OPEN ACCESS Freely available online

S. Agrawal S. Bech, K. Bærentsen, K. De Moor, and S. Forchhammer, "Method for Subjective Assessment of Immersion in Audiovisual Experiences"
J. Audio Eng. Soc., vol. 69, no. 9, pp. 656–671, (2021 September).
DOI: https://doi.org/10.17743/jaes.2021.0013

Method for Subjective Assessment of Immersion in Audiovisual Experiences

SARVESH AGRAWAL,[1,2]* *AES Student Member*, **SØREN BECH,**[1,3] *AES Fellow*,
(sraj@bang-olufsen.dk) (sbe@bang-olufsen.dk)

KLAUS BÆRENTSEN,[4] **KATRIEN DE MOOR,**[5] **AND SØREN FORCHHAMMER**[2]
(klaus@psy.au.dk) (katrien.demoor@ntnu.no) (sofo@fotonik.dtu.dk)

[1]*Bang & Olufsen a/s, 7600 Struer, Denmark*
[2] *Technical University of Denmark, Department of Photonics Engineering, 2800 Lyngby, Denmark*
[3]*Aalborg University, Department of Electronic Systems, 9220 Aalborg, Denmark*
[4]*Aarhus University, Department of Psychology, 8000 Aarhus C, Denmark*
[5]*Norwegian University of Science and Technology, Department of Information Security and Communication Technology, 7491 Trondheim, Norway*

Studying immersion in audiovisual experiences can help technologists deliver engaging and enhanced experiences. As a first step toward this goal this paper details an investigation conducted to establish an experimental paradigm for quantifying immersion and determining the influence of immersive tendency (susceptibility to become immersed) on immersion. A balanced incomplete block design was employed where 21 assessors rated 15 commercially available stimuli (representative of the highest quality encountered in domestic AV applications) without repetitions and simultaneous comparisons. The assessors were instructed to rate immersion on a graphic line scale and document their familiarity with the content. A questionnaire was administered to measure the immersive tendency after the rating experiment. The results show that the assessors can comprehend the description of immersion and follow the experimental protocol. It is found that immersion is a graded experience and the correlation between immersive tendencies and immersion ratings is predominantly statistically insignificant. The experimental paradigm presented in this paper can form the framework for assessing immersion and developing novel methods to thoroughly explore the concept of immersion in audiovisual experiences.

0 INTRODUCTION

The adoption of spatial audio reproduction for domestic audiovisual applications (e.g., MPEG-H audio in televisions [1], hearables with spatial audio rendering capabilities, virtual reality) is on the rise. Coupled with advancements in visual technology that facilitate higher resolution (e.g., higher pixel-density), enhanced color reproduction (e.g., extended color spaces and improved chroma subsampling [2]), and greater dynamic range (e.g., improved brightness regulation) among others, the paradigm for domestic audiovisual experiences is changing swiftly. In this context subjective tests play a vital role in characterizing audiovisual experiences and understanding the influence of the physical properties of the system and the signal on human perception. Additionally the role of human characteristics needs to be better understood in this respect. Affective/hedonic measurements help obtain an overall impression of the experience as they go beyond the senses by accounting for the cognitive factors (e.g., mood, context, expertise, and expectation). However systematic efforts are needed toward establishing a thorough understanding of the factors influencing affective measures and to exploit those insights for enabling enhanced experiences for the users.

Immersion is a cognitive concept that attempts to capture mental engagement in an experience. It has been a central topic in virtual reality and video game studies but is applicable in various domains such as literature and film as well. The subject of immersion has been covered at length

*Correspondence should be addressed to Sarvesh Agrawal sraj@bang-olufsen.dk

in [3–6]. Therefore only the points critical for conducting experimental investigations are discussed in this work. In a study conducted by Agrawal et al. [4], they synthesized the following definition of immersion:

> Immersion is a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world.

Immersion is viewed as a "normal occurrence of focused attention during waking consciousness" [4]. The five factors that can influence immersion were determined to be 1) the system, 2) narrative (content), 3) environmental and contextual conditions, 4) individual factors, and 5) interaction between the individual and experience (significance of the content, alignment of goal and motivation, etc.) [4]. It has been hypothesized that some individuals are more susceptible to experiencing immersion [7] than others. This is attributed to their immersive tendency, which is defined as "an individual's predisposition to get immersed" [4]. Nevertheless the role of the individual in experiencing immersion is, to date, poorly understood and unquantified.

The fundamental assumption underlying the desire to study immersion is that more immersive experiences are preferred. Thus a deeper understanding of immersion and the influencing factors can help technologists enable the delivery of more engaging, adaptable, and enhanced experiences. This is our primary motivation and the ultimate goal of studying immersion.

The challenge with measuring immersion is a lack of a reliable experimental framework. Hence a method for the subjective quantification of immersion while considering the implications for the experimental paradigm must be established and tested as a first step. Due to a lack of knowledge about the nature of immersion, it must be determined whether it is a binary or graded experience. This is critical for developing the conceptual understanding as well as selecting the appropriate scaling procedure. It was decided to concentrate the experimental investigation on the influence of immersive tendency on immersion ratings over other influencing factors since the role of the individual is paramount in experiencing immersion [8]. This paper builds on the work presented in [4] by seeking answers to the following research questions:

RQ1: How can immersion in an audiovisual experience be quantified through subjective testing?
RQ2: Is immersion a binary (all-or-nothing) or graded experience?
RQ3: What is the influence of immersive tendency on immersion ratings?

The rest of this paper is organized as follows. The experimental strategy is explained in Sec. 1. The experimental framework pertaining to RQ1, including the reproduction setup and program material, is presented in Secs. 2 and 3. Data analysis and the empirical findings used to answer RQ2 and RQ3 are detailed in Sec. 4. Sec. 5 discusses the results along with the limitations and avenues for future work. The conclusion is presented in Sec. 6 and supplemental information is provided in APPENDIX A.

1 EXPERIMENTAL STRATEGY

This section states the implications and results from previous studies along with the high-level decisions that guided this investigation. The specific details about the experimental setup and procedure are covered in the following sections.

Current methods for quantifying immersion can be classified as subjective tests (psychometric) or objective tests (behavioral and physiological measures). Zhang [6] provides a detailed discussion of the merits and demerits of these methods. A lack of established links between objective measures and immersion restricts us to subjective measures. Since the goal in this study was to obtain information on the overall immersion in the experience with the least possible complexity, subjective measures were deemed to be more suitable for the task. Please refer to Zhang [6] and Agrawal et al. [4] for a review on the existing experimental paradigms for evaluating immersion.

Questionnaires are the predominant tool for self-reporting immersion by means of quantitative measures. However they can be quite lengthy (greater than 25 questions) and administering them after each experience drastically increases the experimental time. Additionally questionnaires are based on a limited set of pre-defined questions and thus fail to capture any unexpected aspect of the immersive experience [6]. In their study on the experience of immersion in games, Jennett et al. [3] compared the results obtained from a questionnaire developed to assess immersion to that of an individual question on immersion. They demonstrated that "people can reliably reflect on their own immersion in a single question" [3] when asked to grade immersion on a discrete 10-point scale. This is an important result that helped guide the current investigation to be conducted as a rating experiment. Importantly, asking participants to rate the immersion in the experience helps capture unexpected aspects of the immersive experience (unlike questionnaires, the rating task can capture all influencing factors relevant to the individuals) and reduce the workload for assessors, making them more suitable for evaluating multidimensional concepts such as immersion.

Agrawal et al. [4] discussed the implications for designing appropriate experimental paradigms to investigate immersion. Following their recommendation, a graphic line scale [9] (continuous) was selected as the response format with two anchor points—not very immersive and very immersive. The scale (shown in Fig. 1) offers the participants infinite steps (in theory) to indicate their immersion. A lack of numbers and/or verbal anchors (except those toward the ends) reduces bias effects but it makes it difficult for the subjects to use. Nevertheless the subjects adapt to the scale after initial usage and do not require further instructions [9].

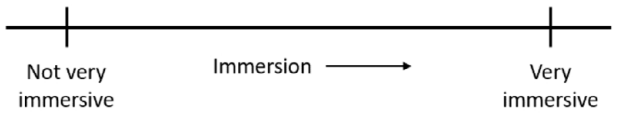


Fig. 1. The graphic line scale used for rating immersion. A similar scale with different anchor points was used for the immersive tendency questionnaire.

Please see [10, 9] to learn more about the scale, its usage, and data analysis.

The assessors were not permitted to switch between stimuli to avoid disrupting the psychological state when occupied in an experience. Further, they were subjected to each stimulus only once to limit the effect of repetition since the effects of repetition and familiarity are not well understood. Distractor tasks were incorporated in the experimental design as it is suspected that individuals may require time between consecutive experiences to return to their initial psychological state (see Sec. 3.2).

