Aesthetic experience and judgment: Exploring three-dimensional arts in real-world, virtual reality, and online settings

by

Doğa Gülhan

Thesis submitted to the Department of Psychology
Royal Holloway, University of London
in partial fulfilment of the requirements for the Degree of
Doctor of Philosophy

2021

Abstract

In three sets of experiments, this thesis aimed to demonstrate the ways of adapting emerging research tools and environments to be used in empirical aesthetics research, particularly by using 3D spatial artworks in-situ, in virtual reality, and in online settings. The main methodological problem addressed was to ask and answer some basic research questions, which are outside the frame-of-reference of the traditional lab-based research. In the broadest sense, the main reasoning behind the experiments was to open research perspectives beyond established experimental paradigms, arguably, in the second oldest discipline of psychology. The original contribution to knowledge is based on three studies exploring different aspects of novel opportunities to explore visual aesthetic experience.

Firstly, in the study of the Mondrian Room in the Albertinum Museum, whilst visitors were engaging with a physical installation and a virtual reconstruction based on Mondrian's design draft entitled "Salon for M. Bienert", their gaze patterns were measured using mobile and VR eye-trackers. Despite the presence of individual differences, overall, the eye-tracking analysis, along with questionnaire results, yielded high similarity between the two settings. Whilst the experiment showed the methodological ability to compare gaze data between physical and virtual environments, results obtained on the similarity of viewing patterns suggested a potential status of VR galleries as ecologically valid proxies to physical galleries.

Secondly, in the study of the Virtual Reality Gallery, participants walked through a virtual gallery after viewing a selection of 3D objects and environments, which do not share much similarity or have an obvious common property. During this period, their eye-movements were recorded. After that, participants provided a set of aesthetic judgments, such as liking, novelty, emotional valence, perceived viewing duration towards spatial artworks. Results showed positive, linear, and mostly moderate correlations between liking and the other perceived judgment attributes with positive connotations; supplementary VR eye-tracking analysis showed diverse viewing strategies of the artworks; and an online follow-up experiment showed converging correlational results compared to the virtual gallery. The results could be interpreted such that the relations between measured aesthetic ratings were similar when the artwork presentation medium was changed from the 3D VR environment to the 2D image.

Lastly, in the study of the Mural in the Compton Verney Art Gallery, a real-world experiment was followed up by an online study. The analysis mainly comprised absolute and area normalized dwell time on aggregated gaze data of gallery visitors collected from

mobile eye-trackers whilst they were viewing an abstract, room-scale mural. Results showed particular viewing trends such as preference on edges and vertices between colour patterns, and a horizontal central tendency. In the in-situ experiment, the results were interpreted such that the artwork's particular topological properties influenced viewing trends. Following on from this observation, parametrically modified variations of the artwork were generated and presented as rendered videos to participants online for aesthetic judgments. These manipulations of bottom-up factors such as colour and edge shaped the subsequent ratings, and participants provided vastly different justifications behind their judgments, along with exploratory online eye-tracking data which turned out to be very noisy. Results underlined the potential of data-driven approaches for the curatorial and design practice.

Based on this work, implications and scalability of the methods developed here such as simulated experiences, novel visualization tools and behavioural measures are outlined, emphasizing their potential in empirical aesthetics.

Declaration of authorship

Chapter 1: Introduction and background. I declare that this chapter is entirely my own work.

Chapter 2: Methods. I declare that this chapter is entirely my own work.

Chapter 3: Mondrian Room in Albertinum Museum. I declare that this chapter as a manuscript is co-authored with both of my supervisors, Johannes Zanker and Szonya Durant. The author contributions are as follows: Conceptualization, DG, JMZ; Methodology, DG, JMZ; Software, DG, JMZ; Formal Analysis, DG, SD; Investigation, DG, SD, JMZ; Writing – original draft, DG; Writing – review & editing, DG, SD, JMZ; Visualization, DG, JMZ; Supervision, SD, JMZ; Project administration, JMZ.

Chapter 4: Virtual Reality Gallery. I declare that this chapter as a manuscript is co-authored with both of my supervisors, Johannes Zanker and Szonya Durant. The author contributions are as follows: Conceptualization, DG, JMZ; Methodology, DG, JMZ; Software, DG; Formal Analysis, DG, SD; Investigation, DG, SD, JMZ; Writing – original draft, DG; Writing – review & editing, DG, SD, JMZ; Visualization, DG, JMZ; Supervision, SD, JMZ; Project management, JMZ.

Chapter 5: Mural in Compton Verney Art Gallery. I declare that this chapter as a manuscript is co-authored with both of my supervisors, Johannes Zanker and Szonya Durant. The author contributions are as follows: Conceptualization, DG, JMZ; Methodology, DG, JMZ; Software, DG, JMZ; Formal Analysis, DG, SD; Investigation, DG, SD, JMZ; Writing – original draft, DG; Writing – review & editing, DG, SD, JMZ; Visualization, DG, JMZ; Supervision, SD, JMZ; Project administration, JMZ.

Chapter 6: Overall discussion. I declare that this chapter is entirely my own work.

Doğa Gülhan

Acknowledgements

This research was supported by a College Studentship offered by Royal Holloway, University of London. I would like to pay my special regards to Johannes Zanker as my primary supervisor and to recognize the invaluable assistance of Szonya Durant as my secondary supervisor. This research would not have been possible without them. I am also very grateful to be a part of the Psychology Department, Royal Holloway, University of London. I would like to extend my deepest gratitude to my family and appreciation to my friends.

Contents

Chapter 1

Introduction and background

This chapter comprising four parts aims to summarise visual empirical aesthetics, by firstly describing the research field, followed by an overview of models and theories concerning aesthetics. After that, to provide the current directions in the field, methodological tools, measures, research questions, and findings from recent studies are presented. Lastly, some crucial research issues are remarked and linked to the presented research. Note that, with the inclusion of otherwise unmentioned arguments, and an overview of the field with a personal perspective on the field, this section is relatively broader and conceptually more scattered, compared to the individual introductory sections of the three main experimental chapters.

1.1. Empirical aesthetics

Aesthetics as a research field refers to the investigations concerning features of (art) objects and spaces, designated meanings by observers such as beauty or quality, related emotional or cognitive mental states, types of experiences, and judgments such as appreciation or pleasure. Aesthetics can be described in countless other ways, depending on the researcher and the expanding nature of arts such as alternative forms of artistic expressions or discourse in arts. A common theme or one main aim frequently sought after of research comprises definitions and functional properties of art, experience, judgment, and beauty; although expanding complexity of aesthetic investigations allows a cascade of new themes. To briefly and partially illustrate the overarching directions of aesthetic research, for the sake of simplicity, the scope of aesthetics can be reduced as a spectrum ranging from theoretical to experimental.

Traditionally, the theoretical research of aesthetics tends to be dominated by philosophy, and often aesthetics is still regarded as a branch of philosophy. Without introducing the glimpses of philosophy and theory in the ancient history of known civilisations, in the Western schools of thoughts, philosophical investigation in aesthetics can be easily traced back to the Ancient Greek philosophy, such as the works of Aristotle and Plato (Gurd, 2012), followed by later classical philosophy, for example, into the works of Plotinus, and scholasticism in the medieval period into the works of Aquinas and Augustine (Beardsley, 1998); whereas in the Eastern schools of thoughts, similarly past works on aesthetics can be seen in many cultures, such as in the Islamic context (Gonzalez, 2001), Indian context (Chakrabarti, 2016), and Chinese context (Li, 1994). Arguably, the philosophy of art truly matured in late modern philosophy, following the Western enlightenment period, around the 18th century and developed ever since (Gaut & Lopes, 2013; Neill & Ridley, 1995), and an enormous amount of philosophical arguments have been discussed extensively. Apart from philosophy, many areas of sciences such as sociology (Guyau, 1889) and anthropology (Boas, 1927; Layton, 1991) contributed mostly by their theory-driven approaches. Along with philosophy, humanities and sciences; art theory and criticism were also always present in various forms such as opinion essays or artist manifestos (Lack, 2017). Lastly, the experimental philosophy of aesthetics (Cova & Réhault, 2019; Torregrossa, 2020), albeit still leaning toward the theoretical end of the spectrum, might be seen as a promising emerging field to expand theoretical research.

On the other hand, the experimental research of aesthetics aiming to measure and quantify the experience and judgment of the observer, often linking those to the attributes of the observed, was developed in the late 19th century in line with visual and auditory psychophysics (Fechner, 1876; Helmholtz, 1863). In a way, empirical aesthetics can be defined as the second oldest branch of psychological sciences, slightly younger than psychophysics. Just to illustrate the historic context, in the early period of the discipline, the emerging technology called photography was debated whether it should be regarded as an art form (Eastlake, 1857). However, further developments were rather slow and scattered, overshadowed by the other flourishing branches of psychology, and often empirical aesthetics was seen as a peculiar and potentially obsolete area of study. Following a relatively long disinterest period until the late 20th century, a revitalisation of empirical research became visible, with advancements from various disciplines such as neuroscience and linguistics, in line with methodological developments such as neuroimaging (Crozier & Chapman, 1984; Jacobsen, 2010). Particularly, behavioural aesthetic science continues to deal with the experimentally observable aspects of aesthetic experience and judgment, by adapting methodologies derived from experimental and cognitive psychology, and sometimes from particular subdisciplines such as vision science or memory research (Palmer et al., 2013; Lindell & Mueller, 2011). Beyond that, computer science and particularly machine learning research are hugely contributing to experimental aesthetic research (Brachmann & Redies, 2017), generally in very pragmatic ways, for example, often aiming to make better predictions about stimulus or behavioural data. Currently, the scope and general aims of experimental aesthetics, computational aesthetics, and (neuro)scientific investigations are constantly updated by researchers, resulting in an excessive amount of dialogues between experimentalists and theorists, yet sometimes both humanities and sciences continue to criticise experimental approaches on aesthetics (Pearce et al., 2016). As always, dialogue between artists and scientists and other stakeholders seems to be a crucial aspect to further the empirical investigations.

Traditionally philosophy, and arguably all other potential research disciplines associated with aesthetic research have distinct frameworks and methodologies, yet they also have the potential to cooperate with each other to consolidate the overall impact of the claims. Parallel to the rise of philosophical interest as an aesthetic turn in recent decades, aesthetics can be described beyond just philosophical aesthetics or philosophy of art, or maybe beyond applied philosophy (van Gerwen, 2015). Although aesthetic(s) is often conceptualized as "a relatively distinct phenomenon requiring its relatively distinct theoretical reflection" (Erjavec, 1999), many emerging concepts (that were previously not counted as part of the art) can be easily investigated under the aesthetic research. Arguably, this "aestheticization of everything" might create a challenging theoretical research context, and forming explanatory philosophical claims may become more difficult. The same concern can be translated upon many types of empirical research as well, for example, proprioception was defined as an aesthetic sense (Montero, 2006), aesthetic preference upon choreography was investigated empirically (Orgs et al., 2013), or visual presentation of a dish on a plate was already a question in empirical aesthetics (Deroy & Spence, 2014). Parallel to the argument of "anything can be art", arguably, many seemingly odd inquiries about aesthetics in its most general sense can be (sometimes controversially) translated into valid research questions.

Whilst meanings and sometimes common-sense definitions of art-related concepts get fuzzier or at the very least highly context-dependent, ways of aesthetic investigation get diverse in line with including researchers among diverse disciplines. Currently, a wide scope of research themes engages researchers from very distinct backgrounds to work on aspects of aesthetic experience (Huston et al., 2015), and this form of comprehensiveness can be seen as a major prominent, defining, yet challenging characteristic of aesthetics research. Drawing the knowledge from several remote fields seems to call for an interdisciplinary (yet not a unified) field as aesthetic sciences, similar to the foundational impulses during the emergence of cognitive sciences (Shimamura & Palmer, 2011). In this sense, assuming that the nature of interdisciplinarity is aiming to link between the theory-driven (philosophical) aspects and the data-driven approaches of aesthetics, and is not heading towards a monstrous amalgamation, the mutually beneficial exchange of knowledge and method between disciplines is a crucial aspect of the current state and the future of aesthetic research.

1.2. Models and theories of aesthetics

An overview of either historically important, recent, or neglected yet stimulating ideas is presented in this section, aiming to underline commonalities and link these models and theories to the theoretical framework of the presented empirical research. This section is a short personal primer on the model/theory building in aesthetics, without a strictly enforced selection criterion on compatibility or relevancy. Besides, although coming from an undoubtedly biased and arbitrary selection, evaluating art theories in full were simply beyond the scope of this research, for example, aiming to summarize Kantian aesthetics in a paragraph may be at least unfair to a Kant scholar. Since theories of art constitute a distinct set of generalised explanations, which is arguably different from the positivist scientific theories (Freeland, 2002), the following selection can be clustered into two parts: pure theories, and theories derived from (or strongly associated with) empirical research. The theories leaning towards the first part are more or less deductive in the sense that they are often developed by intuition and introspection, whereas the theories leaning towards the second part are inductive in the sense that they mainly rely on the cumulative results of some objective measures, aiming for a less speculative argument building. For the sake of simplicity, a model or a theory can be described as a set of arguments, aiming to explain a phenomenon in a generalisable sense, and often has a causal component. It is important to note that, in both instances though, at least personal biases can radically shape the contents of these models and theories.

Because aesthetics was dominantly addressed as a mere study branch of philosophy and linked to art practice and art theory; non-empirical research and resulting theories can be traced back in line with philosophy in the 18th century (Burke, 1767; Hume, 1757; Kant, 1781, 1790), forming the concept of modern aesthetics that was shaped by the European Enlightenment (Ahlberg, 2014), and followed by the evolution of arts and art theories in the 19th and 20th century (Harrison et al., 2001; Harrison & Wood, 2003). There are glimpses of philosophical inquiry in early philosophy as well (Aristotle, 350 BCE | 2014), for example, defining beauty as a relative quality attributed to objects in a relational manner (i.e., an object is more or less beautiful than another), and further describing that beauty subsisting in an object can vary depending on time. It can be heuristically thought that initial philosophical theories were roughly based on judgment-based aspects upon arts, whereas later on attitude and experience-based theories have emerged (Shelley, 2017).

Although historical theories are conceptually valuable, their validity may have been decreased over time, sometimes over centuries, simply because of the cultural progression, and artistic diversification in particular. In terms of cultural progression, for example, overlooked female artists and gender-stereotypes in artworks was only introduced to art theories after the 1970s and can be seen as one of the main directions in feminist aesthetics (Ecker, 1986; Korsmeyer, 2004). In terms of artistic diversification, for example, the emergence of conceptual art (Kosuth, 1991) resulted in theories regarding art pieces without a need to be perceived by sensory input, or art pieces without perceptual properties, or simply non-perceptual art (Shelley, 2003; Carroll, 2004). The key challenge for the theoretical research may be to keep its up-to-date arguments in line with the contemporary art world where for example ML algorithms are generating art (Elgammal et al., 2017; Elgammal et al., 2018), our basic attribution of artist or creator of the artwork is becoming a valid empirical research question (Epstein et al., 2020), and result in an emerging set of issues related to ownership of art production (Eshraghian, 2020).

During the history of philosophy of aesthetics and of art theories, at least one common characteristic is visible: many philosophers and theorists (particularly from the analytical tradition) argue for a form of essentialist account, which often tries to define required features or necessary and sufficient conditions for the artworks and the experience of art. "Necessary and sufficient conditions" is a useful concept and applied extensively in philosophy, from logic to ontology (Brennan, 2017). This concept, however, may not be congruent with arts. The challenge to merely define art(work) was remarked (Ziff, 1953) and some views emerged to reject such necessity and sufficiency as well as to argue for that art as an indefinable open concept (Weitz, 1956). These viewpoints are often described as anti-essentialism, and they were influenced by the concept of family resemblance (Wittgenstein, 1953), which roughly argues that if concepts belonging to a single categorical set can be linked together by some shared commonalities and not necessarily have an all-encompassing property, then it is possible to have two objects in this set who belong together yet do not share a single common feature. Similarly, the hope of "a forthcoming, correct theory of art" is challenging, and the possibility of a true theory might be rejected (Weitz, 1956): an art theory can be just logically impossible, simply because that art itself does not have any necessary and sufficient properties. Weitz's controversial claim was that the aesthetic theories should not try to provide definitions about arts but should be regarded as "recommendations to attend in certain ways to certain features of art". These somewhat overlooked anti-essentialist viewpoints are readdressed in philosophy from time to time, for example, defining art as a cluster concept

(Gaut, 2005), but arguably, similar ways of thinking may be less controversial, or more compatible with the current experimental research.

The wide scope of written materials as art theories and personal insights upon building theories are valuable for many researchers today, but the arbitrary nature of defining anything as a potential theory is very challenging to be directly linked to empirical research. For example, a simple philosophical belief about whether there should be a distinction between the aesthetic and non-aesthetic object (Sibley, 1959) or should not (Iseminger, 1973) seems to be not showing a consensus soon, both among theorists and among experimentalists. The gap between theories of arts and empirical research has also been visible and raised (Makin, 2017), often calling for more dialogue between disciplines as well as further empirical investigations. This gap is especially visible when one is inclined to think that the foundation of aesthetics is philosophy, while empirical investigations are built upon sciences such as cognitive and experimental psychology. Diverging from pure philosophical studies, the main viewpoint of empirical research tends to be that if the aesthetic experience is defined as a cognitive-emotional process, almost always as a part of conscious experience, then on the empirical level, controlled experiments can be conducted, and on the theoretical level, cognitive modelling may be utilised to explain these processes.

The empirical investigation of the aesthetic experience in the laboratory, albeit interrupted through the decades, has a historical background. Fechner, alongside the seminal works on psychophysics (Fechner, 1832, 1860, 1877), contributed hugely to the foundation of empirical aesthetics (Fechner, 1876), and derived methodologies from psychophysics particularly into empirical aesthetics, such as the method of choice, production, and use. The main claim was to demonstrate and quantify relations between stimuli and mental states, and in aesthetics, those were the artworks and measurable aesthetic experiences. It is important to note that the objective, as well as perceptual attributes of stimuli, were underlined in the writings. On the theoretical level, some surprising precursor ideas can be visible, for example, postulating two-level processing of aesthetic appreciation consisting of an initial impression state followed by a continuing evaluation state, which is mostly intact in the contemporary models. Overall, the general argumentation translates into arguing against dualist accounts of the mind-body (but not necessarily arguing for a type of physicalism). Fechner's contributions were and still are often (either weakly or strongly) defended, and his attempts towards modelling aesthetics can be still seen as a very influential anchor point (Ortlieb et al., 2020).

Although Gestalt psychology did not produce complete cognitive models for aesthetic experience, arguments on visual aesthetics and aesthetic qualities were formed. For example, Koffka emphasised the effects of environmental and social factors (apart from the more intrinsic properties of the artwork itself) on the aesthetic judgments between individuals (Koffka, 1935): this brief yet important remark has been addressed again very later on in empirical research (Bullot & Reber, 2013) and philosophy (Porter, 2009). Arnheim, whose works have strong links to the Gestalt school, for example, underlined the disassociation between natural language and vision (favouring the importance of vision over verbal language in visual arts), and roughly described the use of language in arts as a partial and inadequate representation of visual thinking: in a way, equating the (visual) perception and (visual) thinking into the same construct (Arnheim, 1954). At the same time, researchers outside the Gestalt tradition were, for example, approaching the problems of aesthetics from a mathematical perspective (Birkhoff, 1933), or trying to find a formula (Eysenck, 1941) to describe, for example, the relation between aesthetic perception and properties of art objects, which is arguably a reductionist view compared to current perspectives. However, to briefly underline Gestalt's historic importance, the assumptions of Gestalt theory were often transformed into baseline hypotheses during experimental design (McManus et al., 2011); and conceptual implementations of Gestalt theory into visual arts, followed by the discussions and interpretations upon them were visible in the upcoming decades and up to today (Beardsley & Arnheim, 1981; Cupchik, 2007; Spehar & van Tonder, 2017). Apart from the aesthetic experience, Gestalt psychologists' descriptive generalisations influenced theory-building particularly in visual sciences, either directly (Biederman, 1987) or indirectly (Marr, 1982; Gibson, 1979).

Alongside years of model and theory building in empirical aesthetics in the 20th century, it is important to keep in mind that historical changes of visual arts, related to emerging and changing art practices, paradigm-shifting artist statements, such as on Neoplasticism (Mondrian, 1920; van Doesburg, 1923), and countless other radical changes shaped the theories of art in modern art (Chipp, 1968). Seminal assumptions on aesthetics (Fechner, 1876) may be more in line with the scarce number of artistic expressions in Western visual arts in the mid-to-late 19th century. Some recently proposed models, for example, underline the importance of special cases such as post-modern and abstract art and incorporate both bottom-up (as mainly object-driven) and top-down (as mainly subject-driven) aspects into the framework (Redies, 2015). In line with scientific progress, some theories of art already strengthened efforts to implement scientific methods into their arguments, including, for example, underlining the effects of the Zeitgeist on the perception of art in Renaissance and conceptualizing societal effects on individuals' appreciation of visual arts as the concept of "period eye" (Baxandall, 1988), interests on the cultural influences on artworks (Deręgowski, 1984), or emphasis on visual neuroscience and proposal for biologically based theories of art (Zeki, 1999), and suggesting principles or heuristics based on neuroscientific research (Ramachandran & Hirstein, 1999). More recently, incorporating neuroimaging into the aesthetics settings resulted in the foundations of neuroaesthetics as a field, which comes with major promises and challenges (Nadal & Pearce, 2011), along with the usefulness and other questionable aspects of "the era of 'neuroeverything'" (Bassett et al., 2020).

As an attempt to contextualise aesthetics in the framework of neuroscience, it was proposed that the aesthetic experience emerges from the interaction between sensory-motor, emotion-valuation, and meaning-knowledge neural systems (Chatterjee & Vartanian, 2014). Two temporally distinct phases of art perception which are early aesthetic pleasure (activated by early insights) and late aesthetic pleasure (requiring more insights) were also proposed (Consoli, 2015), similar (but in a more detailed manner) to Fechner's original proposal back in 1876. A search for detailed and more unifying models such as the psycho-historical framework (Bullot & Reber, 2013) or the pleasure-interest model of aesthetic liking (Graf & Landwehr, 2015) is an ongoing challenge. Although currently there are methodological suggestions towards a general model of neuroaesthetics (Marin, 2015), other proposed unifying models of aesthetic liking (Leder et al., 2004) and promising research avenues in neuroaesthetics (Cela-Conde et al., 2011; Chatterjee, 2014) and experimental philosophy as an emerging movement in the field approaching philosophical questions with empirical evidence (Kamber, 2011; Monsere, 2015; Liao & Meskin, 2015), more inclusive models are being proposed such as The Vienna Integrated Model of top-down and bottom-up processes in Art Perception (Pelowski et al., 2017), yet more work seems to be needed for such a hard problem.

Compared to the complex characteristics of contemporary visual arts, the currently proposed models aiming to explain aesthetic phenomena might fall short. Because of the necessity of solid interdisciplinary frameworks (Marin, 2015), the development of better models and theories is arguably slow or even avoided, in comparison to advancements in a single research area. Additionally, the research questions previously asked only in the context of "pure" aesthetic experience and traditional art forms are expanded, for example, into the interactive aesthetic experience and towards design products, which further results in model building to incorporate such hands-on experience with artefacts

(Locher et al., 2010). Similarly, the multisensory perception was sometimes discussed in addition to the visual-only or visual-dominant models on aesthetics, such as including haptic perception into such frameworks (Carbon & Jakesch, 2013), since, for example, experimental research on haptic properties of the stimuli underlines this understudied state in the experimental aesthetics (Calbi et al., 2019).

