THE IMMERSIVE AUDIENCE EXPERIENCE EVALUATION TOOLKIT

J. Freeman, J. Lessiter, P. Borden, L. Turner-Brown, L. Kurta & E. Ferrari

i2 media research ltd, c/o Psychology Department, Goldsmiths University of London, UK

ABSTRACT

This paper introduces the Immersive Audience Experience Evaluation Toolkit, an online questionnaire designed to measure audience perceptions of impact and value of different immersive experiences. Initially created and developed in 2017, the Toolkit has been applied to various immersive media formats and contents such as virtual reality (VR), augmented reality (AR), mixed reality (MR), games and screen-based media. With examples of insights generated for creators and funders of immersive experiences, this paper summarises the need for, development and refinement of the Toolkit. Drawing from audience evaluations with 460 people using 10 different immersive contents, the added value of the Toolkit to the creative industry of further exploring aggregated and comparative ratings is also discussed.

INTRODUCTION

Access to immersive experiences

The past five years has seen a surge in immersive technology developments around the world. Consumers can now more easily access virtual and augmented reality hardware at home; from high end headsets requiring high spec PCs to run content (such as the HTC Vive, Oculus Rift, PlaystationVR, and Microsoft Hololens) to smartphone based apps and viewers. Location-based immersive experiences (e.g., exhibitions and galleries) are also gaining popularity in the arts and culture sector. Within this rapidly evolving industry which fuses arts and technologies, a range of genres and applications are available that challenge audience expectations and push traditional boundaries of curated and mediated audience experiences.

Impact of immersive experiences

The psychological impact on audiences of mediated experiences across a range of genres including film and games has been widely explored (e.g., [1] [2]). Their results have implications for the types of self-report audience evaluations considered relevant to immersive content, and give a sense of the range of qualities which may be relevant to impact judgements such as emotional intensity (e.g., [3]), strangeness, vividness (e.g., [4]), a sense of presence (e.g., [5]), and ease of use (of experience, interactions and tasks, e.g., [2]).

In collaborative qualitative research with Nesta for Digital Catapult [6] we explored what makes a curated piece of content "impactful" and "valuable" from the perspective of content funders and content creators of immersive productions, as well as potential end user audiences. The findings, summarised below, revealed insights about the differences and similarities in how impact is judged; key words used to describe impactful experiences; important audience characteristic that might influence impact ratings; and considerations for the interaction design (real world and augmented virtual).

Perspectives on impact

End users, content creators and content funders described similar and different emphases in their views on value and impact dependent on their perspective. For instance, content creators emphasised the level of match/mismatch in achieving their artistic intention from audience evaluations, and funders, in turn, considered it important that content creators met their stated objectives. For funders, audience reach and figures, accessibility and acclaim were also prominent. End users reported appreciating surprise, novelty and satisfying, easy to use experiences. Temporal dimensions of impact and value were highlighted too, whereby audience journeys and appreciation could be extended beyond the immediate experience of the content itself.

Descriptors of impact

A wide range of constructs were frequently cited as relevant to short and long term impact and value across different perspectives. These included the content's perceived significance and relevance or timeliness in relation to the world today (relating to cultural value), and adjectives that described the *overall or general quality of experience* (e.g., it was "good", "powerful"). The creator's intention for the piece of content often included discrete positive and negative emotions, and participants from all groups referred to the *intensity of the experience* (whether intense or relaxing), the *sense of presence* (or "being there"), *engagement*, perceptions of *naturalness*, *comfort*, and *memorability*. Social capital associated with experiencing a particular content (e.g., *shareability*) was also considered relevant to impact, both in terms of extending and enriching the experience for the end user, and also increasing reach and potential revenue for the stakeholders. With regard to the economic value of these types of experience, *willingness to pay* varied depending on the context of use (e.g., home or public installation).

Characteristics of the person

Variables in relation to the person also emerged as having potential to influence the impact of any particular immersive content, whether at home or in a more public space. These included stable as well as more variable *characteristics of the person*. For instance, a person's willingness to suspend disbelief, and their susceptibility or openness to new experiences may influence their tendency or propensity to feel immersed. This might be measured as a personality trait or a state characteristic, for instance, consider the influence of alcohol consumption on immersive experiences.