The considerations for selecting stimuli to conduct investigations on immersion include incorporating stimuli of different narratives, lengths spanning five to twelve minutes, and those that require no prior knowledge of the narrative [4]. To this end a set of stimuli from different genres and of varying lengths were to be selected. An important consideration was choosing narratives that can evoke spatial (absorption in exploration), temporal (curiosity to know what happens next), and emotional (attachment to characters or recollection of memories of emotional relevance) immersion as categorized in [11]. Since the results are dependent on the stimuli, an attempt was made to pick stimuli that cover the entirety of the immersion range. This is particularly important for RQ2.

There are no objective benchmarks or technical specifications that could be used to select the stimuli. Thus an effort was made to pick stimuli from different genres and narratives and those that were believed to elicit different emotions. The stipulation of selecting stimuli that require no prior knowledge of the narrative severely restricted the content universe. Thus it was decided to provide the participants with a short synopsis of the narrative before each experience. These synopses were constructed to include the relevant information required to make sense of the story without disclosing the details about the experience. A list of synopses is provided in APPENDIX A.2.

A non-exhaustive survey of the domestic media landscape hinted that the integration of ultra-high-definition (UHD), high dynamic range (HDR) video, and spatial audio is being actively embraced by streaming platforms, broadcasters, and movie studios. As the paradigm shifts away from high-definition video with stereo audio it was decided to use stimuli representative of this shift. To accommodate the stimuli consisting of spatial audio and UHD visuals, a 7.1.4 audio rendering system was used in conjunction with a domestic television screen capable of supporting UHD HDR visuals (see Sec. 2.2).

A pilot test that employed a randomized complete block design was used to assess the experimental protocol and aid with the selection of stimuli. The test where six subjects graded five stimuli revealed that the assessors experienced fatigue around the 75 min mark. The implications for the experimental paradigm and the need to test a large number of stimuli to determine the nature of immersion indicated that a balanced incomplete block (BIB) design would be the optimal experimental design. Since a simple BIB design would require an impractical number of subjects, precision was traded based on practical experience to reduce the number of participants [12].

A reduced design with 21 participants (blocks) where each participant grades a subset of five stimuli (from a selection of 15) was employed. The stimuli were allocated to the blocks such that each stimulus appeared only twice with every other stimulus in the blocks (detailed table is presented in APPENDIX A.1). Thus only two out of the 21 participants would grade any given pair of stimuli. In total 21 participants graded five stimuli each, yielding 105 total trials for the entire stimulus set. As the sole tool for gauging immersive tendency, an existing questionnaire (Wimter and Singer's ITQ [7]) was used to assess immersive tendency.

2 EXPERIMENTAL SETUP

The choice of program material, physical setup of the reproduction system, and environmental conditions are explained in this section.

2.1 Program Material

A major obstacle in conducting experiments with audiovisual media boasting UHD HDR video and spatial audio is an apparent lack of free or open-licensed content. Thus commercially available content had to be used for this experiment. Content released on Blu-ray for home entertainment that has UHD HDR visuals along with Dolby Atmos or DTS:X[1] audio sufficed for the audiovisual requirements. A major drawback was that the stimuli had to be presented directly from the Blu-ray to each participant due to copyright laws and content protection. The audiovisual signal chain for reproduction is detailed in the next sub-section and illustrated in Fig. 2.

The 15 stimuli selected for this investigation as per the implications mentioned in the preceding section are listed in Table 1. All excerpts were in UHD resolution with either HDR10 or Dolby Vision and were presented as such.[2] The native aspect ratio and chroma sub-sampling were maintained through reproduction. The audio was rendered as 7.1.4 except for the excerpts from *The Revenant* (excerpt C and L in Table 1) that were rendered as 7.1, since they were available in DTS-HD Master Audio format. The audio and video signals were not processed at any stage.

[1]Dolby Atmos and DTS:X trademarks/service marks are the property of their respective owners.

[2]While an attempt was made to select stimuli mastered in 4K/UHD, many of them have been mastered in 2K and upscaled.

Table 1. Audiovisual excerpts used in the experiment.

Excerpt	Content	Genre	UK year of release	Timecode
Example	*Earth: One Amazing Day*	Nature documentary	2018	00:08:50 – 00:16:49
sA	*Mission: Impossible – Fallout*	Action/Adventure	2018	01:12:31 – 01:16:09
sB	*Apocalypse Now – Final Cut*	War/Drama	2019	02:12:45 – 02:20:24
sC	*The Revenant*	Western/Adventure	2016	01:53:09 – 01:58:24
sD	*Fantastic Beasts: The Crimes of Grindelwald*	Fantasy/Adventure	2019	01:34:50 – 01:42:47
sE	*Dynasties: Lion*	Nature documentary	2018	00:16:11 – 00:20:00
sF	*The Darkest Hour*	War/Drama	2018	00:41:09 – 00:48:00
sG	*Murder on the Orient Express*	Mystery/Drama	2018	00:00:53 – 00:08:31
sH	*Braveheart*	War/Drama	2018	00:22:05 – 00:28:36
sI	*Ad Astra*	Sci-fi/Thriller	2020	01:15:23 – 01:21:17
sJ	*Earth: One Amazing Day*	Nature documentary	2018	00:57:50 – 01:02:39
sK	*Spider-Man: Into the Spider-Verse*	Animation/Action	2019	00:02:32 – 00:13:42
sL	*The Revenant*	Western/Adventure	2016	00:02:30 – 00:14:59
sM	*Sicario*	Crime/Action	2018	00:01:00 – 00:12:53
sN	*Earth: One Amazing Day*	Nature documentary	2018	00:47:47 – 00:51:37
sO	*Earth: One Amazing Day*	Nature documentary	2018	00:16:50 – 00:22:35

Note.
1. The copyright for the content used in this experiment is held by the respective parties. No files were copied or stored during any stage of experimentation.
2. The genres have been obtained from IMDb (https://www.imdb.com) and selected to reflect the primary genre of the content.
3. The year of release represents the release on Blu-ray. The actual release year may differ.
4. The length of the excerpts ranges from 4–12 minutes approximately.
5. Please refer to the table in APPENDIX A.2 for the narrative synopsis provided to the participants before each experience.

2.2 Reproduction System

As communicated in the preceding subsection, the content had to be presented live to the assessors due to legal constraints. Thus the protocol required control over the audio and video for smooth transitions to present the excerpts. A Sony UBP-X500 UHD Blu-ray disc player was used to read the discs and feed the audiovisual signal to a Marantz AV7704 audiovisual processor. This processor was used to decode the audio (Dolby Atmos/DTS-HD Master Audio) to 7.1.4 analog audio channels while the video was passed through to the Roland V-600UHD (UHD HDR capable and high-bandwidth digital content protection compliant) video switcher.

Ten Genelec 8340a and one Genelec 7380a speakers were distributed on a hemisphere of 2-m radius around the listener for sound reproduction. The speakers were placed at ± 30, ± 90, and ± 135-degree angles on the horizontal plane at ear height ($\sim$1.15 m) while the subwoofer was placed at -45 degrees on the floor. The elevation loudspeakers were mounted on the ceiling at ± 45 and ± 135 azimuth with a 50-degree elevation. A phantom center was incorporated to accommodate the screen. The placement of speakers was compliant with the guidelines from Dolby for domestic setups [13].

The speakers were time aligned and level calibrated using the Genelec loudspeaker manager (maximum time difference of 0.3 ms and level difference of 0.6 dB). The

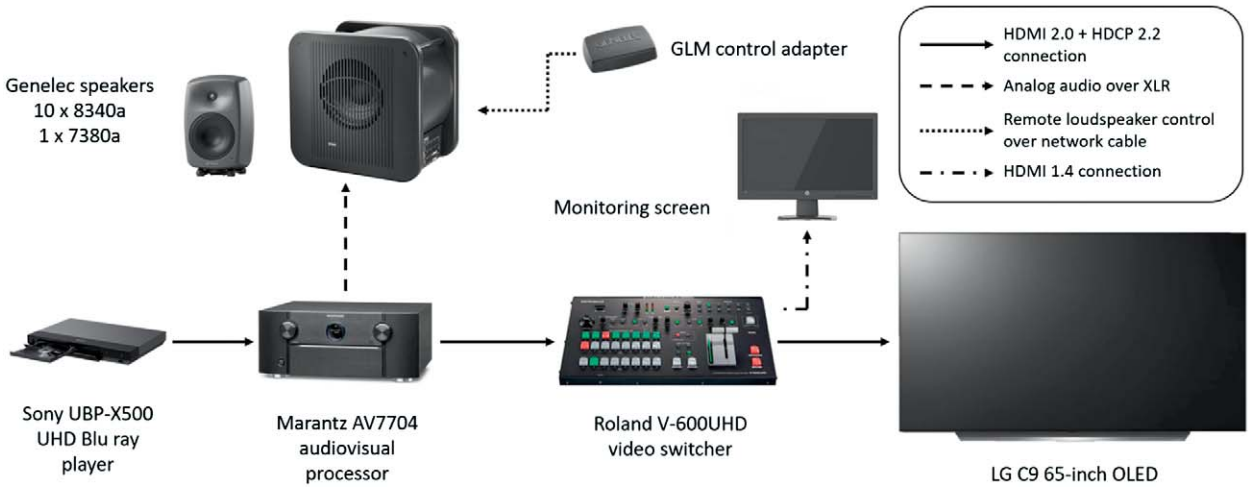


Fig. 2. Signal chain used to present the audiovisual stimuli. All equipment except the LG screen and the speakers were located outside the room.

program material was not loudness normalized. Instead the reproduction level was varied among stimuli such that they were approximately equally loud (determined by ear) at the listening position. Two experienced listeners auditioned all excerpts to confirm that they were at a comfortable listening level and intelligibility was maintained during the quieter sections.