Lastly, some additional remarks beyond the empirical aesthetics can be underlined, to provide stimulating glimpses on wider theoretical issues. Traditional aesthetic theories are under some generic criticisms including the over-subjective nature, argumentation issues including logical fallacies or circular definitions, or incomprehensiveness. Art-specific neurological model and theory building was also criticised within the discipline (Skov & Nadal, 2018). As artists continue to challenge art practice itself, any definitional boundary of arts becomes more ambiguous: as a result, elaborative (cognitive) models have to account for such fuzziness. (In this sense, yet outside the domain of aesthetics, see the introductory paper on fuzzy sets (Zadeh, 1965), a brief editorial on fuzzy models (Bezdek, 1993), and a neuroscientific adaption as fuzzy-trace theory (Reyna, 2012)). Similarly, although beyond the current scope of the empirical aesthetics, a critical analysis on cognitive modelling was pointed out (Datteri & Laudisa, 2014), underlining a major issue in box-and-arrow (BA) models requiring the use of abstract concepts: a cognitive model as neuroscientific mechanism descriptions (NMDs) consisting only direct neural mechanisms such as processing brain regions and functional relation between them may be equally explanatory as BAs. For example, in a potential case for empirical aesthetics, including hundreds of concepts, arrows and boxes into a single page may not be a comprehensive theory of aesthetics at all. Finally, comparing (scientific) theories is in itself a major methodological research question, which has been addressing partially in the philosophy of science (Fine, 1975) with recent interdisciplinary approaches such as incorporating Bayesian framework (Huber, 2008), and some alternative ways of psychological theory testing are emerging, such as ontology-based modelling systems (Hale et al., 2020), although not applied into empirical aesthetics to the best of my knowledge. Theory crisis of psychology is an ever-underlined concept (Eronen & Bringmann, 2021), and particularly to the experimental aesthetics, proposing ambitious yet arguably excessively verbose frameworks are often avoided in other areas of cognitive and experimental psychology: as a notorious toy example, it took about 25 years and a huge set of experiments to revise the model and introduce a fourth component (Baddeley, 2000) to Baddeley's three-part model of working memory (Baddeley & Hitch, 1974). In a way, the inclusion of all mechanisms associated with conscious experience into aesthetic experience starts to resemble answering the hard problem of consciousness (Chalmers, 1997), and theory forming for consciousness, which ranges from arguably relatively more testable theories such as integrated information theory (Tononi et al., 2016) to less testable ones such as field theories of consciousness (Hagelin, 1987; Pockett, 2012).

The valuable philosophical arguments from theorists, coupled with proposed models from experimentalists can be highly useful in research. Although major philosophical arguments and comprehensive theories are not necessarily falsifiable merely by the experiments, partial aspects of models can be further potentially tested: for example, (i) an experimental paradigm using the rapid presentation of artworks (De Winter et al., 2020), similar to the research on gist in scene perception, might be particularly further adapted to target the suggested initial and automatic processing of aesthetic experience, (ii) an expert-novice experiment using a between-group design could test the amount of contribution of knowledge-meaning aspect on the experience, or (iii) an experimental paradigm incorporating imaginary settings can challenge the assumptions about medium besides art gallery and lab. This thesis is not directly aiming to test major aesthetic models and theories, or derivative models and theories from cognitive and experimental psychology. Instead, this thesis inevitably accepted some inductive ideas on aesthetics, and made use of some claims, especially in the experimental design phase: to briefly illustrate; (i) the study of Mondrian Room in Dresden, 3rd thesis chapter, was based on the assumption that the eye-tracking can be defined as a proxy measure of visual attentional allocation during the aesthetic experience, and it can be analysed to assess the similarity of the experience between physical and VR settings, without relying on the subjective responses. (ii) In the study of Virtual Reality Gallery, 4th thesis chapter, it was assumed that if relations exist between perceptual attributes such as liking, novelty, and perceived complexity, then these relations might be observable irrespective of a specific artwork category such as impressionist landscape paintings, i.e., even when these attributions (as judgments) were provided to a wide range of objects and spaces which do not share an obvious common property. On the conceptual level, an anti-essentialist standpoint was assumed to provide methodological flexibility without the constant need for philosophical justification, such as whether a digital reproduction of an installation in VR can be described as an art or not. (iii) In the study of Compton Verney Mural, 5th thesis chapter, if the viewing behaviour towards an abstract room-scale installation was assumed to be mostly driven by the bottom-up factors, then using the measured fixation density maps along with the computed saliency maps, parametric modifications regarding the topological layout of the installation should shift the canonical judgment of liking on those variations.

1.3. Experimental approaches

An overview of prominent research methods and particular findings can be summarised to describe the progress and the status of visual empirical aesthetics. This section presents some examples of recent research from the last ten years, whilst also aims to acknowledge some outdated or otherwise noteworthy studies to provide further historical perspective. Specifically, this section starts by (i) introducing some well-established research tools such as eye-tracking, neuroimaging and other behavioural measures, and emerging tools such as VR; (ii) followed by some of the frequently used and manipulated measures and concepts such as colour, symmetry, novelty, expertise; and lastly, (iii) mentions relatively understudied concepts. Some methodological aspects mentioned here were adapted in the experimental chapters of this thesis, for example, either as a part of lab-based psychophysical tasks on aesthetic judgments or as a part of real-world and VR settings using eye-tracking and complementary questionnaires.

This section only aims to provide an overview of recent research but does not aim for a complete synthesis of visual empirical aesthetics in the traditional sense. Similar to the previous section on models and theories, this section presents methodological aspects, key questions, and findings in a personally biased way, and arguably as a scattered selection. Some of the wider ideas related to these questions can be found in a range of publications, for example, (i) on diversified ways of experimental techniques to collect behavioural data, particularly in genuine settings (Locher, 2011); (ii) on the last decades of a specialized journal, i.e., Empirical Studies of the Arts, with overarching art domains in first half and with a particular emphasis on analysis on music perception in the second half (Greb et al., 2017); (iii) on linking the recent advancements in computational aesthetics to empirical research (Brachmann & Redies, 2017); (iv) on cross-cultural empirical aesthetics, which mainly argues for some observed commonalities or aesthetic universals across cultures (Che et al., 2018); (v) on approaching empirical aesthetic in a more developmental sense, whilst underlining the concept of schemas, which roughly refers to the cognitive patterns or structures related to linking prior knowledge and current experience (Jacobsen & Beudt, 2017); (vi) on a non-standard take about conceptualising aesthetic appreciation whilst underlining the emotional aspects of the experience (Fingerhut & Prinz, 2018); (vii) on a particular information-processing model as a ten-year follow-up (Leder & Nadal, 2014).

In most cases of visual aesthetic experience, an active observer using their eye movements is present for the experience and evaluation after capturing the image contents, irrespective of whether the object or event of a visual art form is physical or virtual. After all, the baseline condition of a generic aesthetic experience can be easily conceptualised as an active exploration, an action-based visual perception: for the sake of simplicity, as an active interaction between the agent, artwork, and environment. The wide context of aesthetic experience involving perception, memory, decision making, and other cognitive-emotional processes are allowing researchers to implement controlled, quasi, or field experiments. Therefore, researchers can include a range of methodologies such as eye-tracking, visual psychophysics, neuroimaging, questionnaire, content analysis, and develop other methods without even a need for participants in line with the long tradition of computational aesthetics (Gips & Stiny, 1975; Sartori et al., 2016). Apart from exploratory research, experimenters can manipulate variables on object-based levels (such as colour, illumination, symmetry), on observer-based levels (such as expert vs. novice studies), on environment-based levels (such as lab vs. real-world comparative studies), among others. Even if the inconsistencies in empirical works were ever-remarked (McWhinnie, 1965), or for example, empirical research was casting doubt on canonical beliefs and observations about visual preference for particular proportions such as golden ratio (McManus, 1980), here, a review of the methodological diversity can be provided to better picture the historical and current state of empirical research. Each of the following paragraphs mentions one key theme (either a particular method, type of measurement, or a frequently readdressed variable of manipulation), and can be read as snippets.

As one of the earliest eye movement experiments, eye-tracking methods were employed on paintings; and therefore the relations between painting properties, individual differences, and instructions given to observers were further investigated (Buswell, 1935), which had methodological similarities and precursor ideas to later seminal works on eye-tracking (Yarbus, 1967). In line with the methodological claims in psychophysics tradition, Buswell's principal contribution can be summarised such that eye-tracking as an objective measure can be targeted to unconscious choices or automatic processes during the aesthetic experience. The recent developments of eye-tracking technology (Kowler, 2011; Liversedge et al., 2011; Holmqvist & Andersson, 2017) revitalise eye-tracking in aesthetic research. Since visual cognition is a crucial aspect of visual arts, often metrics on eye movements (and visual attention as a related concept) are required to grasp relationships between elements of the art object (Rosenberg & Klein, 2015), leading to, for instance, analysing the change in eye-movements based on the effects of subtle lighting in early Renaissance paintings (Leonards et al., 2007), linking between pupil diameter and arousal ratings (Powell & Schirillo, 2011), or investigating the effects of abstraction level of painting on observers' gaze patterns (Pihko et al., 2011). By measuring shifts in eye movement patterns depending on the experimental condition, one might aim to distinguish between top-down and bottom-up contributions of aesthetic experience (Massaro et al., 2012). Further studies linked eye movements to aesthetics in more subtle aspects: for example, the interaction between gaze patterns and expertise were also demonstrated by art students in drawing classes (Ishiguro et al., 2016), asking for further investigations of the role of eye movements in processing aesthetic attributes. Eye-tracking can be also used as an active parameter on aesthetic preferences (Makin et al., 2016). Additionally, real-world empirical examination including mobile eye-tracking studies is a relatively new area of exploration (Heidenreich & Turano, 2011; Zanker et al., 2016, 2017), offering a new and arguably ecologically valid research setting.

In parallel, virtual environments, under the umbrella concept of extended reality (XR), including virtual reality (VR) and augmented reality (AR), are slowly being accessible tools for empirical studies in behavioural vision research (Scarfe & Glennerster, 2015) after their decades-long history in psychophysics (Wright, 1995) alongside with their use in other disciplines, such as in digital archaeology (Terras, 1999). To illustrate the scalability of the research potential, for example, the experience of observers within architectural rooms presented in a VR setup (as a semi-spheric setup instead of a headset), and the correlation relations between room properties (such as openness and brightness) and perceptual ratings (such as interestingness and pleasure) can be investigated (Franz et al., 2005). Note that, since experimental results from the real-world contexts generally differ from the lab-based conditions (Brieber, Nadal, et al., 2015), more ecological-valid experimental designs using either real-world scenarios or VR tools may be sought after (Duchowski, 2017). VR is especially useful to test various experimental conditions which are otherwise impractical to create in real-world settings since generating such conditions only requires easy manipulations on digital 3D models. This flexibility can be also highly useful beyond the aesthetics research, for example, in urban design research the real-world observations or lab-based studies (Wilkins et al., 2018; Burtan et al., 2021) can be further investigated in VR and AR. However, implementations of virtual environments in aesthetics research and design research are arguably relatively scarce.

Neuroimaging has the potential to provide insights through experimental findings (Vartanian & Skov, 2014), but it may be too early to generate major claims on aesthetic experience. For example, empathy was suggested as a key aspect of aesthetic responses, and an arguably radical proposition was that the mirror neurons are the primary neural element for such responses (Freedberg & Gallese, 2007). This particular idea was also challenged by claiming that any activation of the mirror neuron system can be neither necessary nor sufficient for aesthetic appreciation (Casati & Pignocchi, 2007). Additionally, in neuroaesthetics, reverse inference, i.e., explaining a psychological phenomenon based solely on neural activation, may be a pitfall to be avoided (Chatterjee, 2014). Although neuroimaging is the reason that the neuroaesthetics research exists, and it can provide, for example, spatiotemporal yet partial representations of aesthetic experience, judgments or other parts of aesthetic episode in the brain, sometimes, the neural underpinnings of aesthetic experience seem to result in some bold claims and challenging discussions in the field.

Complementary bodily measures were also implemented in research, for example, the bodily reactions of participants whilst they were engaging with artworks, by employing facial electromyography and skin conductance response (Gernot et al., 2018). As their main findings indicated that high emotional contagion resulted in more bodily reactions, the researchers claimed that emphatic response may be a key factor of aesthetic experience. In a similar direction, psychological models valuing emotion as a key aspect of aesthetic experience were proposed (Silvia, 2005), and are providing a justification behind the use of these additional measures. (albeit their relative rarity of use in the field).

Colour and luminance information can be seen as two important aspects of vision, and art perception in particular. To investigate their contribution, for example, two paintings can be merged into one by superimposing the luminance component of either the first or second painting and colour components of both (Anstis et al., 2012): the results indicated that the perceptual resemblance of the superimposed image depends on the luminance component. This "visual analogue of the auditory cocktail party problem" suggests that perceptually, the luminance information is coupled with the congruent colour information while disregarding incongruent colour information.

Symmetry is another prominent object-based measure and often integrated into the research questions, for instance, arising from conceptually associated concepts to the symmetry-asymmetry scale, investigating asymmetries in figurative face depictions in Italian Renaissance art (McManus, 2005). Particularly, the effect of symmetry on liking is a commonly addressed question. For example, following the participants' rating judgments on both fully symmetric and almost symmetrical basic patterns (Gartus & Leder, 2013), the results suggested that a small deviation from fully symmetrical patterns, referred to as broken symmetry, increased perceived complexity, and yet decreased liking. Another study investigated implicit aesthetic responses by using an implicit association test on 2AFC designs on both black-and-white visual patterns and words related to arousal and valence (Bertamini et al., 2013), and their results indicated that observers associate symmetrical patterns with words high in arousal, positive in valence, and simple mathematical expressions. Although the concept of symmetry is studied comprehensively, perceptual symmetry and mathematical symmetry can show discrepancies if the models do not consider, for example, contextual effects (Cohen & Zaidi, 2013). Additionally, research symmetry is sometimes combined with additional measurements apart from rating judgments: for example, evolutionary algorithms coupled with eye-tracking was implemented to measure preference on patterns with types and levels of symmetry (Makin et al., 2016). Therefore, perceptual symmetry of a set of artworks is a potential candidate to be used in a psychophysical task and to be correlated with other measures.

Statistical image properties such as (perceived) complexity, prototypicality, or self-similarity can be assigned to images, design objects, or artworks (Hekkert & van Wieringen, 1990). The relation between (measured) complexity and image pleasingness was generally found famously as an inverted U-curve (Berlyne, 1971), however, recent research showed that this relation may be simply due to an analysis error: as some researchers suggest that averaging all participant data merely generates a non-existing average observer by disregarding subgroups of participants who show, for example, two completely different trends of relations (Güçlütürk et al., 2016). Divergent results depending on the method or analysis technique may call for more elaborate research on the types of complexity and their contribution at varying degrees (Nadal et al., 2010). In another study, self-similarity was also found to be a predictor of aesthetic preference, along with colour measures (Mallon et al., 2014). The inverted U-curve relation is sometimes observed between some other similar properties, such as between the amplitude spectrum of synthetic images and other perceived attributes such as preference, interestingness, perceived complexity (Spehar & Stevanov, 2021). Similar to perceptual symmetry, perceptual complexity can also be easily indexed by observer ratings.

Additionally, another study investigated the contribution of novelty and typicality on the preference of industrial design objects and found that an optimal combination of both can predict the preferred products (Hekkert et al., 2003). Further findings imply that artificial and visually pleasing images (i.e. print advertisement, visual artworks and architectural images) share higher-order image properties, that are distinct from other types of visual stimuli (Braun et al., 2013), but, as the researchers pointed out, the extent of sufficiency or necessity of these properties on aesthetic experience is a hard question to answer. As a very specific concept, detectability of objects in cubist paintings was found to be positively correlated with liking (Muth et al., 2013), suggesting the importance of the ability to assign relatively abstract labels of meanings for aesthetic appreciation.

A crucial aim for many pieces of research seems to link low-, mid-, and high-level vision to aesthetic experience, and promising research can be visible in related research areas such as scene perception, for example, investigating the contrast gain mechanism for the perception of natural scenes (Bex et al., 2007). Researchers also investigate saliency, for example, whether objects in a visual scene predict fixation better than early saliency (Einhauser et al., 2008) or not (Borji et al., 2013). A cross-cultural study implementing a random-phased version (i.e., following a fast Fourier transformation, randomization of sinusoidal components of power spectra, inverse fast Fourier transformation) of images of low- and high-ranked buildings coming from a standardized set of stimuli (which are no longer recognizable as buildings) still showed a difference in aesthetic judgments among participants, implying contribution of low-level vision (an particularly the information about spatial frequency) to aesthetic preference (Vannucci et al., 2014). Note that, vision-dominant research may not argue for the claim that aesthetics should be modelled in terms of (pure) vision, as such a standpoint seems to be somewhat hard to defend due to its reductionist core.

The role of expertise can be seen as an observer-based aspect that may shape judgment (Harel et al., 2011). For example, previous findings suggested that for the assessment of the quality of paintings, experts value originality more than novices (Hekkert & van Wieringen, 1990). Additionally, both the common and discrete properties of experience between art experts and non-experts, by employing a natural grouping task and correspondence analysis was investigated (Augustin & Leder, 2006): apart from many shared characteristics of experience (e.g. similar categorical attributions on artworks), their results indicated that appreciation criteria of experts are related to prior knowledge such as style while criteria of non-experts relied on personal taste and feelings. The effect of the "educated eye" is present in other artistic domains, for example, two groups of students (either from architecture majors or not) showed differentiating gaze patterns during viewing an architectural scene (Lee et al., 2015). Indeed, previous research by implementing fMRI on architectural expertise showed dissociable patterns of activation in areas related to perceptual processing, memory, and reward processing (Kirk et al., 2009).

Time can be inherent to the artwork, particularly for example for visual arts that have explicit temporal narrative sequences such as movies, kinetic art, or interactive art. Time can be also thought of as a crucial variable during the aesthetic experience (even for atemporal or static artworks such as paintings and sculpture in a traditional sense), and linked to, for example, the amount of effortfulness during the aesthetic episodes, leading to temporal misjudgements (Cupchik & Gebotys, 1988). The aesthetic experience as a process has temporal dynamics, which can be investigated and modelled e.g. by using continuous ratings from participants (Brielmann & Pelli, 2017). However, research findings suggested that beauty and pleasure (amplitude) are independent of viewing durations ranging from 1s to 30s (Brielmann et al., 2017). Similarly, another study conducted in museum settings using a mobile eye-tracker showed that aesthetic judgment of paintings does not correlate with fixation duration or viewing time (Heidenreich & Turano, 2011). Note that the average time spent on engaging art may depend on some other aspects such as group size, as larger groups of visitors tend to spend more time per artwork (Smith & Smith, 2001). These results imply that as long as the viewing duration is not very short as in millisecond ranges, approximating the gist perception paradigms in vision sciences, any free-viewing duration or arbitrarily-chosen displaying duration in a single-observer condition would not radically bias any potential experimental design.

Type of tasks as an experimental manipulation can affect various aspects of cognition. For example, the type of scene-related tasks (i.e. memorising, searching for an object, evaluating aesthetic preference) affects viewing patterns and scene memory performance differently (Choe et al., 2017). Beyond the experimental aesthetics, the type of task engaged by the observer (for example, whether the task is scene memorisation, scene search, reading or pseudo-reading) can be predicted based on eye-movements using multivariate pattern classification (Henderson et al., 2013) or linear discriminant algorithm (Kardan et al., 2015). Although some studies defend the rough conceptual premise, that aesthetic evaluation or art perception tend to show differences compared to the daily, mostly pragmatic perception, it is not clear whether or how a person defines their aesthetic experience as a task. Speculatively, a researcher can assume that, for example, artwork viewing is not explicitly a task for the participant, or a very particular type of task that can be compared to other non-aesthetic tasks.

Type of artwork, such as whether it is a representational or abstract painting, may shift aesthetic judgments, since depending on the artwork category, as factors shaping the judgment such may have different weights. For example, aesthetic judgment on two types of art as figurative vs. abstract art and two types of architecture as classic vs. contemporary architecture showed a difference in reaction times in a paradigm using an implicit association test (Mastandrea et al., 2011). Another study found that observer judgment on likeness showed more variability in abstract art compared to representational art (Schepman et al., 2015), partially in line with previous findings using abstract and real-world scenes (Vessel & Rubin, 2010). As a related concept, representations of art pieces such as type of camera-related aspects can also shape perception. In a study using eye-tracking on photographs of landscape types having various openness and heterogeneity, and different view angles of the camera, the researchers found a significant influence of both three factors on gaze patterns (Dupont et al., 2013). In relation to the implementation of the research presented in this thesis, using consistent camera-related parameters in VR settings might be crucial to overcoming any systematic error in eye movements. Additionally, using a survey tool, in a hypothetical research scenario, in which engaging with the original artwork was not feasible, researchers suggested that preference and the value of the experience might depend on the types of substitutes, such as reproduction, digital monitor-based views, or optical mirror-based views (Bertamini & Blakemore, 2019). As a related concept, the environment can be defined as the context of experience, and experiments in laboratory settings may not be compatible with the real-world as in art gallery or museum settings. For example, observers assigned artworks as more arousing, positive, interesting, and liked when viewed in the museum context, compared to on-screen, in lab context (Brieber et al., 2014; Brieber, Leder, et al., 2015), underlining the contextual effects on art experience. However, the ecological validity of the current VR systems in comparison to real-world scenarios are not yet studied extensively.

The use of language might shape many aspects of cognition, most famously underlined by the linguistic relativity hypothesis (Sapir, 1929; Whorf, 1956), at its core, either suggesting strongly that the language determines the cognition, or weakly suggesting that the language shapes the cognition. The influence of language on aesthetic experience can be visible, since even speaking while engaging with artworks can change fixation and gaze movements (Klein et al., 2014). In a study using artworks with various combinations of accompanied descriptions (e.g. either matching or non-matching statements about the artwork), results showed that the ambiguity as indexed by the ratio between matching and non-matching statements shapes aesthetic appreciation (Jakesch & Leder, 2009). Even the title-related manipulations can affect eye movement parameters such as saccade amplitude and distribution of dwell time on AOIs on a painting (Kapoula et al., 2009) implying strong top-down effects of language during artwork viewing.

Similarly, the question of whether the type of titling an artwork (i.e. semantically matching, non-matching, and untitled) affect observer liking and interest was investigated, with a complementary measure using a facial electromyographic recording (Gerger & Leder, 2015): this study indicates that both matching-titled and untitled artworks were liked more, yet interest was not affected, and on the theoretical level, it was suggested that high levels of disfluency and cognitive effort might reduce liking. Other similar empirical findings linked these contextual effects to a "psycho-historical framework", which emphasises historical significance and amount of prior knowledge on the appreciation of arts (Swami, 2013). It is also shown that prior knowledge or top-down expectations can change representations in the visual cortex as measured by fMRI (Kok et al., 2013). Additionally, the inclusion of semantic attributions from the observer (such as simple definitions about the artwork or elaborate descriptions about the aesthetic experience) has also the potential to provide additional insights to the research, and in the pragmatic sense, for example, algorithms for aesthetic analysis in computer vision tends to work better while using similar multimodal inputs (Zhou et al., 2016). As a related concept, understanding and judging the artist's intention can be seen as prior knowledge shaping aesthetics (which cannot be simply explained by a direct attentional bias). In a study, an increase of aesthetic appreciation was observed as a mode of observer's explicit knowledge of intention (de Silva et al., 2014). However, a binary distinction (i.e., whether an observer knows the artist's intention or not) may not be nuanced enough, and further research may ask, for example, whether an observer agrees with the artist's intention or not.

The search for robust frameworks or forms of universal rules remains to be a work in progress in empirical research despite conceptual controversies. For example, a semantic differential method was implemented, and multi-coloured visual stimuli accompanied by adjective pairs were used to investigate universal properties for aesthetics (Fang et al., 2015). Similarly, preference for curvature could be defined as an aesthetic primitive (Gómez-Puerto et al., 2016), referencing previous works assuming stimuli properties that are intrinsically interesting (Latto, 1995). One attempt to make progress about the arguably bolder claims on universality might be to use big data, for example, preference on webpages as indexed by millions of ratings across regions, age groups, and education levels can be analysed and modelled (Reinecke & Gajos, 2014). Along the same lines, the generation of datasets consisting of artworks accompanied by subjective observer ratings is a recent trend, which might contribute to future studies. For example, JenAesthetics Subjective Dataset (Amirshahi et al., 2015) comprises more than a thousand images, with aesthetic ratings from observers, and it is aiming to primarily assist the computational aesthetic researchers: their data on collected ratings were analysed and the results showed various correlations between properties of artworks (e.g. between colour, composition, content and aesthetic scores), but not necessarily argued for some aesthetic universals.