Participants reflected on how their *mood state* and *personal interests* might influence impact evaluations, and that choice of content may vary depending on the person's mood, preferences, expectations and needs from that experience. Within a home use context, stakeholders and end users commented on the more demanding, 'active' nature of setting

up and experiencing immersive, interactive content compared to more traditional, passive lean back leisure experiences.

A person's entertainment, leisure and genre preferences also likely to influence awareness of, access to, preferences for and comfort with different types of immersive content. This has implications for measuring characteristics of the audiences sampled and the context or setting in which they are exposed to the immersive content: evaluations from naturalistic samples (e.g., location-based) may vary from lab-based 'research volunteers' who do not choose their content. Similarly, the research raised questions about whether the novelty of immersive experiences could enhance overall evaluations: as the industry becomes more established and familiar to audiences, perhaps the impacts lessen.

Interaction design

Content creators in the sample were acutely mindful of how successfully their interaction design facilitated intuitive, frictionless experiences. This was particularly relevant when deploying more novel technologies (e.g., gaze tracking) or where audiences may be less confident or comfortable in physically exploring interactive spaces that required user prompted actions to unfold the narrative. This suggests that some behavioural measures might also be relevant to evaluating impact. Getting 'the most' from the environment seemed to reflect a tension between providing the *expected and unexpected*. The necessity of at least some instruction and on-boarding strategies to soften the transition into the immersive narrative were reported by many stakeholders. Nonetheless there was awareness that audience delight often emerged from their self-discovery of agency in the narrative.

Summary

Research suggests that the impact or value of a given piece of immersive content is a complex phenomenon to measure. Perceptions vary across individuals and perspectives depending on the nature of their investment in the piece of content. As such impact is influenced by a wide range of interacting factors including affordances and physical formal properties of the content, the type of content and narrative itself, aspects of the person experiencing the content in the form its presented, and the context in which this occurs. Establishing a unified consistent objective criterion of impact within this emerging industry poses a significant challenge.

Objective

In this paper we introduce the Audience Experience Evaluation Toolkit ('Toolkit'), designed to measure the impact of a given immersive production. The contents, samples and contexts in which the Toolkit has so far been used are described. Using aggregated results provided by 460 audience members across 10 different immersive experiences, we aim to demonstrate the current benefits of using the Toolkit, and speculate on future directions for impact metrics in the creative immersive industries.

THE TOOLKIT

The Toolkit is a self-report measure of impact, developed for the creative immersive industry. It is termed the 'Toolkit' because it contains a set of measurement 'modules' (e.g., relating to qualities of experience, and person, form, content, and context factors) that focus on a limited yet pragmatic range of different variables previously identified as

relevant to the overall evaluation of a given production and of interest to stakeholders [6] [7]. This modular form of evaluation allows for new module topics to be developed and introduced, or removed from the Toolkit, as per the intentions for the evaluation. For instance, the Toolkit does not currently measure personality traits and states of the audience which might increase their 'immersive tendency', nor does it measure objective behaviours (e.g., facial expressions; vocalisations), but these are acknowledged as potentially useful metrics to consider in future research.

Intended to be completed by end user audiences and content creators, the Toolkit aims to quantify the quality of immersive experiences and the characteristics of the people providing evaluations, as well as the affordances of the media form and content.

Before audience testing, content creators indicate the affective experiences they intend to elicit from audiences (e.g., the emotions they would like their content to increase/decrease in audiences), along with form characteristics and content affordances (e.g., content duration, mode/s of interaction, compatible headsets) which can support more refined content comparisons.

Evaluation modules for audiences

The current version of the Toolkit contains several modules for audiences to complete. 'Core' modules apply to all contents evaluated and comprise (a) post-test experience questions and (b) questions about audience characteristics. The 'bespoke' module enables stakeholders to specify a small number of additional questionnaire items to address any concerns or questions they have about audience experiences of their production.