The video feed from the audiovisual processor was received by the video switcher. The primary purpose of integrating the video switcher was to allow for video fades to and from the blank screen. A 65-inch LG C9 organic light emitting diode (OLED) screen was used to present the visuals to the participants. The screen was centered such that it yielded near zero horizontal and vertical viewing angles for an optimal viewing experience. A viewing distance of 2 m (same as the listening distance) was chosen in accordance with the design viewing distance recommended in ITU-R BT.2022 [14] for UHD resolution. Although the screen was capable of producing 700–800 nits, the brightness was limited to approximately 120 nits (explained in the following subsection). System colorimetry measurements taken pre and post-experiment of color coordinates using three levels of gray at three different positions on the screen revealed no substantial drifts warranting attention. The screen settings were tuned such that the chromaticity coordinates were close to the D65 value [2]. Further, two experts fine-tuned the settings for optimum reproduction. No time offset was required between the audio and video.

2.3 Environmental Conditions

The experiment was conducted in a climate-controlled IEC 60268-13 standardized listening room [15] of 6.5 m x 4.1 m x 2.5 m. Acoustically transparent curtains were used to hide the speakers to limit the visual influence. To balance the judder of 24 full frames per second signals due to high OLED panel brightness while harnessing the HDR capabilities of the system, the environmental illuminance was lowered below 10 lux (measured perpendicular to the screen). This reduced the judder by allowing for lower panel brightness. The assessors were alone in the space during the experiences and tasks.

3 EXPERIMENTAL PROCEDURE

3.1 Assessors

Twenty-one assessors participated in the experiment where each assessor was randomly assigned to one of the 21 blocks. The participant pool comprised of 15 males and six females with a mean average age of 37.7 years ($\sigma = 14.28$). The youngest participant was 18 and oldest participant was 59. Since immersion is determined to be a cognitive concept, subjects were not required to be experts on audiovisual assessment. Nevertheless as participants were Bang & Olufsen employees there was a blend of experienced and inexperienced assessors. The experienced participants were from the acoustics development department who have had critical listening and viewing training. The inexperienced assessors did not have any training but may have participated in previous audiovisual tests. The participants reported having normal visual and auditory acuity for their age. All 21 participants completed the experiment.

3.2 Distractor Tasks

The order of presentation of stimuli can bias the results due to the effect of preceding stimulus on the assessment of succeeding stimulus. While randomizing the order of the presentation can help mitigate the bias, there are no recommendations regarding the pause in time between the presentation of stimuli in every block. It is suspected that it takes time for individuals to revert to their base or initial state (their psychological state prior to the experience) after an audiovisual experience. Given the lack of conclusive evidence we decided to incorporate short tasks between the presentation of stimuli. The tasks serve the fundamental purpose of engaging participants to shift their attention away from the previous stimuli while introducing a pause.

Four distractor tasks that require active participation were chosen: an 11-piece LEGO[3] puzzle where the assessors were asked to make a unicorn (no instructions or hints were provided), a memory task (7 x 6 tiles) where the goal was to uncover the tiles to find pairs of animals, a matchstick rearrangement puzzle, and a picture interpretation scenario. The memory and matchstick tasks were played as a game on an iPad 2018. For the picture interpretation task the participants were asked three questions to guide their thoughts—1) What is happening in the picture?2) What do you see that makes you say that?3) What more can you find?

A time limit of four minutes was enforced for each task. The participants were instructed to continue on to the following levels if they finished the memory and matchstick tasks before time. The correct solution was explained to the participants for the matchstick task before progressing. Snapshots from the four tasks are provided in APPENDIX A.4.

3.3 Immersive Tendency Questionnaire

Witmer and Singer's [7] immersive tendency questionnaire (ITQ) has been widely used [16–19] to assess immersive tendency. They reduced the original questionnaire to 19 questions (see Table 3) where most items measure involvement and the ability to focus on an activity. Nevertheless a few modifications were made to the response format used in the reduced questionnaire for this experiment. First, a graphic line scale (same as for the rating experiment) was preferred over the categorical seven-point scale to obtain continuous data and save the participants the effort to learn another scale. Second, the middle word anchor from the categorical seven-point scale used in the original questionnaire was dropped as the distribution of scores may exhibit clustering around the verbal anchors [20]. Finally, where applicable, the "rarely/often" pair of word anchors was used over "never/often," as the words should be perfect antonyms (similar change made in [18]).

[3]The LEGO group owns the intellectual property rights to the LEGO name.

The questionnaire was administered for all 21 participants. The order of questions was randomized for each participant.

3.4 Test Procedure

The test procedure comprised of two phases: the rating experiment and ITQ questionnaire. A single session of approximately 90 minutes was required to complete the experiment.

The participants were asked to confirm their visual and auditory acuity to the best of their knowledge prior to the test. They were then given written and verbal instructions for the experiment. For the rating experiment, the following description of immersion was provided:

> *Immersion, also known as deep mental involvement, can be described as being mentally lost (absorbed) in the experience. Immersion is encountered when the experience is involving and absorbs you mentally by capturing your attention. For example, immersion may be experienced when reading a book, playing video games, watching a movie, etc.*

The assessors were instructed to rate *overall* immersion in the experience on the graphic line scale and report if they had experienced the audiovisual content before (selecting yes/no on the response sheet) for each of the five experiences. Additionally they were encouraged to write their comments about the experience. Before commencing the rating experiment the assessors were subjected to an experience that could be immersive, in an attempt to exemplify the concept of immersion. It was explicitly stated that the example was solely for illustration purposes, may not be immersive for them, and should not be considered a reference for the experiment. The example excerpt (see Table 1) was determined by the first author based on experience from the pilot experiment. It was emphasized that there are no correct answers and that the participants are not compelled to use the entire scale.

A synopsis of the narrative was provided to the participants before each experience. One distractor task from those described in Sec. 3.2 was randomly chosen to be performed between consecutive experiences. The entire test was administered as a pen-and-paper test and the response sheets were changed between experiences to avoid comparisons. Immersive tendency data was collected at the end of the experiment. All data collection was performed within three weeks.

4 DATA ANALYSIS AND RESULTS

Ratings obtained for immersion and the immersive tendency questionnaire were converted to scores in the range of 0 to 15 (up to one decimal point) by measuring the physical distance in millimeters. Visualization of data, results from analysis of variance, post-hoc comparisons between stimuli, analysis of questionnaire data, and correlation analysis are presented in this section.

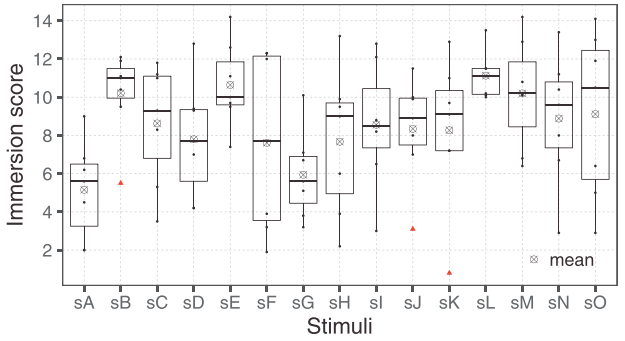


Fig. 3. Box plot of responses for each stimulus where the red triangles indicate outliers.

4.1 Rating Data

Twenty-one subjects rating five experiences each resulted in 105 total observations (seven for each stimulus listed in Table 1). This data is visualized as box plots for each of the 15 stimuli in Fig. 3. Remarkable differences between the mean scores for stimuli can be observed (e.g., sA vs. sL). Some stimuli exhibit high inter-participant agreement (e.g., sB) while some are highly contested (e.g., sF, sO). Three groups of stimuli appear to emerge based on immersion scores: stimuli where assessors experienced relatively low degree immersion (sA, sG), high degree immersion (sB, sE, sL), and all other stimuli. The observations do not indicate any obvious association with genre or the length of the experience. Fig. 3 shows that the participants used approximately the entire range of the scale among the stimuli.

It should be noted that the above plot is based on raw data. The effects related to the participants' use of scale (level, scaling, and disagreement effects [21]) are confounded in the scores. Hence the plot should not be accepted at face value. The scores were not standardized for the purpose of this investigation. Instead estimated marginal means were used rather than marginal treatment means (descriptive means) for estimating the treatment effects in BIB analysis [12]. The detected outliers were inspected using the comments provided by the subjects. In total 7 points were removed from the following analysis, yielding 98 total data points (see APPENDIX A.3 for details).

