OPEN ACCESS Freely available online
R. F. Fela, N. Zacharov, and S. Forchhammer, “Comparison of Full Factorial and Optimal Experimental Design for Perceptual Evaluation of Audiovisual Quality”
J. Audio Eng. Soc., vol. 71, no. 1/2, pp. 4–19, (2023 January/February).
DOI: https://doi.org/10.17743/jaes.2022.0063

Comparison of Full Factorial and Optimal Experimental Design for Perceptual Evaluation of Audiovisual Quality

RANDY FRANS FELA,[1,2,†] *AES Student Member*, **NICK ZACHAROV,**[3,‡] *AES Fellow* **AND**
(ranf@fotonik.dtu.dk) (nzacharov@fb.com)

SØREN FORCHHAMMER[2,*]
(sofo@dtu.dk)

[1]*SenseLab, FORCE Technology, Hørsholm, Denmark*
[2]*Department of Electrical and Photonics Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark*
[3]*Meta Reality Labs., Paris, France*

Perceptual evaluation of immersive audiovisual quality is often very labor-intensive and costly because numerous factors and factor levels are included in the experimental design. Therefore, the present study aims to reduce the required experimental effort by investigating the effectiveness of optimal experimental design (OED) compared to classical full factorial design (FFD) in the study using compressed omnidirectional video and ambisonic audio as examples. An FFD experiment was conducted and the results were used to simulate 12 OEDs consisting of D-optimal and I-optimal designs varying with replication and additional data points. The fraction of design space plot and the effect test based on the ordinary least-squares model were evaluated, and four OEDs were selected for a series of laboratory experiments. After demonstrating an insignificant difference between the simulation and experimental data, this study also showed that the differences in model performance between the experimental OEDs and FFD were insignificant, except for some interacting factors in the effect test. Finally, the performance of the I-optimal design with replicated points was shown to outperform that of the other designs. The results presented in this study open new possibilities for assessing perceptual quality in a much more efficient way.

0 INTRODUCTION

Perceptual evaluation has attracted much attention in the multimedia industry in recent decades and is mainly used for audio, video, and audiovisual systems [1–4]. It usually involves a set of studied system variables with a list of variable levels in terms of their effects on the evaluated perceptual attributes (e.g., quality, preference, intrinsic attributes, etc.). The purpose of conducting perceptual evaluation can range from simply understanding the relationship between variable inputs and outputs, or within variable inputs, to product quality development involving model optimization. In conjunction with a number of variables involved in the evaluation, careful design of experiment (DOE) is a key topic discussed by researchers and practitioners to deal with experimental constraints such as time and cost as well as psychological and physiological effects of involving human subjects.

The full factorial design (FFD)[1] is a traditional experimental design that takes into account all possible combinations of factors affecting the response variables and

[*] To whom correspondence should be addressed, e-mail: sofo@dtu.dk.

[†] This work was performed while the author was with FORCE Technology—SenseLab, and he is now employed at GN Audio A/S (Jabra).

[‡] This work was performed while the author was with FORCE Technology—SenseLab, and he is now employed at Meta Reality Labs.

[1] Please note that FFD is associated in some references with fractional factorial design, which is different from the definition here.

therefore requires the most testing. The completeness of the FFD allows for the examination of main and interaction effects and testing of model curvature, resulting in a more trustworthy empirical model, but at a higher cost and longer duration of experimental work required. Additionally, when a system has additional features beyond the traditional multimedia system that must be evaluated (e.g., haptic and spatial features in extended reality application) and that simultaneously multiply a number of factor levels, FFD becomes less practical.

Traditionally, two common approaches of experimental designs are within-subject and between-subject designs [5]. When evaluating user-perceived multimedia quality (audio, video, and audio-video), FFD within-subject experiments are the predominant approach. This is partly due to the relative ease with which audio stimuli can be acquired/simulated and subsequently presented, compared, and evaluated in real-time [1]. However, the number of stimuli is often very large, which sometimes is in conflict with the time constraint, and bias could arise from the order of stimulus presentation. To tackle this issue, the experimenter used to employ a DOE that has a reduction in the number of design points/testing conditions (e.g., fractional factorial design) [6] and a balanced presentation order across subjects and between subjects (e.g., Balanced Latin Square design) [7, 8].

In contrast, a between-subjects design requires more subjects to evaluate each condition in order to match the statistical power of a within-subjects design, which in turn demands more resources and costs. Furthermore, there is a chance that the collected responses differ in important ways between conditions because different participants provide data only for certain conditions. Several studies have reported the use of a within-subject or between-subject design in the perceptual evaluation of sound [9, 10], video [11, 12], and audiovisual [13, 14]. Recent works have also reported the use of mixed methods for a single experiment, as documented, for example, in [15–18].

Many attempts have been made to find an alternative DOE that can reduce the number of conditions under test while providing sufficient information for the analysis of effect terms. These efforts have been driven primarily by sensory scientists in the food industry [19] but have great potential for application in the multimedia industry [2, 20, 21]. There are two subcategories of DOE techniques. The first involves optimizing a set of experimental conditions to identify a polynomial Response Surface Methodology (RSM). Box-Behnken Design (BBD) and Central Composite Design (CCD) are two experimental designs that belong to the RSM family. The second method is to use optimal experimental designs (OEDs) to create tailored conditions in the design region based on the optimization criteria applied in the model structure to estimate the values of the model parameters.

DOE techniques beyond FFD were originally developed out of necessity, due to the nature of experimental conditions outside the audiovisual domain. For example, when testing in real climatic conditions, agriculture, or complex processes, it may not be possible to test all conditions, so sparse selection of the entire design space may be the only possible approach. This is far from the situation faced in the audiovisual field, where in principle, FFD experiments can be created, even if they are very large. It is therefore of great interest to utilize these techniques to facilitate large-scale audiovisual experiments. Readers are referred to Montgomery [22], Myers et al. [23], Kiefer and Wolfowitz [24], and Pukelsheim [25] for a more detailed discussion of DOE, RSM, and OED, respectively.

Although the evidence is limited, previous studies have reported the use of DOE beyond FFD for perceptual evaluation. RSM was applied by Lorho [20] when investigating listeners' preferences for headphone frequency response equalization played back for music and speech content. In all experiments, a CCD was chosen to obtain a quadratic interaction between center frequency and amplitude in two prominent peaks (3 and 11 kHz).

A shortcoming of using RSM is that the experimental points are more regularly distributed over the design space. While BBD examines only the points that do not have extreme factor combinations, CCD considers only the boundary regions. Another limitation is that these designs can provide a maximum of only three and five levels for BBD and CCD, respectively, for each factor. To compensate for these shortcomings, statistical OED offers a number of advantages, including (i) efficiently filling an irregularly shaped design space, (ii) minimizing the number of runs to what is needed to fit the assumed polynomial model, (iii) accommodating unusual requirements in terms of number of blocks or number of runs per block, and (iv) being able to handle a combination of factor types such as continuous, discrete, categorical, and mixed.

Fela et al. [21] combined the use of FFD and OED in studying the interaction between perceived audio, video, and audiovisual quality in 360 videos with ambisonic played back via head-mounted display (HMD) and multichannel loudspeakers. Whereas FFD was used for unimodal evaluation (audio or video), OED was designed with the D-optimal criterion and coordinate exchange algorithm and used for audiovisual evaluation. Although the perceptual quality models proposed in the study concluded that the combined DOE was able to make accurate predictions for the audiovisual model using both conventional and machine learning approaches [21, 26], there was no validation of the extent to which the D-optimal design differed from the full-factorial design in terms of building predictive models.

Another recent approach to selecting experimental design points is the use of active learning sampling strategies, often referred to as "active sampling." Conceptually, active learning is a similar approach with OED but uses a machine learning technique in which a learning algorithm searches specifically for the data that is most informative to the model, rather than being trained on the entire data set. The application of active learning has been dedicated by previous researchers for active generation of stimulus representations in perceptual evaluation of audio [27], video [28, 29], and audiovisual [30]. However, this method is more developed for pairwise comparison experiments, and its use for a wide application of different evaluation methods is

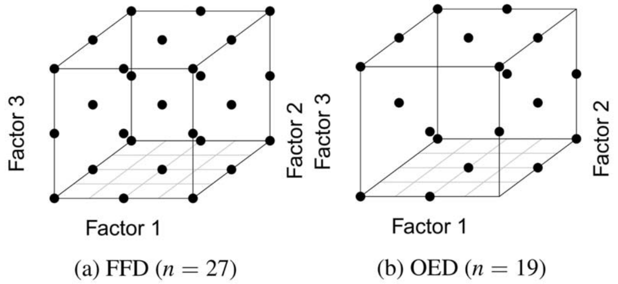


Fig. 1. Illustration of data points located in a design space generated for (a) FFD and (b) OED (n = number of design points).

considered limited. This is because it is non-trivial to find the right query strategy and loop size in active sampling.

Following the above literature, this study set out to investigate the use of OED in audiovisual perceptual evaluation studies. The main research question addressed in this study is how powerful OEDs are, in particular the D-optimal and I-optimal designs with a varying number of data points, in detecting the effects of independent variables on the dependent variable in evaluating the perceptual quality of immersive multimedia content and determining to what extent the results of OEDs are beneficial or detrimental to the evaluation compared to FFD. By introducing an alternative experimental design that reduces the number of trials while retaining statistical properties, this study aimed to provide some important insights into the field of perceptual evaluation in the consumer electronics and broadcasting industry.

The remainder of the paper proceeds as follows: SEC. 1 begins by laying out the theoretical dimensions of experimental design, full factorial experimental design, and randomized OED along with the optimality criterion. SEC. 2 focuses on the materials and methodology used for the experiments conducted in this study. The analysis of the results obtained from the experiments is described in SEC. 3 and discussed in SEC. 4. SEC. 5 portrays the conclusion drawn from the experiments, followed by SEC. 6, which presents the limitations and research outlook of the present study.

1 THEORETICAL BACKGROUND: DOE

The origins of DOEs and OEDs date back to the late 1800s, with major contributions by Peirce [31, 32]. In the early 1900s, Ronald Fisher published two valuable texts on the subjects, including [33] and [34], laying the foundation for today's DOE methods. Nowadays, one can take full advantage of the power of DOE and OED because of the available computing power that allows for easy design, simulation, and optimization—something that 20 years ago was not viable. A conceptual comparison of the generated experimental design in a design space between FFD and OED for three factors with three levels each is shown in Fig. 1, illustrating the potential efficiency gains available through OEDs.[2]

1.1 FFD

When using the FFD, all possible conditions in the design space being evaluated are available, as shown in Fig. 1(a), and it is usually not necessary to interpolate between the levels of each test condition. In perceptual multimedia evaluation, such designs often have high statistical power, allowing the experimenter to model and explore main, two-way and sometimes up to three-way interaction data collected using such FFD. FFD may be feasible if only a few factor levels are considered in the design or if high-throughput experimental facilities are available. Nevertheless, covering a full design space to obtain useful information is often not necessary because the lower-order effects tend to be dominant in most of the case. To deal with this situation, OEDs were introduced to find the numerically optimized experimental design with respect to optimality criterion.

