DOI: 10.61782/fa.2023.0312

forum acusticum 2023

COMPARING ONLINE VS. LAB-BASED EXPERIMENTAL APPROACHES FOR THE PERCEPTUAL EVALUATION OF ARTIFICIAL REVERBERATION

Vincent Martin[1*] **Lorenzo Picinali**[1]

[1] Dyson School of Design Engineering, Imperial College London, United Kingdom

ABSTRACT

A common approach for reproducing room acoustics effects is geometrical acoustics. The accuracy of such an approach is tied, among other variables, to the geometrical accuracy of the simulated room, and to the information regarding the absorption coefficients of its materials. However, from a perceptual standpoint, a model that accounts for all of a room's features would come at a high computational cost and could be redundant. As a result, a compromise can be reached between the perceived quality (e.g. authenticity, immersion, etc.) of the replicated room effect and the model's complexity. The purpose of this study is to look into the perceptual impact of simplifying the room geometry and minimizing the number of materials' absorption coefficients. Two separate experiments were conducted, both based on the MUSHRA methodology: one was run in a controlled lab environment through a Virtual Reality (VR) headset, while the other was run through a web-based interface. This paper focuses on the differences between the two protocols' impact on the results. It appears that the online-based experiment, notwithstanding the lack of control of the playback system and environment, and the participants' likely limited attention, produced minor but substantial differences with the results of the VR experiment.

Keywords: *geometrical acoustics, reverberation, online experiment, virtual reality*

*__Corresponding author__: *vincent.martin@imperial.ac.uk.*

1. INTRODUCTION

Different methods can be used to auralise (i.e. listening to an acoustic signal as if it was played inside a given space, whether measured or simulated) an environment. These fall into two categories: perceptually-motivated (optimized for human perception, for example in [1] [2]) and physically-motivated (aiming at replicating the acoustic soundfield). Geometrical acoustics is a physically-motivated approach that studies sound propagation through the concept of acoustic rays, estimating the room impulse response from the room's geometry and materials' acoustic properties [3]. This type of approach is widely used for auralization of various acoustic spaces and has been implemented by different software [4] [5]. However, it can be rather expensive from a computational point of view, requiring a balance between cost, complexity, and accuracy. We are currently running a study that investigates the perceptual effects of simplifying the information for geometrical acoustic auralization, specifically in binaural rendering. More precisely, we are looking at the impact of reducing the number of surfaces in the overall simulation of the room on the perception of the simulated reverberation.

Two experiments built with the same type of evaluation paradigm were run:

- A lab-based experiment consisting of a VR scenario labeled "Experiment I"
- An online-based experiment labeled "Experiment II"

Online data collection has become a widely used methodology in the behavioural sciences. However, con-

ducting carefully controlled behavioral online experiments introduces a number of new technical and scientific challenges, from the experiment design to the online compatibility, including participant recruitment. A considerable strength of online studies is that they can be easily scaled to large pools of participants, as recruiting larger samples does not require a higher workload. However, unlike lab-based experiments, many concerns about data quality have to be taken into account when preparing an online experiment. Four different aspects can be drastically different in online experiments and impact data quality: attention, comprehension, and reliability [6].

This paper examines the difference between the results of the two experiments and is organized as follows:: the initial section introduces the models and stimuli assessed in both experiments. Then the experimental methods are described. The results of both experiments are then presented, followed by a comparison and discussion of the findings. Finally, the last section provides a conclusion based on the study's results.

2. ROOM MODELS & STIMULI

CAD software was used to create five different variants of a living room model. The "reference" model, which had the most detail, was used as a starting point. Four geometrically reduced (GR) models were then produced by progressively removing larger objects. Each model had a different threshold that determined the smallest allowable surface area (i.e. surfaces smaller than that were eliminated), as detailed in Table 1. The higher the decimation threshold, the fewer polygons in the model, with the GR5 model being a "shoebox" room.

To ensure a more accurate comparison, the absorption coefficients of the materials in all models were automatically adjusted to match the decay profiles and reverberation times (RT60), per frequency band (8 in total), of the reference model using the Eyring formula [7].