4.1.1 Modality of Data

The nature of immersion, i.e., whether it is a graded experience as opposed to being a binary (all-or-nothing) experience, has not been tested empirically before (RQ2). If immersion is a binary experience irrespective of the stimuli, the distribution of immersion scores should by bimodal. Thus, to determine if the distribution of scores is unimodal (graded) or multimodal, Hartigan's dip test statistic (HDS) was calculated and tested for significance using raw immersion scores. HDS was chosen as it is more robust than the bimodality coefficient [22, 23] and the results are reliable and stable [24].

The dip statistic in HDS is the maximum difference between the empirical distribution function and the unimodal

distribution function that minimizes the difference between the two distributions. Thus large HDS values signal departure from unimodality. A uniform distribution is chosen as the unimodal distribution as it is argued that the dip is asymptotically larger for the uniform distribution in comparison to other unimodal distributions [25]. The dip test value obtained from the empirical distribution is compared iteratively to the dip test statistic of bootstrapped samples that are generated randomly from a uniform distribution to compute the *p*-value. Thorough mathematical description for calculating the dip statistic and testing for significance is presented in [25, 26].

The null hypothesis for HDS is that the distribution of data is unimodal. *P*-values lower than the specified significance level indicate multimodality whereas higher values indicate unimodality of data. The average *p*-value (from 100 calculations[4]) was found to be 0.862 ($\sigma = 0.04$) for 98 replicates (bootstrapped samples) at $\alpha = 0.05$. This shows that immersion is a unimodal experience (RQ2).

4.1.2 Analysis of Variance (ANOVA) and Pairwise Comparisons

Analysis of variance was used to quantify the effect of the stimuli on immersion scores after verifying that the assumptions were satisfied. The model for analyzing data from a BIB is identical to that for randomized complete block design [12]. The standard model considers the block (participants) and treatment (stimuli) to be fixed factors. This model was amended to include the subjects as a random factor (random intercepts) to account for the random variability among participants and improve the generalizability of the results to the population. It was hypothesized that familiarity with the content could influence immersion ratings. However familiarity (dichotomous and dummy coded) could not be included as a covariate since the assumptions for analysis of covariance were violated.[5] Thus the final model was limited to the block and treatment effects.[6] The variances for the treatment levels were checked using Levene's test and the trials were independent of each other by design.

Type III analysis of variance with Satterthwaite's method revealed that there was a highly significant effect of the stimuli on immersion scores, $F(14, 74.82) = 3.32$, $p < 0.001$. This is an important finding as it shows that the participants were able to discriminate between the stimuli. The subject factor was found to be marginally significant at $p = 0.099$ for the model fitted with restricted maximum likelihood (REML). Since the participants graded a small subset of stimuli, some blocks could have consisted of stimuli with only high/low immersive potential.[7] Nonetheless it is not certain that the allocation of stimuli to the blocks explains all of the variance since the scaling effects and the random variability among the subjects are contained within the subject factor. A disadvantage is that the interaction between subjects and stimuli could not be investigated due to a lack of sufficient degrees of freedom.

Table 2. Significant ($\alpha = 0.05$) and marginally significant pairwise comparisons between pairs of stimuli.

Contrasts	Estimate	SE	df	t-ratio	P-value
sA - sB	−5.95	1.55	74.1	−3.85	0.018*
sA - sE	−5.83	1.55	76.0	−3.76	0.023*
sA - sL	−5.98	1.49	71.7	−4.00	0.012*
sA - sM	−5.49	1.55	75.4	−3.54	0.045*
sG - sB	−5.28	1.55	72.0	−3.38	0.072
sG - sL	−5.25	1.48	73.6	−3.55	0.045*

*$p < 0.05$

The residuals for the model, $W = 0.99$, $p = 0.710$ were approximately normally distributed as determined by the Shapiro-Wilk test. The distribution of the residuals had skewness of −0.069 ($SE = 0.24$) and the excess kurtosis was −0.22 ($SE = 0.48$). The assumption of homoscedasticity was not violated.

Following the analysis of variance it was of interest to identify the differences between the levels of the stimuli. The least square means were estimated to make pairwise comparisons between stimuli. The degrees of freedom were approximated using the Kenward-Roger method and the *p*-values were adjusted using Tukey's method. Out of the 105 comparisons, five were found to be statistically significant ($\alpha = 0.05$) and one was found to be marginally significant. These are listed in Table 2.

Stimuli sA and sG are found to be statistically different from stimuli for which the assessors reported experiencing a high degree of immersion (sB, sE, sL, sM). These results confirm the observation of three groups of stimuli based on their immersion ratings made in Sec. 4.1.

4.2 Questionnaire Data

Item correlations and the analysis of the overall questionnaire structure are presented in this section before assessing the influence of immersive tendency on immersion ratings in Sec. 4.3. The ratings provided by 21 participants on 19 questions constituted the administration of the immersive tendency questionnaire. The scores with respect to each question across all participants are shown in Fig. 4. The last question (S in Table 3) was retained but not used for analysis since it was categorical.

It can be observed that the responses span the entire range of the scale among the questions. The variance of scores is quite different for the individual questionnaire items. Responses to questions such as F and H show a clear distinction between what appears to be two groups of par-

[4]Increasing the number of calculations to 1,000 yielded similar results.

[5]The covariate and the treatment levels (stimuli) were not independent.

[6]Overall, the model had marginal R^2 value of 0.302 and conditional R^2 value of 0.410. Please refer to [27] for the calculation of the R^2 values.

[7]Immersive potential is the potential of the system or the content to evoke immersion [8].

Table 3. The reduced version of the immersive tendency questionnaire (ITQ) developed by Witmer and Singer [7]. The table lists the questions and corrected item-total correlation coefficients.

	Question	Corrected item-total correlations
A	Do you easily become deeply involved in movies or TV dramas?	0.63
B	Do you ever become so involved in a daydream that you are not aware of things happening around you?	0.44
C	Do you ever have dreams that are so real that you feel disoriented when you awake?	0.34
D	When watching sports, do you ever become so involved in the game that you react as if you were one of the players?	−0.12
E	How good are you at blocking out external distractions when you are involved in something?	0.00
F	Have you ever remained apprehensive or fearful long after watching a scary movie?	0.32
G	Have you ever gotten scared by something happening on a TV show or in a movie?	0.31
H	Do you ever become so involved in a video game that it is as if you are inside the game rather than moving a joystick and watching the screen?	0.65
I	How often do you play arcade or video games? (OFTEN should be taken to mean every day or every two days, on average)	0.41
J	Have you ever gotten excited during a chase or fight scene on TV or in the movies?	0.65
K	How well do you concentrate on enjoyable activities?	0.05
L	Do you ever become so involved in a television program or book that people have problems getting your attention?	0.26
M	How mentally alert do you feel at the present time?	−0.33
N	How physically fit do you feel today?	0.17
O	How frequently do you find yourself closely identifying with the characters in a story line?	0.58
P	When playing sports, do you become so involved in the game that you lose track of time?	0.25
Q	Do you ever become so involved in a movie that you are not aware of things happening around you?	0.55
R	Do you ever become so involved in doing something that you lose all track of time?	0.68
S	What kind of books do you read most frequently? Select one Spy novels Fantasies Science Fiction Adventure Romance novels Historical novels Westerns Mysteries Other fiction Biographies Autobiographies Other non-fiction	...

Note.
1. Correlation coefficients are Pearson correlation coefficients

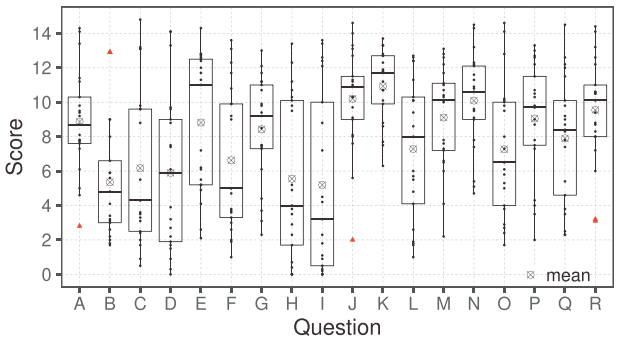


Fig. 4. Box plot representation of responses for each question by every assessor. The triangles indicate outliers. Refer to Table 3 for the questionnaire items.

ticipants. These questions could yield high discrimination for measuring immersive tendency. Unlike for the rating data the outliers could not be investigated as the comments were absent for the questions. Thus they were included in all further analysis.