To summarise, everyday visual perception and perception of art might be two distinct concepts of research (Mamassian, 2008), but basically, a common aim for both might be solving ambiguities such as composition, illumination, or colour; via expected priors or conventions. Generally, vision seems to be shaped by the expectation of the observer as top-down mechanisms, and these stochastic processes may be also visible particularly in the early processing stages (Kok et al., 2013). Using comprehensive experimental designs in terms of visual contents seems to be a potential way of investigating the role of vision in aesthetic experience and judgment. Overall, behavioural studies showed that (i) observers can assign and update meanings, (ii) provide various types of judgements about artworks sometimes intrinsically and sometimes when they are asked to do so, and (iii) show a set of emotional responses to visual stimuli with minimum effort. Even this simplified description of the observer opens the possibility of countless novel studies, by controlling or manipulating many potential variables during the aesthetic experience, whilst recording a handful of behavioural data, and recently in the novel, simulated environments. Moreover, the increasing diversity of the artworld resulting in a variety of novel forms of art requires yet more elaborate studies (Leder et al., 2004), which should be also translated into studies beyond the laboratories into the real world settings and simulated environments, for example, using virtual and augmented reality. Recent developments of technology give artists the possibility to use novel media, ending up, for example, with the formation of digital works as an art form. The discursive debates of contemporary art, either in a physical or a digital or in any other form, produce claims which may need further elaborations. In this sense, experimental aesthetics has the potential to evaluate these theoretical claims empirically (Locher, 2011), and to resolve discrepancies to some extent.

1.4. Concluding remarks

Empirical aesthetics is a conceptually challenging area of study, with debates over various theoretical and experimental issues. This subsection explicitly recites some of those major issues, and the conceptual standpoints of this thesis. It is also a reminder of some conceptual limitations, with individual paragraphs aiming to outline these issues can be read just as personal remarks.

The research areas dealing with arts have distinct approaches in general, and arguably resulting in a translation problem between these areas. Even some frequently used terms and concepts including experience or appreciation may denote completely distinct things depending on the context, and this translation problem may be prone to misinterpretations of findings from one research area to the other researchers outside. To illustrate a particular example, one general aim of neuroimaging in this context is spatially and temporally localise and quantify the neural correlates of some mental states, and the results of neuroimaging studies (e.g. associating the activation changes in the default mode network (DMN) with the aesthetic appeal or the BOLD signal changes in the medial orbitofrontal cortex (mOFC) with the beauty) does not entail a fast and ultimate answer of the aesthetics; or similarly, approaches using machine-learning are not necessarily searching for an ultimate answer. Additionally, various language-related problems such as vagueness, logical fallacies, circular definitions, and incomprehensiveness are visible either explicitly in some existing research, or more implicitly in the general discourse. Therefore, the presented research, whilst accepting main foundational and sometimes very peculiar opinions of the discipline of empirical aesthetics, is aimed to provide brief definitions or descriptions whenever suitable, to minimise such a translation problem and to avoid other obvious language-related problems.

A closely related aspect is a type of disconnection between theories about (visual) arts and aesthetic experience, which can be described in two main ways: firstly, a form of frame of reference between branches of aesthetic theories, or even between individual researchers causes major discrepancies in such a way that two contradictory arguments can, often counter-intuitively, co-exist or have equal validity or soundness. As a toy example, two major arguments of "there are aesthetic universals" and "there are no aesthetic universals" are present in the current experimental literature. In more extreme cases, a theory or even a single argument from, for example, an art theoretician may not be meaningful at all for a neuroscientist, and vice versa. Secondly, most of the time an art theory is not compatible with a positivist scientific theory, sometimes simply due to the different denotations on what a theory is or means. As an example, phrenology as a pseudoscience was once a scientific theory based on some strict ideas about the mind later to be falsified and superseded; whereas neoplasticism was and is an art theory oftentimes based on strict ideas, that might be arguably superseded, but it is not something falsifiable in the sense of a falsifiable scientific theory. Arguably, similar disconnections can be further extended between theoretical studies and empirical research. This research strongly leans towards positivist empirical sciences, and it neither puts forward a single comprehensive theory strongly nor proposes a better one alongside the countless others, although this may sound like a simple escape from theories. Although the experiments presented have an inductive research approach similar to most of the existing body of research, the main conceptual standpoint here is the belief that a general or unified theory is not feasible, or at least beyond the rationale of current research, assuming that these types of theoretical disconnections are lacking adequate resolution in foreseeable time. Here, a weak mysterianism is favoured, roughly assuming that the relation between conscious aesthetic experience and the external world cannot be explained with the existing state of the knowledge and mental faculties (and further debates are well beyond the scope of this thesis).

A common pitfall is the fallacies of oversimplification and exaggeration. On the one hand, especially in empirical research, one recurrent issue is the tendency to oversimplify the components of the aesthetic experience and frame it as something fully deductible into simple arguments, or basic arrow-and-box models. For example, defining the person of interest as a mere observer or a passive subject of the experience; and the object or the event of interest as the most generic example or a representation of art, despite all the potential variations. Arguably, experimental researchers (maybe unwittingly) conceptually create an ideal (and often neurotypical) observer and ideal experimental condition. Although the ideal observer (Geisler, 2011) is a very useful concept particularly in vision research, for example, to define the baseline, expected condition in a highly controlled low-level vision experiment, it is at least questionable to extrapolate the same conceptual framework into the experiments dealing with the aesthetic experience. On the other hand, especially theoretical works might show an opposite trend as exaggeration or overstatement: rejecting the possibility of empirical examination of arts and aesthetic experience altogether and allowing only insights or interpretation about the aesthetic experience. Across research disciplines, this tendency can be sometimes associated with post-positivist and anti-positivist standpoints or spectrums (Derudder & van Meeteren, 2019), which are (in the very rough sense and generally in social sciences) either weakly bringing criticism to the existing data-driven methods of positivism, or completely arguing against the search for objectivity, and leaning more towards interpretations of subjectivity. This thesis argues against both extreme ends of oversimplifications such as "aesthetics is fully deductible", and overstatements such as "aesthetics cannot be examined empirically at all".

Aesthetics research is often dealing with a partially overlapping dichotomy of phenomenal and physical contents (of experience): the phenomenal content is generally corresponding to the qualitative aspects and means that 'what it is like for the person to have the experience'; whereas the physical content is roughly any aspect that can be, at least in theory, fully explained in terms of relations between physical properties. It is assumed that this two-sidedness is almost always present for any kind of aesthetic experience or judgment. In general, the focus of empirical research is the physical content whereas theoretical research puts qualitative aspects in the centre. This thesis only investigates parts of the physical contents of experience, which is measurable to some degree in the physiological or behavioural levels and does not include inaccessible phenomenal experience. Linking between these two types of contents seems to be much more suitable for philosophical investigations, instead of empirical ones.

This thesis only assumes a simple (ontological) framework about the art object, observer, experience, and medium: in one of the most reduced and simplified descriptions, here, this thesis empirically investigates observers' experience of visual arts in a given environment or medium. If we accept this claim, then we can postulate some common-sense descriptions and properties or qualities for these four main constituents: (i) Art (object) is anything that can be denoted or intended as art. It may have objective and often measurable properties such as the contrast or luminance levels, and subjective properties such as having a high value of perceived complexity or brightness, or as being a sculpture. (ii) The observer is the owner or performer of the action involving art (object), and they may have properties such as being a novice on art knowledge or being bored with viewing a painting. (iii) Experience is any form of action. It may have physical, measurable, accessible properties such as gaze location and duration, and phenomenal properties such as 'qualia'. (iv) Medium is the context or mode of experience. It can be real, simulated, or imaginary, and may have objective and subjective properties, similar to an art object. Particular to this research, these four aspects can be further narrowed down: (i) The selected visual art objects included a physical installation artwork, a physical and virtual realisation of the same design idea of an artwork, and a set of 3D digital models either presented in VR or on screen. (ii) Observers were study participants, often asked to perform both an audience or a spectator role (as in free viewing) and a critic role (as in judgment tasks). (iii) Parts of their immediate experience (using types of behavioural data tracking) and their consecutive and elaborate judgments (using types of response tracking) were targeted, recorded, and analysed. (iv) The selected mediums were either physical (i.e., often regarded as genuine or classical settings), virtual, or online contexts.

Lastly, the timeline of the studies in this thesis might be underlined, even though individual experimental chapters progressed in parallel during the PhD research. The studies were not planned as part of a strict consecutive timeline, mainly due to the time flexibility required to learn and apply multiple research skill sets, and partially due to the uncertainty about the collaboration with institutions. In the later stages of the PhD research, the pandemic led to closures of departmental lab spaces as well as galleries and museums for a long time, which partially affected the structure of the thesis. Briefly, (i) the experiment for the study of the Mondrian Room in Dresden, 3rd thesis chapter, was conducted in April 2019. A follow-up, lab-based VR experiment to investigate environmental effects, such as digital avatars, on gaze patterns was planned and the 3D environment was drafted, but the follow-up experiment was cancelled due to the pandemic. (ii) In the study of Virtual Reality Gallery, 4th thesis chapter, the lab-based VR experiment was conducted in May-June 2019, and the online experiment was conducted in June 2019. A more generalised follow-up experiment consisting of 3D fractal artworks and environments was planned and drafted, but eventually cancelled. (iii) The in-situ experiment of the study of Compton Verney Mural, 5th thesis chapter, was conducted in September 2017, the online experiment was conducted in July 2020, and the online eye-tracking experiment was conducted in December 2020. The online experiments were based on a contingency plan, since the prepared and ready-to-run experiment that included both 360° renders and 3D walkable environments was cancelled. (iv) One main additional study candidate was in development as a fourth empirical chapter: the thematic exhibition at the Royal Holloway Picture Gallery, where the pilot eye-tracking data were already collected and the VR draft environment for eye-tracking was created. This additional experimental chapter was also cancelled.

1.5. References

Ahlberg, L.-O. (2014). The invention of modern aesthetics: From Leibniz to Kant. In *Notions of the Aesthetic and of Aesthetics*. Peter Lang. https://doi.org/10.3726/978-3-653-04597-0

Amirshahi, S. A., Hayn-Leichsenring, G. U., Denzler, J., & Redies, C. (2015). JenAesthetics subjective dataset: Analyzing paintings by subjective scores. In L. Agapito, M. M. Bronstein, & C. Rother (Eds.), *Computer Vision—ECCV 2014 Workshops* (Vol. 8925, pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-319-16178-5_1

Anstis, S., Vergeer, M., & Van Lier, R. (2012). Looking at two paintings at once: Luminance edges can gate colors. *I-Perception*, *3*(8), 515–519. https://doi.org/10.1068/i0537sas

Aristotle. (350 B.C.E. | 2014). *Complete works of Aristotle: The revised Oxford translation*. (J. Barnes, ed.) Princeton University Press. (Original work published 350 B.C.E.)

Arnheim, R. (1954). *Art and visual perception: A psychology of the creative eye*. University of California Press.

Augustin, M. D., & Leder, H. (2006). Art expertise: A study of concepts and conceptual spaces. *Psychological Science*, *48*(2), 135–156.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A., & Hitch, G. (1974). *Working memory* (G. Bower, Ed.; Vol. 8, pp. 47–89). Academic Press.

Bassett, D. S., Cullen, K. E., Eickhoff, S. B., Farah, M. J., Goda, Y., Haggard, P., Hu, H., Hurd, Y. L., Josselyn, S. A., Khakh, B. S., Knoblich, J. A., Poirazi, P., Poldrack, R. A., Prinz, M., Roelfsema, P. R., Spires-Jones, T. L., Sur, M., & Ueda, H. R. (2020). Reflections on the past two decades of neuroscience. *Nature Reviews Neuroscience*, *21*(10), 524–534. https://doi.org/10.1038/s41583-020-0363-6

Baxandall, M. (1988). *Painting and experience in fifteenth century Italy: A primer in the social history of pictorial style* (2nd ed). Oxford University Press.

Beardsley, M. C. (1998). *Aesthetics from classical Greece to the present: A short history* (11. reprint). Univ. of Alabama Press.

Beardsley, M. C., & Arnheim, R. (1981). An exchange of views on Gestalt psychology and aesthetic explanation. *Leonardo*, *14*(3), 220. https://doi.org/10.2307/1574279

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. Appleton-Century-Crofts.

Bertamini, M., & Blakemore, C. (2019). Seeing a work of art indirectly: When a reproduction is better than an indirect view, and a mirror better than a live monitor. *Frontiers in Psychology*, *10*, 2033. https://doi.org/10.3389/fpsyg.2019.02033

Bertamini, M., Makin, A., & Rampone, G. (2013). Implicit association of symmetry with positive valence, high arousal and simplicity. *I-Perception*, *4*(5), 317–327. https://doi.org/10.1068/i0601jw

Bex, P. J., Mareschal, I., & Dakin, S. C. (2007). Contrast gain control in natural scenes. *Journal of Vision*, *7*(11), 12. https://doi.org/10.1167/7.11.12

Bezdek, J. (1993). Fuzzy models—What are they, and why? [Editorial]. *IEEE Transactions on Fuzzy Systems*, *1*(1), 1–6. https://doi.org/10.1109/TFUZZ.1993.6027269

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Birkhoff, G. D. (1933). *Aesthetic measure*. Harvard University Press. https://doi.org/10.4159/harvard.9780674734470

Boas, F. (1927). *Primitive art* (Unabridged republ). Dover Publ.

Borji, A., Sihite, D. N., & Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of Einhauser et al.'s data. *Journal of Vision*, *13*(10), 18–18. https://doi.org/10.1167/13.10.18

Brachmann, A., & Redies, C. (2017). Computational and experimental approaches to visual aesthetics. *Frontiers in Computational Neuroscience*, *11*, 102. https://doi.org/10.3389/fncom.2017.00102

Braun, J., Amirshahi, S. A., Denzler, J., & Redies, C. (2013). Statistical image properties of print advertisements, visual artworks and images of architecture. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00808

Brennan, A. (2017). Necessary and sufficient conditions. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/necessary-sufficient/

Brieber, D., Leder, H., & Nadal, M. (2015). The experience of art in museums: An attempt to dissociate the role of physical context and genuineness. *Empirical Studies of the Arts*, *33*(1), 95–105. https://doi.org/10.1177/0276237415570000

Brieber, D., Nadal, M., & Leder, H. (2015). In the white cube: Museum context enhances the valuation and memory of art. *Acta Psychologica*, *154*, 36–42. https://doi.org/10.1016/j.actpsy.2014.11.004

Brieber, D., Nadal, M., Leder, H., & Rosenberg, R. (2014). Art in time and space: Context modulates the relation between art experience and viewing time. *PLOS ONE*, *9*(6), e99019. https://doi.org/10.1371/journal.pone.0099019

Brielmann, A. A., & Pelli, D. G. (2017). Beauty requires thought. *Current Biology*, *27*(10), 1506-1513.e3. https://doi.org/10.1016/j.cub.2017.04.018

Brielmann, A. A., Vale, L., & Pelli, D. G. (2017). Beauty at a glance: The feeling of beauty and the amplitude of pleasure are independent of stimulus duration. *Journal of Vision*, *17*(14:9), 1–12. https://doi.org/10.1167/17.14.9

Bullot, N. J., & Reber, R. (2013). The artful mind meets art history: Toward a psycho-historical framework for the science of art appreciation. *Behavioral and Brain Sciences*, *36*(2), 123–137. https://doi.org/10.1017/S0140525X12000489

Burke, E. (1767). *A philosophical enquiry into the origin of our ideas of the sublime and beautiful*. Dodsley.

Burtan, D., Joyce, K., Burn, J. F., Handy, T. C., Ho, S., & Leonards, U. (2021). The nature effect in motion: Visual exposure to environmental scenes impacts cognitive load and human gait kinematics. *Royal Society Open Science*, *8*(1), 201100. https://doi.org/10.1098/rsos.201100

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology and perception in art*. University of Chicago Press.

Calbi, M., Aldouby, H., Gersht, O., Langiulli, N., Gallese, V., & Umiltà, M. A. (2019). Haptic aesthetics and bodily properties of Ori Gersht's digital art: A behavioral and eye-tracking study. *Frontiers in Psychology*, *10*, 2520. https://doi.org/10.3389/fpsyg.2019.02520

Carbon, C.-C., & Jakesch, M. (2013). A model for haptic aesthetic processing and its implications for design. *Proceedings of the IEEE*, *101*(9), 2123–2133. https://doi.org/10.1109/JPROC.2012.2219831

Carroll, N. (2004). Non-perceptual aesthetic properties: Comments for James Shelley. *The British Journal of Aesthetics*, *44*(4), 413–423. https://doi.org/10.1093/bjaesthetics/44.4.413

Casati, R., & Pignocchi, A. (2007). Mirror and canonical neurons are not constitutive of aesthetic response. *Trends in Cognitive Sciences*, *11*(10), 410. https://doi.org/10.1016/j.tics.2007.07.007

Cela-Conde, C. J., Agnati, L., Huston, J. P., Mora, F., & Nadal, M. (2011). The neural foundations of aesthetic appreciation. *Progress in Neurobiology*, *94*(1), 39–48. https://doi.org/10.1016/j.pneurobio.2011.03.003

Chakrabarti, A. (Ed.). (2016). *The Bloomsbury research handbook of Indian aesthetics and the philosophy of art*. Bloomsbury Academic, an imprint of Bloomsbury Publishing Plc.

Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory* (1. issued as an Oxford University Press paperback). Oxford Univ. Press.

Chatterjee, A. (2014). *The aesthetic brain: How we evolved to desire beauty and enjoy art*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199811809.001.0001

Chatterjee, A., & Vartanian, O. (2014). Neuroaesthetics. *Trends in Cognitive Sciences*, *18*(7), 370–375. https://doi.org/10.1016/j.tics.2014.03.003

Che, J., Sun, X., Gallardo, V., & Nadal, M. (2018). Cross-cultural empirical aesthetics. In *Progress in Brain Research* (Vol. 237, pp. 77–103). Elsevier. https://doi.org/10.1016/bs.pbr.2018.03.002

Chipp, H. B. (1968). *Theories of modern art: A source book by artists and critics*. Univ. of California Press.

Choe, K. W., Kardan, O., Kotabe, H. P., Henderson, J. M., & Berman, M. G. (2017). To search or to like: Mapping fixations to differentiate two forms of incidental scene memory. *Journal of Vision*, *17*(12), 8. https://doi.org/10.1167/17.12.8

Cohen, E. H., & Zaidi, Q. (2013). Symmetry in context: Salience of mirror symmetry in natural patterns. *Journal of Vision*, *13*(6), 22–22. https://doi.org/10.1167/13.6.22

Consoli, G. (2015). Creativity and aesthetic evaluation. Two proposals to improve the model of aesthetic dis/fluency. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01520

Cova, F., & Réhault, S. (Eds.). (2019). *Advances in experimental philosophy of aesthetics*. Bloomsbury Academic.

Crozier, W. R., & Chapman, A. J. (1984). *Cognitive processes in the perception of art.* North Holland ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co. http://site.ebrary.com/id/10276408
Cupchik, G. C. (2007). A critical reflection on Arnheim's Gestalt theory of aesthetics. *Psychology of Aesthetics, Creativity, and the Arts*, *1*(1), 16–24. https://doi.org/10.1037/1931-3896.1.1.16
Cupchik, G. C., & Gebotys, R. (1988). The experience of time, pleasure, and interest during aesthetic episodes. *Empirical Studies of the Arts*, *6*(1), 1–12. https://doi.org/10.2190/5YN3-J3P8-FWHY-UDB3
Datteri, E., & Laudisa, F. (2014). Box-and-arrow explanations need not be more abstract than neuroscientific mechanism descriptions. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00464
de Silva, O., Crilly, N., & Hekkert, P. (2014). Intentions and the aesthetics of artifacts. *Proceedings of IAEA 2014*, 148–152.
De Winter, S., Vissers, N., Bossens, C., & Wagemans, J. (2020). Instantaneous art? Investigating Frank Stella's Moroccan Paintings with a short-exposure experiment. *Art and Perception*, *8*(2), 121–157. https://doi.org/10.1163/22134913-bja10001
Deręgowski, J. B. (1984). *Distortion in art: The eye and the mind*. Routledge & K. Paul.
Deroy, O., & Spence, C. (2014). Can you find the golden ratio in your plate? *Flavour*, *3*(1), 5. https://doi.org/10.1186/2044-7248-3-5
Derudder, B., & van Meeteren, M. (2019). Engaging with "urban science". *Urban Geography*, *40*(4), 555–564. https://doi.org/10.1080/02723638.2019.1585138
Duchowski, A. (2017). *Eye tracking methodology: Theory and practice* (3rd ed.). Springer International Publishing.
Dupont, L., Antrop, M., & Van Eetvelde, V. (2013). Eye-tracking analysis in landscape perception research: Influence of photograph properties and landscape characteristics. *Landscape Research*, *39*(4), 417–432. https://doi.org/10.1080/01426397.2013.773966
Eastlake, E. (1857). Photography. *The London Quarterly Review*, *101*, 442–468.
Ecker, G. (Ed.). (1986). *Feminist aesthetics*. Beacon Press.
Einhauser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 18–18. https://doi.org/10.1167/8.14.18
Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative adversarial networks, generating 'art' by learning about styles and deviating from style norms. *ICCC*.
Elgammal, Ahmed, Mazzone, M., Liu, B., Kim, D., & Elhoseiny, M. (2018). The shape of art history in the eyes of the machine. *ArXiv:1801.07729 [Cs]*. http://arxiv.org/abs/1801.07729
Epstein, Z., Levine, S., Rand, D. G., & Rahwan, I. (2020). Who gets credit for AI-generated art? *IScience*, *23*(9), 101515. https://doi.org/10.1016/j.isci.2020.101515
Erjavec, A. (1999). Aesthetics as philosophy. *Filozofski Vestnik*, *20*(2). https://ojs.zrc-sazu.si/filozofski-vestnik/article/view/4062
Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 174569162097058. https://doi.org/10.1177/1745691620970586
Eshraghian, J. K. (2020). Human ownership of artificial creativity. *Nature Machine Intelligence*, *2*(3), 157–160. https://doi.org/10.1038/s42256-020-0161-x
Eysenck, H. J. (1941). The empirical determination of an aesthetic formula. *Psychological Review*, *48*(1), 83.
Fang, S., Muramatsu, K., & Matsui, T. (2015). Experimental study of aesthetic evaluation to multi-color stimuli using semantic differential method. *Transactions of Japan Society of Kansei Engineering*, *14*(1), 37–47. https://doi.org/10.5057/jjske.14.37
Fechner, G. T. (1832). *Repertorium der Experimentalphysik: Enthaltend eine vollständige Zusammenstellung der neuern Fortschritte dieser Wissenschaft*. Verlag von Leopold Voss.
Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf und Hartel.
Fechner, G. T. (1876). *Vorschule der Aesthetik*. Breitkopf & Härtel Verlag.
Fechner, G. T. (1877). *In Sachen der Psychophysik*. Breitkopf und Hartel.
Fine, A. (1975). How to compare theories: Reference and change. *Noûs*, *9*(1), 17. https://doi.org/10.2307/2214339

Fingerhut, J., & Prinz, J. J. (2018). Wonder, appreciation, and the value of art. In *Progress in Brain Research* (Vol. 237, pp. 107–128). Elsevier. https://doi.org/10.1016/bs.pbr.2018.03.004

Franz, G., von der Heyde, M., & Bülthoff, H. H. (2005). An empirical approach to the experience of architectural space in virtual reality—Exploring relations between features and affective appraisals of rectangular indoor spaces. *Automation in Construction*, *14*(2), 165–172. https://doi.org/10.1016/j.autcon.2004.07.009

Freedberg, D., & Gallese, V. (2007). Motion, emotion and empathy in esthetic experience. *Trends in Cognitive Sciences*, *11*(5), 197–203. https://doi.org/10.1016/j.tics.2007.02.003

Freeland, C. A. (2002). *But is it art? An introduction to art theory* (Pbk. ed). Oxford University Press.