Room Impulse Responses (RIRs) of the reflected components (i.e. without the direct path) were estimated for different source-listener positions using ray tracing within *CATT Acoustics* software, and exported as Spatial Room Impulse Responses (SRIRs) in third order Higher Order Ambisonics (HOA) format. Scattering coefficients were taken into account in all models except for the shoebox, which contained only specular reflections. In addition to the simulated SRIRs, a set of "anchor" SRIRs were generated by applying a 2.5kHz low-pass filter to the reference SRIRs. This was done in order to create

Model	Polygons	Removed surfaces
Reference	590	None
GR2	406	$<0.1m^2$
GR3	241	$<0.4m^2$
GR4	66	$<0.4m^2$ & furnitures
GR5 (shoebox)	6	All surfaces

Table 1. Geometrically-reduced models evaluated in this study. The number of polygons and the amount of them removed from the geometry are reported.

a set of reference impulse responses with reduced high-frequency content. For all models, the direct sound was convolved with a publicly available Head-Related Impulse Responses (HRIRs) [8], and added to the SRIRs.

The resulting HOA signals were decoded into binaural format using the so-called "virtual speakers" paradigm [9]. This was done employing the same approach as [10], therefore spatialising separately the direct path (rendered through direct convolution with HRTFs) and the reflected one (rendered through virtual loudspeakers and HOA), and using the *3DTI Toolkit* [11]. The levels of direct sound and reverberation were adjusted to maintain a consistent direct-to-reverberant ratio (DRR) across the different models' RIR. The binaural signals were recorded and used in an online-based version of the experiment, while these signals were processed in real time for the lab-based version using head tracking.

Two different sets of anechoic recordings were used in this study (the same as those used in [10]):

- A music recording of a performance of "Take Five" by Paul Desmond, consisting of three dry recordings: piano, drum kit, and saxophone. Each source was rendered in a different position (see Figure 1). This 3-channel recording has been cropped to a length of 7 seconds.
- A speech recording from the Music for Archimedes collection [12] of a single female speaker, cropped at a length of 5 seconds.

The requirements for a space with reasonably diverse acoustic properties drove the choice of this single measured and simulated environment for the experiment. The room is composed of a living room and an open kitchen space separated by a bar. It contains elements such as wooden floors and carpeted areas, tiled and plasterboard

walls, two separate ceiling sections, the kitchen area with hard and reflective surfaces, and the living room area with absorbing surfaces and elements (e.g. bookshelves, sofa, etc.). The speech stimuli had a single source location, while the music sources had three separate locations. Three listener positions were also chosen in various sections of the space and at various distances from the sources (see Figure 1).

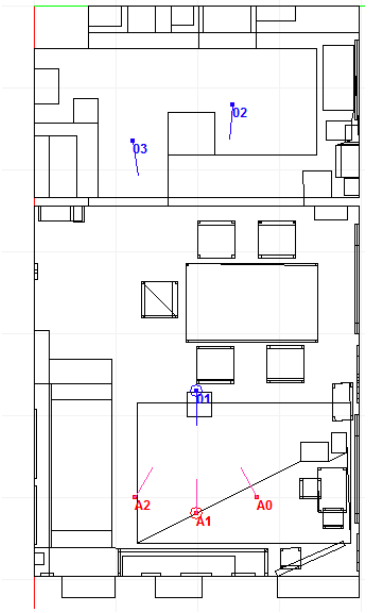

Figure 1. Floor plan of the room used for the experiment, the room consists of an open-plan kitchen, and a living room space. They are only separated by a bar. The blue dots (1-3) show the three different listener positions, while the lines represent the direction in which binaural recordings were made for the online experiment. The positions of the sound sources are represented by the red dots (A1-3). The A0 position is assigned to the speech recording. The piano recording is emitted from the A0 position for the music stimuli, the drums from A1, and the saxophone from A2.

3. METHODS

The following section outlines the protocol devised for each experiment. Initially, the shared features of both experiments are presented. Subsequently, the distinct characteristics of each experiment version are elaborated.