4.2.1 Item Correlations

The aim of the questionnaire is to measure immersive tendency. Thus if every item in the questionnaire contributes to immersive tendency they should be correlated with each other and the entire scale as a whole. Corrected item-total correlations were calculated to assess the internal consistency of the questionnaire. Pearson correlation coefficients between each question in the questionnaire and total of the scale (sum of scores for all questions except the one being correlated to) are presented in Table 3. The coefficients could not be checked for statistical significance [28] because the assumptions were violated. Items D, E, K, and M show negative or no correlation with the total scale. The inter-item correlation matrix also showed similar results and the average inter-item correlation was found to be $\bar{r} = 0.14$. These results point to a lack of internal consistency

in the ITQ. While it is common practice to calculate coefficient (Cronbach's) alpha as a reliability estimate using the average inter-item correlation for questionnaire data, it was skipped in this study.[8]

4.2.2 Questionnaire Structure and Analysis

Determining the statistical structure of the questionnaire based on the responses serves multiple purposes: uncover the central ideas of the questionnaire, help in deciding if the scores should be summed together, and provide information for the continuous development of the tool. Hierarchical clustering was used to explore the relationship between the questionnaire items. Three clusters were obtained by performing agglomerative hierarchical clustering (AHC) on the questionnaire items. The number of clusters was selected using the elbow method and silhouette method (silhouette coefficient = 0.17) [31]. The dissimilarity between clusters was calculated using the Euclidean distances between them. Ward's minimum variance criterion was used for clustering, which yielded an agglomerative coefficient of 0.66.

The result of AHC is shown in Fig. 5, where the dissimilarity between clusters is represented on the y-axis and the color bar shows Witmer and Singer's [7] subscales. The first cluster consists of questions regarding concentration and general well-being. Questions on games are captured by the second cluster. The final cluster includes questions related to emotions and physical effect. Comparing the classification of the questions to the subscales in the original questionnaire, it is observed that the results are largely consistent. However we label the final cluster as emotional and physical involvement instead of involvement.

Witmer and Singer [7] summed the scores of all items to arrive at the ITQ total score. This may not be appropriate in the presence of stable clusters. To this end multiscale bootstrap resampling [32–34] was applied to the questionnaire data to assess the stability of the clusters. This method uses bootstrapped samples of varying sample sizes and applies hierarchical clustering to all bootstrapped sample replicates. Clustering results are used to calculate the bootstrap probabilities that are in turn used to calculate the approximately unbiased *p*-values for all levels of clustering.

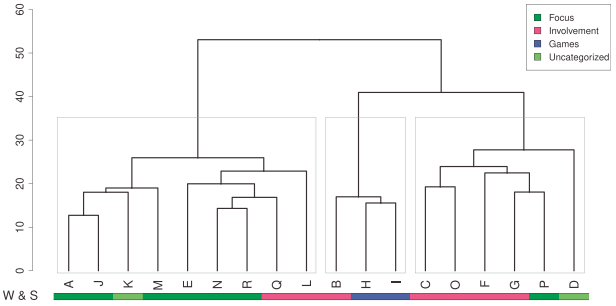


Fig. 5. Dendrogram produced from performing agglomerative hierarchical clustering on the immersive tendency questionnaire items. The height represents the Euclidean distances (dissimilarity) between the clusters and the alphabets along the y-axis refer to the questions in Table 3. The legend points to the color bar at the bottom that shows the classification of questions in the three subscales found by Witmer and Singer [7].

For 1,000 bootstrap replicates[9] the null hypothesis that the clusters do not exist could not be rejected for any level of clustering at $\alpha = 0.05$. Thus all subsequent analyses were conducted using cumulative immersive tendency scores as performed in the original work [7]. This implies that all questions have equal weighting in determining immersive tendency (which is seldom the case).

4.3 Influence of Immersive Tendencies on Immersion Ratings

It was stated that some individuals may be more susceptible to experiencing immersion than others in the introduction. A test was conducted to evaluate if this susceptibility has an influence on immersion ratings (RQ3). The hypothesis for this investigation was that as immersive tendency ratings increase (as measured by the ITQ), the immersion ratings should increase as well. Thus we were interested in the presence of a monotonic relationship between the ITQ score and immersion ratings. Kendall's rank order correlation, also known as Kendall's τ, was used for this purpose. The values for Kendall's τ range from -1 to $+1$, where a value of $+1$ indicates a perfectly monotonic association and -1 indicates complete disagreement between the the variables. There is no monotonic relationship if the value is found to be 0.

Kendall's τ between ITQ total scores and immersion scores was calculated for each of the 15 stimuli. These are stated in Fig. 6. There is a statistically insignificant correlation for most stimuli. Correlation values for stimuli D and J (positive and negative, respectively) were found to be statistically significant. The null hypothesis for the test: true τ values equal to 0, could not be rejected for 13 out of the 15 stimuli. Assuming that the ITQ captures immersive tendency accurately and the participants' use of scale does not differ between the rating phase and questionnaire, it can be inferred that there is a minimal influence of immersive tendency on immersion ratings in the present study.

[8]Coefficient alpha is a single test administration reliability estimate that is of interest when "error factors associated with the use of different items are of interest" [29]. Importantly it is dependent on data and not the measurement tool. Determining the reliability of a measure (here, the questionnaire) is a complex issue and the relationship between reliability of a measure and coefficient alpha has not been thoroughly established [30]. Nonetheless coefficient alpha equals reliability for a single factor model under the assumptions of uncorrelated errors, tau-equivalence, and unidimensionality [30]. These assumptions can be checked using structural equation modeling methods when sufficient data is available. Since it is known that immersive tendency is a multidimensional concept and inter-item correlations are drastically different (items D, K, E, and M are negatively correlated to most other items) signaling a potential lack of tau-equivalence, Cronbach's alpha has not been calculated.

[9]The number of replicates was increased from 1,000 to 100,000 without any considerable change in the result.

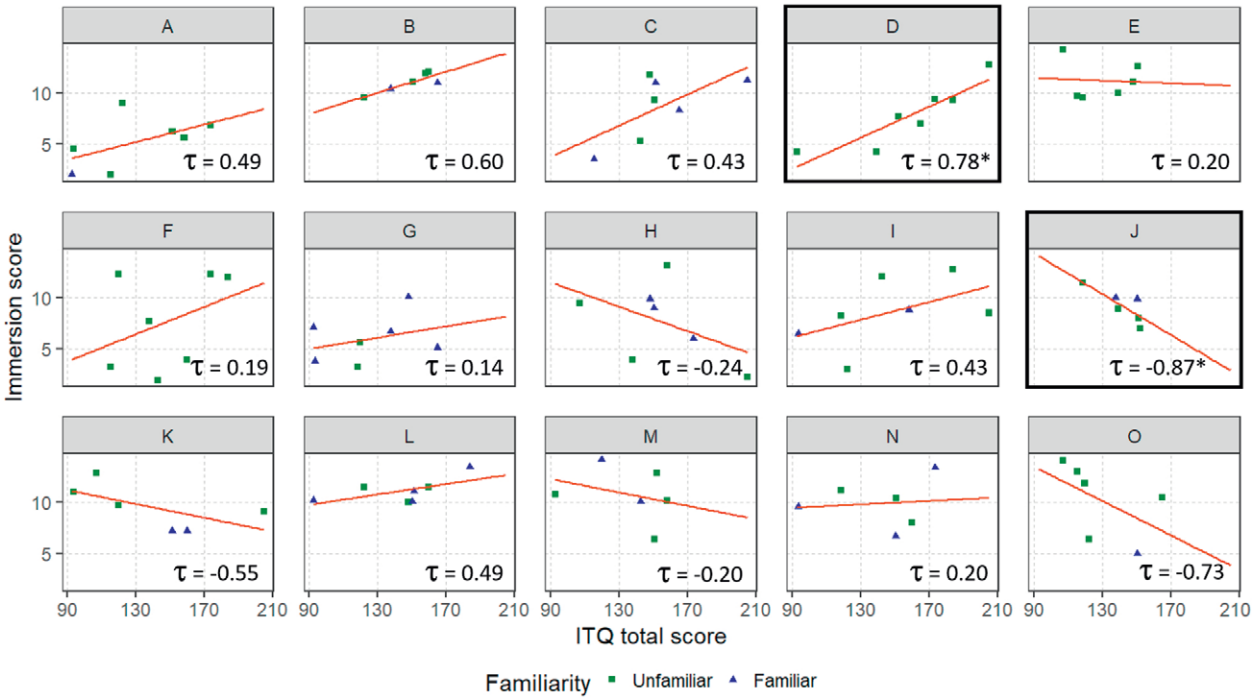


Fig. 6. Kendall's rank order correlation (Kendall's τ) between the immersive tendency total score and the immersion ratings for each stimulus. The regression line is plotted only to aid the reader. Correlations for stimulus D and J were found to be statistically significant. Assessors' familiarity with the content is depicted using different shapes.

It is important to note that this result does not account for familiarity with the content. Since familiarity was not included in the analysis of variance model, its role remains unquantified. Fig. 6 shows the subjects' familiarity with the content. A visual inspection of the plots does not reveal obvious patterns with regards to familiarity.