To understand the experimental design space in DOE and its relation to the regression model, suppose that the experimental design matrix $\mathbf{D}$ represents all experimental settings n of each experimental factor predictor m, where x_{nm} denotes the observation of the nth setting of the mth factor predictor.

$$\mathbf{D} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}. \tag{1}$$

Suppose that the goal of a perceptual evaluation is to model the dependence of a particular response $\mathbf{Y}$ given by test assessors for each trial run. Adding matrix $\mathbf{D}$ to linear regression model yields

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \tag{2}$$

where design matrix $\mathbf{D}$ now has $m + 1$ columns with the elements of the first column are all ones. This matrix is called model matrix $\mathbf{X}$. In Eq. (2), the nth perceptual responses of Y_n can be calculated by the model matrix $\mathbf{X}$, a vector of unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ and a vector of random errors $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$.

1.2 OED

As illustrated in Fig. 1(b), a sparse sampling of the design space is considered compared to the FFD. The OED leverages the concept of interpolation between the presented conditions to estimate the untested factor levels based on mathematical curve fitting principles. Conceptually, OED attempts to reconstruct the results of FFD without testing all conditions. Whereas the FFD is uniquely defined by the combination of all factors and their interactions, there are a very large number of potential OEDs to choose from, some of which sparsely cover the design space. A word of caution is in order at this point because not all OEDs are equal, and the experimenter should be aware of the tradeoffs involved in using OEDs. Compared with FFD, OEDs provide fewer conditions for assessors and as a result will have less statistical power.

[2] This design was constructed using JMP Pro 15.

Additionally, other techniques can be used to create an efficient OED. This may result in a design that is not fully balanced [e.g., the top layer in Fig. 1(b)] or certain factors that are confounded or aliased. It is up to the experimenter to define which levels of factors and interactions are critical to the experiment and should be examined after data collection. This is usually based on prior knowledge of the field and is part of the OED optimization process. For example, in audio-video experiments, the literature often assumes that all main factors and two-way interactions must be included. In rare cases, the experimenter may wish to include specific and carefully selected three-way interactions. Ideally, the experimenter can use prior knowledge of interactions that are usually insignificant or uninteresting and can be excluded from the OED optimization process, resulting in a more efficient design. Once excluded from an OED, experiments should not attempt to analyze such effects because they are likely to be aliased. Based on a good understanding of the domain, the experimenter can design efficient and robust OEDs.

A mathematical theory of optimum designs was first proposed by Smith [35] for a series of single-factor polynomial models. The first extended ideas were presented by Kiefer [36] at a meeting of the Royal Statistical Society. In 1943, Wald [37] proposed the criterion of maximizing the determinant of the information matrix $\mathbf{I} = \mathbf{X}^T\mathbf{X}$, which later was known as the D-optimality criterion named by Kiefer and Wolfowitz [24] when it was extended for general use in regression models. Whereas most studies have focused on using the D-optimality criterion to obtain precise parameter estimates, recent years have seen increased interest in the prediction-based optimization criterion. The latter goal is consistent with the I-optimal design, which focuses on accurate predictions by minimizing the average prediction variance, as discussed in Haines [38].

Applying the ordinary least-squares (OLS) estimation to Eq. (2) produces the optimum parameters, $\widehat{\beta}$, in relation to the information matrix of $\mathbf{X}$, $(\mathbf{I} = \mathbf{X}^T\mathbf{X})$.[3]

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3}$$

Thus, the variance-covariance matrix can be calculated for optimum parameters $\widehat{\beta}$ and yields

$$\mathbf{C} = var(\widehat{\beta}) = \widehat{\sigma}^2_\varepsilon(\mathbf{X}^T\mathbf{X})^{-1}, \tag{4}$$

where $\widehat{\sigma}^2_\varepsilon$ is the approximation of mean-squared model error or an estimation of the irreducible noise in the system. Finally, the variance can be calculated in the predicted fitting function by

$$var(\widehat{y}(\mathbf{x})) = \widehat{\sigma}^2\mathbf{f}^T(\mathbf{x})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}). \tag{5}$$

Eqs. (3)−(5) are the basic equations used in optimal designs, conveying the idea that optimization can be achieved by manipulating either the determinant of the information matrix $\mathbf{I}$ or the variance-covariance matrix.

The D-optimality criterion was developed by utilizing the determinant of the information matrix $|\mathbf{X}^T\mathbf{X}|$, which corresponds to covariance of the parameter estimates, to measure the overall uncertainty [39]. It is called D-optimal design when the objective is either to maximize $|\mathbf{X}^T\mathbf{X}|$ or minimize the determinant of $(\mathbf{X}^T\mathbf{X})^{-1}$.

$$\mathbf{D}_{\text{opt}} = \arg\min_{\text{D}}\left[\det\left(\widehat{\sigma}^2_\varepsilon(\mathbf{X}^T\mathbf{X})^{-1}\right)\right]. \tag{6}$$

If the objective is to increase the accuracy of the model prediction by minimizing the average variance, I-optimal design is more appropriate and can be expressed as follows.

$$\mathbf{I}_{opt} = \frac{\int_\chi \mathbf{f}(\mathbf{x})\left((\mathbf{X}^T\mathbf{X})^{-1}\right)\mathbf{f}^T(\mathbf{x})\mathrm{dx}}{\int_\chi \mathrm{dx}}, \tag{7}$$

where $\mathbf{x}$ represents a vector of predictors in an available design space χ and $\mathbf{f}(\mathbf{x})$ is a Jacobian matrix of model parameters. Therefore, the integrand represents the transferring error of model prediction from the fitting coefficients.

Another optimality criterion is G-optimality, which looks for design points that minimize the highest predicted variance in the design space. According to [40], limiting the maximum predicted variance is associated with an increase in the predicted variance by more than 90%. Therefore, the study was restricted to D- and I-optimal designs. In addition to these optimality criteria, many other criteria that support relatively different design objectives exist and have been discussed in Atkinson et al. [41].

The following are some significant benefits of optimal design: 1) it can cover all terms (e.g., x_1x_2, quadratic x_1^2, and higher-order terms $x_2^3x_3^2$) in linear statistical models; 2) depending on the number of terms included in the model function, it allows constructing a relatively small design; 3) the sparse sampling of optimal design makes it possible to augment the data if the original design was not optimal; and 4) depending on the goal of DOE, a variety of optimality criterion can be defined. In this paper, the applicability of both I-optimality and D-optimality to audiovisual perceptual evaluation is investigated.

2 MATERIALS AND METHOD

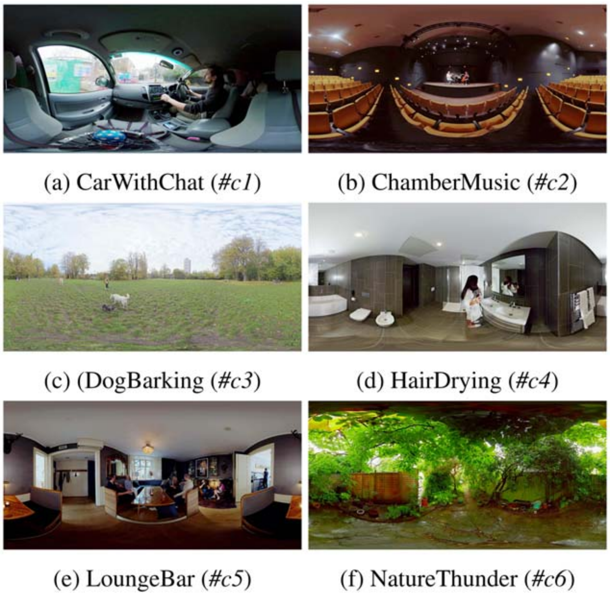
The following steps were taken to carry out the study provided in this article: (i) an FFD experiment was conducted to obtain baseline data; (ii) two existing optimality criteria with different numbers of additional and replicated points were applied to the OED so that twelve designs were run for simulation; (iii) based on the simulation performance, four OEDs were selected for the experiment; and (iv) the performance of the FFD and the selected OED techniques with respect to the model parameter estimates in the experimental measurements was discussed.

2.1 Stimuli

Six audiovisual source materials (SRCs) were selected from the higher-order ambisonic sound scene repository (HOA-SSR) database,[4] a public database consisting of 360 videos and higher-order ambisonic audio of 20-s length

[3] Some authors prefer to use $\mathbf{M}$ to denote the information matrix $\mathbf{X}$.

[4] https://bit.ly/HOA-SSR-Dataset.



(a) CarWithChat (*#c1*) (b) ChamberMusic (*#c2*)

(c) (DogBarking (*#c3*) (d) HairDrying (*#c4*)

(e) LoungeBar (*#c5*) (f) NatureThunder (*#c6*)

Fig. 2. The audiovisual HOA-SSR dataset used in the study.

each [42]. All video scenes were captured using an Insta360 Pro2, a 360 camera consisting of a spherical array of six lenses; were in raw format; and specified in YUV422 color format, 8K resolution (7680x3840), 8 bits, and 30 fps. The sound was recorded using an em32 Eigenmike microphone array that capture up to fourth-order ambisonic recording. The audio materials from the database were in Ambisonic B-format AmbiX (25 channels, 48 kHz, and 24 bits) with ambisonic channel number ordering and semi-normalized 3D normalization.

The equirectangular preview images of 360 videos used in this study are shown in Fig. 2. The selection of audiovisual stimuli was based on spatio-temporal features of video and characteristic of audio [e.g., with a content of speech (*#c1*), music (*#c2*), impulsive sound (*#c3, #c6*), white noise (*#c4, #c6*), and speech with background noise (*#c5*)].

2.1.1 Encoding and Decoding

The video SRCs were converted from raw YUV422 format to playable YUV420 format before the video encoding process. The processed video sequences were created by using libx265 (H.265/HEVC) in FFmpeg in four quantization parameters (QP; 0, 22, 28, 34) and three different video resolutions (1920x1080, 3840x1920, and 6144x3072). QP controls how much spatial detail of the image is retained in each frame of the video. Each value in QP represents the step size of the quantizer during video compression. The higher QP is, the higher the quantization step size, which compresses the image details, decreases the bit rate, and results in more distortion and some loss of quality. All audio SRCs were encoded into four different bit rates/channels (16, 32, and 64 kbps; pulse-code modulation/reference) using the AAC-LC encoder in FFmpeg. Ambisonic audio files were decoded using the All-Round Ambisonic Decoding algorithm as proposed in [43] to 26 multichannel speaker setups, which follows the standard in [44].