3.1 Conditions & paradigm in both experiments

During a trial, participants were requested to assess the similarity between stimuli and a reference stimulus. To conduct the evaluation, a MUSHRA methodology (ITU-R BS.1534 [13]) was used in a double-blind listening test. The reference stimulus was produced using the model with the highest geometrical precision, while an "anchor" stimulus was generated using the model described in subsection 2. The participants rated the similarity of each stimulus on a scale from 0 to 100, with 100 indicating complete similarity with the reference stimulus. The stimuli were classified into each combination of stimulus type (speech, music), listener position (1, 2, 3) resulting in a total of 6 trials for each participant. The trials were presented in random order in both experiments.

3.2 Experiment I: VR-based protocol

3.2.1 Design & tools

The MUSHRA testing report UI, consisting of audio playback and rating controls, was integrated into a basic VR scenario created using *Unity*, which was displayed on an *Oculus Quest 2* VR headset (illustrated in Figure 2). The visuals were generated using 360° images (non-stereoscopic) taken from the actual room, providing participants with an interactive visual reference of the room from the correct perspective and with the correct number of visual sources in the form of loudspeakers. The VR headset provided real-time head tracking, which was transmitted via Open Sound Control (OSC) to the *3DTI Tune-In Toolkit* Test Application, hosted on a separate laptop, allowing for binaural rendering that responded to head movements.

The sound level was calibrated so that the loudness of the anechoic speech signal spatialized at 1 meter was 60dB (LAeq), as per ISO 3382-3 guidelines. The signals were reproduced through a pair of *Sennheiser HD650* headphones.

3.2.2 Participants & Procedure

Twenty participants (13 males, 7 females) were enrolled; all participants confirmed that they had no hearing impairment, and filled out a questionnaire that included questions about their gender, age, and audio experience, before taking part in the experiment.

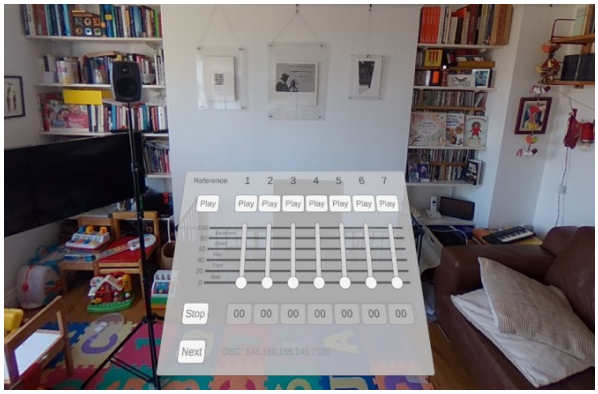

Figure 2. Interface displayed to the participant wearing the VR headset in the lab-based experiment. The trial corresponds here to listener position 1 and a speech signal. A 360° picture with a single visual sound source is presented to the participant.

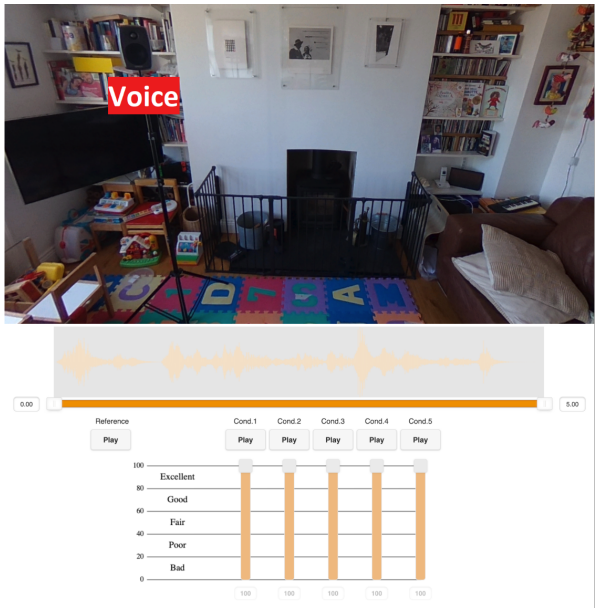

Figure 3. Interface displayed to participants in the online experiment. The trial corresponds here to listener position 1 and a speech signal. A fixed picture with a single visual sound source is presented to the participant.