Post-experience questions first ask participants to rate their immediate overall impression of impact using five global adjectives relating to experience (e.g., good, powerful), six items relating to cultural value, three engagement items derived from the SOPI [5] and three relating to overall emotional intensity, and positive and negative emotional intensity, separately. Finally, participants are asked if they would be willing to pay for the experience, and if so, the optimal price point is estimated [8].

The post-experience items are followed by questions about participant demographics and about their engagement and experience with arts, culture and VR/AR. These support interpretation of results given the audience composition for each content tested.

Scoring the Toolkit

Global experience, cultural value, engagement, and emotional intensity variables from the post-experience metrics are scored and presented:

- as mean absolute ratings for the audience sample (e.g., an average rating of X out of 100). This can be provided for the individual items (e.g., "Good") and the metric (e.g., Global experience) by computing a mean scale score for the composite items, and
- as proportions of the audience/sample giving a score of 75 or higher (‘high’ impact scorers). For engagement, their score on a 5-point scale is multiplied by 20 to enable comparable data.

For Willingness to pay, participants are asked to indicate price points at which the experience they just had would be considered: too expensive; too cheap; quite expensive

but not out of the question; and a bargain. The data are graphed using the standard presentation format for Van Westendorp's price sensitivity meter [8], in which cumulative frequencies for each of the price categories e.g. 'too expensive', 'too cheap' are plotted, with the data for 'too cheap' and 'bargain' inverted to produce a graph with intersecting lines. The intersections are interpreted as providing an indication of different price points and a range of acceptable costs, including the optimal price point.

Table 1. Core Toolkit modules completed by end user audiences

Measures	**# items**	**Item examples**	**Scale**
Post-experience			
Global experience	5	The experience was... "NOT/Good" "NOT/Powerful"	1 (NOT) to100
Cultural value	6	The experience was... "NOT/An interesting idea" "NOT/Thought provoking"	1 (NOT) to100
Engagement	3	To what extent do you dis/agree...? "I lost track of time"	1 (strongly disagree) to 5 (strongly agree)
Emotion intensity/valence	3	How intense were any emotions you experienced? "NOT/Intense"	1 (NOT) to100
Willingness to pay	5	Would you be willing to pay for the experience you just had? At what price would you begin to think the experience is...? "Too expensive to consider" "A bargain, great buy for the money"	Yes/No Free response
Audience characteristics			
Demographics	10	e.g., Age, Gender, Ethnicity	variable
Arts/Cultural/VR experience	4	In the last 12 months have you visited any of the following in the UK for recreational/ entertainment and or educational purposes?	Tick any of 10 visits

IMMERSIVE CONTENTS AND SAMPLES

The results presented here are based on Toolkit data from 10 content evaluations across 460 people. Results for contents 1-3 have been published elsewhere [6], and contents 4-8 are pending publication [7].

The 10 contents varied in duration from 3 to 60 minutes, and included single and multi-user experiences. In some instances, audience members adopted different 'roles'. Some experiences used 360 video, whilst others used entirely computer generated imagery. Some required extensive interaction to unfold the narrative, whilst others afforded limited navigation. All pieces of content were suited to location-based presentations and some

were designed for use with one or more types of VR headset (most commonly the HTC Vive and Oculus Rift).

A range of characteristics are relevant to interpreting the results of each content's impact scores, a selection of which are shown in Table 1 for the 10 contents reported here. Characteristics include content affordances such as experience duration, level of interaction, and one, two or multi-person experience. Aspects of the context/ environment, and recruitment strategy for the evaluation also have potential to skew key characteristics of each sample with potential to influence the impact ratings. For instance, some contents were presented within the context of another experience (as part of a museum experience), expected to comprise mostly keen audiences, whilst other pieces of content were tested with unpaid, mostly student, research participants.