Gartus, A., & Leder, H. (2013). The small step toward asymmetry: Aesthetic judgment of broken symmetries. *I-Perception*, *4*(5), 361–364. https://doi.org/10.1068/i0588sas

Gaut, B. (2005). The cluster account of art defended. *The British Journal of Aesthetics*, *45*(3), 273–288. https://doi.org/10.1093/aesthj/ayi032

Gaut, B., & Lopes, D. (Eds.). (2013). *The Routledge companion to aesthetics* (Third Edition). Routledge, Taylor & Francis Group.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*(7), 771–781. https://doi.org/10.1016/j.visres.2010.09.027

Gerger, G., & Leder, H. (2015). Titles change the esthetic appreciations of paintings. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00464

Gernot, G., Pelowski, M., & Leder, H. (2018). Empathy, Einfühlung, and aesthetic experience: The effect of emotion contagion on appreciation of representational and abstract art using fEMG and SCR. *Cognitive Processing*, *19*(2), 147–165. https://doi.org/10.1007/s10339-017-0800-2

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Gips, J., & Stiny, G. (1975). An investigation of algorithmic aesthetics. *Leonardo*, *8*(3), 213. https://doi.org/10.2307/1573240

Gómez-Puerto, G., Munar, E., & Nadal, M. (2016). Preference for curvature: A historical and conceptual framework. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00712

Gonzalez, V. (2001). *Beauty and Islam: Aesthetics in Islamic art and architecture*. I.B. Tauris in association with the Institute of Ismaili Studies, London ; Distributed in the U.S. by St. Martin's Press.

Graf, L. K. M., & Landwehr, J. R. (2015). A dual-process perspective on fluency-based aesthetics: The pleasure-interest model of aesthetic liking. *Personality and Social Psychology Review*, *19*(4), 395–410. https://doi.org/10.1177/1088868315574978

Greb, F., Elvers, P., & Fischinger, T. (2017). Trends in empirical aesthetics: A review of the journal Empirical Studies of the Arts from 1983 to 2014. *Empirical Studies of the Arts*, *35*(1), 3–26. https://doi.org/10.1177/0276237415625258

Güçlütürk, Y., Jacobs, R. H. A. H., & Lier, R. van. (2016). Liking versus complexity: Decomposing the inverted U-curve. *Frontiers in Human Neuroscience*, *10*, 112. https://doi.org/10.3389/fnhum.2016.00112

Gurd, S. (2012). The origins of aesthetic thought in ancient Greece: Matter, sensation, and experience. *Comparative Literature*, *64*(3), 325–330. https://doi.org/10.1215/00104124-1672979

Guyau, J.-M. (1889). *L'art au point de vue sociologique*. ENS Éditions. https://doi.org/10.4000/books.enseditions.5837

Hagelin, J. S. (1987). *Is consciousness the unified field? A field theorist's perspective*. Maharishi International University.

Hale, J., Hastings, J., West, R., Lefevre, C. E., Direito, A., Bohlen, L. C., Godinho, C., Anderson, N., Zink, S., Groarke, H., & Michie, S. (2020). An ontology-based modelling system (OBMS) for representing behaviour change theories applied to 76 theories. *Wellcome Open Research*, *5*, 177. https://doi.org/10.12688/wellcomeopenres.16121.1

Harel, A., Ullman, S., Harari, D., & Bentin, S. (2011). Basic-level categorization of intermediate complexity fragments reveals top-down effects of expertise in visual perception. *Journal of Vision*, *11*(8), 18–18. https://doi.org/10.1167/11.8.18

Harrison, C., & Wood, P. (Eds.). (2003). *Art in theory, 1900-2000: An anthology of changing ideas* (2nd ed). Blackwell Pub.

Harrison, C., Wood, P., & Gaiger, J. (Eds.). (2001). *Art in theory, 1815-1900: An anthology of changing ideas*. Blackwell Publishers.

Heidenreich, S. M., & Turano, K. A. (2011). Where does one look when viewing artwork in a museum? *Empirical Studies of the Arts*, *29*(1), 51–72. https://doi.org/10.2190/EM.29.1.d

Hekkert, P., Snelders, D., & Wieringen, P. C. W. (2003). 'Most advanced, yet acceptable': Typicality and novelty as joint predictors of aesthetic preference in industrial design. *British Journal of Psychology*, *94*(1), 111–124. https://doi.org/10.1348/000712603762842147

Hekkert, P., & van Wieringen, P. C. W. (1990). Complexity and prototypicality as determinants of the appraisal of cubist paintings. *British Journal of Psychology*, *81*, 483–495.

Helmholtz, H. von. (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Friedrich Vieweg & Sohn Verlag.

Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PLoS ONE*, *8*(5), e64937. https://doi.org/10.1371/journal.pone.0064937

Holmqvist, K., & Andersson, R. (2017). *Eye tracking: A comprehensive guide to methods, paradigms, and measures* (2nd edition). CreateSpace.

Huber, F. (2008). Assessing theories, Bayes style. *Synthese*, *161*(1), 89–118. https://doi.org/10.1007/s11229-006-9141-x

Hume, D. (1757). *Of the standard of taste*.

Huston, J. P., Nadal, M., Mora, F., Agnati, L. F., & Cela Conde, C. J. (Eds.). (2015). *Art, aesthetics, and the brain* (First edition). Oxford University Press.

Iseminger, G. (1973). Aesthetic judgments and non-aesthetic conditions. *Analysis*, *33*(4), 129. https://doi.org/10.2307/3327616

Ishiguro, C., Yokosawa, K., & Okada, T. (2016). Eye movements during art appreciation by students taking a photo creation course. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01074

Jacobsen, T. (2010). Beauty and the brain: Culture, history and individual differences in aesthetic appreciation. *Journal of Anatomy*, *216*(2), 184–191. https://doi.org/10.1111/j.1469-7580.2009.01164.x

Jacobsen, T., & Beudt, S. (2017). Stability and variability in aesthetic experience: A review. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00143

Jakesch, M., & Leder, H. (2009). Finding meaning in art: Preferred levels of ambiguity in art appreciation. *Quarterly Journal of Experimental Psychology*, *62*(11), 2105–2112. https://doi.org/10.1080/17470210903038974

Kamber, R. (2011). Experimental philosophy of art. *The Journal of Aesthetics and Art Criticism*, *69*(2), 197–208. https://doi.org/10.1111/j.1540-6245.2011.01461.x

Kant, I. (1781). *Kritik der reinen Vernunft*. Verlag von Johann Friedrich Hartknoch.

Kant, I. (1790). *Kritik der Urteilskraft*.

Kapoula, Z., Daunys, G., Herbez, O., & Yang, Q. (2009). Effect of title on eye-movement exploration of Cubist paintings by Fernand Léger. *Perception*, *38*(4), 479–491. https://doi.org/10.1068/p6080

Kardan, O., Berman, M. G., Yourganov, G., Schmidt, J., & Henderson, J. M. (2015). Classifying mental states from eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1502–1514. https://doi.org/10.1037/a0039673

Kirk, U., Skov, M., Christensen, M. S., & Nygaard, N. (2009). Brain correlates of aesthetic expertise: A parametric fMRI study. *Brain and Cognition*, *69*(2), 306–315. https://doi.org/10.1016/j.bandc.2008.08.004

Klein, C., Betz, J., Hirschbuehl, M., Fuchs, C., Schmiedtová, B., Engelbrecht, M., Mueller-Paul, J., & Rosenberg, R. (2014). Describing art – An interdisciplinary approach to the effects of speaking on gaze movements during the beholding of paintings. *PLOS ONE*, *9*(12), e102439. https://doi.org/10.1371/journal.pone.0102439

Koffka, K. (1935). *Principles of Gestalt Psychology*. Harcourt, Brace and Company.

Kok, P., Brouwer, G. J., van Gerven, M. A. J., & de Lange, F. P. (2013). Prior expectations bias sensory representations in visual cortex. *Journal of Neuroscience*, *33*(41), 16275–16284. https://doi.org/10.1523/JNEUROSCI.0742-13.2013

Korsmeyer, C. (2004). *Gender and aesthetics: An introduction*. Routledge.

Kosuth, J. (1991). *Art after philosophy and after: Collected writings, 1966-1990*. MIT Press.

Kowler, E. (2011). Eye movements: The past 25years. *Vision Research*, *51*(13), 1457–1483. https://doi.org/10.1016/j.visres.2010.12.014

Lack, J. (Ed.). (2017). *Why are we 'artists'? 100 world art manifestos*. Penguin Books.

Latto, R. (1995). The brain of the beholder. In R. L. Gregory, J. Harris, P. Heard, & D. Rose (Eds.), *The artful eye* (pp. 66–94). Oxford University Press.

Layton, R. (1991). *The anthropology of art* (2nd ed). Cambridge University Press.

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, *95*(4), 489–508. https://doi.org/10.1348/0007126042369811

Leder, H., & Nadal, M. (2014). Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode - Developments and challenges in empirical aesthetics. *British Journal of Psychology*, *105*(4), 443–464. https://doi.org/10.1111/bjop.12084

Lee, S., Cinn, E., Yan, J., & Jung, J. (2015). Using an eye tracker to study three-dimensional environmental aesthetics: The impact of architectural elements and educational training on viewers' visual attention. *Journal of Architectural and Planning Research*, *32*(2), 145–167.

Leonards, U., Baddeley, R., Gilchrist, I. D., Troscianko, T., Ledda, P., & Williamson, B. (2007). Mediaeval artists: Masters in directing the observers' gaze. *Current Biology*, *17*(1), R8–R9. https://doi.org/10.1016/j.cub.2006.11.046

Li, Z. (1994). *The path of beauty: A study of Chinese aesthetics*. Oxford University Press.

Liao, S., & Meskin, A. (2015). Aesthetic adjectives: Experimental semantics and context-sensitivity. *Philosophy and Phenomenological Research*, *94*(2), 371–398. https://doi.org/10.1111/phpr.12217

Lindell, A. K., & Mueller, J. (2011). Can science account for taste? Psychological insights into art appreciation. *Journal of Cognitive Psychology*, *23*(4), 453–475. https://doi.org/10.1080/20445911.2011.539556

Liversedge, S. P., Gilchrist, I. D., & Everling, S. (Eds.). (2011). *The Oxford handbook of eye movements*. Oxford University Press.

Locher, P. (2011). Contemporary experimental aesthetics: State of the art technology. *I-Perception*, *2*(7), 697–707. https://doi.org/10.1068/i0449aap

Locher, P., Overbeeke, K., & Wensveen, S. (2010). Aesthetic interaction: A framework. *Design Issues*, *26*(2), 70–79. https://doi.org/10.1162/DESI_a_00017

Makin, A. (2017). The gap between aesthetic science and aesthetic experience. *Journal of Consciousness Studies*, *24*(1–2), 184–213.

Makin, A., Bertamini, M., Jones, A., Holmes, T., & Zanker, J. M. (2016). A gaze-driven evolutionary algorithm to study aesthetic evaluation of visual symmetry. *I-Perception*, *7*(2), 204166951663743. https://doi.org/10.1177/2041669516637432

Mallon, B., Redies, C., & Hayn-Leichsenring, G. U. (2014). Beauty in abstract paintings: Perceptual contrast and statistical properties. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00161

Mamassian, P. (2008). Ambiguities and conventions in the perception of visual art. *Vision Research*, *48*(20), 2143–2153. https://doi.org/10.1016/j.visres.2008.06.010

Marin, M. M. (2015). Crossing boundaries: Toward a general model of neuroaesthetics. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00443

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Massaro, D., Savazzi, F., Di Dio, C., Freedberg, D., Gallese, V., Gilli, G., & Marchetti, A. (2012). When art moves the eyes: A behavioral and eye-tracking study. *PLOS ONE*, *7*(5), e37285. https://doi.org/10.1371/journal.pone.0037285

Mastandrea, S., Bartoli, G., & Carrus, G. (2011). The automatic aesthetic evaluation of different art and architectural styles. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(2), 126–134. https://doi.org/10.1037/a0021126

McManus, I. C. (1980). The aesthetics of simple figures. *British Journal of Psychology*, *71*(4), 505–524. https://doi.org/10.1111/j.2044-8295.1980.tb01763.x

McManus, I. C. (2005). Symmetry and asymmetry in aesthetics and the arts. *European Review*, *13*(S2), 157–180. https://doi.org/10.1017/S1062798705000736

McManus, I. C., Stöver, K., & Kim, D. (2011). Arnheim's Gestalt theory of visual balance: Examining the compositional structure of art photographs and abstract images. *I-Perception*, *2*(6), 615–647. https://doi.org/10.1068/i0445aap

McWhinnie, H. J. (1965). A review of some research on aesthetic measure and perceptual choice. *Studies in Art Education*, *6*(2), 34–41. https://doi.org/10.1080/00393541.1965.11650608

Mondrian, P. (1920). Le néo-plasticisme: Principe général de l'equivalence plastique. *Éditions de l'effort moderne*. https://doi.org/10.11588/DIGLIT.25420

Monsere, A. (2015). Experimental philosophy and intuitions on what is art and what is not. *Teorema*, *34*(3), 159–175.

Montero, B. (2006). Proprioception as an aesthetic sense. *The Journal of Aesthetics and Art Criticism*, *64*(2), 231–242. https://doi.org/10.1111/j.0021-8529.2006.00244.x

Muth, C., Pepperell, R., & Carbon, C.-C. (2013). Give me Gestalt! Preference for Cubist artworks revealing high detectability of objects. *Leonardo*, *46*(5), 488–489. https://doi.org/10.1162/LEON_a_00649

Nadal, M., Munar, E., Marty, G., & Cela-Conde, C. J. (2010). Visual complexity and beauty appreciation: Explaining the divergence of results. *Empirical Studies of the Arts*, *28*(2), 173–191. https://doi.org/10.2190/EM.28.2.d

Nadal, M., & Pearce, M. T. (2011). The Copenhagen neuroaesthetics conference: Prospects and pitfalls for an emerging field. *Brain and Cognition*, *76*(1), 172–183. https://doi.org/10.1016/j.bandc.2011.01.009

Neill, A., & Ridley, A. (1995). *The philosophy of art: Readings ancient and modern*. McGraw-Hill.

Orgs, G., Hagura, N., & Haggard, P. (2013). Learning to like it: Aesthetic perception of bodies, movements and choreographic structure. *Consciousness and Cognition*, *22*(2), 603–612. https://doi.org/10.1016/j.concog.2013.03.010

Ortlieb, S. A., Kügel, W. A., & Carbon, C.-C. (2020). Fechner (1866): The Aesthetic Association Principle—A commented translation. *I-Perception*, *11*(3), 204166952092030. https://doi.org/10.1177/2041669520920309

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual Review of Psychology*, *64*(1), 77–107. https://doi.org/10.1146/annurev-psych-120710-100504

Pearce, M. T., Zaidel, D. W., Vartanian, O., Skov, M., Leder, H., Chatterjee, A., & Nadal, M. (2016). Neuroaesthetics: The cognitive neuroscience of aesthetic experience. *Perspectives on Psychological Science*, *11*(2), 265–279. https://doi.org/10.1177/1745691615621274

Pelowski, M., Markey, P. S., Forster, M., Gerger, G., & Leder, H. (2017). Move me, astonish me… delight my eyes and brain: The Vienna Integrated Model of top-down and bottom-up processes in Art Perception (VIMAP) and corresponding affective, evaluative, and neurophysiological correlates. *Physics of Life Reviews*, *21*, 80–125. https://doi.org/10.1016/j.plrev.2017.02.003

Pihko, E., Virtanen, A., Saarinen, V.-M., Pannasch, S., Hirvenkari, L., Tossavainen, T., Haapala, A., & Hari, R. (2011). Experiencing art: The influence of expertise and painting abstraction level. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00094

Pockett, S. (2012). The electromagnetic field theory of consciousness: A testable hypothesis about the characteristics of conscious as opposed to non-conscious fields. *Journal of Consciousness Studies*, *19*(11–12), 191–223.

Porter, J. I. (2009). Is art modern? Kristeller's 'Modern system of the arts' reconsidered. *The British Journal of Aesthetics*, *49*(1), 1–24. https://doi.org/10.1093/aesthj/ayn054

Powell, W. R., & Schirillo, J. A. (2011). Hemispheric laterality measured in Rembrandt's portraits using pupil diameter and aesthetic verbal judgements. *Cognition & Emotion*, *25*(5), 868–885. https://doi.org/10.1080/02699931.2010.515709

Ramachandran, V. S., & Hirstein, W. (1999). The science of art: A neurological theory of aesthetic experience. *Journal of Consciousness Studies*, *6*(6–7), 15–51.

Redies, C. (2015). Combining universal beauty and cultural context in a unifying model of visual aesthetic experience. *Frontiers in Human Neuroscience*, *09*. https://doi.org/10.3389/fnhum.2015.00218

Reinecke, K., & Gajos, K. Z. (2014). Quantifying visual preferences around the world. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 11–20. https://doi.org/10.1145/2556288.2557052

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in Fuzzy-trace theory. *Judgment and Decision Making*, *7*(3), 332–359.

Rosenberg, R., & Klein, C. (2015). The moving eye of the beholder: Eye tracking and the perception of paintings. In J. P. Huston, M. Nadal, F. Mora, L. F. Agnati, & C. J. C. Conde (Eds.), *Art, Aesthetics, and the Brain* (pp. 79–108). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199670000.003.0005

Sapir, E. (1929). The status of linguistics as a science. *Language*, *5*(4), 207. https://doi.org/10.2307/409588

Sartori, A., Culibrk, D., Yan, Y., Job, R., & Sebe, N. (2016). Computational modeling of affective qualities of abstract paintings. *IEEE MultiMedia*, *23*(3), 44–54. https://doi.org/10.1109/MMUL.2016.20

Scarfe, P., & Glennerster, A. (2015). Using high-fidelity virtual reality to study perception in freely moving observers. *Journal of Vision*, *15*(9:3), 1–11. https://doi.org/10.1167/15.9.3

Schepman, A., Rodway, P., Pullen, S. J., & Kirkham, J. (2015). Shared liking and association valence for representational art but not abstract art. *Journal of Vision*, *15*(5), 11. https://doi.org/10.1167/15.5.11

Shelley, J. (2003). The problem of non-perceptual art. *The British Journal of Aesthetics*, *43*(4), 363–378. https://doi.org/10.1093/bjaesthetics/43.4.363

Shelley, James. (2017). The concept of the aesthetic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/aesthetic-concept/

Shimamura, A. P., & Palmer, S. E. (Eds.). (2011). *Aesthetic science: Connecting minds, brains, and experience*. Oxford University Press.

Sibley, F. (1959). Aesthetic concepts. *The Philosophical Review*, *68*(4), 421. https://doi.org/10.2307/2182490

Silvia, P. J. (2005). Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology*, *9*(4), 342–357. https://doi.org/10.1037/1089-2680.9.4.342

Skov, M., & Nadal, M. (2018). Art is not special: An assault on the last lines of defense against the naturalization of the human mind. *Reviews in the Neurosciences*, *29*(6), 699–702. https://doi.org/10.1515/revneuro-2017-0085

Smith, J. K., & Smith, L. F. (2001). Spending time on art. *Empirical Studies of the Arts*, *19*(2), 229–236. https://doi.org/10.2190/5MQM-59JH-X21R-JN5J

Spehar, B., & Stevanov, J. (2021). Expressive qualities of synthetic textures. *Psychology of Consciousness: Theory, Research, and Practice*. https://doi.org/10.1037/cns0000241

Spehar, B., & van Tonder, G. J. (2017). Koffka's Aesthetic Gestalt. *Leonardo*, *50*(1), 53–57. https://doi.org/10.1162/LEON_a_01020

Swami, V. (2013). Context matters: Investigating the impact of contextual information on aesthetic appreciation of paintings by Max Ernst and Pablo Picasso. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(3), 285–295. https://doi.org/10.1037/a0030965

Terras, M. (1999). *A virtual tomb for Kelvingrove: Virtual reality, archaeology and education*. *7*. https://doi.org/10.11141/ia.7.5

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Torregrossa, C. (2020). A defence of experimental philosophy in aesthetics. *Inquiry*, *63*(8), 885–907. https://doi.org/10.1080/0020174X.2017.1385527

van Doesburg, T. (1923). Von der neuen Ästhetik zur materiellen Verwirklichung. *De Stijl*, *6*(1), 10–14.

van Gerwen, R. (2015). Aesthetics as first philosophy. *Aesthetic Investigations*, *1*(1).

Vannucci, M., Gori, S., & Kojima, H. (2014). The spatial frequencies influence the aesthetic judgment of buildings transculturally. *Cognitive Neuroscience*, *5*(3–4), 143–149. https://doi.org/10.1080/17588928.2014.976188

Vartanian, O., & Skov, M. (2014). Neural correlates of viewing paintings: Evidence from a quantitative meta-analysis of functional magnetic resonance imaging data. *Brain and Cognition*, *87*, 52–56. https://doi.org/10.1016/j.bandc.2014.03.004

Vessel, E. A., & Rubin, N. (2010). Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of Vision*, *10*(2:18), 1–14. https://doi.org/10.1167/10.2.18

Weitz, M. (1956). The role of theory in aesthetics. *The Journal of Aesthetics and Art Criticism*, *15*(1), 27. https://doi.org/10.2307/427491

Whorf, B. L. (1956). *Language, thought, and reality: Selected writings*. The MIT Press.

Wilkins, A. J., Penacchio, O., & Leonards, U. (2018). The built environment and its patterns: A view from the vision sciences. *Journal of Sustainable Design and Applied Research in Innovative Engineering of the Built Environment*, *6*(1), 1–9. https://doi.org/10.21427/D7VV5G

Wittgenstein, L. (1953). *Philosophical investigations*. Basil Blackwell.

Wright, R. H. (1995). *Virtual reality psychophysics: Forward and lateral distance, height, and speed perceptions with a wide- angle helmet display*.

Yarbus, A. L. (1967). *Eye movements and vision*. Springer. https://doi.org/10.1007/978-1-4899-5379-7

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zanker, J., Stevanov, J., & Holmes, T. (2017). Mobile eye tracking in the Royal Academy: Analysing the interaction with abstract paintings. *40th European Conference on Visual Perception (ECVP 2017)*, 212. http://journals.sagepub.com/page/pec/collections/ecvp-abstracts/index/ecvp-2017

Zanker, J., van der Vliet, S., Stevanov, J., & Holmes, T. (2016). Mobile Eye tracking in the Gallery: Square murals by Sol LeWitt. *39th European Conference on Visual Perception (ECVP) 2016 Barcelona*, *45*, 1–383. https://doi.org/10.1177/0301006616671273

Zeki, S. (1999). *Inner vision: An exploration of art and the brain*. Oxford University Press.

Zhou, Y., Lu, X., Zhang, J., & Wang, J. Z. (2016). Joint image and text representation for aesthetics analysis. *Proceedings of the 24th ACM International Conference on Multimedia*, 262–266. https://doi.org/10.1145/2964284.2967223

Ziff, P. (1953). The task of defining a work of art. *The Philosophical Review*, *62*(1), 58. https://doi.org/10.2307/2182722

Chapter 2

Methods

In three parts of this section, I present the general rationale, research framework, and aims; followed by the overview of the general method (i.e., participants, stimulus and material, design, procedures, and data analysis); and additional experimental principles on validity, reliability, reproducibility, and scalability.

2.1. Rationale, research framework, and aims

To provide a methodological framework from which the studies are designed, the overarching thesis rationale, research framework, and aims can be summarised.

2.1.1. Rationale

The diversified ways of artistic expression in new media and rapid technological developments create an expanding intersection between arts and sciences. Particular forms of art production, often described under the umbrella term of digital art (Paul, 2016) (with related concepts such as electronic-, computer-, algorithm-, generative-, virtual-art) are emerging and evolving (but also see post-digital art (Berry & Dieter, 2015) for a further discussion). With the increased accessibility of tools on extended reality, behavioural data, and machine learning; these tools are integrated into the creative industry as novel experiences, into contemporary art practice as art pieces, and into business as publicity, marketing and advertising material. However, in behavioural research, these concepts and tools call for either updating existing research methods or forming novel ones, particularly in empirical aesthetics, and more broadly in vision science (Caixinha & Nunes, 2017; Serre, 2019; Çöltekin et al., 2020). Currently, some researchers are exploring these alternative ways of conducting experiments by going beyond traditional practice in lab settings, with particular promises on the clinical and applied sciences (Maggio et al., 2019; Zhao et al., 2019; Siu et al., 2020). However, a thorough and actively updated search of the relevant literature yielded that many basic research questions in empirical aesthetics, related to the behavioural aspects in real- and simulated environments, are still in need of investigation and they still require an accumulation of research outputs.