5 DISCUSSION

There is interest for comprehending immersion to harness the technical capabilities of audiovisual systems for delivering engaging experiences. A lack of well-developed techniques for behavioral and physiological measures limits us to subjective methods for now. Although questionnaires have been the tool of choice for subjective assessment they require substantial time commitment, increase the workload for the assessors, and fail to capture all aspects of the immersive experience, which makes them less suitable for assessing immersion. An appropriate experimental method for quantifying immersion with low complexity remains undefined. Establishing an experimental paradigm compliant with the implications will facilitate the understanding of the topic beyond theoretical conceptualizations, which could be beneficial in delivering enhanced experiences. To this end a rating experiment was conducted to quantify immersion in audiovisual experiences using commercially available stimuli with spatial audio and ultra-high-definition video.

An important caveat when conducting rating experiments is ensuring that the idea to be investigated is understood as intended. This is particularly challenging for an encompassing cognitive concept like immersion. The results from the experiment show that the description of immersion can be communicated and the participants are able to follow the experimental protocol. The analysis of variance and subsequent pairwise comparison results show that the assessors were able to discriminate between the different levels of stimuli. Five pairs of stimuli were found to be statistically significantly different, exemplifying that even with limited precision there are obvious differences in the levels of immersion experienced by the participants. The assessors did not report any difficulty in understanding the provided description of immersion or the experimental protocol during experimentation. These findings indicate that immersion can be measured using a unidimensional scale by asking assessors to reflect on the overall immersion in the experience as performed in the current experiment, thereby testing the methodology presented in [4] to answer RQ1. This is in-line with Jennett et al.'s [3] findings.

Immersion has always been conceptualized as a graded experience as opposed to being an all-or-nothing (binary) experience in the literature [4]. Nevertheless empirical tests to explicitly test this assumption have not been conducted to the best of our knowledge. Result of Hartigan's dip test and the normally distributed residuals from ANOVA show that distribution of immersion scores is unimodal, suggesting that immersion is a graded experience (RQ2). This reinforces the conceptual understanding of the topic. It should be noted that this interpretation of the results is based on the entire stimulus set and there may be some stimuli (e.g.,

sF) that may elicit responses whose distribution is more bimodal. While this cannot be explored in the current study due to a limited number of observations per stimulus, increasing the data points can add clarity. Such stimuli can be valuable in understanding the role of the individual and the narrative for experiencing immersion.

The possibility of having different degrees of immersion implies that comparisons can be made between experiences. Thus the influence of the physical parameters of rendering systems on immersion can be quantified, for example. Furthermore the current result suggests that the nature of the experience might well be one of the distinguishing factors between the concepts of immersion and flow (see [4]) since flow is an optimal, all-or-nothing experience [3].

The final research questions (RQ3) sought quantification of the influence of immersive tendencies on immersion ratings. Thirteen of the 15 stimuli showed non-significant Kendall's rank-order correlation, signaling an absence of a monotonic relationship between the ITQ total score and immersion ratings. However it should be noted that other relationships may be present. A couple of reasons could have led to the results. First, the assumption of scale usage being constant is fair but fails to account for the interactions between the subject and the question being answered. Second, the theoretical basis for the selection of items on the questionnaire is unclear. One can argue that assigning equal weights to all questions when there is a lack of internal consistency may have skewed the results. Error factors associated with the passing of time should be assessed either by test-retest or parallel single administrations of the ITQ [29].

Witmer and Singer's ITQ [7] is the most prevalent questionnaire for the purpose, but it is not the only option. Weibel et al. [35] determined the relationship between the ITQ and personality traits (Big Five), which can be helpful in providing an alternate perspective on immersive tendencies. Finally, this experiment traded data points per stimulus in order to test a large number of stimuli. Although Kendall's τ is quite robust for small sample size, increasing the number of observations for each stimulus will help the ability to detect differences.

An important aspect of this study was to assess the validity of the questionnaire and aid the continuous development of the tool. A lack of internal consistency for the ITQ is not particularly problematic as the idea being measured by the questionnaire can be multidimensional. However, since the questionnaire has been reduced to achieve internal consistency [7], the theoretical foundation for including uncorrelated and negatively correlated items needs to be scrutinized. The clustering of questions is surprisingly similar to Witmer and Singer's [7] a priori classification. Nonetheless a study with 220 participants suggested a two-factor solution of emotional involvement and absorption [35].

It should be noted that the questionnaire results presented in Sec. 4 are based on inherently noisy data. The number of participants was limited and the true structure might be notably different than what is presented. Structural equal modeling techniques such as confirmatory factor analysis can be used to verify the results when data from a sufficient number of participants is available. Alternatively, for non-inferential analysis, exploratory factor analysis could also be used to determine the sub-groups.

5.1 Limitations

The experimental method tested in this work has several limitations. Post-experience evaluation with lengthy stimuli is prone to inaccurate recall and the recency effect. Moreover participants may judge their immersion in the scene based on a limited number of memorable moments. The measurement does not reveal the temporal variation in immersion and is limited to overall assessment. The goal of the new method was to improve on the shortcomings of questionnaires by reducing the experimental time. However inclusion of distractor tasks adds to the experimental time and their usefulness is unclear. The current experimental method draws heavily from theoretical implications proposed by Agrawal et al. [4] that have not been tested and thus remain unoptimized. For instance participant reliability cannot be evaluated since the number of trials for stimuli is limited to one.

5.2 Future Work

A number of avenues for developing a deeper insight on the subject of immersion can be explored. Foremost, future work should aim to test and validate the method presented in this work. The theoretical implications for the experimental paradigm and the selection of stimuli must be tested to optimize the method. Eventually, developing objective measures using physiological signals can be beneficial for quantifying immersion without adding cognitive load on the participants, learning about temporal changes in immersion, and adapting the experience to maximize immersion. It is crucial to develop an understanding from a psychological perspective to understand the qualitative differences between immersive experiences. Sensory analysis techniques such as free choice profiling may provide inspiration for understanding the core ideas of immersion from an assessor's point of view.

Here the experiment was focused on domestic audiovisual experiences. Nevertheless the method described here is application agnostic. It would be interesting to adapt the method for virtual and augmented reality applications and explore the relationship between presence and immersion. Future work should look to quantify the influence of the physical parameters on immersion. For example audio spatialization can be easily controlled compared to other factors and may prove to be helpful in choosing the appropriate audio configuration to deliver engaging experiences. Lastly the assumption that users prefer more immersive experiences is at the heart of the quest to study immersion. This assumption must be tested to avoid surprises in the later stages.

6 CONCLUSION

The aim of the experiment conducted in this study was to establish an experimental method for quantifying immer-

sion in audiovisual experiences that can form the framework for future investigations. Current results show that immersion can be evaluated post-experience on a unidimensional scale much like preference and liking. The assessors are able to comprehend the description of immersion and reflect on their overall immersion when asked to rate immersion in an experience. It is found that immersion is not a binary (all-or-nothing) concept and there can indeed be different levels of immersion in an experience. Kendall's rank-order correlations between the questionnaire and immersion ratings are largely insignificant, revealing that there is minimal influence of immersive tendency on the experience of immersion.

Future work should aim at validating the methodology tested in this work. Developing objective measures, understanding the relationship between presence and immersion, adapting the presented method for virtual reality, and mapping the relationship between preference and immersion should be investigated going forward. Determining the influence of physical parameters of the system on immersion can open new avenues for augmenting audiovisual experiences.

7 ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 765911. The authors would like to thank Georgij Engkjær-Trautwein for his efforts in determining the distractor tasks and valuable discussions. Jens Rahbek, Poul Erik Trabjerg, Jesper Pedersen, and Geoff Martin are thanked for helping with the facilities and experimental setup.

8 REFERENCES

[1] S. Meltzer and A. Murtaza, "First Experiences With the MPEG-H TV Audio System in Broadcast," *SET Int. J. Broadcast Eng.*, vol. 4, p. 6 (2018). https://doi.org/10.18580/setijbe.2018.6.

[2] ITU-R, "Parameter Values for Ultra-High Definition Television Systems for Production and International Programme Exchange," *Recommendation BT.2020-2* (2015 Oct.).

[3] C. Jennett, A. L. Cox, P. Cairns, et al., "Measuring and Defining the Experience of Immersion in Games," *Int. J. Human-Comp. Studies*, vol. 66, no. 9, pp. 641–661 (2008 Sep.). https://doi.org/10.1016/j.ijhcs.2008.04.004.

[4] S. Agrawal, A. Simon, S. Bech, K. Bæntsen, and S. Forchhammer, "Defining Immersion: Literature Review and Implications for Research on Audiovisual Experiences," *J. Audio Eng. Soc.*, vol. 68, no. 6, pp. 404–417 (2020 Jun.). https://doi.org/10.17743/jaes.2020.0039.

[5] N. C. Nilsson, R. Nordahl, and S. Serafin, "Immersion Revisited: A Review of Existing Definitions of Immersion and Their Relation to Different Theories of Presence," *Human Tech.*, vol. 12, no. 2, pp. 108–134 (2016 Nov.).