3.3 Experiment II: an online-based perceptual evaluation

3.3.1 Design & tools

The experiment task was designed with *webMUSHRA* [14] and hosted on a local server. The use of the *webMUSHRA* allows MUSHRA testing to be carried out within a web browser while being compliant with the ITU-R Recommendation [13].

3.3.2 Participants & Procedure

Twenty participants (12 males, 8 females) were recruited for this experiment; prior to the test, participants were given screen instructions to use the whole scale provided to evaluate the stimuli, and were informed that the test required the use of headphones. Participants were also asked to use *Sennheiser HD650* headphones if they had access to them, and to adjust the volume to a comfortable level (a test signal was provided for this). The interface displayed to participants during the test is shown in Figure 3. In this version of the experiment, no head-tracking is used. A fixed picture corresponding to the receiver position is displayed with the MUSHRA interface (see Figure 3).

4. RESULTS

The design defined two experiments that were evaluated by two different groups of 20 participants each. The same auditory stimuli were evaluated in both experiments. Therefore, two separate analyses are presented in this section. The results of both experiments are reported consecutively.

The analyses used for both experiments are of the same nature. In each experiment, inferential analysis was performed through a repeated measures analysis of variance (RM-ANOVA). The ratings of each participant on each stimulus were considered the dependent variable. For each experiment, the RM-ANOVA was conducted with MODEL (6 models), STIMULUS (2 types of stimulus) and POSITION (3 listener positions) as within-subject factors. A significance value of $\alpha = 0.05$ was used. The effects of the factors quantified for Experiment I are reported in Table 2 and for Experiment II in Table 3.

The main difference between both experiments highlighted by these initial statistical analyses is that the effect of the stimulus type is significant in Experiment II and not

Effects	*df*	*F*	*p-value*
Model	5	93.2	<0.001
Position	2	16.5	<0.001
Stimulus type	1	0.563	0.463
Model * Position	10	5.28	<0.001
Model * Stimulus	5	2.122	0.135
Model * Stimulus * Pos.	10	2.58	0.006

Table 2. Within subject effects quantified by the RM-ANOVA applied to the ratings of the GR models in Experiment I (VR). Two effects have been shown to be not significant: the main effect of the stimulus type and the cross-effect of the model used and the stimulus type.

Effects	*df*	*F*	*p-value*
Model	5	212	< 0.001
Position	2	5.69	0.008
Stimulus type	1	6.54	0.021
Model * Position	10	7.03	< 0.001
Model * Stimulus	5	4.76	< 0.001
Model * Stimulus * Pos.	10	2.15	0.024

Table 3. Within-subject effects quantified by the RM-ANOVA applied to the ratings of the GR models in Experiment II (Online).

in Experiment I.

Ratings associated with GR models in Experiment I and Experiment II are shown, respectively, in Figure 4 and 5. Descriptive analysis illustrates that models with a larger amount of geometrical detail obtained higher ratings. The highest rating is consistently obtained with the reference, and the lowest with the anchor.

For each type of model and experiment, post-hoc tests were run to highlight differences between models' ratings in the different positions, or stimulus types. A Bonferroni correction was applied to these post-hoc test results. The models in a shaded area (Figure 4 and 5 are not significantly different from each other, according to the results of the post-hoc tests.

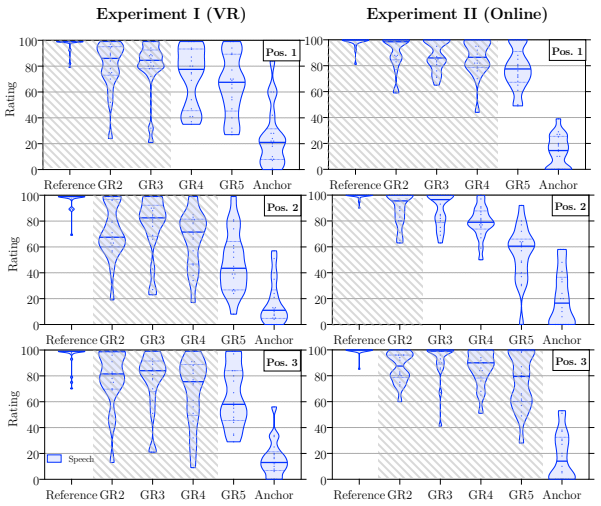

Figure 4. Results from Experiment I (VR, left) and Experiment II (Online, right) for speech stimuli represented by violin plots, which show the probability density of the data and the mean (horizontal line). Separated per listener position (top to bottom). The grey area represents a portion of the models whose ratings are not significantly different from each other according to post-hoc tests.