In the broadest sense, the main reasoning behind this body of research is to adapt some of these emerging tools into study designs to answer basic research questions on aesthetic experience, judgment, and interaction with arts, in the arguably ecological valid contexts, either in-situ settings or close proxies of those settings. The key problem to be tested is methodological, aiming to answer the question of "how the empirical aesthetics research can benefit from the simulated experiences, novel visualisation tools and behavioural measures". This thesis can be seen as a single attempt to try out some possible directions, to assess these emerging concepts' methodological validity, research potential, and the extent of their usefulness. Following on from that, the primary aim is to explore the feasibility of implementing those emerging research tools into previously unfeasible research scenarios, by taking experimental work into environments where individuals can interact with arts. Apart from the empirical contribution, speculated aspects on practical use-cases, both in individual and general discussion chapters, aimed to offer a conceptual ground to explore pathways into future directions.

2.1.2. Research framework

The current research exploring visual aesthetic experience and judgments was conducted in three linked experimental settings, which were not in a hierarchical order, or mutually exclusive, but complementary to each other. (i) In real-world settings, using 3D artworks as case studies, quantitative measures based on eye movements were developed, accompanied by supplementary questionnaires. (ii) In virtual reality settings, two types of approaches were present: (ii-a) In an art gallery, visual exploration patterns and findings from a real-world installation were compared with a VR reconstruction, and advantages and disadvantages of using VR were assessed. (ii-b) In the lab, perceptual measures of judgment on 3D visual art objects and spaces were developed. (iii) In online settings, participants' judgments on manipulated 3D artworks, after generating variations of artworks as a part of further experimental design, were explored, responses from a wider and potentially novel audience were received, and the similarity of findings between lab-VR and online-screen settings was checked.

2.1.3. Research aims

To provide the scope of the empirical chapters, aims and research questions of the three individual chapters can be introduced:

The first empirical chapter (i.e., Mondrian Room in Dresden as the 3rd thesis chapter) took place in an art gallery, utilising simultaneous use of mobile eye-tracking for an installation artwork and VR eye-tracking for the digital reconstruction, both derived from the same design draft from Piet Mondrian. The main aims were to quantitatively compare observers' exploration of artwork (as part of the aesthetic experience) within an art installation in physical and virtual instantiations, and to assess the similarity between physical and VR contexts.

The second empirical chapter (i.e., Virtual Reality Gallery as the 4th thesis chapter) took place in the lab-based VR and online settings, utilising VR eye-tracking for a set of 3D artworks and judgment tasks. The main aim was to incorporate various commonly used judgment types for visual arts from previous studies including liking and novelty, and to include relatively unusual types of judgments such as liking from a third-person point-of-view and perceived one's own viewing duration.

The third empirical chapter (i.e., Compton Verney Mural as the 5th thesis chapter) was carried out firstly in an art gallery and followed by in online settings; by utilising mobile eye-tracking for a room-scale abstract installation in the gallery, and judgment tasks on variations of the same installation online. The main aims were to investigate (i) viewing trends such as the distribution of dwell time as cumulative fixation duration and fixation count on compositional features of the installation in a gallery, (ii) participants' aesthetic judgments on variations of the original artwork, their reasonings behind those preferences, participants' insights about the digital art viewing, and (iii) the feasibility of implementing current online methods and measures of eye-tracking.

2.2. Overview of general methods

Since the empirical chapters have similar method subsections, an overview of general methods is provided here, aiming to explain the basic approaches, methodological considerations, and an overall justification behind the choices of methods. The aspects are aligned with the standard quantitative methods in human behavioural research: in this sense, the guidelines are based on general reference resources on experimental design and statistics (Field & Hole, 2003; Bourne, 2017; Argyrous, 2011) and specialized references on, for example, eye tracking (Holmqvist & Andersson, 2017; Duchowski, 2017), online experiments (Sauter et al., 2020), VR using Unity (Linowes, 2018), data visualisation (Wilke, 2019), or general scientific writing (Katz, 2009). Note that specific details are presented in the methods sections of the individual empirical chapters.

2.2.1. Participants

Participants were recruited through convenience sampling. Depending on the research setting, they were either museum visitors interested in participating, or mostly university students and members of staff interested to be part of a VR experiment, or participants who signed up for the experiment from an online study recruitment platform (Prolific.co). Convenience sampling was chosen because it is time-efficient and requires little logistics for in-situ and VR settings. One major drawback was the potentially biased sampling, which might lead to a low power to identify differences of subgroups, but it was aimed to be countered to some extent by relatively large online follow-up samples. The participants were naive to the specific aims of the experiments. The briefing before the experiment and debriefing afterwards always explicitly mentioned the details on the given research on aesthetic experience and judgment. No intentional deception was introduced into the research. In terms of anonymity, confidentiality, and privacy, the best available practices were sought, in line with the guidelines of the General Data Protection Regulation (GDPR) by the European Union (Voigt & von dem Bussche, 2017). All participants provided either written informed consent for the in-situ or lab-based studies, or informed consent for the online studies, before the experiments. The collected and open access shared data did not contain any identifiable information about the participants. All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed under the ethical guidelines and regulations of the Declaration of Helsinki. Whenever appropriate, either a priori or sensitivity analyses were conducted for the estimations depending on the research context. It should be noted that the standard power or sample size estimations were not always well suited especially for the eye-tracking experiments, as sample size could not always be pre-determined due to, for example, the lack of meta-analysis indicating effect sizes, scarcity of closely related previous empirical works, and the novel nature of the presented studies. In this sense, these aspects might cast doubt on some assumptions for those presented estimation analyses.

2.2.2. Stimulus and material

All studies used only 3D artworks, such as physical installations or a set of artworks presented in VR. As almost all prior research relied on 2D artworks such as paintings (or often, representations of them as digital photographs displayed on monitors), the ability to demonstrate the inclusion of installation art and presenting spatial 3D objects and spaces in VR were the main drives for the stimulus selection criteria. Nevertheless, in online experiments, 2D renders either as videos from the first-person point of view, or as static 2D images to collect more reliable online eye-tracking data were presented, because of the low feasibility of implementing online VR studies at the time of conducting the experiments, and because of the lack of personal VR-HMD adoption. The judgment tasks were designed to answer particular research questions, and the designed questionnaires were treated as supplementary measures and were not a major part of the studies. Note that because of the scarcity of the well-accepted measures, such as art-related questionnaires and surveys, they were mentioned whenever relevant, but were not adapted to be re-used in this research.

Software for stimulus preparation and experimental design mainly include 3D modelling software (SketchUp, Blender, Maya), game engine and scripting language (Unity and C#), 3D rendering software (Lumion), 2D image editors (Adobe Creative Cloud and Affinity Suite), programming language for lab-based task creation (Matlab with Psychtoolbox), eye-tracking software (Tobii Pro Lab and Tobii Pro VR Analytics), online experimental software (Qualtrics, Labvanced, Google Forms). These selected software are often regarded as the standards in their respective software categories, frequently used by the researchers, designers, game developers, and are often well-documented due to their user-base. But note that other existing alternative software could have been suitable for the same procedures, for example, there are tens of alternative 3D render pipelines to create 360-degree images of an environment. The main hardware available during the data collection were HTC Vive HMD with integrated Tobii eye trackers for the VR eye-tracking experiments, Tobii Glasses Pro 2 for the mobile eye-tracking, and VR-ready PCs.

2.2.3. Design

All studies were designed as within-subject experiments, mainly because no specific hypothesis about particular subgroups was set (unlike, for example, expert-novice studies in empirical aesthetics). For the eye-tracking data, often the main dependent variables (DVs) were absolute and area-normalised dwell times as the cumulative fixation durations in seconds (and also absolute and normalised fixation count in supplementary analyses). The main independent variables (IVs) were sets of areas of interests (AOIs) as indexed by corresponding 3D geometry of artworks, and the research setting (as in VR vs. real-world). For the data coming from aesthetic judgments, judgment ratings were either DVs where the IV was, for example, the artwork variation, or judgment ratings were covariates in correlational analyses. Data derived from questionnaire responses were only analysed using descriptive statistics, such as reporting modes of the responses.

2.2.4. Procedures

All in-situ or lab-based experiments shared the same overall flow of a basic procedure: they started with a briefing and instructions phase and received written informed consent from the participant (whereas in the online experiments, participants received informed consent). After that, and following the standard eye-tracking calibration process, participants were engaging with visual arts whilst their eye movements were recorded, and sometimes it was followed by judgment tasks, and then they were asked to respond to an exit-questionnaire. In general, the experiments were aimed to create a research setting, resembling as close as the genuine experience of art. Due to the within-subjects design, to overcome carryover effects, appropriate randomisation was introduced, such as in the order of experimental conditions or the order of the stimuli presentation.

2.2.5. Data analysis

For the data analysis, selected software were MatLab, R (and jamovi/JASP as GUI-based tools based on R-packages), and rarely Mathematica and OriginPro, simply due to the practicality. Similar to the software selection for stimulus preparation, this selection of scripting for analysis was trivial in the sense that the available software was used at any given point, and all the analysis presented could be alternatively conducted using just any of these languages, or any other language of choice such as Python. Throughout the experimental chapters, relevant data plotting followed by proper selection of statistical data analyses was sought to the best of found knowledge. Whenever possible, the raw data were shown either in plots or made available in data repositories. Note that all data were primary data collected specifically for the presented studies, as it was feasible to do so for these small-scale experiments, and no secondary data (e.g., large-scale, collected by others, and readily available data) were used for the analysis. Also, since all empirical chapters were designed as quantitative research, qualitative data analysis approaches such as content analysis, discourse analysis, or interpretative phenomenological analysis were disregarded. In this thesis, the main reasoning behind favouring quantitative analysis over qualitative analysis, or over mixed methods (by combining both qualitative and quantitative research approaches) was mainly to avoid introducing another level of complexity in the empirical chapters: These chapters were primarily investigating (often embodied) aesthetic experiences using emerging tools such as in-situ or VR eye-tracking, whose data output can be analysed quantitatively but arguably, cannot be merged easily with additional qualitative data.

Data collected with eye-trackers, judgment tasks, and questionnaires were treated as quantitative data, and analysed using well-established frequentist methodologies of inference. In brief, these methods include either parametric or non-parametric types of RM-ANOVA, linear mixed-effects model, correlation, regression; and also, basic descriptive statistics such as reporting means and variances of a given continuous variable, mode of Likert-scale data, or word-frequencies to describe the answers provided on open-ended questions in the questionnaire. The assumptions underlying the choice of analysis were guided by the established workflows and commonly used approaches in frequentist statistics: For example, (i) when the assumptions were violated for a particular analysis (e.g., normality of residual distributions or homoscedasticity for the analysis of variance), then another suitable method (e.g., Friedman test as a non-parametric alternative) was used; (ii) when the data were treated as ordinal in a correlational design, then the relation between variables were measured using Spearman's correlation coefficient as an appropriate alternative, instead of Pearson's correlation coefficient which requires both variables to be continuous (i.e., either ratio or interval, but not ordinal); (iii) when some datapoints were missing in a repeated measures design, then a linear mixed-effects model was treated as an equivalent analysis, where each data point was treated as a single observation, instead of excluding participant's all data in an analysis of variance.

The main reason to select the frequentist inference over alternatives, such as Bayesian statistics or decision theory derived primarily from game theory, was simply because of that the overwhelming majority of the existing research seems to embrace frequentist approaches. By using the frequentist approach as the common practice, it was easier to convey the results to the audience, including potential readers and reviewers in the field. It is worth noting that some advantages of Bayesian analysis were underlined by researchers. Additionally, particularly to compare the Frequentist approach with the Bayesian one, for the presented experiments, both methods would lead to a similar inference, similar uncertainty levels, but the interpretation would be slightly different. As a mere statistical exercise, although without further reporting, parts of the frequentist results were further re-analysed using Bayesian counterparts. To illustrate, for example, a sample comparison from a single case in VR gallery chapter, for the liking rating difference between two levels of artwork spatiality, where the frequentist RM-ANOVA showed probability value of $p_{\text{Bonferroni}} = 0.00009$ (roughly interpreted such that the probability of obtaining the results is 0.009% whilst assuming that the null hypothesis is correct), and a Bayesian counterpart ANOVA showed a BF = 855.350 (roughly interpreted such that given the data, the alternative hypothesis is 855 times as likely as the null hypothesis): as a gist of both, the outcomes can be described as strong evidence against the null hypothesis.

Although the further discussion on issues related to data analysis is outside the reach of this section, note that, some recent criticism suggests, for example, abandoning null hypothesis testing altogether (McShane et al., 2019), or seeking alpha-level justifications instead of blanket values (Lakens et al., 2018), or proposing alternative statistical approaches including the avoidance of p-values and preference of reporting 95% CIs (Cumming, 2014). Besides, it is important to underline that the Bayesian framework can be also a supplement to the frequentist framework, and obviously, the Bayesian framework is much more capable in terms of statistical designs, and also not a simple alternative form of frequentist approach (e.g., there is not a direct cross-tabulation between p-value and BF). Besides the theory-driven resources on introducing the Bayesian statistics (Bolstad, 2007), more recently, advocates aim to make the analysis methods more approachable and accessible (Wagenmakers, Love, et al., 2018; Wagenmakers, Marsman, et al., 2018). Lastly, as all the data are made available on repositories, the alternative approaches to data analysis are open for anyone interested enough to re-analyse.

2.3. Additional experimental principles

2.3.1. Experimental validity

The validity roughly refers to the amount of (conceptual) overlap between the phenomenon which the experimenter aims to measure, and the measure itself, and many subcategories of validity are proposed by researchers (Kaplan & Saccuzzo, 2017). In particular, for example, ecological validity can be described as the extent of generalisability of the lab-based observations and results into the real-world. (Schmuckler, 2001). Here, in-situ and VR experiments can be described as having a higher ecological validity since they were either real-life art experiences or close proxies of them. Similarly, the online settings can also be described as an ecologically valid setting, in the sense of online art experience (for example, by engaging with artworks via websites), but not in the sense of experience in physical spaces such as art galleries. Many researchers often precisely control or manipulate the visual stimulus, and all the experimental protocols to create otherwise identical conditions across participants. This approach with well-justified reasonings especially in line with the psychophysical tradition mainly aimed to increase the internal validity, and partially aimed as a way of minimising unwanted effects of the potential mediator or moderator variables. Here, to approximate the experimental conditions that facilitate genuine aesthetic experience is somewhat contradicting the precise protocol control. Indeed, as a general trade-off in behavioural research, some research settings were far from highly controlled conditions in the classical sense, therefore the internal validity was minimised in the sense of stimulus control and environmental factors. For example, compared to regular lab settings where inactive participants view representations of artworks on a monitor whilst their securely seated and their heads are fastened upon a chin-rest, in-situ experiments were more prone to confounding factors. In this sense, both Compton Verney Mural and Mondrian Room in Dresden studies carried out in real-world settings can be defined more or less as a case study under minimally controlled conditions (compared to lab-based experiments), and therefore had relatively less internal validity. The generalisability of findings beyond the experiment to a wider sample and stimulus, which is often associated with external validity, were partially addressed in lab-based VR and online experiments. Arguably, in this thesis, whilst aiming for a high ecological validity, to some possible extent, the experimental designs were aimed to balance ecological validity and rigid control.

2.3.2. Reliability of measurements

The concept of reliability is associated with the consistency and reproducibility of a measure, given the particular research method and context (Brysbaert & Rastle, 2009). Although sometimes linked to validity, such as sometimes denoted as a required prior step for validity, reliability is a distinctive concept compared to validity, and the amount of reliability does not necessarily imply validity. Reliability can be classified into categories, with classical types such as (i) inter-rater, (ii) test-retest, and (iii) inter-method reliability, and so on. However, the presented studies were not suitable to be directly related with these classical reliability types, since (i) no rater was present in the data analysis, (ii) no new measure was developed to be reused (such as a new survey on aesthetics as an instrument), and (iii) no assumption was made whether to obtain similar results with varied methods and instruments. If the reliability can be conceptualized as a spectrum, then the researcher can aim for higher reliability. Improving the reliability is not always a straightforward path, and even the same research setting can affect two measures differently: For example, whether the experiment's stimulus was a participant's first artwork that they engaged with in their gallery visit or last could undoubtedly influence their responses in the supplementary exit-questionnaire (based on e.g., potential mood and feeling changes during their visit), but arguably in the same context, the gaze patterns are less prone to such influencing factors (under the assumption that when engaging with abstract installation artworks, the bottom-up processes may be the principal driving force of the viewing patterns). Additionally, it is not easy to compare the level of reliability between distinct measures (such as eye-tracking and questionnaire). To some extent, the reliability was aimed to be enhanced in experimental design, for example, by conducting VR-only studies at the lab to minimize the influence of external factors that are inherently present in-situ scenarios; or to some extent, by aiming to minimize the primacy and recency effects during judgment tasks, by allowing participants to view all artworks or artwork variations at the same time and to provide ratings on continuous scales for the main data (instead of Likert-type scales or binary judgments). Note that eye-tracking alone can only provide partial answers, and aesthetic judgments are prone to fluctuate even for the same participant, whose decision-making process is not something fixed, and can be either affected by the study design, by context, or just change throughout the viewing. In this sense, whenever possible, whilst the reliability was aimed to be improved, the studies also often incorporated multiple methods of measurements simultaneously (such as using eye-tracking and questionnaire) with varying degrees of reliability into the same experiment.

2.3.3. Reproducibility

The open science movement at its core defends the idea that all aspects of the research should be transparent and accessible to all, wherever possible (Stodden et al., 2014; Christensen et al., 2019). Here, in terms of reproducibility in the sense of replicable and repeatable experiments, individual methods along with supplementary information were provided in detail. The anonymized data were made publicly available on the Open Science Framework, and individual links to the separate repositories are referred in chapters: specific hyperlinks for empirical chapters were www.osf.io/bgtpy (for the Mondrian Room in Dresden as the 3rd thesis chapter), osf.io/ec46q (for the Virtual Reality Gallery as the 4th thesis chapter), and osf.io/m7nk8/ (for the Compton Verney Mural as the 5th thesis chapter). To distribute the research outputs to a wide range of potential readers without any barriers in the sense of accessibility, all empirical chapters as individual papers are aimed to be published in open access format. Additionally, the reproducibility crisis in psychology (and potentially in many other disciplines) is echoed by many researchers (Open Science Collaboration, 2015), and a growing tendency is the practice of preregistration, although it comes with some criticism. The studies presented here were not preregistered mainly due to the exploratory nature of the experiment sets and analyses (which are not always easily compatible with the practice, compared to confirmatory analyses), and under the assumption that the lack of preregistration is not a de-facto or perceived state of less credible research. Nevertheless, I am aware that the lack of preregistration could be seen as an unanticipated weakness of the methods.

2.3.4. Novelty and scalability of methods

Compared to the existing body of research in empirical aesthetics, some unique aspects of the presented approach should be underlined: briefly, (i) utilising eye-tracking both in-situ and VR settings, and relying on quantitative eye-tracking data analysis to compare the similarity of viewing patterns between a physical and a virtual recreation of a historically important artwork design; (ii) the use of 3-Dimensional artworks as stimuli to investigate the relationship between aesthetic judgments across a wide range of objects and the inclusion of understudied types of judgments such as perceived viewing duration to check temporal distortion during artwork viewing; and (iii) using in-situ eye-tracking data analysis for an installation artwork as a primer to forming parametrically modified artwork variations, to follow up the aesthetic judgments and eye-tracking on those variations in online settings. In terms of scalability, although the methodological aspects seem to require either a wide skill set or interdisciplinary teams for the future, it is feasible to extend the presented workflows into broader research areas in vision sciences and psychology, and practice-led contexts such as data-driven museum practice, digital cultural heritage, architectural design, and creative industry.

2.3.5. Epilogue

Firstly, the reasoning behind the alternative thesis formatting can be mentioned: Since (i) each main study was not strictly in successive order, but complementary to each other, (ii) each set of experiments as a single empirical chapter was more or less contained as individual research on its own, and (iii) the thesis was not based on a single, rigorous hypothesis testing consisting of sets of experiments, an alternative thesis format is thought to be more suited instead of a monograph format.

Additionally, some aspects of the presented studies can fall into the categories of descriptive, exploratory, or observational research; built upon basic scientific curiosity, but some methodological peculiarities were present throughout the empirical chapters. To highlight a conceptual one on the approach to the hypothesis testing, and to illustrate in a specific example, in the Mondrian Room in Dresden as the 3rd chapter, the expectation could have been to observe similar eye movement from museum visitors between two conditions (i.e., whilst viewing physical and VR installation art) using a dwell time analysis based on areas of interests (AOIs). In the formal phrasing, this translates into conceptually defending a null hypothesis over an alternative hypothesis, which is arguably an indisputable statistical mistake. One potential reason behind this depiction may be related to the conventions in certain types of behavioural research, such that a remarkable portion of the experimental psychological circle seems to reduce the aim of an experiment into rejecting a null hypothesis, thus often induces a highly restrictive way of thinking on researchers. In a way, this resonates with the (highly criticised) philosophical movement of logical positivism which roughly aims to translate all statements and propositions into strict logical structures. Obviously, given the meaningfully designed experiments, many theoretical claims can be converted into testable and context-dependent hypotheses. However, and as an arguably highly controversial claim, not all empirical research should be in favour of an alternative hypothesis, or even should explicitly have to state one.

Following on from that, established practices should not force every researcher to be a conformist, and unconventional methodological aspects should not be perceived as false.

Lastly, the single largest obstacle of this research is the effects of the global and (at the time of writing) ongoing pandemic: besides the obvious difficulties, at a minimum, either a total closure or restricted access to both lab spaces and cultural institutions for long periods has been impeding the progress of this project, in terms of cancellation of some already prepared or intended experiments. The impacts of these obstacles were partially aimed to be minimised by conducting additional online experiments and presenting the feasible yet unrealisable experiments in the discussion sections.

2.4. References

Argyrous, G. (2011). *Statistics for research* (3rd ed). SAGE.

Berry, D. M., & Dieter, M. (2015). *Postdigital aesthetics: Art, computation and design*. Palgrave Macmillan.

Bolstad, W. M. (2007). *Introduction to Bayesian statistics* (2nd ed). John Wiley.

Bourne, V. (2017). *Starting out in methods and statistics for psychology: A hands-on guide to doing research* (1st ed.). Oxford University Press.

Brysbaert, M., & Rastle, K. (2009). *Historical and conceptual issues in psychology*. Pearson/Prentice Hall.

Caixinha, M., & Nunes, S. (2017). Machine learning techniques in clinical vision sciences. *Current Eye Research*, *42*(1), 1–15. https://doi.org/10.1080/02713683.2016.1175019

Christensen, G. S., Freese, J., & Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science*. University of California Press.

Çöltekin, A., Lochhead, I., Madden, M., Christophe, S., Devaux, A., Pettit, C., Lock, O., Shukla, S., Herman, L., Stachoň, Z., Kubíček, P., Snopková, D., Bernardes, S., & Hedley, N. (2020). Extended reality in spatial sciences: A review of research challenges and future directions. *ISPRS International Journal of Geo-Information*, *9*(7), 439. https://doi.org/10.3390/ijgi9070439

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Duchowski, A. (2017). *Eye tracking methodology: Theory and practice* (3rd ed.). Springer International Publishing.

Field, A. P., & Hole, G. (2003). *How to design and report experiments*. Sage.

Holmqvist, K., & Andersson, R. (2017). *Eye tracking: A comprehensive guide to methods, paradigms, and measures* (2nd edition). CreateSpace.

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, & issues* (9th ed.). Cengage Learning.

Katz, M. J. (2009). *From research to manuscript: A guide to scientific writing*. Springer.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Linowes, J. (2018). *Unity virtual reality projects: Create interactive VR projects with Unity 2018*. PACKT Publishing Limited.