[6] C. Zhang, "The Why, What, and How of Immersive Experience," *IEEE Access*, vol. 8, pp. 90878–90888 (2020 May). https://doi.org/10.1109/access.2020.2993646.

[7] B. G. Witmer and M. J. Singer, "Measuring Presence in Virtual Environments: A Presence Questionnaire," *Pres. Teleop. Virt. Environ.*, vol. 7, no. 3, pp. 225–240 (1998 Oct.).

[8] S. Agrawal, A. Simon, S. Bech, K. Bærentsen, and S. Forchhammer, "Defining Immersion: Literature Review and Implications for Research on Immersive Audiovisual Experiences," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), paper 10275.

[9] H. Stone, R. N. Bleibaum, and H. A. Thomas, "Chapter 6 - Descriptive Analysis," in H. Stone, R. N. Bleibaum, and H. A. Thomas (Eds.), *Sensory Evaluation Practices*, pp. 233–289 (Academic Press, Cambridge, MA, 2012), 4th ed. https://doi.org/10.1016/B978-0-12-382086-0.00006-6.

[10] H. T. Lawless and H. Heymann, *Sensory Evaluation of Food* (Springer, New York, NY, 2010).

[11] M.-L. Ryan, *Narrative as Virtual Reality: Immersion and Interactivity in Literature and Electronic Media* (Johns Hopkins University Press, Baltimore, MD, 2003).

[12] J. Lawson, *Design and Analysis of Experiments With R* (Chapman and Hall/CRC, London, England, 2015).

[13] "Dolby Atmos® Home Theater Installation Guidelines," Dolby Laboratories, San Francisco, CA (2018 Dec.).

[14] ITU-R, "General Viewing Conditions for Subjective Assessment of Quality of SDTV and HDTV Television Pictures on Flat Panel Displays," *Recommendation BT.2022* (2012 Aug.).

[15] IEC, "Sound System Equipment - Part 13: Listening Tests on Loudspeakers," *Standard 60268-13* (1998 Mar.).

[16] C. J. Jerome and B. G. Witmer, "Human Performance in Virtual Environments: Effects of Presence, Immersive Tendency, and Simulator Sickness," *Proc. Human Fact. Erg. Soc. Ann. Meeting*, vol. 48, no. 23, pp. 2613–2617 (2004 Sep.). https://doi.org/10.1177/154193120404802302.

[17] K. J. Kim, E. Park, S. S. Sundar, and A. P. del Pobil, "The Effects of Immersive Tendency and Need to Belong on Human-Robot Interaction," in *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 207–208 (Boston, MA) (2012 Mar.).

[18] J. Hou, Y. Nam, W. Peng, and K. M. Lee, "Effects of Screen Size, Viewing Angle, and Players' Immersion Tendencies on Game Experience," *Comp. Human Behav.*, vol. 28, no. 2, pp. 617–623 (2012 Mar.). https://doi.org/10.1016/j.chb.2011.11.007.

[19] C. D. Murray, J. Fox, and S. Pettifer, "Absorption, Dissociation, Locus of Control and Presence in Virtual Reality," *Comp. Human Behav.*, vol. 23, no. 3, pp. 1347–1354 (2007 May). https://doi.org/10.1016/j.chb.2004.12.010.

[20] S. Zielinski, F. Rumsey, and S. Bech, "On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451 (2008 Jun.).

[21] P. B. Brockhoff, "Statistical Testing of Individual Differences in Sensory Profiling," *Food Qual.*

Pref., vol. 14, no. 5–6, pp. 425–434 (2003 Jul.–Sep.). https://doi.org/10.1016/S0950-3293(03)00007-7.

[22] R. Pfister, K. A. Schwarz, M. Janczyk, R. Dale, and J. B. Freeman, "Good Things Peak in Pairs: A Note on the Bimodality Coefficient," *Front. Psychol.*, vol. 4 (2013). https://doi.org/10.3389/fpsyg.2013.00700.

[23] J. B. Freeman and R. Dale, "Assessing Bimodality to Detect the Presence of a Dual Cognitive Process," *Behav. Res. Methods*, vol. 45, pp. 83–97 (2013). https://doi.org/10.3758/s13428-012-0225-x.

[24] Y.-J. Kang and Y. Noh, "Development of Hartigan's Dip Statistic With Bimodality Coefficient to Assess Multimodality of Distributions," *Math. Prob. Eng.*, vol. 2019, p. 17 (2019 Dec.). https://doi.org/10.1155/2019/4819475.

[25] J. A. Hartigan and P. M. Hartigan, "The Dip Test of Unimodality," *Ann. Stat.*, vol. 13, no. 1, pp. 70–84 (1985 Mar.).

[26] P. M. Hartigan, "Algorithm AS 217: Computation of the Dip Statistic to Test for Unimodality," *J. Royal Stat. Soc. (Appl. Stat.)*, vol. 34, no. 3, pp. 320–325 (1985).

[27] S. Nakagawa and H. Schielzeth, "A General and Simple Method for Obtaining R^2 From Generalized Linear Mixed-Effects Models," *Methods Ecol. Evol.*, vol. 4, no. 2, pp. 133–142 (2013 Feb.).

[28] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R* (SAGE, Thousand Oaks, CA, 2012).

[29] J. M. Cortina, "What Is Coefficient Alpha? An Examination of Theory and Applications," *J. Appl. Psych.*, vol. 78, no. 1, pp. 98–104 (1993).

[30] T. Raykov and G. A. Marcoulides, "Thanks Coefficient Alpha, We Still Need You!" *Educ. Psych. Meas.*, vol. 79, no. 1, pp. 200–210 (2019 Feb.). https://doi.org/10.1177/0013164417725127.

[31] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *Multidis. Sci. J.*, vol. 2, no. 2, pp. 226–235 (2019 Jun.). https://doi.org/10.3390/j2020016.

[32] H. Shimodaira, "An Approximately Unbiased Test of Phylogenetic Tree Selection," *Syst. Biol.*, vol. 51, no. 3, pp. 492–508 (2002 May). https://doi.org/10.1080/10635150290069913.

[33] R. Suzuki and H. Shimodaira, "Pvclust: An R Package for Assessing the Uncertainty in Hierarchical Clustering," *Bioinformatics*, vol. 22, no. 12, pp. 1540–1542 (2006 Jun.). https://doi.org/10.1093/bioinformatics/btl117.

[34] R. Suzuki and H. Shimodaira, "An Application of Multiscale Bootstrap Resampling to Hierarchical Clustering of Microarray Data: How Accurate Are These Clusters?" in *Proceedings of the Fifteenth International Conference on Genome Informatics* (Yokohama, Japan) (2004).

[35] D. Weibel, B. Wissmath, and F. W. Mast, "Immersion in Mediated Environments: The Role of Personality Traits," *Cyberpsych. Behav. Soc. Netw.*, vol. 13, no. 3, pp. 251–256 (2010 Jun.). https://doi.org/10.1089/cyber.2009.0171.

APPENDIX A

A.1 Balanced Incomplete Block (BIB) Design

Table 4. Allocation of stimuli to the experimental blocks for the BIB design used in the study.

Block	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
1	D	J	M	N	O
2	A	B	H	I	M
3	A	C	E	F	O
4	C	E	G	H	L
5	B	D	E	J	K
6	B	C	D	G	O
7	F	G	K	M	O
8	C	F	I	J	M
9	A	D	G	L	M
10	A	C	J	K	L
11	C	D	H	I	K
12	B	F	G	H	J
13	H	J	L	N	O
14	D	E	F	I	L
15	B	F	K	L	N
16	A	G	I	K	N
17	A	B	I	L	O
18	B	C	E	M	N
19	E	G	I	J	N
20	A	D	F	H	N
21	E	H	K	M	O

Note.
1. The prefix 'S' used for representing the stimuli is dropped in this table for clarity. These are the same excerpts listed in Tables 1 and 5.

A.2 Narrative Synopses

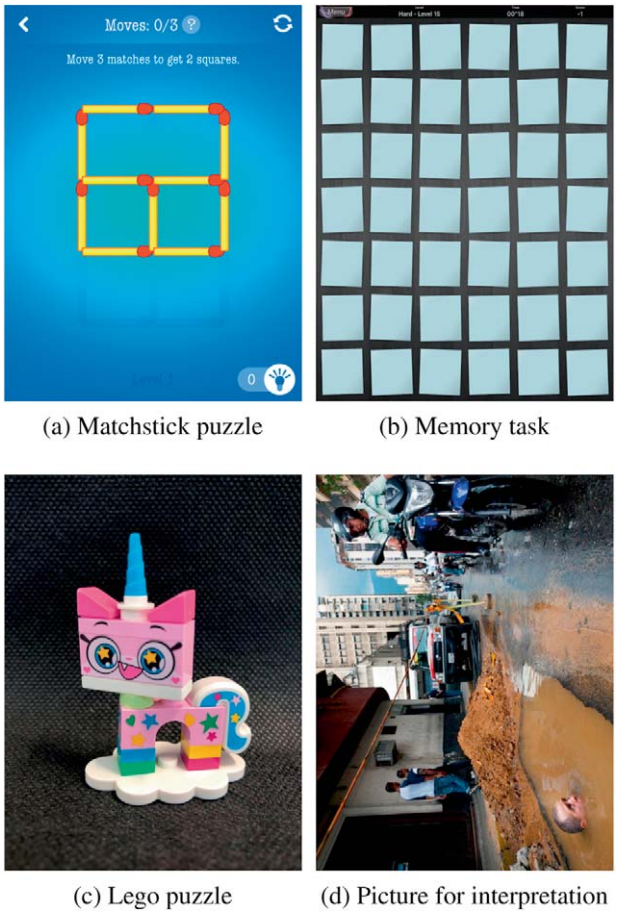
A list of narrative synopses for the 15 stimuli is presented in Table 5.