2.2 FFD Experiment

The FFD experiment was conducted with two purposes that are the collected response was treated as the data baseline (i) to be used in design simulation and (ii) to act as a reference model when comparing different OED models. In the design of experiment, the factor levels can be summarized as six sample clips, four video QPs, three video resolutions, and four audio bit rates, resulting in 288 total runs for each assessor. Twenty assessors were invited to participate in the experiment conducted at FORCE Technology SenseLab. They were 12 males and eight females, ranging in age from 22 to 37 years (mean = 27.9, SD = 4.0), with different nationalities, most of whom were postgraduate students. All assessors were not hard of hearing or visually impaired and met the selection criteria based on a systematic screening process as described in [45].

The experiment was conducted in a standardized listening room, which complies with the acoustic requirements of EBU 3276 [46] and ITU-R BS.1116-3 [47] and allows the experiments to be conducted with audio (listening test) and audiovisual systems. SenseLabOnline 4.2 [48] was used as the user interface and conducted double-anonymized and randomized trials. The participant sat on a swivel chair located in the acoustic sweet spot and used a pad controller to perform the test. The audio stimuli were reproduced through a 26-channel system with Genelec 8040A loudspeakers calibrated to 65–73 dB for the most comfortable loudness depending on the samples and measured in listening position. Visual stimuli were displayed using a Samsung Odyssey+ HMD, which has a screen resolution of 1,440 × 1,600 per eye, a horizontal field of view of 110°, and 90 Hz refresh rate.

Multiple stimulus rating was used to generalize the common multiple stimulus rating method used in SAMVIQ [49] and MUSHRA [50] for intermediate video and audio quality, respectively. A training session was included prior to the experiment to familiarize the assessor with the protocol, system, user interface, pad controller, and stimuli. None of the results in the training session were included in the analysis. During the experiment, each assessor was asked to rate his or her overall perceived quality of a combination of impaired audio and video stimulus on a continuous rating scale ranging from 0 to 100, divided into five categories (Bad, Poor, Fair, Good, and Excellent). The number of stimuli on each trial was limited to seven to match the number of objects that an average person can hold in short-term memory according to Miller's Law [51]. In order to avoid simulator sickness effect, the user interface was displayed virtually on the head-mounted display (HMD), and the rating can be made continuously without taking the HMD off. In the middle of the session, the system will automatically stop every 20 min for a short break. At the end of the experiment, a total of 5,760 data points had been collected.

2.3 Simulated OED

Twelve OEDs (*sim#1−sim#12*) were simulated in JMP Pro15, consisting of D-optimal and I-optimal experimental designs with minimal settings (*sim#1 & sim#2*) and with

Table 1. Simulated experimental design for D-optimal and I-optimal designs (*sim#1 - sim#12*) derived from the FFD experiment.

Design	Type	Replication	Additional points	Total points
FFD	FFD	···	···	288
sim#1	D-Opt	···	···	120
sim#2	I-Opt	···	···	120
sim#3	D-Opt	···	24	144
sim#4	I-Opt	···	24	144
sim#5	D-Opt	24	···	144
sim#6	I-Opt	24	···	144
sim#7	D-Opt	···	60	180
sim#8	I-Opt	···	60	180
sim#9	D-Opt	60	···	180
sim#10	I-Opt	60	···	180
sim#11	I-Opt	···	48	168
sim#12	I-Opt	48	···	168

either additional or replicated points (*sim#3−sim#12*) (see Table. 1). The design with additional points means that a set of new test conditions was added to the design space, whereas replicated points mean that a part of current test conditions in minimal settings were replicated to reach the limit of the design space. The total number of data points in each simulation was a multiple of six to allow implementation in the perceptual evaluation with a maximum of seven stimuli (six plus one reference) in a row.

It should be noted that the main concept of OED is to reduce the number of test conditions for the experimental trial, which also results in a reduction of statistical power. In the case of the simulation DOE, this could also increase the risk of a Type I error. To avoid Type I error, a 95% confidence interval is considered sufficient in a perceptual evaluation study, and statistical power was also maintained when selecting the simulated designs of *sim1−sim12* to not less than 0.7 for up to two-way interactions.

In this study, the design simulation was conducted to find answers to the first three research questions below:

RQ1: How well do the simulated OEDs perform compared to FFD in terms of predicted variance?

RQ2: What is the impact of adding and replicating data points on the performance of the simulated OEDs?

RQ3: Are there any performance differences between simulated OEDs and FFD?

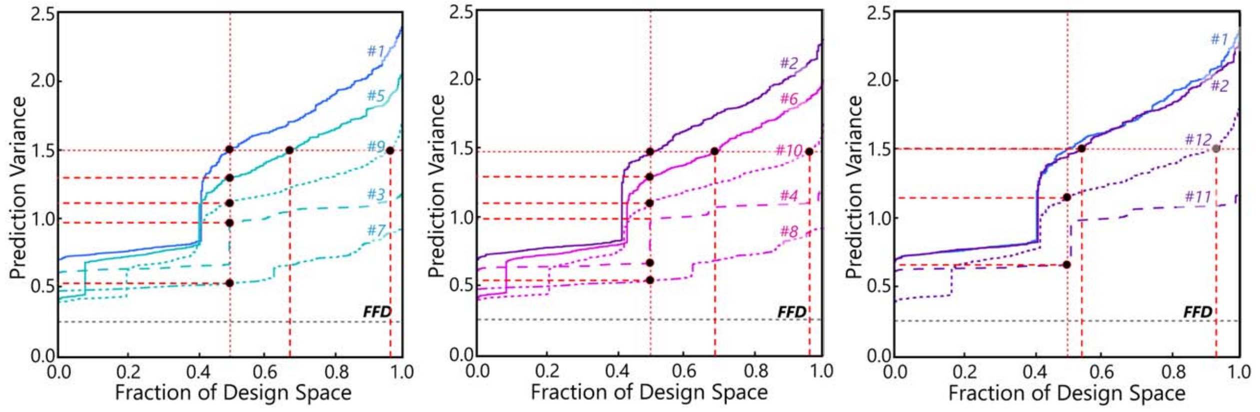
The first two questions (RQ1 and RQ2) were examined using a fraction of design space (FDS) plot [52], which allows for evaluation of the FDS over which the relative predictive variance is below a certain value. The FDS plots for *FFD* and *sim#1−sim#10* are shown in Figs. 3(a) and 3(b) for the D-optimal and I-optimal designs, respectively. The slope of the curve indicates how quickly the design reaches the maximum value of the predictive variance, with a value closer to horizontal being preferred. Here, the performance of the simulated designs is evaluated against the FDS plots by measuring two parameters. The first parameter is a value of the prediction variance in 50% of the design space ($PV_{FDS50\%}$), and the second parameter is a portion of the FDS when the prediction variance is equal to 1.5 ($FDS_{PV=1.5}$), which is a threshold observed in simulated OEDs with minimal settings. A horizontal and vertical dashed line within the FDS plot in Fig. 3 correspond to these two evaluation parameters. The values of $PV_{FDS50\%}$ and $FDS_{PV=1.5}$ for each design are shown in Table 2.

The ideal condition is that the designs have low predictive variance. The lower the FDS profile, the better the performance of the design. As can be seen in the Fig. 3, in comparison with FFD that has a low and stable predicted variance over the FDS, predicted variance of OED increases as the FDS progresses (RQ1). Additionally, as summarized in Table 2, the performance varies between the OEDs, and the FDS between the D-optimal and I-optimal designs has a small difference but exist. From Table 2, it can be seen that the I-optimal design variants generally perform better than the D-optimal design in the same setting. The highest difference in $PV_{FDS50\%}$ was found between *sim#3* and *sim#4* with a score difference of 0.308. It is also clear that both replication and the additional design points (*sim#3−sim#10*) contribute to the lower prediction variance compared to the minimal settings (*sim#1 & sim#2*). Additionally, the same number of data points added to the design space contributes more to the reduction in prediction variance than point replications (RQ2).

The fact that the predictive variances of I-optimal are lower than those of D-optimal, as presented in this study, supports the finding made in a previous study that I-optimal provides lower integrated variance and thus increases the accuracy of predictive models, which corresponds to one of the ultimate goals of quality prediction [53]. Regarding the proportion of FDS at certain thresholds for prediction variance, it is suggested that $FDS > 80\%$ to explore a significantly better response in a real experiment. For an experiment that completes a series of development and may contain noise from unknown variables, $FDS > 95\%$ is recommended [54]. Accordingly, *sim#1, sim#2, sim#5,* and *sim#6* are not theoretically recommended for laboratory experiments because $FDS_{PV=1.5} < 70\%$. On the other hand, *sim#9* and *sim#10* achieve $FDS_{PV=1.5} \geq 95\%$, but the number of 180 data points is relatively large. For the above reasons, two more I-optimal designs were included in the simulation (*sim#11−sim#12*), each with 168 data points (trade-offs between 144 and 180 data points). The simulation showed that $FDS_{PV=1.5} \geq 90\%$, which was considered sufficient for this study.

The final question (RQ3) was evaluated by a design using FFD data as input for the response variables in simulated OEDs. OLS was modeled for all simulated OEDs. Two-factor interactions and a three-factor interaction (*resolution* × *QP* × *bit rate*) were added as model terms. Analysis of variance (ANOVA) was examined to determine whether the response variable changed as a function of the level of the independent variable. The hypothesis test in ANOVA was as follows:

- $\boldsymbol{H_0}$ **:** there is no difference between the group means.



(a)

(b)

(c)

Fig. 3. FDS plot between simulated OEDs compared with FFD respectively for (a) D-optimal design, (b) I-optimal design, and (c) OEDs designs selected for laboratory experiments.

- $\boldsymbol{H_1}$ **:** at least one group differs significantly from the overall mean of the independent variable.