Table 1. Example characteristics of the contents tested

Content #	Sample size	Approx. duration (mins)	Single/multi person immersion	% audience aged < 45 yrs	% audience with VR/AR experience in past 12 mo	Test location /access
1	N=54	20	Single	81%	65%	Uni lab/private
2	N=57	10	Single	77%	53%	Uni lab/private
3	N=57	6	Single	73%	61%	Uni lab/private
4	N=36	50	Multi	86%	69%	Theatre/ public
5	N=36	12	Multi	65%	50%	Museum/public
6	N=61	5	Single	90%	64%	Uni/private labs
7	N=39	30	Single	87%	72%	Uni lab/private
8	N=30	30	Multi	100%	33%	Private install
9	N=34	60	Single	64%	62%	Private install
10	N=56	3	Single	73%	84%	Conference/public

RESULTS AND INSIGHTS

Comparing contents' evaluation ratings

Toolkit results give audience ratings for a given piece of content tested individually, and relatively compared to the other individual but anonymised contents. These show its absolute and comparative strengths and weaknesses. These can be produced for any of the core measures to illustrate results for the overall measures and composite items. Figure 1 shows the overall ratings for the measures of global experience, cultural value, engagement, and emotional intensity for each of the 10 anonymised contents, the blue icon is used to denote the content being reported, relative to the bank of other contents.

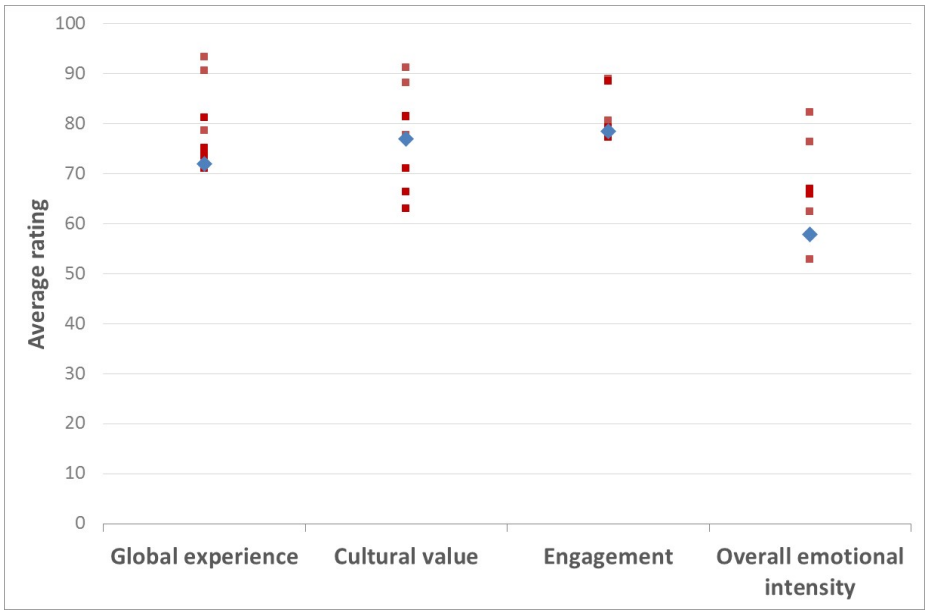

Figure 1- Distribution of impact and value ratings across the 10 tested contents

These types of chart also reveal the overall range of scores (min, max) across the bank of contents, and how they are distributed, for instance, whether there seem to be distinct clusters of ratings.

Exploring relationships across aggregated results

Using a similar representation to Figure 1 each content's relative position in terms of an audience characteristic can also be illustrated. For instance, audiences for each content can be compared in terms of propensity to engage within the past 12 months with arts and culture visits and activities. Analyses reported elsewhere from earlier phases [6, 7] have shown how sample characteristics are likely to have influenced evaluation ratings. For instance, higher emotional intensity ratings have been associated with respondents who are older, female, and with less experience with VR. Furthermore, audiences classified as engaged with either vr/tech or arts/culture were found to give higher ratings of cultural value and emotional intensity compared with audiences who were engaged with *both*, or *neither,* perhaps suggesting a 'benefit of the doubt' effect for audiences engaged with just one element of this arts/tech media fusion. Relationships between these variables can be explored in more depth and with more power, as the bank of (cross-content) samples increase.