Maggio, M. G., Maresca, G., De Luca, R., Stagnitti, M. C., Porcari, B., Ferrera, M. C., Galletti, F., Casella, C., Manuli, A., & Calabrò, R. S. (2019). The growing use of virtual reality in cognitive rehabilitation: Fact, fake or vision? A scoping review. *Journal of the National Medical Association*, *111*(4), 457–463. https://doi.org/10.1016/j.jnma.2019.01.003

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Paul, C. (Ed.). (2016). *A companion to digital art*. John Wiley & Sons Inc.

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences*, *10*(4), 251. https://doi.org/10.3390/brainsci10040251

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy*, *2*(4), 419–436. https://doi.org/10.1207/S15327078IN0204_02

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, *5*(1), 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

Siu, A. F., Sinclair, M., Kovacs, R., Ofek, E., Holz, C., & Cutrell, E. (2020). Virtual reality without vision: A haptic and auditory white cane to navigate complex virtual worlds. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376353

Stodden, V., Leisch, F., & Peng, R. D. (Eds.). (2014). *Implementing reproducible research*. CRC Press, Taylor & Francis Group.

Voigt, P., & von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR)*. Springer International Publishing. https://doi.org/10.1007/978-3-319-57959-7

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wilke, C. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures* (First edition). O'Reilly Media.

Zhao, Y., Cutrell, E., Holz, C., Morris, M. R., Ofek, E., & Wilson, A. D. (2019). SeeingVR: A set of tools to make virtual reality more accessible to people with low vision. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300341

Chapter 3

Mondrian Room in Albertinum Museum

Similarity of gaze patterns across physical and virtual versions of an installation artwork

Doga Gulhan* · Szonya Durant · Johannes M. Zanker

Department of Psychology, Royal Holloway, University of London, UK

*doga.gulhan@rhul.ac.uk

3.1. Abstract

An experiment was conducted to compare museum visitors' gaze patterns using mobile eye-trackers, whilst they were engaging with a physical and a virtual reality (VR) installation of Piet Mondrian's Neo-plasticist room design. Visitors' eye movements produced approximately 25,000 fixations and were analysed using linear mixed-effects models. Absolute and area-normalized dwell time analyses yielded mostly non-significant main effects of the environment, indicating similarity of visual exploration patterns between physical and VR settings. One major difference observed was the decrease of average fixation duration in VR, where visitors tended to more rapidly switch focus in this environment with shorter bursts of attentional focus. The experiment demonstrated the ability to compare gaze data between physical and virtual environments as a proxy to measure the similarity of aesthetic experience. Similarity of viewing patterns along with questionnaire results suggested that virtual galleries can be treated as ecologically valid environments that are parallel to physical art galleries.

3.2. Introduction

Empirical aesthetics emerged in the late 19th century (Fechner, 1876; Helmholtz, 1863), and was roughly contemporaneous with foundations of experimental psychology and psychophysics. Following on, pioneering eye-tracking research (Buswell, 1935; Yarbus, 1967) was also asking questions on how observers engage with artworks visually, whilst another line of inquiry was aiming to capture the external world via photography. Now, half a century after the earliest computer-generated 3D movies (Noll & Hill, 1965) and VR headsets (Sutherland, 1968), recent developments in computational power and VR eye-tracking (Clay et al., 2019) allow researchers to conduct experiments in accurate recreations of 3D environments: this includes our work presented here, which made use of a unique opportunity to compare visual exploration of a single installation artwork alongside its VR reconstruction in a museum space, which can lead the way to evaluating the validity of virtual and online arts experience.

In line with psychophysical approaches, visual artworks have been generally treated as controllable or categorizable stimuli (Locher, 2013; Vessel & Rubin, 2010). Similarly, and on a more theoretical level, approaches to conceptualise aesthetics "from below" on the basis of visual information processing assume that perception of artworks relies mostly on bottom-up processing with minimum influence from top-down processing. Generally, bottom-up (data-based) processing frames visual perception as a stimulus-driven and direct phenomenon, whereas top-down (knowledge-based) processing underlines the influence of past experience and prior knowledge on visual perception, and describes it as a more indirect, inference-making process (Gregory, 1995; Goldstein & Brockmole, 2017), although this arguably imprecise dichotomy calls for further conceptual refinements (Rauss & Pourtois, 2013). Following criticism against some reductionist approaches on aesthetics (Machotka, 1995) and evidence from less restrictive experiments, detailed models of aesthetic experience were suggested (Wagemans, 2011; Bullot & Reber, 2013; Leder & Nadal, 2014; Menninghaus et al., 2015), often embracing both the universality of aesthetic experience as a result of low level visual neural dynamics and the diversity of aesthetic experience as a result of contextual and personal factors (Nadal & Chatterjee, 2019). A similar inclusive viewpoint argues for influences of both bottom-up and top-down processing on (overt) visual attention and attention-related tasks (Carrasco, 2011; Theeuwes, 2010), which can be linked to eye movements in general, and fixation-related metrics in particular.

Although aesthetic experience, as with other related concepts, is a highly debatable topic in itself by both theorists and experimentalist (Carroll, 2002; Chatterjee & Vartanian, 2014; Makin, 2017; Iseminger, 2005), prone to circular definitions; at its simplest, it refers to a particular state of mind whilst engaging with a denoted artwork, a context-dependent spatiotemporal episode of the conscious experience. Additionally, the concept of attention can be linked to visual perception and more particularly to the visual aesthetic experience (Nanay, 2010, 2015), since it can be described as a mechanism that shapes (e.g. selects, concentrates, distributes) the information received from the scene; although often linking the theoretical works and empirical findings related to attention into a coherent description is an ongoing challenge (Ferretti & Marchi, 2020). Simply assuming aesthetic experience to be a set of highly complex cognitive-emotional processes involving attentional mechanisms, eye movements can be seen as a reflection of both underlying bottom-up and top-down processes.

Until recently, the only feasible way of implementing eye-tracking in a study was by restricting it to laboratory environments, where generally the only stimulus option was the reproduction of artworks instead of originals. In line with the development of mobile eye trackers, a paradigm-shift for empirical aesthetic research was stepping outside of the well-controlled laboratory environments into real-world where observers engage with works of art in their original forms (Pelowski et al., 2018). In a previous pilot experiment, for example, we investigated the eye movements of gallery visitors whilst they were engaging with a room-scale installation (Gulhan & Zanker, 2019): this installation was later recreated virtually with a set of variations based on the topological properties and observed gaze patterns, and these variations were used in an online eye-tracking experiment in a 2D view. To further illustrate the research potential, researchers have investigated (i) interaction between gaze patterns and abstract paintings in a gallery (Zanker et al., 2017) and potential implementation of scan path analysis using support-vector machine algorithms to classify paintings based on fixation sequences (Stevanov et al., 2019), (ii) use of mobile eye-tracking analysis on abstract and representational paintings in a museum (Heidenreich & Turano, 2011), (iii) effects of bottom-up factors (as indexed by saliency maps derived from paintings) and top-down factors (as manipulated by the information about paintings provided to the participants, who were allowed to view the same paintings again) between children and adults, whilst viewing Van Gogh paintings (Walker et al., 2017), (iv) interaction between speaking and fixation patterns and various gaze metrics (Klein et al., 2014), (v) difference in exploration strategies among wheelchair and non-chair users in museums (Tymkiw & Foulsham, 2019), (vi) amount of attentional shift between museum content itself and a supplementary tablet containing information on that content (Guntarik et al., 2018), (vii) whether fixation duration can predict aesthetic choice (Isham & Geng, 2013), among others. One commonality across these diverse studies is their emphasis on the necessity of fieldwork in empirical aesthetics, aiming to measure aesthetic experience and judgments in genuine settings.

Presenting arts online and more recently virtually was a huge step forward for accessibility of cultural heritage. Although VR has been used previously in pioneering works (Heilig, 1962) and research (Fisher et al., 1987), there is currently a growing interest in both consumer-grade and research-grade VR solutions. One particular reasoning behind this interest seems to be the experimental research potential to employ freely moving participants in virtual environments (Scarfe & Glennerster, 2015; Wilson &

Soranzo, 2015; Rothacher et al., 2018). Also, the accessibility to modelling software and game engines provides ease and widely accessible tools to create novel immersive environments (Steinicke, 2016; Tricart, 2017; Pangilinan et al., 2019). Additionally, following the development on eye movement analysis in 3D space collected from digital simulated environments (Duchowski et al., 2002; Duchowski, 2017), the recent emergence of VR headsets capable of eye-tracking offers a completely new opportunity to step beyond the conventional lab and into the in-situ context. A relatively unexplored area with emerging experimental design guidelines as well as some ethical concerns (Madary & Metzinger, 2016; Miller et al., 2020), nevertheless VR holds exciting promise for empirical aesthetic research. There is some previous research comparing museum and laboratory settings as well as original and reproduction artworks (Brieber et al., 2015), and similarly, investigating preference towards types of substituted representations of artworks (Bertamini & Blakemore, 2019), or targeting emotional experience using mobile EEG to further develop a classifier based on the data recordings from a real and virtual museum (Marín-Morales et al., 2019). In line with recent experimental results underlining the observed contextual differences (particularly between lab-based and real-world conditions), aesthetics research in laboratories resembling the genuine contexts of aesthetic experience as much as possible was proposed (Carbon, 2020). However, direct comparative research between the art galleries and arguably their closest proxy, immersive environments, (particularly for 3D arts and based on eye-tracking) is still missing. As the VR environment develops into a valid and comparable setting to physical galleries and museums, a direct comparison between VR and in-situ environments seems to be crucial to enable future use of VR: if the potential upcoming research favours the similarity between two settings, then immersive environments can be treated as both highly controllable and ecological valid research settings.

Here we focus on the work of Piet Mondrian, whose abstract paintings are prominent examples of the De Stijl art movement. In most of his late works, Mondrian radically restricted compositional features of artworks following the art movement of Neoplasticism, by using only horizontal and vertical lines and three primary colours red, blue, and yellow, along with black, white, and grey. This abstraction epitomising purity and sparseness of lines and colours lends itself to straight-forward mathematical descriptions to aid quantitative approaches. In this sense, reproduction of Mondrian paintings and quasi-Mondrians as manipulated versions of originals have been used as stimuli for empirical aesthetics, arguably due to the artist's historical significance as well as the low-level compositional features, offering clear and easily modifiable geometric structures as experimental stimuli. For example, researchers investigated whether computer-generated synthetic Mondrians were preferred more compared to originals (Noll, 1966; McManus et al., 1993), whether original or rotated, oblique orientations of Mondrian's paintings were preferred and whether eye-movement patterns were similar across orientation conditions (Plumhoff & Schirillo, 2009), and whether aesthetic preference towards Mondrian paintings was correlated with measured pupil size of participants (Johnson et al., 2010). Incorporation of other variables in relation to Mondrian's work has also been a prominent research theme, such as asking whether liking of original Mondrians was mediated by personality factors like openness to experience (Swami & Furnham, 2012). Recently, a distinct example of Mondrian's work, a room design proposal commissioned by Ida Bienert in the early 20th century but never realised (Troy, 1980) has drawn attention from researchers, along with an art-historic curiosity. Using variations of scale physical and digital models, it was argued that the room-scale artwork proposal conflicted with strict neoplasticist ideals, because of perspective distortions in retinal projections, which are exacerbated by changes of viewpoints in the room (Stevanov & Zanker, 2020). Following on from that, our test case mainly aimed to measure observers' visual exploration as indexed by eye-trackers inside 1:1 scale physical and virtual versions of this particular design proposal. In terms of art research, our approach can be seen both as a behavioural experiment in a physical gallery, and as a comparative study aimed to investigate whether a virtual installation would be a suitable proxy for a physical installation.

The main aim of the present study was to compare observers' gaze patterns (as constituents of the aesthetic experience) within an art installation in physical and virtual instantiations. The physical installation created by the artist Heimo Zobernig, and the VR reconstruction developed by our team were temporarily exhibited in the Albertinum Museum in Dresden, Germany. Having a full day of access to a flagship exhibition on historical milestones of abstract art in an internationally renowned museum (Dalbajewa et al., 2019) to quantitatively analyse looking behaviour in a live gallery setting and to assess the similarity between physical and VR contexts was a unique opportunity for us. In this case study, we collected both implicit measures as ocular responses using a mobile eye-tracker for the physical installation and using a VR eye-tracker for a virtual reconstruction, and supplementary explicit measures as questionnaire responses. Specific ocular responses related to fixations, such as dwell time and fixation count, are usually associated with overt attention, visual attention guidance, and other related (unconscious) cognitive processes (Geisler & Cormack, 2011). On the other hand, questionnaire responses as a part of psychological testing were considered to reflect decision making and other (conscious) cognitive processes (Kaplan & Saccuzzo, 2017). As a result, to the best of our knowledge, we present the first direct comparison of quantitative measures capturing core aspects of visual aesthetic experience of an installation artwork in both physical and virtual embodiment in its museum context. Given the strong topological similarity between physical and VR installations, we expect very similar visual exploration behaviours between the two contexts. In this sense, the non-directional alternative hypothesis can be formulated such that the visual exploration patterns as indexed by absolute and area-normalized fixation duration regarding sets of area of interests (AOIs) during the viewing of a static abstract installation between in-situ and VR condition are different, whereas the null hypothesis as the default state can be formulated that there is no difference between the two contexts.

3.3. Methods

3.3.1. Participants

Museum visitors were approached at the exhibition entrance and invited to take part using opportunity sampling during regular visiting hours. They took part in the study voluntarily. All participants provided written informed consent prior to the experiment. Thirty-one museum visitors (21 females, 9 males, M_{Age} = 49.23 years, SD_{Age} = 18.25 years, R_{Age} = 20-79 years) participated in the study. All participants reported to have normal or corrected-to-normal vision, in the sense that they viewed both stimuli in the same conditions as if they were viewing other artworks of the exhibition: participants could use their contact lenses for both settings, or wear their glasses in the VR headset and a corrective lens was added to the wearable eye tracker whenever needed. Although no explicit vision status measure, such as a visual acuity or contrast sensitivity test, was not implemented; a screening questionnaire comprising eleven items were provided, aiming to link any unusual eye-tracking data (such as calibration failure or frequent lack of fixation detection) to the vision condition (such as recent laser surgery), and potentially to exclude the participant data (see Supplementary Fig. S1 for the screening form and exit questionnaire, English version). All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki.

3.3.2. Stimulus and material

The physical stimulus, the installation artwork, a spatial appropriation of the Mondrian's room design by the artist Heimo Zobernig (heimozobernig.com), was exhibited in the Albertinum Museum in Dresden, Germany (skd.museum) as a part of the exhibition entitled Future Spaces: Kandinsky, Mondrian, Lissitzky and the abstract-constructivist avant-garde in Dresden 1919-1932. The digital stimulus, the VR reconstruction based on the same room design, was developed by our team using a modelling software called SketchUp (sketchup.com) and a game engine called Unity (unity.com). Both stimuli could be described as faithful interpretations of Mondrian's room design from 1926 entitled Design Draft of Salon for Madame Bienert. Both versions had minor adjustments compared to the original design draft of Mondrian, as an artistic statement by Zobernig in the case of the installation (see Fig. 1a-b and Supplementary Fig. S2a), or as a reconstructive decision for the VR implementation to match it to the physical layout of the original room, the Damenzimmer in Ida Bienert's villa in Dresden, for which it was designed (see Fig. 1c-d and Supplementary Fig. S2b). Briefly, Zobernig produced the artwork as an interpretation, which deliberately did not try to exactly reproduce Mondrian's commissioned watercolour painting of the design, which furthermore did not match the actual, physical dimensions of the room in the Bienert Villa. Our team's VR reconstruction was based on Mondrian's composition combined with physical room measurements taken by us, and in line with Neoplasticist rules, slightly adjusting the design such as to fit into the actual room layout, including positions of walls, windows, and doors. In this sense, we did not aim to create identical architectural constructs, but compare the aesthetic experience in two very similar environments inspired by the same design idea. The adjustment of the VR design in accordance with the actual room dimensions led to VR dimensions of 499 by 494 cm, with a height of 360 cm, whereas the dimensions of the physical installation were 483 by 510 cm with a height of 385 cm, which followed Mondrian's design sketch. Both the physical and VR installation incorporated monochromatic coloured patches for the room surfaces, and two main natural white lighting sources as an ambient light and as a surface light coming from the ceiling inside of the room, with no further controls for the similarity of the colour saturation and luminance. The outer environment surrounding the installation was a static grey scene in VR without any additional digital audio or digital avatars in the scene, whereas the physical installation was situated in the large museum space, right next to our VR installation (compare Fig. 1a-b and Fig. 1c-d). In both settings, some background noise from the visitors were inevitably present in the gallery space. Hardware and software used to record was a wearable eye-tracker (Tobii Pro Glasses 2) via Tobii Pro Glasses

Controller, and a VR headset (HTC Vive with integrated Tobii eye tracker) via an executable file built by using Unity with Tobii Pro VR Analytics, a software package to enable data collection. In addition, an exit questionnaire was completed by participants, which included four items on basic demographic information, eighteen rating items as five-point Likert-scale that implement both positive and negative scoring on interest and opinions about art, and four open-ended questions as feedback (see Supplementary Fig. S1 for questions from the rating-scale questionnaire).

3.3.3. Design

The study was designed as a within-subjects experiment, since each participant was expected to take part in both physical and VR conditions. To overcome carryover effects, the order of visiting the physical and virtual room was counterbalanced as the participants were pseudo-randomly assigned to either physical-first or VR-first conditions such that the half of the participants viewed the physical installation first, and the other half viewed the VR version first. The main conceptual justification for encountering two versions of the room design was the assumption that the forms of (top-down) effects due to the temporal order of viewing tend to cancel each out, albeit potentially introducing some noise to the data. Main dependent variables, both as absolute and area-normalised values, were dwell time defined as cumulative fixation duration, number of fixations, and average fixation duration on particular regions of the room, and all measured using the eye-trackers. Main independent variables were artwork media type as a binary variable labelled either as physical or virtual, and sets of AOIs as indexed by corresponding 3D geometry of the rooms, such as six surfaces of the cuboid room, three pieces of furniture, and six colours on room panels (see Fig. 2 for an overview of the AOI mapping). In line with the null hypothesis, no main difference is expected in terms of absolute and area-normalized dwell time, depending on the artwork media type and sets of AOIs. Note that although it was not explicitly recorded, we expected noise in data from a type of re-exposure effect or forms of order effects such as fatigue, boredom, or practice effects, but aimed to minimize them using the counter-balancing. Additionally, most (if not all) participants engaged with the artworks for the first time during the experiment, at least for the first time during the data collection day: this assumption can be also supported by that none of the participants mentioned about a previous viewing of the artwork, and a majority of them were not familiar with the artist as indexed by their response on the questionnaire.

3.3.4. Procedure

After participants were given the written and oral instructions, they completed consent forms and screening forms about vision status at a reception desk, and they went through the experiment consisting of three steps. For one set of participants, they firstly were equipped with the mobile eye-tracking glasses and explored the physical artwork as long as they wanted by themselves. Researchers did not accompany participants, but since the study was conducted in a public museum space where other visitors had rights to view the artwork, there were a few instances of an additional visitor inside the installation, but only for brief intervals: there were five instances where another visitor was present simultaneously with a participant. For those five participants, the duration of co-presence was approximately 85 s in total, corresponding roughly to the 3% of the experiment's whole eye-tracking recording duration inside the physical installation. Secondly, participants were equipped with the VR headset and explored the digital artwork, again without any time limitation. In this phase, a researcher always accompanied the participant simply to handle the cables between the headset and the computer. For the other set of participants, the order of viewing was reversed. For the physical installation, the participants always entered and exited the room using the same door opening on the South Wall. Due to the technical limitation of the physically walkable area in the VR version, participants started the VR experience at the centre of the VR room, and all were instructed to face forward towards the same wall at the start. Lastly, they completed a brief exit questionnaire either in German or in English depending on their preferred language at the reception desk (see Supplementary Fig. S3a for an overview of the procedure, and Supplementary Fig. S3b for a view from the gallery space). Note, eye-tracking calibrations were executed prior to the data collection, separately for the VR and mobile eye-tracker for each participant, to ensure the reliability of gaze data (in terms of precision and accuracy). The default in-built calibrations provided by the Tobii software (i.e., five-point for the VR and single-point for the mobile) were expected to provide similar level of quality (but comparatively noisier than the data obtained in a lab setting from a research-grade screen-based eye tracker). No further interim recalibration, which is an occasionally used practice for the eye-tracking drift correction for longer experimental designs, was carried out, because the installation was treated as a single stimulus, and the recording duration was already short.

3.3.5. Data analysis

Three sets of data were formed: from the questionnaire, mobile eye-tracking, and VR eye-tracking. All participants completed the physical part of the experiment; however, due to a data-saving error on the mobile eye-tracker data during the study, seven recordings could not be recovered, resulting in twenty-four valid recordings from eye-tracking glasses, instead of thirty-one. One participant did not participate in the VR part of the experiment due to discomfort caused by the screen brightness, resulting in thirty valid recordings from the VR headset. All participants completed the exit questionnaire, resulting in thirty-one respondents. Analysis of the questionnaire included descriptive statistics indicating the frequency distribution of rating responses on a Likert-scale.

The workflow from the raw recording data to the statistical analysis are explained in detail in Supplementary Fig. S4. The main difference in workflow between mobile and VR was the existence of an interim manual coding for the fixation locations, realized by the lead author (also see Supplementary Fig. S5 for a reliability comparison). Following on from that, the AOI mapping (as illustrated in Fig. 2) was the basis for five comparisons of eye-tracking data, related to the sparse spatial layout and composition style of Mondrian's design, generating subsets of the data that were split out into separate levels for the data analysis: (i) Room elements consisted of three levels: interior cube surfaces, furniture, and regions representing the outside vista as door and window openings. (ii) Colour types were split into two levels: luminance- and chrominance-type, such that whether the colour had luminance-information only as white, grey, black, or had chroma-information as red, blue, yellow. (iii) Individual colours: all possible colours in the artwork produced six levels. (iv) Cube surfaces contained six levels: ceiling, floor, along with north (N), east (E), south (S), and west (W) walls, where the directions of walls were approximately based on the original room location in Dresden. (v) Furniture comprised three levels: bed, cupboard, and bookcase. Note that minor topological differences between the physical and VR versions resulted in slightly different surface areas (see Supplementary Fig. S6 for visible surface areas corresponding to each set of AOIs).

Five main types of eye-tracking variables were analysed which were either reported as main results or summarized as supplementary results: (i) Absolute dwell time was defined as the cumulative fixation duration per AOI. (ii) Area-normalized dwell time aimed to measure a type of fixation density or attentional density, after accounting for sampling at chance, to correct for the relative size of areas. It was calculated as cumulative fixation duration multiplied by the given AOI area in percentage, such that any given AOI was expressed as a fraction of the total area of all the AOIs. (iii) Fixation count was the number of individual fixations on a given AOI. Unless there is a significant difference in the average fixation duration metric (see below), fixation count tends to be highly correlated with absolute dwell time. (iv) Area-normalized fixation count was the normalized metric using area size of an AOI as a fraction, as above. (v) Average fixation duration was the mean duration of fixations on any given AOI, and a derivative metric since it was calculated as absolute dwell time divided by the number of fixations. This derived metric allowed comparison of how often the gaze is relocated between different image regions and viewing conditions (refer to Supplementary Fig. S7 for an overall table of mean and standard errors of the all above-mentioned measures, and also see Supplementary Fig. S8 for time to first fixation comparing the physical and VR settings for five comparisons, visualized as boxplots). Note that our explicitly presented results were mainly based on absolute and normalized dwell time to prevent analytical redundancy, since we expected very similar results for the potential analyses based on absolute and normalized fixation counts (see Supplementary Fig. S9 for the correlation table indicating the strength of the relation between dwell time and fixation count). Lastly, because the free-viewing introduced a difference in viewing time between conditions and across participants, viewing percentage can be calculated (see Supplementary Fig. S10). Eye-tracking data for absolute and area-normalized dwell time were analysed using linear mixed-effects model (LME), which is equivalent to repeated measures analysis of variance (RM ANOVA). The reason not to run RM ANOVAs was that missing data from a single condition entail deletion of all data from the participant, whereas in LME each data point is treated as a single observation without participant exclusion. The main software packages used for the data analysis were MatLab, R, jamovi, Mathematica (mathworks.com, r-project.org, jamovi.org, wolfram.com); software for the data visualization were Unity, SketchUp, Lumion, Adobe CC (unity.com, sketchup.com, lumion.com, adobe.com).