5. DISCUSSION

Our overall aim is to investigate the trade-off between computational requirements and perceptual accuracy of binaural auralisations. The focus of this paper is on the comparison of the results between two versions/conditions of the same experiment, a VR lab-based one and an online web-based one.

The evaluation paradigm based on MUSHRA is highly inspired by a study by Engel et al. [10]. The purpose of this study was to investigate the trade-off between computational complexity and perceived quality in binaural Ambisonics-based reverb. Here, the MUSHRA paradigm has been demonstrated to be efficient in detecting small perceptual changes, such as spatial aspects of the reverberation.

Results of both perceptual experiments show that for position 1 no significant differences can be found between the reference and the GR2-GR3-GR4 models. At this distance, with the direct-to-reverberation energy ratio being lower, no significant differences between models could be

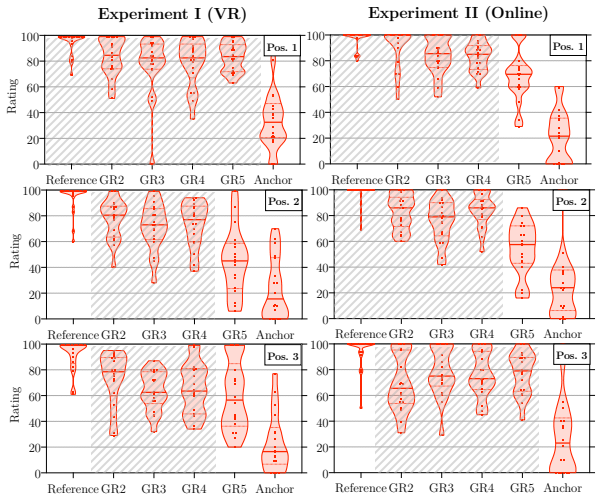

Figure 5. Results from Experiment I (VR, left) and Experiment II (Online, right) for music stimuli represented by violin plots, which show the probability density of the data and the mean (horizontal line). Separated per listener position (top to bottom). The grey area represents a portion of the models whose ratings are not significantly different from each other according to post-hoc tests.

detected by participants, and this was the same for the VR and Online testing conditions.

Results from Experiment I (VR), for positions 2 and 3, show that removing small surfaces (GR2, $< 0.1\text{m}^2$) has a perceptual impact and ratings are significantly different from the reference and GR5. Removing furniture and small surfaces (GR4, all furniture and surfaces $< 0.4\text{m}^2$) does not have a significant impact. These findings align with a study by Abd Jalil *et al.* [15], which suggests that removing small surfaces has little effect on acoustic parameters in open-plan office rooms. As long as a room's larger surfaces are presented, geometrical acoustics can provide acceptable auralisation results.

Results from Experiment II (Online) for positions 2 and 3 show smaller differences between the reference and the geometrically-reduced models. More precisely, the statistical analyses revealed less consistency, in Experiment II, in the similarity between models' ratings across positions and stimulus types (see Figure 4 and 5). For instance, in Position 2, no significant differences are found between the ratings associated with the GR2 model and the reference. In Experiment II, in position 3, no significant differences between the shoebox model and the other reduced models are identified.

It can be argued that the presence of dynamic cues could have led more easily to the detection of slight differences between models in the VR experiment. Therefore, while participants did not distinguish the reference from the first GR model in Experiment II (online), participants of Experiment I could.

Therefore, despite including identical stimuli, results from these experiments can lead to slightly different conclusions. Results from Experiment II (online) hint that geometrically-reduced models could be perceptually similar to the reference model, while Experiment I shows that from a certain distance, these models are always significantly different.