A.3 Outliers

On the basis of the comments provided by the participants for each stimulus, seven data points were eliminated from the immersion ratings. Outliers for stimuli B and K were from the same participant (a visual expert) who could only focus on the visual quality. Other stimuli graded by them consisted of holistic views and hence were not removed. The outlier for stimulus J was reported to be too unrealistic and exaggerated. Participant 1 communicated that they disliked film music in nature documentaries and could not focus on the experience. Thus the two lowest scores for N and O were removed. The second lowest score for stimulus M was removed as the participant said that they could not watch the violence and had closed their eyes periodically during the experience. Finally stimulus E had the lowest score eliminated as the participant was familiar with the stimulus in the same experimental setting as they had participated in the pilot experiment.

A.4 Distractor Tasks

Snapshots from the four distractor tasks are provided in Fig. 10.



(a) Matchstick puzzle (b) Memory task

(c) Lego puzzle (d) Picture for interpretation

Fig. 7 The images above are a snapshot of each of the four tasks described in Sec. 3.2. a) Level 1 from the second pack of *Matchsticks* ∼ *Free Puzzle Game with Matches*; b) Level 15 of *Memory* • *Classic*; c) an 11-piece LEGO® puzzle; d) the picture provided to the participants for interpretation (obtained from an online article from the *New York Times* found at https://www.nytimes.com/2016/09/22/learning/40-intriguing-photos-to-make-students-think.html).

Table 5. A list of narrative synopses provided prior to each experience.

Excerpt	Content	Narrative synopsis
Example	*Earth: One Amazing Day*	...
sA	*Mission: Impossible – Fallout*	Ethan Hunt and Ilsa are undercover agents who have worked together in the past. They are on separate but related missions. They aren't fully aware of the other's mission.
sB	*Apocalypse Now – Final Cut*	A Captain from the US military is sent to kill a Colonel of the US military who has gone rogue. The story takes place during the Vietnam war.
sC	*The Revenant*	Glass is a scout who has been separated from his crew. He is alone in the North American wilderness during winter when his crew gets a clue that he may be alive.
sD	*Fantastic Beasts: The Crimes of Grindelwald*	Yusuf explains why he wants to kill Credence.
sE	*Dynasties: Lion*	Red is young male lion who grew up in Kenya's Maasai Mara.
sF	*The Darkest Hour*	Germany has invaded Belgium and France, pushing the English army into the English Channel. Winston Churchill, the prime minister of England, is getting ready to address the nation for the first time.
sG	*Murder on the Orient Express*	Hercule Poirot is a detective in Jerusalem.
sH	*Braveheart*	The story is from 13th century Scotland when the king of England claimed the throne of Scotland and the English lords started oppressing the Scottish.
sI	*Ad Astra*	Roy is an astronaut who is on a mission to destroy the Lima project.
sJ	*Earth: One Amazing Day*	The rainforests of Ecuador are home of some of the most special and intriguing animal species.
sK	*Spider-Man: Into the Spider-Verse*	Miles is a teenager who studies at a boarding school in New York, loves spiderman and making graffiti.
sL	*The Revenant*	Frontiersmen from the colonized parts of the New World (modern day USA) have ventured west of the Missouri river. Three of the men are out hunting in the woods while the rest prepare for their departure from the area.
sM	*Sicario*	Kate is a law enforcement officer who runs a kidnap response team in the Phoenix area.
sN	*Earth: One Amazing Day*	The tropical islands off the coast of Panama are host to colorful forests and the three-toed sloth.
sO	*Earth: One Amazing Day*	Bamboo is the fastest growing plant in the world. However, it is not very nutritious.

THE AUTHORS


Sarvesh R. Agrawal


Søren Bech


Klaus Bærentsen


Katrien De Moor


Søren Forchhammer

Sarvesh Agrawal was born and raised in Mumbai, India. He moved to the United States in 2014 and received a B.S. in audio production with a minor in entertainment technology from Middle Tennessee State University (MTSU) in 2016. Following a brief stay at MSE Audio Group in Kansas, he moved to New York, where he attained a graduate degree in architectural acoustics from Rensselaer Polytechnic Institute (RPI) in 2018. Sarvesh joined Bang & Olufsen in 2018 as a research fellow and has been pursuing a Ph.D. from the Department of Photonics Engineering at the Technical University of Denmark (DTU) as an early stage Marie Curie fellow. Psychoacoustics, perceptual evaluation of sound, and subjective assessment of multimodal experiences are his primary research interests. Currently Sarvesh is investigating the concept of immersion in domestic audiovisual environments.

•

Søren Bech received an M.Sc. and Ph.D. from the Department of Acoustic Technology (AT) of the Technical University of Denmark. From 1982–1992 he was a research fellow at AT studying perception and evaluation of reproduced sound in small rooms. In 1992 he joined Bang & Olufsen, where he is Director of Research. In 2011 he was appointed Professor in Audio Perception at Aalborg University and he is Adjunct Professor at Surrey University (GB) and McGill University (CAN). He is a Fellow of the Acoustical Society of America and Audio Engineering Society (AES). He is past Governor and Vice-President of the AES and now serves as associate technical editor of the AES Journal. He has been vice-chair of the International Telecommunication Union working group 10/3. In 2006 he and Dr. Zacharov published the book *Perceptual Audio Evaluation – Theory, Method and Application* (Wiley and Sons). His research interests include psychoacoustics and in particular human perception of reproduced sound in small and medium-sized rooms. Other interests include experimental procedures and statistical analysis of data from sensory analysis of audio and video quality.

•

Klaus B. Bærentsen is associate professor in the department of Psychology at the University of Aarhus, Denmark. He received his M.Sc. and Ph.D. in psychology from the department of Psychology at the University of Aarhus. During the period 1983 to 1995 he was an assistant professor engaged in various research projects concerning man-machine interaction and was teaching at the department of Psychology at the University of Aarhus. In 1995–2000 he worked at Bang & Olufsen as a specialist in user interaction technology and concurrently held various temporary and part-time engagements at the University of Aarhus, the University of Aalborg, and Grundfos A/S. In 2000–2001 he was an assistant professor at the MR Research Center at Skejby Hospital doing fMRI research. In 2001 he was appointed associate professor in the department of Psychology at the University of Aarhus, doing research and teaching on neuroscientific psychology, man-machine interaction, and general cognitive psychology. His interests span a wide range of phenomena related to human life activity: consciousness and personality, encompassing motivation, sensory-motor control, perception, and cognition and its neurological foundation, and cultural-historical development of societal activity and technology, as well as synergetics, complex self-organizing systems, classical yoga, and Buddhist psychology.

•

Katrien De Moor is associate professor in the department of Information Security and Communication Technology at NTNU, mainly focusing on socio-technical approaches in ICT research. She is co-Editor-in-Chief of the multidisciplinary journal "Quality and User Experience" (Springer) and affiliated researcher at the Research Group for Media, Innovation and Communication Technologies (Ghent University, Belgium). She received her Ph.D. degree in Social Sciences from Ghent University (2012) with a thesis on bridging gaps in Quality of Experience research and its challenges. She has been a visiting researcher at several institutions, including the University of Eindhoven, TU Berlin, and NTNU (as Marie Curie Postdoctoral Fellow). Katrien is passionate about user research and user involvement in user-centric innovation processes. Her main research interests and activities are centered around human/user experiences with technology (Quality of Experience, User Experience, User Engagement, immersive experiences. . .), related methodological challenges (ecological validity, user diversity. . .) and ethical implications (e.g., data privacy, human agency, power dynamics in design processes, and ecological footprint of ICT). She has published her work in a range of international, peer-reviewed journals, conferences, and books, acts as a reviewer for several international journals, and has served on several program committees of international conferences and workshops. She is one of the 20 founding members of the Young Academy of Norway.

•

Søren Forchhammer received an M.Sc. degree and Ph.D. degree from the Technical University of Denmark, Lyngby. Currently he is a Professor with DTU Fotonik, Technical University of Denmark, where he has been since 1988. He is Head of the Coding and Visual Communication Group at DTU Fotonik. He is Coordinator of the EU MSCA ITN RealVision. He is flagship lead in the DNRF CoE SPOC. 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, two-dimensional information theory, and visual communications.