The results were analyzed using the overall performance of the ANOVA model considering all factors combined at a high level and using the effect tests to further examine the interaction of the independent variables. The summary of the analysis for the FFD and simulated OEDs is presented in Table 2. Generally, the model goodness of fit of each design is relatively high ($\geq$0.97), with the root-mean-square error ranging from 2.35 to 3.71, and the mean value is about 47.0. The ANOVA results show that the models constructed from the FFD and simulated OEDs have a significant relationship between the dependent and independent variables with a p <0.05 in a 95% confidence interval. Additionally, the influence of each factor and factor interaction on the model was analyzed in an effect test, which was presented in the form of F values and significance signs. For a single term effect, video resolution (x_1) has the highest F value in all models, followed by audio bit rate (x_3) and video QP (x_2). Meanwhile, the sample material (x_4) has the lowest F value for all models and makes an insignificant contribution to the model in *sim#1−sim#3*. However, this condition may also be affected by the low variation in the sample. Increasing the number of samples could improve the significance of this term for the model.

A standard D-optimal design (*sim#1*) has three insignificant terms and a relatively low F value compared with other designs. In the same settings, the D-optimal designs have a lower significance effect than the I-optimal designs (e.g.,

Table 2. Comparison of FDS score, model fit, and ANOVA results between FFD and simulated OEDs.

Parameters	FFD	sim#1	sim#2	sim#3	sim#4	sim#5	sim#6	sim#7	sim#8	sim#9	sim#10	sim#11	sim#12
Data points	288	120	120	144	144	144	144	180	180	180	180	168	168
Fraction of Design Space													
$PV_{FDS50\%}$	0.260	1.509	1.448	0.975	0.667	1.292	1.290	0.538	0.540	1.119	1.109	0.584	1.149
$FDS_{PV=1.5}$	1.00	0.50	0.54	1.00	1.00	0.67	0.69	1.00	1.00	0.95	0.96	1.00	0.91
Summary of fit													
R^2	0.984	0.994	0.996	0.994	0.992	0.993	0.995	0.989	0.990	0.996	0.994	0.989	0.994
Adjusted R^2	0.977	0.975	0.982	0.984	0.978	0.981	0.987	0.978	0.980	0.991	0.988	0.976	0.987
RMSE	3.67	3.65	3.30	3.09	3.61	3.26	2.75	3.52	3.45	2.35	2.65	3.71	2.76
Mean of response	47.24	46.53	47.64	47.02	47.23	47.22	46.53	47.44	47.87	47.38	46.06	47.81	46.71
ANOVA													
F -value	132.08	52.23	71.57	93.86	7.05	8.85	121.28	88.98	95.47	211.20	156.05	73.93	141.11
p -value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Effect Test (F -value)													
x_1	2429.87	763.54	973.68	1694.37	1233.12	1162.46	1757.50	1517.74	1588.97	2928.44	2494.14	1262.84	1832.69
x_2	805.33	293.45	355.56	569.00	382.87	394.87	564.49	519.59	524.95	952.24	744.88	437.68	697.82
x_3	1179.45	441.00	517.90	763.39	614.27	612.29	766.51	77.89	84.21	1462.37	1222.37	651.45	874.40
x_4	5.27	1.37^{ns}	1.94^{ns}	2.35^{ns}	3.86∗	1.97∗	1.07∗	4.18∗	3.83∗	5.39	2.42∗	2.88∗	5.34
$x_1 \times x_2$	25.52	4.73	11.29	16.83	12.48	11.14	17.26	13.07	13.41	32.90	19.17	14.46	25.32
$x_1 \times x_3$	83.47	3.19	27.08	51.36	38.74	29.33	44.05	46.92	51.09	77.70	65.16	41.81	55.48
$x_1 \times x_4$	3.35	2.56∗	6.25	9.59	7.47	2.84∗	7.09	5.96	6.46	14.19	9.39	5.78	8.72
$x_2 \times x_3$	12.80	9.94	9.73	15.04	11.35	15.32	18.46	19.02	2.55	23.41	23.30	12.73	26.42
$x_2 \times x_4$	5.04	1.94^{ns}	2.31∗	4.29	1.91∗	3.53	3.41	4.54	4.80	7.72	3.63	2.26∗	7.54
$x_3 \times x_4$	6.28	2.22∗	2.09∗	3.58	2.61	2.33∗	4.85	4.06	3.53	7.20	5.02	3.65	4.05
$x_1 \times x_2 \times x_3$	4.77	1.70^{ns}	2.37∗	4.68	2.07∗	2.34∗	3.24	3.28	3.38	8.26	4.04	2.65	5.05

$x_1 =$ video resolution, $x_2 =$ video QP, $x_3 =$ audio bit rate, $x_4 =$ sample, (∗) 0 .01 $\leq p$ -value ≤ 0.05 , ns = not significant

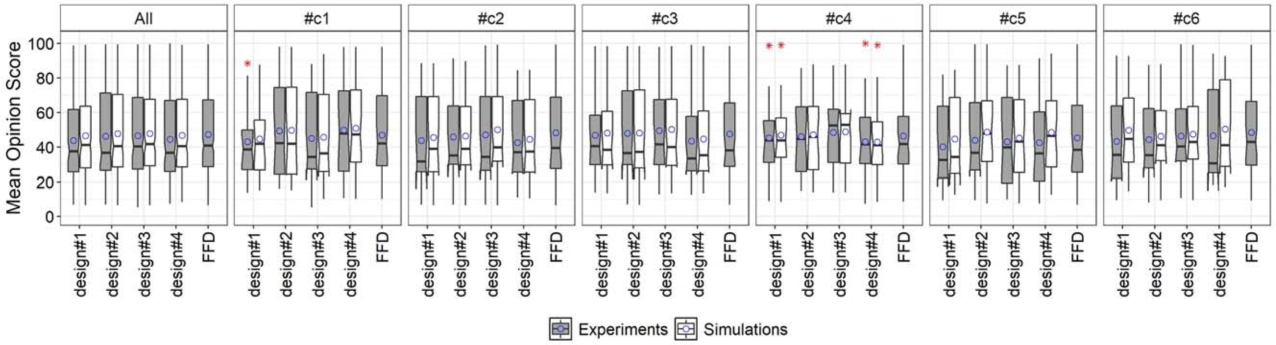


Fig. 4. Box and whisker plot of all investigated designs. The notches, horizontal lines inside the boxplot, circles, and stars represent 95% CI, median values, mean values, and outliers, respectively.

sim#1 vs *sim#2, sim#3* vs *sim#4, sim#5* vs *sim#6*). With respect to the number of data points, either providing additional data points or replicating improves the significance of individual terms in the model (e.g., *sim#1*, *sim#3*, and *sim#7* vs *sim#2*, *sim#6*, and *sim#10*).

Considering (i) the trade-offs between the number of data points to be tested in each design and the design performance indicated by the ANOVA results as well as the number of significant parameters in the effect test and (ii) the interest to test the D-optimal and the I-optimal with minimal settings (120 data points) and modified settings, four OEDs (*sim#1*, *sim#2*, *sim#11*, and *sim#12*) were selected for the laboratory experiment. It should be noted that minimal settings in OEDs do not necessarily give the best performance because of the reduction of more than 50% of the data points, which may reduce statistical power. However, this is part of the interest in evaluating how the minimal setting might differ from the modified setting and experimental data. In the following, these four OEDs will be referred to as *design#1*, *design#2*, *design#3*, and *design#4*.

2.3.1 OED Experiments

The same assessors of the FFD experiment were invited to participate in four OED experiments. However, only 19 assessors (12 males, seven females; mean age = 28.1, SD = 3.8) were able to participate in the experiments due to unavailability.[5] Because of the extensive work with the previous FFD experiment, an interval of approximately 1 month was scheduled between the FFD and OED experiments to avoid fatigue and memory effects between the two experiments. The experiments lasted two to three sessions on different days, conducted by each assessor. The presentation of each experimental design and the stimulus were randomized for each assessor so that, for example, one had the order of experimental *designs #1*, *#4*, *#2*, and *#3*, and others had a completely different order. Finally, the remaining experimental setups were the same as for the FFD experiment, as SEC. 2.2 described.

2.3.2 Analysis

The experimental data were analyzed to answer two more research questions addressed in this study, by including:

- RQ4: Is there a difference between the results of the simulated and experimental OEDs? If so, to what extent?
- RQ5: Is there a difference between the results of FFD and experimental OEDs? If so, to what extent?

The OLS regression model was used to estimate the relationship between one or more independent variables and a dependent variable. A summary of model fit and ANOVA is reported. The ANOVA was used at a 95% confidence level ($p < 0.05$) to assess the significance of each variable and its interaction based on its F and p values. The larger the F value and the lower the p value, the greater the evidence that there is a difference between group means. These analyses were used to make comparisons between FFD and experimental OEDs and within experimental OEDs.

3 RESULTS

3.1 Differences Between Simulated and Experimental OEDs

Box and whisker plots showing the distribution of mean opinion score (MOS) of perceived audiovisual quality of all simulated and experimental designs are shown in Fig. 4. Overall, the range of data for each design is relatively similar in the lower and upper ranges (between 10 and 99). The quartile range of *design#1* is shorter compared with the other designs. The mean of all designs has a relatively similar range between 40 and 50, with the experimental results being slightly lower than the simulations for both the mean and median. Nevertheless, the differences are not significant, as shown by the overlapping confidence intervals. According to Fig. 4, there is no difference between simulation and experiment for the data distributed in each design. This is also evidenced by a calculated multiple independent t-test, as shown in Table 3, which shows p values >0.05, indicating that there is no signifi-

[5] Two out of 20 assessors were unavailable, and one expert assessor was added.

Table 3. Multiple *t*-test comparison between experiment and simulation for each and overall designs.

Design	Group1	Group2	n1	n2	*t*	df	*p*
All	Experiment	Simulation	576	576	–1.35	1149	0.177^{ns}
design#1	Experiment	Simulation	120	120	–0.940	238	0.348^{ns}
design#2	Experiment	Simulation	120	120	–0.456	238	0.649^{ns}
design#3	Experiment	Simulation	168	168	–0.497	334	0.620^{ns}
design#4	Experiment	Simulation	168	168	–0.832	334	0.406^{ns}

ns = not significant

cant difference between experiment and simulation for each design (RQ4).

3.2 Differences Between FFD and Experimental OEDs

Table 4 shows a summary of OLS and ANOVA results of FFD and OEDs in order to answer RQ5. The means range from 43.71 to 47.24, with FFD and *design#1* having the largest and smallest means, respectively. The model fits show that the coefficient of determination R^2 is higher than 0.97 for all models, demonstrating good prediction accuracy. *design#1* has the largest R^2 (0.984) and the smallest RMSE (3.12) among all designs, whereas the adjusted R^2 has the largest value for *design#4* (0.983) and RMSE = 3.31. In linear models, R^2 tends to be overestimated because it always increases with the number of independent variables in the model. Therefore, the adjusted R^2 or so-called "corrected goodness-of-fit" attempts to correct for this overestimation by determining the percentage of variance in the target field that is explained by the inputs.

