric statistical tools are Binomial tests or χ^2 tests for majority analysis on counts (to check whether a majority of preference ratings for one factor-level is significant). As non-parametric counterparts to t-tests and ANOVA, the Mann-Whitney U, Kruskal-Wallis and Friedman tests exist. Rank-order analysis for finding total transitive preference orders between pairs are provided by the zeta method [3] and Thurstone's law of comparative judgement [12]. For more thorough investigations on the underlying structure of the data and to find a linear combination of variables that explains how factors contribute to effects an exploratory factor analysis using generalised linear models or logit models should be considered.

Although we obtain repeated measures, for analysis we regard response data as independent (we drop subject-specific data). Our rationale is that although within-subject variability may be useful to explain effects, we have to sacrifice some of the data quality to compensate for the unbalanced design, in particular the large number of empty cells. Hence, the repeated measures design is just for convenience' sake of obtaining more data from each assessor. We could as well just collect a single response per assessor, which may be more adequate for web-based self-studies.

Ignoring subject-specific data for audiovisual experiments is reasonable because we are interested in general observations which are independent from individual assessors abilities, expectations or perceptual limits. Conclusions from audiovisual quality experiments are expected to apply to a broad spectrum of end-users. Hence we regard all assessors as relatively homogeneous. If a specific target group is of interest in a study, then assessors should be representative for that group.

4. METHOD VALIDATION

To validate the usability and reliability of our R/PC method we performed a simple quality assessment study. Purpose of the study was to obtain two data sets, one with a conventional pair comparison method based on a full factorial experiment design (F/PC) and a second data set with R/PC . We first explain the design of our example study and present an initial analysis of our findings afterwards.

4.1 Example Study

As a simple example of an audiovisual study, we examined the visibility of different video quality reductions in relation to already existing impairments. Our quality impairments originate in a loss of image fidelity between five different operation points, which have been created using different fixed quantisation parameters (QP) for encoding.

In this study, we focus on three factors that are assumed to have main effects, namely the original quality level, the amplitude of a quality change and the content type. These factors can mutually influence the user's perception. For example, the same amplitude in quality changes may be perceived differently depending on the original quality and the direction of the quality change. Interactions may also exist between the change amplitude and the content.

To test different kinds of content with varying detail and motion, we selected six 8 second long clips (200 frames) without scene cut from different genres (see table 1). All clips were downscaled and eventually cropped from their original resolution to 480x320 pixel in order to fit the screen size of our display devices. We used x264 to encode the original clip in constant quantiser mode so that the same amount of signal distortion was added over all frames in a test sequence. Since the visibility of quality impairments is not linearly related to the size of QP, we selected a set of five QPs with

logarithmically distributed values. In a pilot study, the corresponding QPs (10, 25, 34, 38, and 41) have been verified to yield perceptual differences. With five quality levels we can create $\binom{5}{2} = 10$ unique combinations of contrast pairs that have quality change amplitudes between 1 to 4 and five equal reference pairs per content type. In total, we created 120 contrast pairs in both orders and 30 (25%) equal reference pairs. In our example study, a F/PC test session lasted for 60 min while a R/PC test session lasted for only 12 min.

The clip pairs were displayed to assessors on an iPod touch which has a 3.5-inch wide-screen display and 480x320 pixel resolution at 163 pixels per inch. Display and voting were performed on the same device using a custom quality assessment application. The experiment was carried on in a test room at Oslo university. Overall, 49 participants (45% female) at an age between 19 and 39 performed the experiment. Among the participants, 34 people (50% female) were paid assessors who performed both the F/PC test and R/PC test while 15 participants (40% female) are volunteers who performed only the R/PC test. Half of the participants who did both tests, performed the R/PC method first, while the other half did the F/PC test first. During all test sessions the participants were free to choose a comfortable watching position and to adjust the watching distance. They were also free to decide when and for how long they needed a break.