That content-aggregated data analyses showed variation in ratings depending on audience and content characteristics also means that production teams can better

interpret their results in the light of their sample (e.g., types of skews) and content offerings/affordances, particularly where ratings are lower or higher than expected.

Exploring influence of physical parameters on audience evaluation ratings

Aggregating data across contents as the Toolkit is used more widely would enable an increasingly powerful search and criterion-based comparison tool. For instance, if a piece of content is VR-based, the production team might wish to compare their audience evaluation with *only* those gathered from contents also coded as VR. By coding contents by other parameters permits an exploration of other patterns within the data across the bank of content evaluations. For instance, the trend of evaluation ratings by content duration is shown in Figure 2. Durations across the contents ranged from 3 to 60 minutes (180 - 3600 seconds). In general, the longer the immersive experience, the greater the evaluated impact, particularly for cultural value, emotional intensity and global experience. A similar but weaker relationship is observable for Engagement ratings across duration.

There were notable exceptions to this pattern however, indicating that some short experiences could command relatively high audience ratings. For instance, one 3-minute content was given particularly high ratings of cultural value.

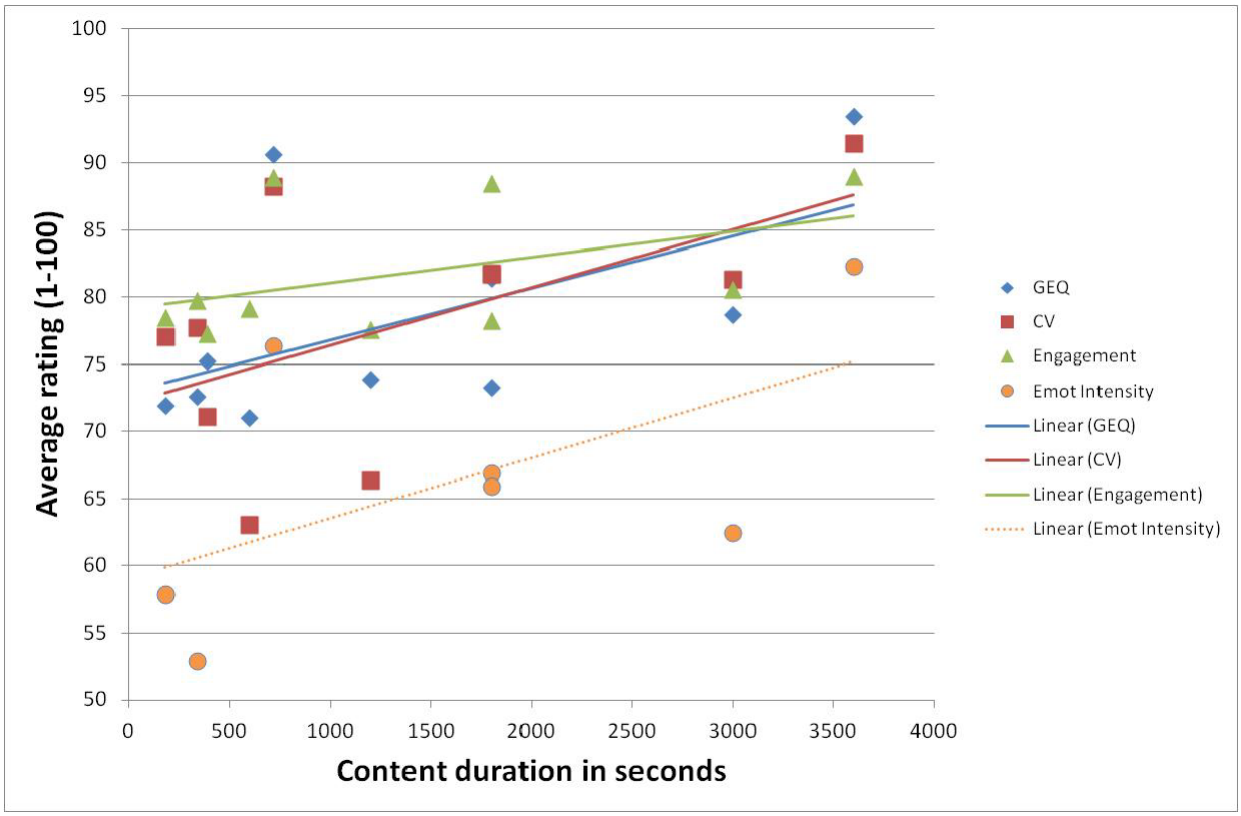

Figure 2. Relationship between content duration and impact ratings

CURRENT AND FUTURE APPLICATIONS

The outputs of the Toolkit, some of which have been described here, aim to support and guide decision making of stakeholders and content development teams working in the