Based on the results of the ANOVA, all models show strong evidence against the null hypothesis illustrated by the *F* and *p* values, implying that H_0 is rejected and H_1 may be accepted. The I-optimal design with 48 repeated points (*design#4*) has the largest model *F* value according to FFD with $F = 132.8$, followed by *design#1* (73.06), *design#2* (64.44), and *design#3* (54.46). In terms of the effect test, video resolution (x_1) has the highest effect in the single interaction, followed by audio bit rate (x_3), video QP (x_2), and sample material (x_4). It can be clearly seen that the effect of the sample material is rather small compared with the other factors ($F < 6.0$). This result supports Lorho's [20] earlier study that the effect of the sample material is significant in the main effect but may vary in the interaction effects.

Likewise, in this study, sample material (x_4) remains significant for the model, even though this factor does not contribute significantly in the factor interaction terms, as indicated by the low *F* value and the higher *p* value, e.g., ($x_1 \times x_4$) in *design#1*, ($x_2 \times x_4$) in *design#1* and *design#3*, and ($x_3 \times x_4$) in *design#2*. The fact that these interaction effects are not significant indicates that the joint variability between the sample material (x_4) and corresponding factors is greater than other factors. The three-factor interaction ($x_1 \times x_2 \times x_3$) is not significant for *design#2* and *design#3*. The

Table 4. Comparison of model fit and ANOVA results between FFD, simulated OEDs, and experimental results.

Parameters	FFD		design#1		design#2		design#3		design#4	
Data points	288		120		120		168		168	
Summary of fit										
R^2	0.984		0.996		0.995		0.985		0.992	
Adjusted R^2	0.977		0.982		0.980		0.967		0.983	
RMSE	3.67		3.12		3.59		4.41		3.31	
Mean of response	47.24		43.71		46.23		46.53		44.53	
ANOVA	*F* -value	*p* -value	*F* -value	*p* -value	*F* -value	*p* -value	*F* -value	*p* -value	*F* -value	*p* -value
Model	132.08	<0.0001	73.06	<0.0001	64.44	<0.0001	54.46	<0.0001	103.42	<0.0001
Effect Test (*F* -value)										
x_1	2429.87	<0.0001	997.06	<0.0001	895.96	<0.0001	931.01	<0.0001	1168.38	<0.0001
x_2	805.33	<0.0001	454.82	<0.0001	332.02	<0.0001	319.72	<0.0001	553.36	<0.0001
x_3	1179.45	<0.0001	571.60	<0.0001	439.87	<0.0001	459.89	<0.0001	645.28	<0.0001
x_4	5.27	0.0001	5.31	0.0016	5.52	0.0013	3.32	0.0092	5.78	0.0001
$x_1 \times x_2$	25.52	<0.0001	13.74	<0.0001	9.22	<0.0001	12.07	<0.0001	24.58	<0.0001
$x_1 \times x_3$	83.47	<0.0001	38.66	<0.0001	32.81	<0.0001	33.81	<0.0001	39.07	<0.0001
$x_1 \times x_4$	3.35	<0.0001	1.89	0.0923^{ns}	5.77	0.0001	3.72	0.0004	4.00	0.0002
$x_2 \times x_3$	12.80	<0.0001	15.88	<0.0001	8.93	<0.0001	11.02	<0.0001	22.07	<0.0001
$x_2 \times x_4$	5.04	<0.0001	1.99	0.0585^{ns}	2.12	$0.0432*$	1.44	0.1504^{ns}	4.15	<0.0001
$x_3 \times x_4$	6.28	<0.0001	3.09	0.0052	1.40	0.2161^{ns}	2.74	0.0021	4.49	<0.0001
$x_1 \times x_2 \times x_3$	4.77	<0.0001	2.80	0.0077	1.51	0.1626^{ns}	1.29	0.2168^{ns}	3.94	<0.0001

$x_1 =$ video resolution, $x_2 =$ video QP, $x_3 =$ audio bit rate, $x_4 =$ sample, ($*$) 0 .01 $\leq p$ -value $\leq$ 0.05 , ns = not significant

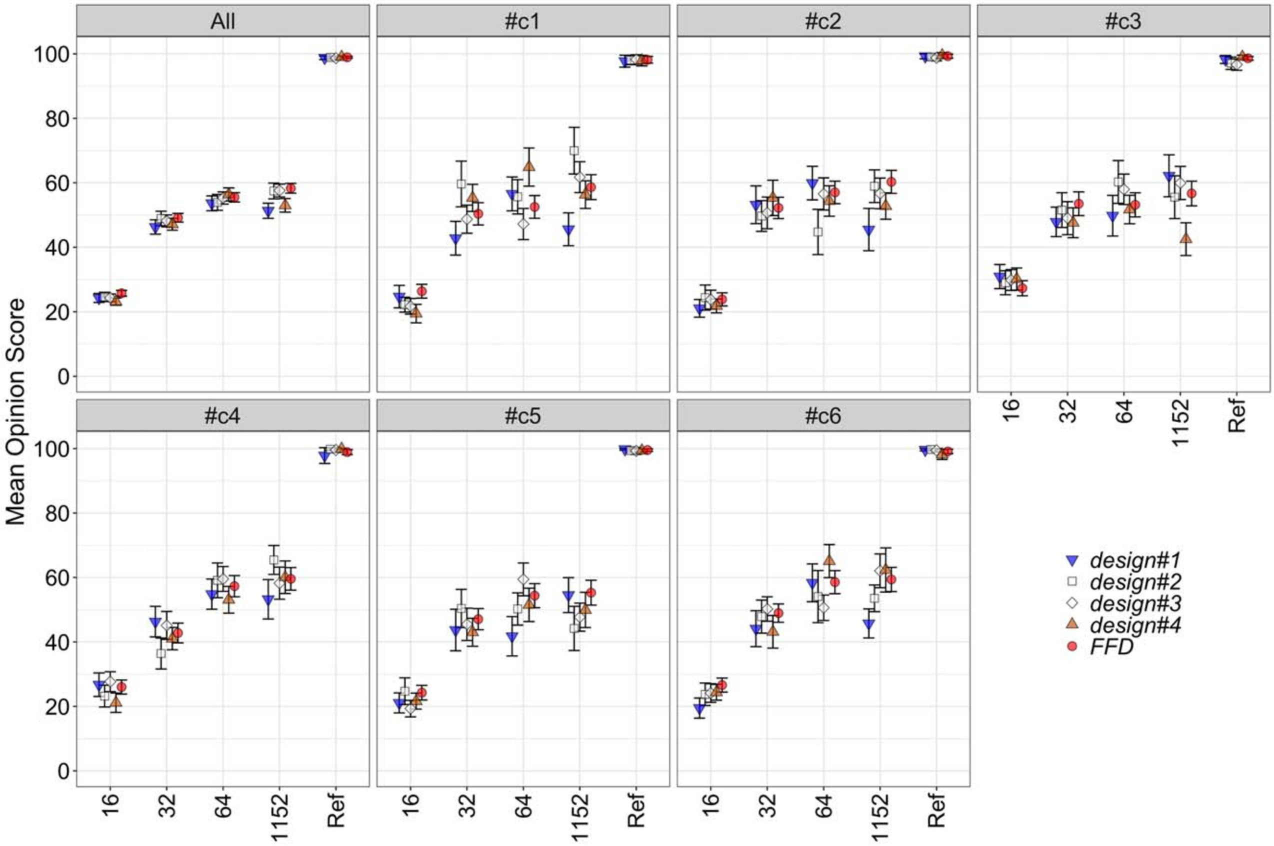


Fig. 5. Comparison of MOS-CI between FFD and four OEDs in terms of the effect of audio bit rates.

smallest *F* value for video resolution (x_1) and audio bit rate (x_3) in *design#2* compared with the others may be the reason for this observation, similar to the sample materials (x_4) in *design#3*, which have the smallest value among the other designs. These insignificant interactions may be due to unique conditions or design points generated by each OED algorithm.

Thus, it shows that among all OEDs experimented in this study, the I-optimal design with repetition *design#4* can produce the design that produces a significant effect for all terms ($p < 0.05$). Note that this reasoning should not neglect the influence of unknown factors, such as the practical aspect of experimental design and expertise of the selected assessors.

3.3 Analysis of *design#1* and *design#4*

As summarized in Table 4, there are competing performances between *design#1* and *design#4*. Although *design#1* has the best fitting performance in terms of a high adjusted R^2, *design#4* is the design with the most significant effect among the others. In Figs. 5–7, the effect of the audio-video encoding parameters on the perceived quality score is presented in terms of the MOS with a 95% confidence interval (MOS-CI) and compared between the designs. Examination of the overlapping confidence interval led to two important discussion points related to the performance comparison between *design#1*, *design#4*, and FFD; these are (i) the MOS-CI distance between *design#1* or *design#4* and FFD and (ii) the performance of *design#1* and *design#4* in detecting the influence of encoding parameters on MOS.

Fig. 5 shows the effect of audio bit rate on MOS averaged across video resolutions and QPs. In FFD, MOS increases as the bit rate is increased, with the significant difference varying by sample clip. In contrast, the result of MOS varies as a function of clips, with OEDs unable to detect the difference between 64 kbps/channel and 1,152 kbps/channel in most cases. *design#4* shows better performance compared with *design#1*, except for *#c3*. Between these two OEDs, *design#4* has a smaller gap with FFD in overall performance compared to *design#1*. For most occurrences at each clip, *design#4* performs better than *design#1* except at *#c1* 64 kbps, *#c3* 1,152 kbps, and *#c6* 64 kbps.