Genre	Content	Detail	Motion
Animation	BigBuckBunny	3.65	1.83
Cartoon	South Park	2.75	0.90
Docu	Earth 2007	3.64	1.61
Movie	Dunkler See	1.85	0.58
News	BBC News	2.92	0.69
Sports	Free Ride	3.32	1.90

Table 1: Sequences used in the experiments. Detail is the average of MPEG-7 edge histogram values over all frames [9] and Motion is the MPEG-7 Motion Activity [6], i.e., the standard deviation of all motion vector magnitudes.

a) Full factorial Pair Comparison

Unique Subjects:	34
Unique Pairs:	150
Unique Responses:	5100
Resp/Subj (min/mean/max):	150 / 150 / 150
Resp/Pair (min/mean/max):	34 / 34 / 34

b) Randomised Pair Comparison

Unique Subjects:	49
Unique Pairs:	150
Unique Responses:	1470
Resp/Subj (min/mean/max):	30 / 30 / 30
Resp/Pair (min/mean/max):	4 / 9.8 / 19

Table 2: Grand totals and statistics for the two data sets in our example study.

4.2 Fatigue and Learning Effects

In order to assess learning and fatigue effects in the 60 minutes long F/PC test, we created a measure of accuracy by coding preference responses as correct, neutral or incorrect. For equal reference pairs, neutral and correct responses were equivalent. Learning and fatigue effects were explored separately, with both reference and contrast pairs.

We expected fatigue effects to become evident already after the first ten minutes, so we divided an experiment session into five equal duration groups, each consisting of 12 min (30 pairs). We also expected the impact of fatigue to be more prominent for video pairs with a fairly visible quality difference, hence video contrasts with one level quality difference were excluded from the analysis. A Pearson chi-square was run for all remaining contrast pairs, but no effects were uncovered ($\chi^2(8)$=11.18, ns). Due to the binary nature of the response categories for the equal reference pairs, a Cochran-Mantel-Haenszel chi-square was used for this analysis. It revealed that response accuracy was conditional of duration (χ^2=4.60(1), p>.05), thus indicating that neutral and incorrect responses varied across one or more duration groups. Binomial tests were applied to further explore this relationship. We found that neutral responses were more frequent in the final compared to the first duration group (S=43, p>.05), otherwise there were no differences in neutral or incorrect responses.

Learning effects were expected to be most relevant where quality differences were hard to spot; hence the analysis included only contrast pairs with one level quality difference. These were grouped according to content repetition, so that five content repetition groups were created based on how many times a video pair with the same content had previously been presented. However, Pearson chi-square revealed no variation according to the number of repetitions ($\chi^2(8)$=5.21, ns). Neither did Cochran-Mantel-Haenszel chi-square reveal any differences for the equal reference pairs (χ^2=3.76(1), ns).

The significant difference in neutral responses for equal reference pairs could indicate that assessors are suffering from fatigue. With more neutral responses towards the end of the experiment, a plausible proposal might be that they become more careless with responses when tired. However, such an effect should perhaps present itself earlier. Another plausible proposal is that the difference is not due to fatigue, but to learning. Although completely randomised, on average an assessor would have observed the presented video contents several times when embarking on the final 30 pairs. Thus the increase in neutral responses may represent an improved ability to detect the absence in difference between equal reference pairs. The current analyses do not provide sufficient data to form a conclusion, but they do suggest that responses change during the course of a F/PC test.

4.3 Reliability

Based on the two data sets we gathered using F/PC and R/PC we did some initial comparative analysis. We are interested whether an investigator using different statistical procedures on either data set would be able to find similar results. Hence, we first looked at the correlation between both data sets and second we tried to fit a linear model to the data in order to find factors which influence main effects.

For the correlation analysis we first calculated the arithmetic mean and the median of all responses per pair. Then we calculated Pearson, Spearman and Kendall correlation coefficients as displayed in table 3. All coefficients were significant below the p<0.01% level.

Metric	CC	SROCC	$\mathcal{T}$
mean	0.974	0.970	0.857
median	0.961	0.965	0.931

Table 3: Correlation between R/PC and F/PC data sets. CC - Pearson Product-Moment Correlation Coefficient, SROCC - Spearman Rank-Order Correlation Coefficient, $\mathcal{T}$ - Kendall's Rank Correlation Coefficient.

Despite the fact that responses in the R/PC data set are very unbalanced (min = 4, max = 19 responses for some pairs, see table 2) and that the total unique responses collected with our R/PC method are only <1/3 of the total F/PC responses, there is still a very strong correlation between both data sets. This supports the assumption that random pair selection may become a useful and robust alternative to full factorial designs for audiovisual quality assessment. However, further analysis is needed to find the minimal number of required assessors and responses.