Moreover, a larger inter-subject variance is observed in Experiment I (VR) when compared with Experiment II (Online), which could be explained considering the additional information provided to the listener through the head-tracking aspect of the VR protocol, as well as the more controlled experimental conditions, which allowed listeners to better focus on the task. Different strategies could have been used by participants, for example, using different head movements and dynamic visual aspects, resulting in higher inter-subject variance.

6. CONCLUSION

In this paper, the trade-off between perceived quality and computational complexity was explored for ray-traced binaural auralisation, specifically looking at geometry reduction/simplification. Two different experimental conditions were compared, one lab-based and one web-based.

It was predicted that the geometry of the reproduced room may be slightly reduced without impacting the perception of its reverberation. Yet, even small geometric simplifications from the reference seemed to have a significant perceptual impact in the lab-based experiment, while for the web-based condition results showed some differences. Under the hypothesis that results for Experiment I are more reliable, they reveal that all reduced models (except for the shoebox one) had similar ratings, but were still significantly different from the reference. This suggests that geometry reduction has a perceptual impact in terms of resulting in an identifiable difference from the reference, but such a difference is not identifiable anymore between versions that are geometrically-reduced to different extents. However, results from both experiments

are partly conflicting, notably on the similarity between the reference model and reduced models. Results from the online experiment did not lead to the same conclusions, being significantly more similar across the different geometrically-reduced conditions and the reference.

This study suggests that results from online experiments implying the perception of reverberation should be interpreted carefully, especially when dynamic aspects can be assessed only in the lab.

7. ACKNOWLEDGMENTS

This study was made possible by support from SONICOM (`www.sonicom.eu`), a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017743.

8. REFERENCES

[1] T. Wendt, S. van de Par, and S. D. Ewert, "Perceptually plausible acoustics simulation of single and coupled rooms," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3178–3178, 2016.

[2] C. Kirsch, T. Wendt, S. Van De Par, H. Hu, and S. D. Ewert, "Computationally-efficient simulation of late reverberation for inhomogeneous boundary conditions and coupled rooms," *Journal of the Audio Engineering Society*, vol. 71, no. 4, pp. 186–201, 2023.

[3] B.-I. Dalenbäck, "Whitepaper: What is geometrical acoustics (ga)?," technical report, CATT, 2021.

[4] D. Schröder and M. Vorländer, "Raven: A real-time framework for the auralization of interactive virtual environments," in *Forum acusticum*, pp. 1541–1546, Aalborg Denmark, 2011.

[5] C. L. Christensen, "Odeon, a design tool for auditorium acoustics, noise control and loudspeaker systems," in *Proceedings of Reproduced Sound 17: Measuring, Modelling or Muddling*, pp. 137–144, 2001.

[6] P. Eyal, R. David, G. Andrew, E. Zak, and D. Ekaterina, "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, pp. 1–20, 2021.

[7] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[8] F. Brinkmann, A. Lindau, S. Weinzierl, M. Müller-Trapet, R. Opdam, M. Vorländer, *et al.*, "A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations," *Journal of the Audio Engineering Society*, vol. 65, no. 10, pp. 841–848, 2017.

[9] F. Zotter, M. Frank, and H. Pomberger, "Comparison of energy-preserving and all-round ambisonic decoders," *Fortschritte der Akustik, AIA-DAGA,(Meran)*, 2013.

[10] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, and L. Picinali, "Perceptual implications of different ambisonics-based methods for binaural reverberation," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 895–910, 2021.

[11] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3d tune-in toolkit: An open-source library for real-time binaural spatialisation," *PloS one*, vol. 14, no. 3, p. e0211899, 2019.

[12] V. Hansen and G. Munch, "Making recordings for simulation tests in the archimedes project," *Journal of the Audio Engineering Society*, vol. 39, no. 10, pp. 768–774, 1991.

[13] I. BS, "1534-3,"method for the subjective assessment of intermediate quality level of audio systems,"," *International Telecommunication Union, Geneva, Switzerland*, 2015.

[14] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[15] N. A. Abd Jalil, N. C. Din, N. Keumala, and A. S. Razak, "Effect of model simplification through manual reduction in number of surfaces on room acoustics simulation," *Journal of Design and Built Environment*, vol. 19, no. 3, pp. 31–41, 2019.