In Fig. 6, the effect of QPs on MOS is shown as an average of audio bit rates and video resolutions. Similar to Fig. 5, the performance of the OEDs varies depending on the encoding parameters (video QP) and sample clips. From the overall performance, it is clear that *design#1* has the lowest MOS compared with the other designs and that *design#4* has the most overlap MOS with FFD. Finally, the effect of video resolution on MOS average of other encoding parameters is shown in Fig. 7. In contrast to the previous two subfigures, the OEDs can discriminate very well between video resolution and sample clips and show a similar trend to FFD in all cases. Moreover, the trend is

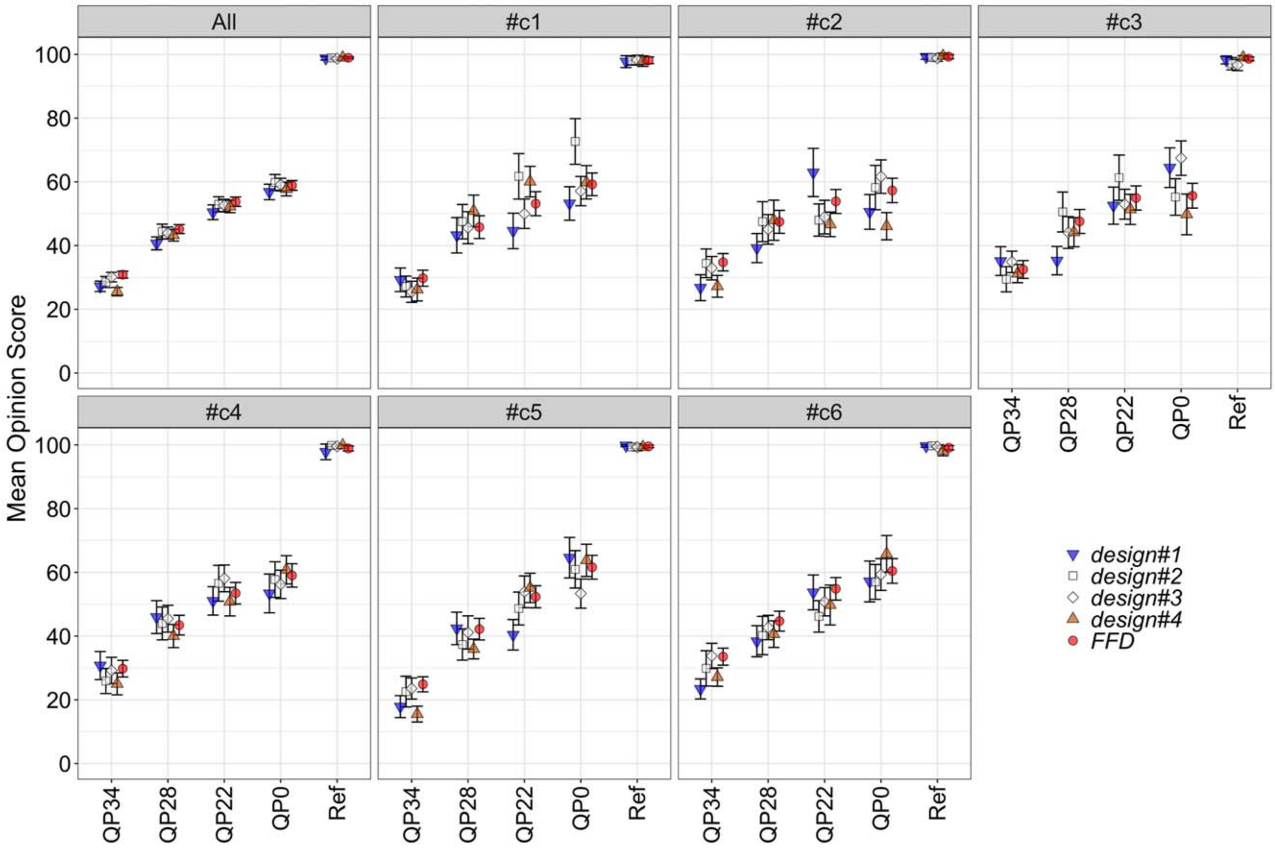


Fig. 6. Comparison of MOS-CI between FFD and four OEDs in terms of the effect of video QPs.

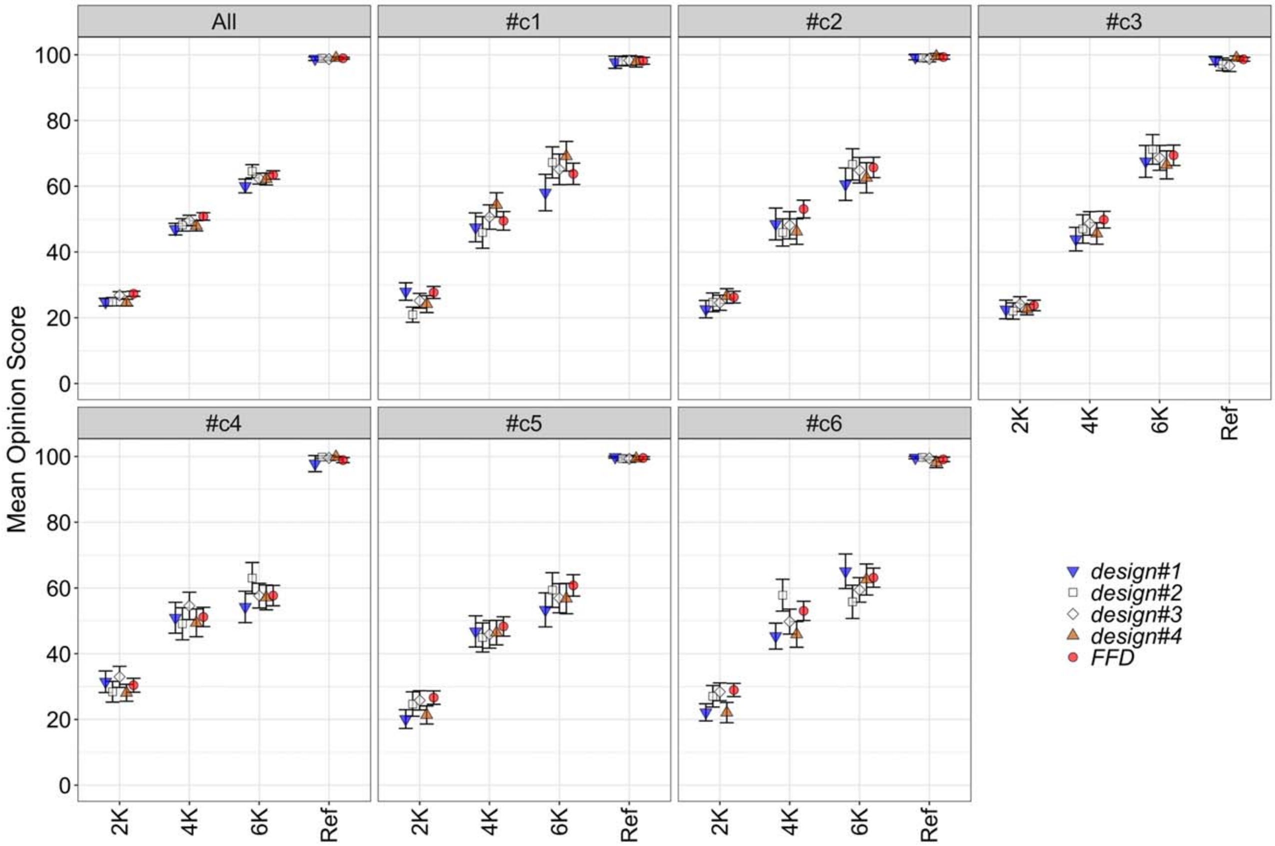


Fig. 7. Comparison of MOS-CI between FFD and four OEDs in terms of the effect of video resolutions.

very close to a linear distribution for clips overall and for *#c3*. *design#4* slightly outperforms *design#1*, except in the case in which *design#1* overlaps more with FFD, such as *#c2* 4K.

4 DISCUSSION

There is an interest in understanding the perceived quality of multimedia and how individual factors interact in constructing overall perceived quality. In addition to numerous efforts on the algorithm development to propose predictive metrics, perceptual assessment by using human subjects is critical because (i) technological development can augment human perception, especially in terms of multimodal assessment, e.g., adding additional factors and modalities can contribute to confounding factors and thus affect the overall judgement, and (ii) with careful experimental design, this is considered the Excel method to validate the accuracy of predictive metrics. However, the traditional DOE, which is commonly used in this field, such as FFD, may quickly become infeasible as the number of factors increases in multimodal assessment in the next-generation multimedia system. With these considerations, this study was motivated to find alternative designs that can generally reduce the experimental burden while maintaining the statistical criterion.

An important caveat in conducting an OED for perceptual evaluation is that the experimenter should have an idea of the goal of the experiment in order to choose an appropriate optimality criterion. Although OED is based on a single mathematical derivation [see Eqs. (3−5)], the criterion applied to OED is related to the experimental objective in terms of change in variance. The D-optimality criterion aims to minimize the determinant of the variance-covariance matrix corresponding to a subset of β and can be useful when the interest is only on a subset of the parameters. Whereas the D-optimality criterion minimizes the average variance of the parameter estimates, the I-optimality criterion looks for experimental designs that minimize the average variance of the prediction and thus can be advantageous in experiments aimed at model prediction [53]. To evaluate the robustness of a constructed experimental design, the FDS plot shows the fraction of the experimental design space in which the relative prediction variance is below a certain value [55]. It is desirable to have a large FDS with low values for the prediction variance [52].

When selecting the optimality criterion, design simulation can be explored first. One way to perform the design simulation is to use the FFD dataset from the previous experiment or public datasets as input data for the response variables. In the study presented here, the FFD experiment was used to generate the data for the simulation and to demonstrate the performance of OED and FFD. Therefore, performing FFD prior to the OED experiment should not be considered a practical step for future work. Another way to perform simulation is to use available data sets of similar experiments, but this is not always the case because each DOE is often unique. Another alternative is to use simulation techniques such as Monte Carlo, which are able to simulate the design by adding random noise to factors and predictions for the model [56, 57]. A notable finding of the simulation strategy adopted in this study was that the difference between the model fit parameters of FFD and the simulated designs was relatively small and that the significance of the main and factor interaction effects could vary depending on the design and number of data points included (RQ1−RQ3). The more data points, the fewer the number of insignificant effects. It should be noted that these results apply only to this case and that a different simulation strategy may result in a different trend.

One of the main interests in the simulation field is to validate the results obtained from simulation with experimental data. The reason for this is that during the experiment, additional noise may occur because of unknown variables, and the data may differ compared to the simulation. If the data obtained from the experiment is not significantly different from that of the simulation, it means that the simulation strategy is considered valid and can be proposed for future testing. As shown in Fig. 4 and Table 3, the authors' efforts have satisfied the above condition. This results in the another finding of this study, that the simulation strategy is accurate when simulating real experimental data (RQ4). Nevertheless, this finding does not necessarily apply to the general cases of perceptual evaluation. Instead, it can be applied to omnidirectional multimedia formats if some specific considerations are made, such as the encoding parameters and evaluation method used in the study.

The final research question (RQ5) aimed to quantify the difference between FFD and the four OEDs studied. Although OED can save between 41.6% (*design#3* and *design#4*) and 58.3% (*design#1* and *design#2*) in experimental effort compared to FFD, the trade-offs should be highlighted. From Table 4, it is evident from the factor interaction between $x_1 \times x_4$ that the F values of OEDs are significantly lower than FFD. Moreover, some OEDs have factor interactions with non-significant p values. Figs. 5–7 shows the deficiencies of OEDs in detecting the influence of encoding parameters and sample clips on MOS.