In our second validation step we compared the results of fitting a generalised linear model (GLM) to both data sets. We used a binomial distribution with a logit link function and modelled the main effects original quality level (Q-max), amplitude of quality change (Q-diff) and content type (content), but no interaction effects. As table 4 shows, all main effects are significant, although the significance is lower in the R/PC case which was to be expected. Again, it is plausible to argue for a sufficient reliability of the R/PC method.

	Factor	Df	Dev	R.Df	R.Dev	P(>χ^2)
f/pc	Q-diff	4	718.43	5095	3505.5	< 2.2e-16
	Q-max	4	54.31	5091	3451.2	4.525e-11
	content	5	34.39	5086	3416.8	1.995e-06
r/pc	Q-diff	4	236.18	1465	1085.2	< 2.2e-16
	Q-max	4	20.48	1461	1064.8	0.0004007
	content	5	16.94	1456	1047.8	0.0046084

Table 4: Deviance analysis for a simple GLM considering main factor effects.

5. CONCLUSION

In multimedia system design the search for optimal solutions is often exploratory, necessitating large numbers of experimental factors which makes full-factorial studies excessively long and draining. In the current paper, we have presented Random Pair Comparison as a practical and economic method for exploratory quality assessment. R/PC provides the possibility of investigating numerous factors, while maintaining the freedom of both experimenters and assessors. We provided first evidence that R/PC is a robust assessment method suitable for finding main effects at reasonable costs.

However, R/PC comes at the expense of higher computational costs for randomisation and data analysis. Violations of normality, uneven response distributions and greater error variance complicate the statistical analysis. In future studies, we aim to further explore non-parametric tests and establish a robust statistical procedure for analysing data generated by R/PC .

One important question remains unanswered so far: what is the minimal number of assessors and responses required to achieve stable results and how much can R/PC really reduce the costs of a study. An answer is not simple since statistical results depend on many factors. In our example study we were able to find significant results with only 29% of responses and costs. A thorough statistical analysis and more data from studies using R/PC will provide further insights.

Acknowledgement

The authors would like to thank Beata Dopierala and Stian Friberg for their help in organising the example study as well as the numerous volunteer participants. This work was sponsored by the Norwegian research council under the Perceval project.

6. REFERENCES

[1] Chen, K.-T., Wu, C.-C., Chang, Y.-C., and Lei, C.-L. A crowdsourceable QoE Evaluation Framework for Multimedia Content. In *MM '09: Proc. of the ACM Intl. Conference on Multimedia* (New York, NY, USA, 2009), ACM, pp. 491–500.

[2] Chi, C., and Lin, F. A Comparison of Seven Visual Fatigue Assessment Techniques in Three Data-Acquisition VDT Tasks. *Human Factors 40*, 4 (1998), 577–590.

[3] International Telecommunications Union. *Report 1082-1. Studies toward the unification of picture assessment methodology*, 1990.

[4] International Telecommunications Union. *ITU-T P.910. Subjective video quality assessment methods for multimedia applications*, 1999.

[5] International Telecommunications Union - Radiocommunication sector. *ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television picture*, 2002.

[6] Jeannin, S., and Divakaran, A. MPEG-7 Visual Motion Descriptors. *IEEE Trans. on Circuits and Systems for Video Technology 11*, 6 (Jun 2001), 720–724.

[7] Kirk, R. E. *Experimental Design: Procedures for Behavioral Sciences*, 2 ed. Wadsworth Publishing, 1982.

[8] Miller, J., Ruthig, J., Bradley, A., Wise, R., H.A., P., and J.M., E. Learning Effects in the Block Design Task: A Stimulus Parameter-Based Approach. *Psychological Assessment 21*, 4 (2009), 570–577.

[9] Park, D. K., Jeon, Y. S., and Won, C. S. Efficient use of Local Edge Histogram Descriptor. In *Proc. of ACM workshops on Multimedia* (2000), pp. 51–54.

[10] Pinson, M., and S.Wolf. Comparing subjective video quality testing methodologies. In *SPIE'03* (2003).

[11] Sheskin, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures*, 2 ed. Chapman & Hall, 2000.

[12] Thurstone, L. L. A law of comparative judgment. *Psychological Review 101*, 2 (1994), 266–70.

[13] Yu, C., Klein, S. A., and Levi, D. M. Perceptual learning in contrast discrimination and the (minimal) role of context. *J. Vis. 4*, 3 (3 2004), 169–182.