Prior to evaluating our results, it is important to appreciate that we aimed to analyse a unique case study, mainly to compare the similarity of eye movement patterns between a physical and a virtual version of an art installation, and therefore our presented results and conclusions were limited within the confines of the real-world experimental conditions, rather than general and definitive. Therefore, the results could not be boldly generalized to the wider population and to wider forms of artworks; this limitation holds true for almost all research in empirical aesthetics. Additionally, as a general disclaimer, due to the small sample size in the traditional sense, this research has potentially low power, which in turn increase the probability of incorrectly failing to reject the null hypothesis and minimise the likelihood of reproducibility of results presented.

3.4. Results

3.4.1. Questionnaire

The eighteen Likert-scale rating questions can be clustered into four categories, and the most frequent response from five scale points as the mode of the data can be regarded as the most informative value (see Supplementary Fig. S11 for a summary of questionnaire responses). Since the 5-point Likert-scale ratings were treated as ordinal data, the mean cannot be obtained, and for some items the median cannot be easily found (for example, for item c3 about the artistic activity and hobby, the median falls between "seldom" and "sometimes" response categories), whereas the mode can be always calculated. Overall, the questionnaire results indicated that the participants were highly educated, from diverse age groups, mostly regular museum visitors, were split into two on particular judgments to compare two settings, open to VR experience, but only for shorter periods of viewing times. Note that the questionnaire was purely aimed at gaining more insight about participants' views such as their overall attitudes towards visual arts and the experiment, or familiarity with the artwork. The main research question was about eye movements during the aesthetic experience, and not qualitative differences in the experience itself, and therefore particular assessment tools based on self-report (e.g., about art expertise (Specker et al., 2020), aesthetic emotions (Schindler et al., 2017) or quality of VR use (Kourtesis et al., 2019)) were not sought to be administered. On the conceptual level, we did not form our main hypothesis on the grounds of the confounding variables, for example, whether the amount of previous knowledge about the artist or an art period has a mediator effect on the relation between dwell time and sets of AOIs (see Supplementary Fig. S12 for some initial exploration of these relationships).

3.4.2. Initial visualizations: Dwell time as heatmaps

Since both physical and VR versions were visited by the same participant group, an initial one-to-one qualitative comparison was possible, using heatmaps to visualise the amount of dwell time on any given point for physical and virtual environments. Note that both eye-trackers collected gaze data with approximately 100 Hz sampling rate and used the same algorithms to detect fixations. An exemplar dwell time heatmap pair from a single participant for both conditions and from two diagonal viewpoints of the room can be seen in Figure 3. An initial qualitative evaluation indicated some similarities between overall viewing patterns and some specific differences such as response to furniture between physical and VR environments. Additionally, individual differences between participants were apparent: for example, some participants spent relatively more time in the artworks, resulting in longer total dwell times and fixation counts, given the free-viewing condition of the experimental design. Some variability was inevitably present in individual patterns of preference, for example, for particular colours or walls (also further refer to the open-data directory to view individual heatmaps of participants: osf.io/bgtpy). Overall in both conditions, hotspots of attention as indicated by densely fixated regions seemed to be preferentially located on coloured patches of red, blue, yellow, as well as the furniture (see Supplementary Fig. S13 for gaze data validity measures).

3.4.3. Total viewing time and fixations

The statistical significance testing was carried out by using a general linear model. Main reported descriptive values were mean (M) and standard error of the mean (SEM) for any given analysis. The total time a participant spent in the room as indexed by the measured duration between entering into the installation and exiting from the installation was approximately two minutes in the physical environment and three minutes in the virtual environment ($M_{Physical}$ = 118.50 ±15.29s, M_{VR} = 172.70 ±19.51s) on average, reaching a difference with statistical significance ($F_{(1, 52)} = 4.43$, $p = .040$, $\eta_p^2 = .078$), and shown in Figure 4a. The viewing duration was comparatively longer than findings in the previous studies where the average viewing duration for 2D artworks such as paintings tends to be around 30 seconds in a museum context (Carbon, 2017; Smith et al., 2017), but rather similar to another research where the viewing durations for two distinct 3D installations were around two and four minutes (Pelowski et al., 2018). In line with this total viewing duration difference, dwell time was relatively shorter in physical environment compared to virtual environment ($M_{Physical}$ = 76.02 ±11.59s, M_{VR} = 97.68 ±11.54s), but did not reach statistical significance (as $F_{(1, 52)} = 1.71$, $p = .197$, $\eta_p^2 = .032$), and shown in Figure 4b. Similarly, total fixation count was smaller in physical-environment compared to virtual environment ($M_{Physical}$ = 322.08 ±42.32**,** M_{VR} = 579.13 ±69.53), reaching statistical significance (as $F_{(1, 52)} = 8.82$, $p = .005$, $\eta_p^2 = .145$), and shown in Figure 4c. Lastly, average fixation duration was substantially longer in physical environment compared to virtual environment ($M_{Physical}$ = 226.25 ±6.56ms**,** M_{VR} = 171.31 ±5.11ms), reaching statistical significance (as $F_{(1, 52)} = 45.05$, $p < .001$, $\eta_p^2 = .464$), and shown in Figure 4d.

Taken together, these initial data analyses showed that participants seemed to be slightly more engaged with the virtual installation compared to the physical installation, which might be attributed to several differences between the two conditions, including a possible novelty effect of VR as suggested by the questionnaire data showing that most participants had not used VR previously, and the presence of other visitors in the physical installation, among other possible distractions. The substantial difference between average fixation durations suggested a shift in terms of general viewing strategy: visitors seemed to be rapidly scanning the VR environment with shorter intervals of attentional focus as reflected by shorter fixations, compared to the physical environment.

3.4.4. Spatial distribution of area-normalized dwell time

Each comparison of area-normalized dwell time (described as attentional density by accounting for sampling at chance to correct for the relative size of areas, and calculated as cumulative fixation duration multiplied by the given AOI area in percentage) was analysed using a separate linear mixed-effects model: for comparison 1 on room elements, all AOIs were used in the analysis, including AOIs belonging to outside areas visible through window and door openings. For comparison 2 and 3 on colour types and individual colours, every coloured surface was used in the analysis including AOIs on furniture and six faces of the cube, but not outside areas. For comparison 4 on cube surfaces, both furniture and outside were excluded from the analysis, as they were not part of the walls. For comparison 5 on furniture, only AOIs on three pieces of furniture was used in the analysis. Note that the cupboard as one of these furniture had four additional AOIs as its frame or profile in VR condition compared to the physical installation, since the cupboard was constructed as a 3D object in VR but rendered as only a 2D flat surface in the physical installation. Also note that each individual rectangular panel of the room was defined as a single AOI, and then they were combined into sets for a given analysis: for example, all four blue panels as four distinct AOIs in the room constituted blue-condition for the comparison 2 and 3 on colour types and individual colours, all panels on the ceiling constituted ceiling-condition for the comparison 4 on cube surfaces, etc. Post hoc comparisons using t-tests were Bonferroni corrected; significance level, denoted by *a*, was set to .05; and Bonferroni-corrected *p*-values as observed, unadjusted *p*-values multiplied by the number of comparisons made were reported for determining significance for all results. Area-normalized dwell times comparing the physical and virtual environment without normalization of area covered by AOIs, are shown as boxplots in Figure 5a-e. Lastly, note that the spatial distribution of absolute dwell time can be further seen as a supplementary analysis in Supplementary Fig. S14.

Comparison 1 on room elements: A significant difference between room elements was found, and a difference was observed between environments and in terms of an interaction: $F_{(2, 35.3)} = 30.64$, $p < .001$; $F_{(1, 30.1)} = 7.14$, $p = .012$; and $F_{(2, 80.2)} = 10.34$, $p < .001$ respectively. In terms of room elements, the area-normalized dwell time on furniture ($M_{Furniture}$ = 185.70 ±17.69s/%) was higher than both for surfaces ($M_{Surfaces}$ = 71.60 ±7.46s/%) and outside ($M_{Outside}$ = 115.00 ±13.09s/%): $t_{(28.7)} = 7.79$, $p < .001$; and $t_{(28.6)} = 5.11$, $p < .001$, respectively. In terms of environments, the area-normalized dwell time in VR (M_{VR} = 148.64 ±13.04s/%) was longer than the physical installation ($M_{Physical}$ = 93.50

±9.07s/%). After normalizing for the area sizes, overall, the density of visual attention was highest in VR compared to in the physical environment, in line with the similar but non-significant trend observed for absolute dwell time. Here, the difference between VR and physical environment reached a statistical significance, mainly due to increased weighting of furniture and outside for the analysis, and also due to minor area size differences between physical and VR versions mentioned previously. Similarly, since the surface areas of furniture and outside were relatively smaller than room elements, the area-normalization changed the trend between room elements such that visitors attended significantly more densely on furniture of the installation compared to surfaces or outside, irrespective of environments. When the interaction was broken down by focusing on the two types of environment to check whether environmental differences exist for any level of room elements, normalized dwell time for furniture and outside were significantly higher in VR compared to the physical environment (p = .041, p = 0.011, respectively), whereas the difference was not present for surfaces (p > .05). When the interaction was broken down by focusing on the room elements to check how room elements differences have an effect differently for VR and physical environment, some trend changes were also visible, such as the normalized dwell time difference between surface_{VR} and outside_{VR} was significant (p < .001), but the dwell time difference between surface_{VR} and $\text{outside}_{Physical}$ was not significant as (p > .05), suggesting that for some pairs, the amount of dwell time difference was dependent on the environment. Note that since the surface areas for both furniture and outside were relatively higher in VR condition, here, the area-normalization enhanced the dwell-time difference between environments and in terms of an interaction, whereas no significant difference was observed for absolute dwell time (compare Fig. 5a and Supplementary Fig. S14a).

Comparison 2 on colour types: A significant difference between colour types was found, but no difference was observed between environments or in terms of an interaction: $F_{(1, 35.1)} = 42.701, p < .001$; $F_{(1, 30.5)} = 0.862, p > .05$; and $F_{(1, 52.4)} = 1.134, p > .05$, respectively. The area-normalized dwell time on chroma-containing areas (M_{Chroma} = 153.00 ±16.40s/%) was higher than luminance-only areas ($M_{Luminance}$ = 72.70 ±7.02s/%): $t_{(35.1)} = 6.54, p < .001$. Again, this shift of trends compared to absolute dwell time was a result of the relatively small area size of chroma-containing areas. Overall, in both environments, the density of visual attention was highest at red, blue, and yellow colours compared to black, grey and white.

Comparison 3 on individual colours: A significant difference between individual colours was found, but no difference was observed between environments or in terms of an interaction: $F_{(5, 53.4)} = 9.840$, $p < .001$; $F_{(1, 30.6)} = 1.300$, $p > .05$; and $F_{(5, 209.00)} = 1.250$, $p > .05$; respectively. After normalizing for the area sizes, overall, in both environments visitors attended most densely on red ($M_{Red} = 184.69 \pm 29.30$s/%), and least on grey ($M_{Grey} = 55.60 \pm 5.57$s/%).

Comparison 4 on cube surfaces: A significant difference between cube surfaces was found, no difference was observed between environments, and an effect was present in terms of an interaction: $F_{(5, 61.00)} = 8.894$, $p < .001$; $F_{(1, 30.8)} = 0.232$, $p > .05$; and $F_{(5, 234.2)} = 2.648$, $p = .024$, respectively. After normalizing for the area sizes, overall in both environments the density of visual attention was highest on east-wall ($M_{East\text{-}Wall} = 112.00 \pm 16.20$s/%), and lowest on the ceiling ($M_{Ceiling} = 27.50 \pm 4.26$s/%). The trend stayed the same compared to absolute dwell time, since the walls were roughly the same size within the cuboid rooms. When the interaction was broken down by focusing on the two types of environment to check whether environmental differences exist for any level of cube surfaces, all six post hoc comparisons yielded non-significant results (where all $p > .05$). The interaction was only pronounced, when the interaction was broken down by focusing on the cube surfaces to check how levels of cube surfaces have an effect differently for VR and physical environment: in this approach, some trend changes were visible, such as the normalized dwell time difference between ceiling$_{VR}$ and south-wall$_{VR}$ was significant ($p < .001$), but the normalized dwell time difference between ceiling$_{VR}$ and south-wall$_{Physical}$ was not significant ($p > .05$), suggesting that for some paired cube surfaces, the amount of dwell time difference was dependent on the environment.

Comparison 5 on furniture: A significant difference between types of furniture was found, but no difference was observed between environments or in terms of an interaction: $F_{(2, 31.09)} = 6.28$, $p = .005$; $F_{(1, 30.05)} = 3.95$, $p > .05$; and $F_{(2, 71.53)} = 0.26$, $p > .05$, respectively. In terms of furniture and after normalizing for the area sizes, the bed attracted highest attentional density ($M_{Bed} = 252.61 \pm 25.12$s/%) compared to the bookcase ($M_{Bookcase} = 155.51 \pm 17.75$s/%), and compared to the cupboard ($M_{Cupboard} = 175.02 \pm 29.52$s/%): $t_{(28.29)} = 3.43$, $p = .006$; and $t_{(28.48)} = 2.50$, $p = .043$, respectively. Note that since the surface areas for furniture were slightly different between VR and physical conditions such as in VR condition the cupboard had four additional AOIs and therefore had more surface area, here, the area-normalization diminished the dwell-time difference between environments ($p = .058$), whereas a significant difference ($p = .009$) had been observed for absolute dwell time (compare Fig. 5e and Supplementary Fig. S14e).

3.5. Discussion

The main research objective was to develop a methodology to assess the active exploration patterns of visual arts experience, and more specifically, to make a first step towards exploring the effects of an artwork's presentation medium as physical or virtual on this experience. We have focused on one example artwork, and tested a limited number of participants, and indisputably, future work should draw on more targeted and possibly larger samples and a wider spectrum of artwork. Nevertheless, we are dealing with a large data set consisting of approximately 25,000 fixations, each of which represents a single, albeit small and relatively unconscious decision about the artwork. As it stands, our case study also aimed to demonstrate that empirical approaches can contribute in a meaningful way to the understanding of art appreciation and its delivery through different media. To the best of our knowledge, this is the first case of a direct and quantitative experiment to compare real-world aesthetic experience side-by-side with its VR counterpart. A major empirical justification of this research can be linked to communicating historic and contemporary visual arts to a remote audience (Hoang & Cox, 2018; Parker & Saker, 2020; Puig et al., 2020; Checa & Bustillo, 2020), especially in the context of novel trends in presenting arts to remote audience in the wake of the COVID-19 pandemic.

Our main conclusion following the overall results was that when engaging with a spatial art installation derived from the Mondrian's design, participants showed predominantly similar viewing patterns on average in both physical and virtual environment, as indexed by gaze data from eye trackers. Our assessment on the similarity was the interpretation of the absolute and area-normalized dwell time analysis, which showed mostly non-significant main effects of environment and a lack of significant pairwise differences between the physical and VR versions for any significant interactions, except for absolute dwell time on some furniture elements (but also note that the furniture elements occupied only about ten percent of the surface area of the whole installation, and the most prominent design difference between physical and virtual installation was also present for furniture, in particular for the cupboard was a 3D piece in the virtual installation, but a 2D projection in the physical installation). In line with our expectations, our findings favour the null hypothesis, since no major difference was observed for the visual exploration patterns between in-situ and VR condition. It is important to briefly restate that, in general, the null results do not necessarily mean the lack of an effect or a difference, and they might be prone to over-interpretation: therefore, the findings can be described as preliminary evidence in need of further research and converging results.

Potential drivers for some particular gaze trends should be considered: (i) Irrespective of the viewing context, chroma-containing colours attracted higher visual attention densely compared to the luminance-only colours, as indexed by normalized dwell time. If the abstract nature of the installation and the minimum amount of semantic information available to the observer in this environment indicate that the participants' visual attention was mostly driven by the bottom-up factors, then we can argue for that even an elementary saliency map based on colour or contrast should have a strong effect on the difference on attended locations (see Supplementary Fig. S15 for exemplar saliency maps generated using Itti algorithm (Itti et al., 1998) and using histogram contrast). (ii) The most prominent semantic information available was the types of furniture. This was only true, if a participant was able to attribute objecthood status to those rather atypical furniture elements in the room. In this sense, object-based attention as a higher-level cognitive process and often studied along with scene perception and semantically-driven saliency maps associated with it can further help to explain some observed behaviours: for example, as a specific AOI set of furniture, the cupboard in the physical condition had the least amount of dwell time. Although it mainly consisted of yellow and black coloured patches, the cupboard was a flat 2D surface in the physical setting but not in the VR, which might reduce participants' ability to recognize the flat surface as a piece of furniture, and therefore potentially diminishes the object-based attentional guidance. (iii) In terms of six surfaces, although ceiling attracted the least amount of attentional density as indexed by normalized dwell time, no statistically significant difference was observed between four cardinal walls (N-E-S-W). This non-significant effect on the cardinal directions was also present in a similar, previous pilot study (Gulhan & Zanker, 2019), where we had utilised a mobile eye-tracker within another abstract installation consisting of coloured patches of parallelograms, covering all four walls of a gallery room. Additionally, a related observation from the present experiment, in both the physical and VR conditions, was that whilst participants were moving within the installation, the participants tended to not rotate themselves continuously, and did not form any number of full rotational circles in either clockwise or anti-clockwise directions. Put differently, the cumulative sum of a participant's rotation on the axial plane parallel to the floor was almost always $\pm 180°$ in the physical installation since they entered and exited the installation from the same door; and very often within the range of $\pm 180°$ in the VR. We speculate that this observed behaviour of self-restriction on rotation might have an equalising effect on the distribution of visual scans on cardinal directions, and therefore on the normalized dwell time corresponding to the cardinal directions. Although the raw data recorded from simple gyroscopes in eye-trackers without precise motion tracking are not suitable for comparison, general movement of participants, such as gait dynamics, might be prone to change depending on the exposure to the environment (Burtan et al., 2021). Here, as an anecdotal observation, participants naive to the VR tended to move more carefully or relatively slowly, compared to the physical world, and a major factor might be the lack of visual bodily cues in the VR (also see Supplementary Fig. S16 for exemplar motion trajectories in VR).

In terms of experimental validity, many (if not all) empirical research in vision science has to make an inevitable trade-off between internal and external validity: internal validity roughly refers to the strength of the link between research findings and design of the study, and it can be increased for example by minimizing confounding variables and presenting well-controlled stimuli. On the other hand, external validity is related to the generalisability of the findings beyond the selected artificial stimulus, testing environment, or group of participants in the research. As a related concept, ecological validity often refers to the generalisability of the findings to the real-world settings (Bourne, 2017). Here, we favoured ecological validity: although collecting gaze data using mobile and VR eye-trackers inside 1:1 scale physical and virtual versions of the artwork in a counter-balanced order from the same group of museum visitors aimed to preserve internal validity to some extent, our testing environment was far away from artificial laboratory conditions, where for example strict control of participant's viewing distance to a well-calibrated monitor accompanied with desktop-grade eye-tracker with higher sampling rate is often regarded as a procedural norm. On the other hand, art galleries and museums can be described as ecologically valid conditions where visitors' behaviour can be measured (Carbon, 2020), and these physical conditions are not well tested so far for VR.

Given the overwhelmingly similar pattern of eye movements in the two different environments, our results would suggest that using VR would be described as a suitable proxy for the aesthetic experience in gallery and museum settings. Describing eye-movements as an indicator and one of the few directly measurable components of aesthetic experience during artwork viewing is a common assumption behind many previous studies: often, researchers utilize eye-tracking as a meaningful tool to compare conditions or participant groups to answer their research questions, for example, (i) whether figure paintings and landscape paintings induce dissociable gaze patterns (Massaro et al., 2012), (ii) whether expert and non-expert participants in visual arts form different oculomotor measures (Francuz et al., 2018), or (iii) whether the overlap of museum visitors' viewing pathways on two paintings can be indexed and compared (Balbi et al., 2016). Additional measures can also be incorporated in studies and researchers can ask, for example, whether motion-capture alongside the eye-tracking during viewing a figurative sculpture by museum visitors who are trained dancers or non-dancers can be a feasible metric for aesthetic and kinesthetic experience (Wiseman et al., 2019). Here, we used fixation maps and derived metrics such as dwell time per AOI as one potential way of comparing physical and VR museum contexts, and our main justification behind this is that the conceptualization of fixation maps (Wooding, 2002b) allows us to quantify the similarity of eye movement traces (Wooding, 2002a). Note that we fully acknowledge that aesthetic experience as a highly complex process cannot be reduced to eye-movements, but nevertheless maintain that eye-movement metrics can be an essential measure to compare the interaction of viewers with an artwork.

However, the assumption of the ecological validity of VR still needs more rigorous test cases to become a generalizable argument. We compared some of the basic measures that may be used to relate to aesthetic experience in terms of attentional engagement. Apart from mostly comparable results on absolute and area-normalized dwell times, visitors spent relatively more time in the virtual environment compared to the physical environment. More specifically, the main eye-tracking results showed that in both conditions, (i) participants visually explored in all directions as all surfaces of the installation except for the ceiling, (ii) preferred coloured parts of the installation over the non-coloured parts as indexed by area-normalized dwell time, and (iii) often revisited the same location as indexed by fixation counts on a given AOI. Results from the exit-questionnaire indicated overall positive feedback from participants, and provided a comparison between physical and virtual artworks, where participants were generally split equally into two towards favouring either physical- or virtual-versions on various evaluations. Since the perception and judgment of art are highly subjective, individual differences both in terms of gaze patterns and questionnaire responses were inevitably present. Overall, our findings suggest that in the test case presented here, the virtual presentation of the artwork did not radically change the observers' visual exploration.

Recently, a comparative study between physical and virtual settings for an art gallery was investigated, with a focus on using EEG and ECG to classify emotion recognition and type of environment (Marín-Morales et al., 2019): relevant to our results, participants' self-assessment ratings on arousal and valence were part of their study, and almost no difference was found between physical and virtual contexts for eight art pieces, except for valence rating on a single art piece. In another study conducted to compare VR-museum and 2D computer monitor settings, no difference towards artworks' perceived quality and artistic quality was found, although the aesthetic experience of paintings was described as more intense in VR (Janković et al., 2019). Similarly, virtual environments can enhance memorability to some degree: for example, one study investigated active and passive view of spherical, 360° movie clips (such that whether the viewpoint of footage is dependent to head-orientation of the participant or not) involving Rubens and Nicolas paintings displayed via a head-mounted display (HMD), and their findings indicated that viewers' impression on paintings were described as more powerful and realistic in the active viewing condition (Hine & Tasaki, 2019). Another study compared the memory recall and recognition between 360° pictures displayed on HMD and on a tablet, and their results favoured the VR-display over tablet-display (Ventura et al., 2019). Apart from forms of enhancing effects of VR, the presentation medium of artwork seemed to induce minimal change on observers' experience.

Looking further afield than virtual art galleries, researchers have compared different examples of VR environments with their corresponding contexts to validate the feasibility of using VR as an empirical research tool: for example, comparing user experience in physical and virtual buildings in terms of architectural research showed that user ratings were mostly not affected between the two conditions, although some difference was present in atmospheric ratings such as boredom, attractiveness, and invitingness (Kuliga et al., 2015). In another study, the perceived spaciousness of a room in VR was investigated, replicating the main findings of its counterpart experiment in a physical room (Meagher & Marsh, 2015). Similarly, comparing participants' evaluations such as perceived pleasantness, interest, excitement, complexity, and satisfaction between physical and virtual interiors in terms of architectural and lighting design yielded no significant differences (Chamilothori et al., 2019). In a rather different research area, using measures of perceived presence, attitude towards a video game, memory recall and recognition of brand placement in a 2D, 3D, and VR version resulted in higher levels of presence in VR context, whereas attitude towards the video game and recognition of the brands was not changed (Roettl & Terlutter, 2018). Overall, the indication of VR as a valid context for behavioural research seems to be echoed by many researchers.