Regarding the OED comparison, although *design#1* contributes to the largest R^2 and lowest RMSE values, it does not capture the significant effects of the two-factor interaction of $(x_1 \times x_4)$ and $(x_2 \times x_4)$. Additionally, the corrected R^2 value, called the adjusted R^2, should be considered a more valid method for evaluation because it can avoid overestimation of the model prediction. Based on this fact, *design#4* is expected to outperform other designs, including *design#3*, which has the same number of data points. It also appears that all effect terms in *design#4* are significant ($p < 0.005$) for the model that provides the best results for FFD. The advantages of using *design#4* are shown in Figs. 5–7, where it can approximate the FFD results very well for most conditions. This reflects the next findings that (i) the I-optimal design with 40% repeated points (120 standard points + 48 repeated points) can make a significant contribution that is relatively similar to FFD and (ii) the repetition of data points makes a greater contribution to significant effect terms than additional data points in the I-optimal design.

It is argued that the aforementioned shortcomings are due to the specific test conditions (data points to be tested) that are unique to each OED. For example, such a design may be dominated by high audio bit rates, low QP, and high video resolution in one sample clip and a different condition in another sample clip. In practice, this results in some trials being dominated by a set of stimuli that have a low perceptual distance, making the assessment process even more difficult and easily failing the test assessor. The solution can be found before the design is created by specifying how flexibly the values of a particular factor can be changed or by specifying a weighting value for a particular factor if it is known. Although these solutions are possible, they are not trivial and require prior knowledge of the experiments, which is beyond the scope of this study.

5 CONCLUSION

This paper presented the results of a comparison between the classical FFD and OEDs in the application of audiovisual perceptual assessment in an omnidirectional media format. The study was conducted by running 12 DOE simulations from FFD experimental data and selecting four designs for laboratory experiments to answer RQ1−RQ5. The primary goal of this study was to gain knowledge with empirical evidence on the effectiveness of OEDs for audiovisual perceptual assessment.

The current results show that the data distribution resulting from all OEDs in this study is experimentally valid for simulated data. ANOVA showed that the variation between experiment and simulation is represented by a large difference in F value. This is an indication that the experiment must be carefully designed when using OEDs. Nevertheless, the proportion of significant terms between simulation and experiment is similar, e.g., the fact that minimal design of D-optimal (*design#1*) contains more insignificant terms compared with I-optimal (*design#2*). Moreover, the study empirically shows that a number of insignificant effect terms can be reduced by adding or replicating a number of data points in the design space, as shown in *design#3* and *design#4*. It is also concluded that I-optimal design with point replication in *design#4* has the largest F value (103.42), can provide all significant factor and factor interactions in the effect test, and is therefore considered as the best OED among other OEDs observed in the study.

6 LIMITATION AND RESEARCH OUTLOOK

This work was exploratory and limited to understanding the potential of OEDs rather than specifically outperforming FFDs. Instead, OEDs was compared to each other to analyze how close a significance test of each term in each design is to the FFD. This work was also limited to the minimal OED settings and basic modification by additive or duplicative points. As for the experiment, the assessments were considered lengthy because each assessor was asked to participate multiple times, which may lead to inaccurate recall and memory effects. However, the decision to use the same group of assessors was made to ensure that the assessors have a similar level of experience in omnidirectional media evaluation. Additionally, the use of assessors with the same level of experience requires extensive training, and the use of assessors with different levels of experience will affect statistical results [1].

For the experiments conducted in this study, omnidirectional (360°) audiovisuals were used because they are considered a borderline case between traditional multimedia and extended reality systems (virtual, augmented, and mixed reality), in which the number of factors and factor levels in DOE can be easily increased. However, there is no doubt that the OED presented in this study can be used for any experiment with a multimedia system as long as it meets the requirements of the OED (e.g., the test factor should be either categorical or numerical). Moreover, for the quality defined in this work, the compression paradigm was used, which is due to the effect of some basic audio-video compression parameters in the omnidirectional multimedia format [58, 59]. Undoubtedly, other parameters can be included, and there are several factors that can affect the perceived quality, such as the type of loudspeaker setup, signal processing algorithms, head-related transfer function when testing headphones, audiovisual presentation, test conditions, etc., which are currently not the subject of this study but are highly recommended for future studies.

Future work should aim to test a different experimental design depending on the application, where optimization paradigms such as RSM, Plackett-Burman experimental design, OED, and active learning method can be used. Another line of research is the adaptation of OED for different applications, either with a single or multimodal evaluation for the domestic experience or even for extended reality applications (e.g., virtual, augmented, or mixed reality). Additionally, the evaluation metrics may also vary for each aspect of quality of experience. Finally, some of these mentioned designs require either numerical only or categorical independent factors, which limits OED to some applications of perceptual assessment studies with nominal factors.

7 ACKNOWLEDGMENT

This research was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 765911 RealVision. The authors convey acknowledgment to FORCE Technology, Bang & Olufsen, Demant, GN Store Nord, Sonova, WSA, and industrial partners who created the 360 audiovisual datasets under the HOA-SSR joint project funded by the Danish Ministry of Science and Technology, and they acknowledge the XRHub Team for the great help in dealing with technicalities, filming, and field recording. Finally, the authors appreciate all anonymous subjects who participated in the audiovisual test described in this paper.

8 REFERENCES

[1] S. Bech and N. Zacharov, *Perceptual Audio Evaluation−Theory, Method and Application* (Wiley, Chichester, UK, 2006).

[2] N. Zacharov, *Sensory Evaluation of Sound* (CRC Press, Boca Raton, FL, 2018).

[3] S. Winkler, *Digital Video Quality: Vision Models and Metrics* (Wiley, Chichester, UK, 2005).

[4] B. Belmudez, *Audiovisual Quality Assessment and Prediction for Videotelephony* (Springer, Cham, Switzerland, 2015). https://doi.org/10.1007/978-3-319-14166-4.

[5] G. Charness, U. Gneezy, and M. A. Kuhn, "Experimental Methods: Between-Subject and Within-Subject Design," *J. Econ. Behav. Organ.*, vol. 81, no. 1, pp. 1–8 (2012 Jan.). https://doi.org/10.1016/j.jebo.2011.08.009.

[6] G. E. P. Box and J. S. Hunter, "The 2^{k-p} Fractional Factorial Designs," *Technometrics*, vol. 3, no. 3, pp. 311–351 (1961 Aug.). https://doi.org/10.2307/1266725.

[7] E. J. Williams, "Experimental Designs Balanced for the Estimation of Residual Effects of Treatments," *Aus. J. Sci. Res.*, vol. 2, no. 2, pp. 149–168 (1949 Dec.). https://doi.org/10.1071/CH9490149.

[8] H. J. MacFie, N. Bratchell, K. Greenhoff, and L. V. Vallis, "Designs to Balance the Effect of Order of Presentation and First-Order Carry-Over Effects in Hall Tests," *J. Sens. Stud.*, vol. 4, no. 2, pp. 129–148 (1989 Sep.). https://doi.org/10.1111/j.1745-459X.1989.tb00463.x.

[9] L. Mion, G. D'Incà, A. De Götzen, and E. Rapanà, "Modeling Expression With Perceptual Audio Features to Enhance User Interaction," *Comput. Music J.*, vol. 34, no. 1, pp. 65–79 (2010 Mar.). https://www.jstor.org/stable/25653531.

[10] S. Schmidt, S. Zadtootaghaj, S. Wang, and S. Möller, "Towards the Influence of Audio Quality on Gaming Quality of Experience," in *Proceedings of the 13th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 169–174 (Online) (2021 Jun.). https://doi.org/10.1109/QoMEX51781.2021.9465450.

[11] R. Haakma, D. Jarnikov, and P. v. d. Stok, "Perceived Quality of Wirelessly Transported Videos," in P. v. d. Stok (Ed.), *Dynamic and Robust Streaming in and Between Connected Consumer-Electronic Devices*, Philips Research Book Series, vol. 3, pp. 213–239 (Springer, Dordrecht, The Netherlands, 2005). https://doi.org/10.1007/1-4020-3454-7_9.

[12] A. Singla, W. Robitza, and A. Raake, "Comparison of Subjective Quality Test Methods for Omnidirectional Video Quality Evaluation," in *Proceedings of the IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6 (Kuala Lumpur, Malaysia) (2019 Sep.). https://doi.org/10.1109/MMSP.2019.8901719.

[13] A. Molnar, "Content Type and Perceived Multimedia Quality in Mobile Learning," *Multimed. Tools Appl.*, vol. 76, no. 20, pp. 21613–21627 (2016 Nov.). https://doi.org/10.1007/s11042-016-4062-2.

[14] D. Johnston, H. Egermann, and G. Kearney, "Measuring the Behavioral Response to Spatial Audio Within a Multi-Modal Virtual Reality Environment in Children With Autism Spectrum Disorder," *Appl. Sci.*, vol. 9, no. 15, paper 3152 (2019 Aug.). https://doi.org/10.3390/app9153152.

[15] A. Moulson and H. Lee, "The Influences of Microphone System, Video, and Listening Position on the Perceived Quality of Surround Recording for Sport Content," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), paper 10309.

[16] S.-H. Yao, C.-L. Fan, and C.-H. Hsu, "Towards Quality-of-Experience Models for Watching 360° Videos in Head-Mounted Virtual Reality," in *Proceedings of the 11th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3 (Berlin, Germany) (2019 Jun.). https://doi.org/10.1109/QoMEX.2019.8743198.

[17] K. Brunnström and M. Barkowsky, "Balancing Type I Errors and Statistical Power in Video Quality Assessment," in *Proceedings of the IS&T International Symposium on Electronic Imaging: Human Vision and Electronic Imaging*, vol. 2017, no. 14, pp. 91–96 (Burlingame, CA) (2017 Jan.). https://doi.org/10.2352/ISSN.2470-1173.2017.14.HVEI-122.

[18] T. Walton, M. Evans, D. Kirk, and F. Melchior, "A Subjective Comparison of Discrete Surround Sound and Soundbar Technology by Using Mixed Methods," presented at the *140th Convention of the Audio Engineering Society* (2016 May), paper 9592.

[19] M. Giovanni, "Response Surface Methodology and Product Optimization," *Food Technol.* , vol. 37, no. 11, pp. 41–45 (1983 Jun.).

[20] G. Lorho, "Subjective Evaluation of Headphone Target Frequency Responses," presented at the *126th Convention of the Audio Engineering Society* (2009 May), paper 7770.