creative immersive industries, through the provision of a short and pragmatic but meaningful, easy to use industry specific tool that gives absolute and comparative evaluation data.

Currently, result outputs of the Toolkit for production teams summarise the research highlights, audience characteristics, within-content strengths/weaknesses, anonymised cross-content comparisons, provide answers to their bespoke questions and finish with qualitative feedback from the comments section of the Toolkit. Results have implications for formative as well as final evaluations, marketing the experience, understanding impact on different types of audiences, and reputation. For instance, outputs could support commissioners at exhibition locations to facilitate more commissions; funders to evidence value their funding has created; distribution platforms to guide pricing; and investors to support the case for additional funding (for the same or different projects). For the production's creative teams, receiving insights about the strengths and weaknesses about the production is a resource to guide development and evidence impact. The Toolkit could provide a powerful measure that tracks change in impact and value, not only across contents and audiences, but also over time and from different perspectives, incorporating new data such as return on investment, and relationship to prestigious awards. As audiences become more familiar with this type of experience, it will allow the industry to explore trends in value as they develop.

We are currently developing the Toolkit and associated methodologies (e.g., passive objective indicators of audience impact) with funding from the UKRI's (UK Research and Innovation) Audience of the Future within the Performance Demonstrator led by the Royal Shakespeare Company. Developments include a Self-Service facility for the Toolkit with feedback report, which aims to support a fast turnaround of results to production teams and other stakeholders. We are also using the Toolkit across multiple productions, both within the Performance Demonstrator and outside it. Over the coming year we will evaluate the experiences of over 100,000 audience members of new creative and cultural immersive experiences. This will provide ample opportunity to dig deeper into the data to better understand relationships between impact ratings, audience characteristics and the technical properties, affordances, and styles of immersive contents. Looking ahead, our key priority is to distil the essence of impact to create the shortest possible metric whilst maintaining its robustness and validity.

REFERENCES

[1] Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., … & Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. Computers & Education, 94, 178-192

[2] Hu, J., Janse, M., & Kong, H. J. (2005). User experience evaluation of a distributed interactive movie. In HCI International. (http://www.idemployee.id.tue.nl/j.hu/publications/HCII2005_icecream.pdf)

[3] Ravaja, N., Salminen, M., Holopainen, J., Saari, T., Laarni, J., & Järvinen, A. (2004, October). Emotional response patterns and sense of presence during video games: Potential criterion variables for game design. In Proceedings of the third Nordic conference on Human-computer interaction (339-347). ACM.

[4] Suzuki, K., Roseboom, W., Schwartzman, D. J., & Seth, A. K. (2017). The Hallucination Machine: A Deep-Dream VR platform for Studying the Phenomenology of Visual Hallucinations. bioRxiv, 213751.

[5] Lessiter, J., Freeman, J., Keogh, E., & Davidoff, J. (2001). A cross-media presence questionnaire: The ITC-Sense of Presence Inventory. Presence: Teleoperators & Virtual Environments, 10(3), 282-297.

[6] Nesta and i2 media research for Digital Catapult (June, 2018). Evaluating Immersive User Experience and Audience Impact. https://www.digicatapult.org.uk/news-and-views/publication/audience-immersive-report/

[7] i2 media research for Digital Catapult and Arts Council England (2019, in preparation). CreativeXR and the audience experience

[8] Van Westendorp, P (1976). NSS-Price Sensitivity Meter (PSM)- A new approach to study consumer perception of price. Proceedings of the ESOMAR Congress.