[21] R. F. Fela, N. Zacharov, and S. Forchhammer, "Towards a Perceived Audiovisual Quality Model for Immersive Content," in *Proceedings of the 12th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6 (Athlone, Ireland) (2020 May). https://doi.org/10.1109/QoMEX48832.2020.9123134.

[22] D. C. Montgomery, *Design and Analysis of Experiments* (Wiley, Hoboken, NJ, 2017), 9th ed.

[23] R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (Wiley, Hoboken, NJ, 2016), 4th ed.

[24] J. Kiefer and J. Wolfowitz, "Optimum Designs in Regression Problems," *Ann. Math. Stat.*, vol. 30, no. 2, pp. 271–294 (1959 Jun.). https://doi.org/10.1214/aoms/1177706252.

[25] F. Pukelsheim, *Optimal Design of Experiments*, Classics in Applied Mathematics (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2006).

[26] R. F. Fela, N. Zacharov, and S. Forchhammer, "Perceptual Evaluation of 360 Audiovisual Quality and Machine Learning Predictions," in *Proceedings of the IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6 (Tampere, Finland) (2021 Oct.). https://doi.org/10.1109/MMSP53017.2021.9733677.

[27] T. Baumann, "Ranking and Comparing Speakers Based on Crowdsourced Pairwise Listener Ratings," in B. Weiss, J. Trouvain, M. Barkat-Defradas, J. J. Ohala (Eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics, pp. 263–279 (Springer, Singapore, 2021). https://doi.org/10.1007/978-981-15-6627-1_14.

[28] Y. Jiang, Q. Xu, W. Zhang, and Q. Huang, "Active Sampling for Subjective Video Quality Assessment," in *Proceedings of the IEEE 4th International Conference on Multimedia Big Data (BigMM)*, pp. 1–5 (Xi'an, China) (2018 Sep.). https://doi.org/10.1109/BigMM.2018.8499064.

[29] J. Li, R. K. Mantiuk, J. Wang, S. Ling, and P. Le Callet, "Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation," in *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)* (Montreal, Canada) (2018 Oct.). https://doi.org/10.48550/arXiv.1810.08851.

[30] S. Ma, Z. Zeng, D. McDuff, and Y. Song, "Active Contrastive Learning of Audio-Visual Video Representations," *arXiv preprint arXiv:2009.09805* (2020). https://doi.org/10.48550/arXiv.2009.09805.

[31] C. S. Peirce, "A Theory of Probable Inference," in C. S. Pierce (Ed.), *Studies in Logic by Members of the Johns Hopkins University*, pp. 126–181 (Little, Brown and Co., Boston, MA, 1883). https://psycnet.apa.org/doi/10.1037/12811-007.

[32] C. S. Peirce, "Note on the Theory of the Economy of Research," *Oper. Res.*, vol. 15, no. 4, pp. 643–648 (1967 Aug.). https://doi.org/10.1287/opre.15.4.643.

[33] R. A. Fisher, "The Arrangement of Field Experiments," in S. Kotz and N. L. Johnson (Eds.), *Breakthroughs in Statistics*, Springer Series in Statistics, pp. 82–91 (Springer, New York, NY, 1992). https://doi.org/10.1007/978-1-4612-4380-9_8.

[34] R. A. Fisher, "Design of Experiments," *Br. Med. J.*, vol. 1, p. 554 (1936 Mar.). https://doi.org/10.1136/bmj.1.3923.554-a.

[35] K. Smith, "On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance They Give Towards a Proper Choice of the Distribution of Observations," *Biometrika*, vol. 12, no. 1/2, pp. 1–85 (1918 Nov.). https://doi.org/10.2307/2331929.

[36] J. Kiefer, "Optimum Experimental Designs," *J. R. Stat. Soc., B: Stat. Methodol.*, vol. 21, no. 2, pp. 272–319 (1959 May). https://www.jstor.org/stable/2983802.

[37] A. Wald, "On the Efficient Design of Statistical Investigations," *Ann. Math. Stat.*, vol. 14, no. 2, pp. 134–140 (1943 Jun.). https://www.jstor.org/stable/2235815.

[38] L. M. Haines, "The Application of the Annealing Algorithm to the Construction of Exact Optimal Designs for Linear−Regression Models," *Technometrics*, vol. 29, no. 4, pp. 439–447 (1987 Nov.). https://doi.org/10.1080/00401706.1987.10488272.

[39] V. V. Fedorov, *Theory of Optimal Experiments* (Academic Press, New York, NY, 1972).

[40] M. Rodriguez, B. Jones, C. M. Borror, and D. C. Montgomery, "Generating and Assessing Exact G-Optimal Designs," *J. Qual. Technol.*, vol. 42, no. 1, pp. 3–20 (2010 month). https://doi.org/10.1080/00224065.2010.11917803.

[41] A. C. Atkinson, A. N. Donev, and R. D. Tobias, *Optimum Experimental Designs, With SAS*, Oxford Statistical Science Series, vol. 34 (Oxford University Press, New York. NY, 2007).

[42] R. F. Fela, A. Pastor, P. L. Callet, et al., "Perceptual Evaluation on Audio-Visual Dataset of 360 Content," *arXiv preprint arXiv:2205.08007* (2022). https://doi.org/10.48550/arXiv.2205.08007.

[43] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820 (2012 Oct.).

[44] ITU-R, "Multichannel Sound Technology in Home and Broadcasting Applications," *ITU-R Recommendation BS.2159-7* (2015 Feb.).

[45] R. F. Fela, N. Zacharov, and S. Forchhammer, "Assessor Selection Process for Perceptual Quality Evaluation of 360 Audiovisual Content," *J. Audio Eng. Soc.*, vol. 70, no. 10, pp. 824–842 (2022 Oct.). https://doi.org/10.17743/jaes.2022.0037.

[46] EBU, "Listening Conditions for the Assessment of Sound Programme Material: Monophonic and Two–Channel Stereophonic," Tech. Rep. 3276 (1998 May), 2nd ed.

[47] ITU-R, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," *ITU-R Recommendation BS.1116-3* (2015 Feb.).

[48] G. Le Ray and J. Khalid, "SenseLabOnline: Combining Agile Data Base Administration With Strong Data Analysis," in *Proceedings of the R User Conference (useR!)*, vol. 10, paper 38 (Albacete, Spain) (2013 Jul.).

[49] F. Kozamernik, V. Steinmann, P. Sunna, and E. Wyckens, "SAMVIQ—A New EBU Methodology for Video Quality Evaluations in Multimedia," *SMPTE Motion Imaging J.*, vol. 114, no. 4, pp. 152–160 (2005 Apr.). https://doi.org/10.5594/J11535.

[50] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems," *ITU-R Recommendation BS.1534-3* (2015 Oct.).

[51] G. A. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits On Our Capacity for Processing Information." *Psychol. Rev.*, vol. 63, no. 2, pp. 81–97 (1956 Mar.). https://psycnet.apa.org/doi/10.1037/h0043158.

[52] A. Zahran, C. M. Anderson-Cook, and R. H. Myers, "Fraction of Design Space to Assess Prediction Capability of Response Surface Designs," *J. Qual. Technol.*, vol. 35, no. 4, pp. 377–386 (2003 Oct.). https://doi.org/10.1080/00224065.2003.11980235.

[53] B. Jones and P. Goos, "I-Optimal Versus D-Optimal Split-Plot Response Surface Designs," *J. Qual. Technol.*, vol. 44, no. 2, pp. 85–101 (2012 Apr.). https://doi.org/10.1080/00224065.2012.11917886.

[54] P. Whitcomb, "FDS−A Power Tool for Designers of Optimization Experiments," *Stat-Teaser* (2008 Sep.).

[55] A. Ozol-Godfrey, C. M. Anderson-Cook, and D. C. Montgomery, "Fraction of Design Space Plots for Examining Model Robustness," *J. Qual. Technol.*, vol. 37, no. 3, pp. 223–235 (2005 Jul.). https://doi.org/10.1080/00224065.2005.11980323.

[56] J. Sall, A. Lehman, M. Stephens, and S. Loring, *JMP Start Statistics: A Guide to Statistics and Data Analysis Using JMP* (SAS Institute, Cary, NC, 2017), 6th ed.

[57] M. J. Anderson, and P. J. Whitcomb, *DOE Simplified: Practical Tools for Effective Experimentation* (CRC press, Boca Raton, FL, 2017).

[58] ISO/IEC, "Information Technology — Coded Representation of Immersive Media (MPEG-I) — Part 2: Omnidirectional Media Format," *Standard 23090-2:2021* (2021 Jul.).

[59] M. M. Hannuksela and Y.-K. Wang, "An Overview of Omnidirectional MediA Format (OMAF)," *Proc. IEEE*, vol. 109, no. 9, pp. 1590–1606 (2021 Sep.). https://doi.org/10.1109/JPROC.2021.3063544.

THE AUTHORS

Randy Frans Fela

Nick Zacharov

Søren Forchhammer

Randy Frans Fela has engineering physics background from Gadjah Mada University (B.Eng.) and Bandung Institute of Technology (M.Sc.) with a focus on acoustics and spatial audio evaluation in immersive multimedia. His Ph.D. project was funded by European Union Marie Skłodowska-Curie Innovative Training Networks (MSCA ITN) RealVision and hosted by FORCE Technology and the Technical University of Denmark. He is currently working as a Perceptual Audio Engineer in the R&D team at GN Audio A/S (Jabra) to continue and expand his interest in optimal experimental design, perceptual evaluation, and perceptual prediction metrics.

•

Nick Zacharov [D.Sc. (Tech.), M.Sc., B.Eng. (Hons.), C.Eng., FAES] is a lead technology manager of perceptual audio evaluation at Meta Reality Labs, focusing on audio quality research. With an academic background in electroacoustics, acoustics, and signal processing, Nick has broad industrial experience in the audio profession spanning from mobile phone audio to augmented or virtual reality devices and professional studio monitor design. Nick is the co-author of *Perceptual Audio Evaluation–Theory, Method and Application* and editor/co-author of the book *Sensory Evaluation of Sound*. He has been an active member of the Audio Engineering Society and has more than 90 publications and patents to his name.

•

Prof. Søren Forchhammer received an M.Sc. and Ph.D. degree from the Technical University of Denmark, Kgs. Lyngby. Currently, he is a Professor with DTU Electro, Technical University of Denmark, where he has been since 1988. He is Head of the Coding and Visual Communication Group at the Department of Electrical and Photonics Engineering. He is currently Coordinator of the European Union Marie Skłodowska-Curie Innovative Training Networks (MSCA ITN) RealVision. His research interests include source coding, image and video coding, processing of image and video, processing for image displays, quality of coded multimedia data, multi-camera and light field images and video, communication theory, 2D information theory, and visual communications.