Eye-tracking and oculomotor data as a tool for aesthetic research, albeit useful, must be used with caution (Nayak & Karmakar, 2019). Correlation between preference and gaze data such as total dwell time and first fixation on one hand implied the feasibility of using eye-tracking metrics as an indicator of observers' aesthetic judgment (Holmes & Zanker, 2012), on the other hand, observers' ability to acquire the gist of a painting rather impressively in sub-second duration regime (Locher et al., 2008) might suggest a type of redundancy of gaze, and the prediction potential of fixation parameters towards aesthetic value has been also challenged (Isham & Geng, 2013). In our study, we described observers' eye-tracking data both as a measure of visual interest and as a similarity measure of aesthetic experience, assuming similar visual input to the observer leads to similar aesthetic experience. Linking oculomotor responses to aesthetic judgment more directly might require additional sources of data such as continuous aesthetic ratings (Isik & Vessel, 2019) or eye movement recording synchronized with event-related potentials (Fudali-Czyż et al., 2018).

Total viewing time is inherently linked to the fixation count and dwell time (i.e., total fixation duration), but not necessarily to the average fixation duration: whilst it is logical to think that the increase of the viewing time is often linearly translated into the increase of fixation count and dwell time; generally, no radical change is expected in terms of average fixation duration. Although there might be various (and potentially coexisting) factors behind, including the novelty of the VR experience, our finding of a significant shortening of average fixation duration in VR compared to physical installation might indicate that VR introduces a change of attitude towards aesthetic appreciation, since the intention to positively appreciate a set of paintings results as a greater number of fixations and lower average fixation duration, compared to intention to negative appreciation (Park et al., 2016). Alternatively, this difference found between the two conditions might be interpreted as an effect of authenticity: although in our case both conditions were reconstructions of the original artwork presented in two different media, the potential effects of originality (such that whether an artwork is original, copy, or fake) on observer rating and gaze behaviour have been noted previously (Locher et al., 2015), therefore it may be possible that visitors might have presupposed the VR condition a less authentic version of the artwork. Similarly, a potential arousal effect induced by the novelty remarked by the participants, might be a factor accounting for the observed difference, since outside the aesthetics research, changes in arousal states are often linked to the changes in various gaze metrics such as average fixation duration (Simola et al., 2015), pupil size (Bradshaw, 1967) or saccadic velocity (Di Stasi et al., 2013). Additionally, compared to the mobile eye-tracking, the VR eye-tracking is, in theory, more robust to the challenging conditions such as rapid head movements and change of environmental illumination: these might affect the fixation detection algorithms, and partially account for the average fixation duration differences.

Previous research also indicated that Mondrian's abstract painting entailed a high amount of visual search as indexed by, for example, the number of saccades compared to other paintings (Sharma & Chakravarthy, 2013). If we were to denote dwell time on AOI sets as an indicator of visual search, then our results suggest that physical and VR condition also resulted in mostly similar visual search strategies during the visual exploration. Speculatively, particular differentiating trends between two settings in general viewing such as average fixation duration (or albeit nonsignificant, total dwell time), might be linked to the current state of the VR. VR was perceived as a novel experience by the participants during the experiment, and this novelty might be linked to, for example, spending more time in the virtual installation. In time, the resemblance between physical and virtual galleries is only expected to get higher, and along with potentially diminished novelty effects, more comparative general results are expected in future studies.

Although promising results and valuable insights were acquired, comparing physical and virtual art spaces is still in its early stages, and our research was not aiming to provide fully comprehensive answers and explanations. Conducting a comparative experiment using two parallel, equally valid reconstructions models in a museum setting can be seen as a unique opportunity, but our findings on the similarity of gaze patterns for only one single, very specific example of an abstract art installation, with a particular population sample, does not justify bolder conclusions and generalizations about the validity of VR-context, especially without further behavioural measurements. First and foremost, most of our participants are regular art gallery and museum visitors, but many are not familiar with VR. Therefore, the extent and amount of some visitors' mental state of surprise especially during VR condition, or their awareness of wearing the mobile eye-tracker or the VR headset, and the possible influence of these aspects on exploration patterns remains unclear. A training phase for both wearable devices in future experiments might reduce novelty effects and the remaining discrepancy between conditions to some extent. It is clear that there is an enormous potential for more comprehensive work, both in a variety of methods and in the scope of arts presented. For example, to increase the inter-stimulus consistency, a rigorous photogrammetric workflow consisting of 3D imaging laser scanner in conjunction with readings from colorimeter measurement can be utilized to be the base of the virtual counterpart of any given static installation, preferably followed by the colour calibration processes of a VR-HMD, which would also require additional psychophysical testing. The methodological workflow might also include comparing gaze patterns with body motions indexed by gyroscopes during the experiment (Bonnet et al., 2019); or alternatively, a change of experimental design might allow for precise control for motion and viewing duration, at a cost of reduced freedom (see Supplementary Fig. S17). The concept of peripersonal space (Bufacchi & Iannetti, 2018) might also help to develop a more comprehensive theoretical perspective. In future research, it would be useful to compare a complete exhibition between physical and virtual environments, instead of comparing just a single artwork. For the physical condition, a complete exhibition as a set of selected artworks in a dedicated gallery space might be provided. For the virtual environment condition, a well-controlled exact digital replica of the physical exhibition might be created, and ideally, use of an untethered HMD with inside-out position tracking might allow visitors to walk within the virtual exhibition without any constraints or without relying on alternative ways of VR locomotion such as teleportation. Additionally, an augmented reality (AR) version of the same exhibition would allow a ternary comparison between physical, VR, and AR conditions. Interacting with artworks as stimuli might allow for asking more fine-tuned research questions, related to memorability (Damiano & Walther, 2019; Krokos et al., 2019) or effects of haptic feedback and visual cues to depth information (Harris et al., 2019). 3D saliency maps as extensions of 360° saliency maps (John et al., 2019) might be investigated to describe the extent of bottom-up influence of the environment on gaze behaviour. In terms of further data analysis, investigation of temporal dynamics (Wu et al., 2014; Marlow et al., 2015; Brielmann et al., 2017) might provide more in-depth results, using tools such as temporal scan path analysis, or adapting methods from graph theory and related fields (see Supplementary Fig. S18 as a primer in such directions). As we step inside the world of virtual museums and gallery spaces, current directions of VR in terms of artistic expression, digital heritage, and empirical research remains wide open. Despite the need for more comprehensive future studies, our research can be seen as an important and promising starting point for comparing aesthetic experience between virtual and physical environments.

Acknowledgements

The authors would like to thank Tim Holmes and the technology company Tobii for developing an initial prototype of VR eye-tracking implementation, Jasmina Stevanov for her contribution to the initial 3D reconstruction of the Mondrian room models, the Albertinum Museum and Dresden State Art Collections for providing generous hosting and advisors including staff support, Sabine Zanker for logistics and participant management, Royal Holloway University of London (RHUL) for the scholarship of Doga Gulhan, and the Department of Psychology (RHUL) for additional funding specifically allocated for this research.

Ethical statement

All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki. Research Ethics Committee (REC) Project ID: 1396, Year: 2018, User ID: PDJT007.

Funding statement

We received no funding for this study.

Data availability

All anonymised data are accessible via Open Science Framework for anyone who would like to re-analyse the data or run any form of additional analyses: osf.io/bgtpy

Competing interests

The authors declare no competing interests.

Author contributions

Conceptualization, DG, JMZ; Methodology, DG, JMZ; Software, DG, JMZ; Formal Analysis, DG, SD; Investigation, DG, SD, JMZ; Writing – original draft, DG; Writing – review & editing, DG, SD, JMZ; Visualization, DG, JMZ; Supervision, SD, JMZ; Project administration, JMZ.

Timeline

The experiment was conducted in April 2019. Please also refer to the last paragraph of the introduction (p. 31), 1st thesis chapter, for a detailed overall timeline and the effects of the pandemic on the experiments.

3.6. References

Balbi, B., Protti, F., & Montanari, R. (2016). Driven by Caravaggio through his painting. In C. Sennersten & Y. Khaluf (Eds.), *COGNITIVE 2016: The 8th International Conference on Advanced Cognitive Technologies and Applications* (pp. 72–76).

Bertamini, M., & Blakemore, C. (2019). Seeing a work of art indirectly: When a reproduction is better than an indirect view, and a mirror better than a live monitor. *Frontiers in Psychology*, *10*, 2033. https://doi.org/10.3389/fpsyg.2019.02033

Bonnet, C. T., Davin, T., Hoang, J.-Y., & Baudry, S. (2019). Relations between eye movement, postural sway and cognitive involvement in unprecise and precise visual tasks. *Neuroscience*, *416*, 177–189. https://doi.org/10.1016/j.neuroscience.2019.07.031

Bourne, V. (2017). *Starting out in methods and statistics for psychology: A hands-on guide to doing research* (1st ed.). Oxford University Press.

Bradshaw, J. (1967). Pupil size as a measure of arousal during information processing. *Nature*, *216*(5114), 515–516. https://doi.org/10.1038/216515a0

Brieber, D., Leder, H., & Nadal, M. (2015). The experience of art in museums: An attempt to dissociate the role of physical context and genuineness. *Empirical Studies of the Arts*, *33*(1), 95–105. https://doi.org/10.1177/0276237415570000

Brielmann, A. A., Vale, L., & Pelli, D. G. (2017). Beauty at a glance: The feeling of beauty and the amplitude of pleasure are independent of stimulus duration. *Journal of Vision*, *17*(14:9), 1–12. https://doi.org/10.1167/17.14.9

Bufacchi, R. J., & Iannetti, G. D. (2018). An action field theory of peripersonal space. *Trends in Cognitive Sciences*, *22*(12), 1076–1090. https://doi.org/10.1016/j.tics.2018.09.004

Bullot, N. J., & Reber, R. (2013). The artful mind meets art history: Toward a psycho-historical framework for the science of art appreciation. *Behavioral and Brain Sciences*, *36*(2), 123–137. https://doi.org/10.1017/S0140525X12000489

Burtan, D., Joyce, K., Burn, J. F., Handy, T. C., Ho, S., & Leonards, U. (2021). The nature effect in motion: Visual exposure to environmental scenes impacts cognitive load and human gait kinematics. *Royal Society Open Science*, *8*(1), 201100. https://doi.org/10.1098/rsos.201100

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology and perception in art*. University of Chicago Press.

Carbon, C.-C. (2017). Art perception in the museum: How we spend time and space in art exhibitions. *I-Perception*, *8*(1), 1–15. https://doi.org/10.1177/2041669517694184

Carbon, C.-C. (2020). Ecological art experience: How we can gain experimental control while preserving ecologically valid settings and contexts. *Frontiers in Psychology*, *11:800*, 1–14. https://doi.org/10.3389/fpsyg.2020.00800

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. https://doi.org/10.1016/j.visres.2011.04.012

Carroll, N. (2002). Aesthetic experience revisited. *The British Journal of Aesthetics*, *42*(2), 145–168. https://doi.org/10.1093/bjaesthetics/42.2.145

Chamilothori, K., Wienold, J., & Andersen, M. (2019). Adequacy of immersive virtual reality for the perception of daylit spaces: Comparison of real and virtual environments. *LEUKOS*, *15*(2–3), 203–226. https://doi.org/10.1080/15502724.2017.1404918

Chatterjee, A., & Vartanian, O. (2014). Neuroaesthetics. *Trends in Cognitive Sciences*, *18*(7), 370–375. https://doi.org/10.1016/j.tics.2014.03.003

Checa, D., & Bustillo, A. (2020). Advantages and limits of virtual reality in learning processes: Briviesca in the fifteenth century. *Virtual Reality*, *24*(1), 151–161. https://doi.org/10.1007/s10055-019-00389-7

Clay, V., König, P., & König, S. U. (2019). Eye tracking in virtual reality. *Journal of Eye Movement Research*, *12*(1:3), 1–18. https://doi.org/10.16910/JEMR.12.1.3

Dalbajewa, B., Wagner, H., Biedermann, H., Dehmer, A., & Wagner, M. (Eds.). (2019). *Visionary Spaces: Kandinsky, Mondrian, Lissitzky and the Abstract-Constructivist Avant-Garde in Dresden 1919-1932*. Sandsteing Verlag.

Damiano, C., & Walther, D. B. (2019). Distinct roles of eye movements during memory encoding and retrieval. *Cognition*, *184*, 119–129. https://doi.org/10.1016/j.cognition.2018.12.014

Di Stasi, L. L., Catena, A., Cañas, J. J., Macknik, S. L., & Martinez-Conde, S. (2013). Saccadic velocity as an arousal index in naturalistic tasks. *Neuroscience & Biobehavioral Reviews*, *37*(5), 968–975. https://doi.org/10.1016/j.neubiorev.2013.03.011

Duchowski, A. (2017). *Eye tracking methodology: Theory and practice* (3rd ed.). Springer International Publishing.

Duchowski, A., Medlin, E., Cournia, N., Murphy, H., Gramopadhye, A., Nair, S., Vorah, J., & Melloy, B. (2002). 3-D eye movement analysis. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 573–591. https://doi.org/10.3758/BF03195486

Fechner, G. T. (1876). *Vorschule der Aesthetik*. Breitkopf & Härtel Verlag.

Ferretti, G., & Marchi, F. (2020). Visual attention in pictorial perception. *Synthese*. https://doi.org/10.1007/s11229-020-02873-z

Fisher, S. S., McGreevy, M., Humphries, J., & Robinett, W. (1987). Virtual environment display system. *Proceedings of the 1986 Workshop on Interactive 3D Graphics*, 77–87. https://doi.org/10.1145/319120.319127

Francuz, P., Zaniewski, I., Augustynowicz, P., Kopiś, N., & Jankowski, T. (2018). Eye movement correlates of expertise in visual arts. *Frontiers in Human Neuroscience*, *12:87*, 1–13. https://doi.org/10.3389/fnhum.2018.00087

Fudali-Czyż, A., Francuz, P., & Augustynowicz, P. (2018). The effect of art expertise on eye fixation-related potentials during aesthetic judgment task in focal and ambient modes. *Frontiers in Psychology*, *9:1972*, 1–9. https://doi.org/10.3389/fpsyg.2018.01972

Geisler, W. S., & Cormack, L. K. (2011). *Models of overt attention*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199539789.013.0024

Goldstein, E. B., & Brockmole, J. R. (2017). *Sensation and perception* (10th ed.). Cengage Learning.

Gregory, R. L. (1995). Black boxes of artful vision. In R. L. Gregory, P. F. Heard, J. P. Harris, & D. Rose (Eds.), *The artful eye* (pp. 5–27). Oxford University Press.

Gulhan, D., & Zanker, J. (2019). Exploring artwork in situ: Empirical aesthetics making use of mobile eye tracking. *Perception*, *48(S1)*, 23–24. https://doi.org/10.1177/0301006618824879

Guntarik, O., Garcia, J. E., Howard, S. R., & Dyer, A. G. (2018). Traces: Mobile eye tracking captures user sensory experience in an outdoor walking tour environment. *Leonardo*, *51*(2), 163–164. https://doi.org/10.1162/LEON_a_01571

Harris, D. J., Buckingham, G., Wilson, M. R., & Vine, S. J. (2019). Virtually the same? How impaired sensory information in virtual reality may disrupt vision for action. *Experimental Brain Research*, *237*, 2761–2766. https://doi.org/10.1007/s00221-019-05642-8

Heidenreich, S. M., & Turano, K. A. (2011). Where does one look when viewing artwork in a museum? *Empirical Studies of the Arts*, *29*(1), 51–72. https://doi.org/10.2190/EM.29.1.d

Heilig, M. L. (1962). *Sensorama simulator* (United States Patent No. US3050870A).

Helmholtz, H. von. (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Friedrich Vieweg & Sohn Verlag.

Hine, K., & Tasaki, H. (2019). Active view and passive view in virtual reality have different impacts on memory and impression. *Frontiers in Psychology*, *10:2416*, 1–8. https://doi.org/10.3389/fpsyg.2019.02416

Hoang, T. N., & Cox, T. N. (2018). Alternating reality: An interweaving narrative of physical and virtual cultural exhibitions. *Presence: Teleoperators and Virtual Environments*, *26*(4), 402–419. https://doi.org/10.1162/PRES_a_00307

Holmes, T., & Zanker, J. M. (2012). Using an oculomotor signature as an indicator of aesthetic preference. *I-Perception*, *3*(7), 426–439. https://doi.org/10.1068/i0448aap

Iseminger, G. (2005). Aesthetic experience. In J. Levinson (Ed.), *The Oxfod Handbook of Aesthetics* (Vol. 1, pp. 99–116). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199279456.003.0005

Isham, E. A., & Geng, J. J. (2013). Looking time predicts choice but not aesthetic value. *PLOS ONE*, *8*(8), e71698. https://doi.org/10.1371/journal.pone.0071698

Isik, A. I., & Vessel, E. A. (2019). Continuous ratings of movie watching reveal idiosyncratic dynamics of aesthetic enjoyment. *PLOS ONE*, *14*(10), e0223896. https://doi.org/10.1371/journal.pone.0223896

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Janković, D., Jevremović, V., & Carbon, C.-C. (2019). Visual art in the digital age: About the art experience in VR vs ordinary displayed museum contexts. *7th Visual Science of Art Conference*. 7th Visual Science of Art Conference, Leuven, Belgium.

John, B., Raiturkar, P., Le Meur, O., & Jain, E. (2019). A benchmark of four methods for generating 360 degree saliency maps from eye tracking data. *International Journal of Semantic Computing*, *13*(03), 329–341. https://doi.org/10.1142/S1793351X19400142

Johnson, M. G., Muday, J. A., & Schirillo, J. A. (2010). When viewing variations in paintings by Mondrian, aesthetic preferences correlate with pupil size. *Psychology of Aesthetics, Creativity, and the Arts*, *4*(3), 161–167. https://doi.org/10.1037/a0018155

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, & issues* (9th ed.). Cengage Learning.

Klein, C., Betz, J., Hirschbuehl, M., Fuchs, C., Schmiedtová, B., Engelbrecht, M., Mueller-Paul, J., & Rosenberg, R. (2014). Describing art – An interdisciplinary approach to the effects of speaking on gaze movements during the beholding of paintings. *PLOS ONE*, *9*(12), e102439. https://doi.org/10.1371/journal.pone.0102439

Kourtesis, P., Collina, S., Doumas, L. A. A., & MacPherson, S. E. (2019). Validation of the virtual reality neuroscience questionnaire: Maximum duration of immersive virtual reality sessions without the presence of pertinent adverse symptomatology. *Frontiers in Human Neuroscience*, *13*, 417. https://doi.org/10.3389/fnhum.2019.00417

Krokos, E., Plaisant, C., & Varshney, A. (2019). Virtual memory palaces: Immersion aids recall. *Virtual Reality*, *23*(1), 1–15. https://doi.org/10.1007/s10055-018-0346-3

Kuliga, S. F., Thrash, T., Dalton, R. C., & Hölscher, C. (2015). Virtual reality as an empirical research tool: Exploring user experience in a real building and a corresponding virtual model. *Computers, Environment and Urban Systems*, *54*, 363–375. https://doi.org/10.1016/j.compenvurbsys.2015.09.006

Leder, H., & Nadal, M. (2014). Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode - Developments and challenges in empirical aesthetics. *British Journal of Psychology*, *105*(4), 443–464. https://doi.org/10.1111/bjop.12084

Locher, P. (2013). Empirical investigation of the elements of composition in paintings: A painting as stimulus. In P. P. L. Tinio & J. K. Smith (Eds.), *The Cambridge Handbook of the Psychology of Aesthetics and the Arts* (pp. 221–242). Cambridge University Press. https://doi.org/10.1017/CBO9781139207058.013

Locher, P., Krupinski, E. A., Mello-Thoms, C., & Nodine, C. F. (2008). Visual interest in pictorial art during an aesthetic experience. *Spatial Vision*, *21*(1), 55–77. https://doi.org/10.1163/156856808782713762

Locher, P., Krupinski, E., & Schaefer, A. (2015). Art and authenticity: Behavioral and eye-movement analyses. *Psychology of Aesthetics, Creativity, and the Arts*, *9*(4), 356–367. https://doi.org/10.1037/aca0000026

Machotka, P. (1995). Aesthetics: If not from below, whence? *Empirical Studies of the Arts*, *13*(2), 105–118. https://doi.org/10.2190/ADEW-NRGC-85CF-8Q9D

Madary, M., & Metzinger, T. K. (2016). Real virtuality: A code of ethical conduct. Recommendations for good scientific practice and the consumers of VR-technology. *Frontiers in Robotics and AI*, *3*(3), 1–23. https://doi.org/10.3389/frobt.2016.00003

Makin, A. (2017). The gap between aesthetic science and aesthetic experience. *Journal of Consciousness Studies*, *24*(1–2), 184–213.

Marín-Morales, J., Higuera-Trujillo, J. L., Greco, A., Guixeres, J., Llinares, C., Gentili, C., Scilingo, E. P., Alcañiz, M., & Valenza, G. (2019). Real vs. immersive-virtual emotional experience: Analysis of psycho-physiological patterns in a free exploration of an art museum. *PLOS ONE*, *14*(10), e0223881. https://doi.org/10.1371/journal.pone.0223881

Marlow, C. A., Viskontas, I. V., Matlin, A., Boydston, C., Boxer, A., & Taylor, R. P. (2015). Temporal structure of human gaze dynamics is invariant during free viewing. *PLOS ONE*, *10*(9), e0139379. https://doi.org/10.1371/journal.pone.0139379

Massaro, D., Savazzi, F., Di Dio, C., Freedberg, D., Gallese, V., Gilli, G., & Marchetti, A. (2012). When art moves the eyes: A behavioral and eye-tracking study. *PLOS ONE*, *7*(5), e37285. https://doi.org/10.1371/journal.pone.0037285

McManus, I. C., Cheema, B., & Stoker, J. (1993). The aesthetics of composition: A study of Mondrian. *Empirical Studies of the Arts*, *11*(2), 83–94. https://doi.org/10.2190/HXR4-VU9A-P5D9-BPQQ

Meagher, B. R., & Marsh, K. L. (2015). Testing an ecological account of spaciousness in real and virtual environments. *Environment and Behavior*, *47*(7), 782–815. https://doi.org/10.1177/0013916514525039

Menninghaus, W., Wagner, V., Hanich, J., Wassiliwizky, E., Kuehnast, M., & Jacobsen, T. (2015). Towards a psychological construct of being moved. *PLOS ONE*, *10*(6), 1–33. https://doi.org/10.1371/journal.pone.0128451

Miller, M. R., Herrera, F., Jun, H., Landay, J. A., & Bailenson, J. N. (2020). Personal identifiability of user tracking data during observation of 360-degree VR video. *Scientific Reports*, *10*(1), 17404. https://doi.org/10.1038/s41598-020-74486-y

Nadal, M., & Chatterjee, A. (2019). Neuroaesthetics and art's diversity and universality. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(3), e1487. https://doi.org/10.1002/wcs.1487

Nanay, B. (2010). Attention and perceptual content. *Analysis*, *70*(2), 263–270. https://doi.org/10.1093/analys/anp165

Nanay, B. (2015). Aesthetic attention. *Journal of Consciousness Studies*, *22*(5–6), 96–118.

Nayak, B. K., & Karmakar, S. (2019). Eye tracking based objective evaluation of visual aesthetics: A review. In F. Rebelo & M. M. Soares (Eds.), *Advances in Ergonomics in Design* (Vol. 777, pp. 370–381). Springer International Publishing. https://doi.org/10.1007/978-3-319-94706-8_41

Noll, A. M. (1966). Human or machine: A subjective comparison of Piet Mondrian's "Composition with Lines" (1917) and a computer-generated picture. *The Psychological Record*, *16*(1), 1–10. https://doi.org/10.1007/BF03393635

Noll, A. M., & Hill, M. (1965). Computer-generated three-dimensional movies. *Computers and Automation*, *14*(11), 20–23.

Pangilinan, E., Lukas, S., & Mohan, V. (2019). *Creating augmented and virtual realities: Theory and practice for next-generation spatial computing*. O'Reilly Media.

Park, H., Lee, S., Lee, M., Chang, M.-S., & Kwak, H.-W. (2016). Using eye movement data to infer human behavioral intentions. *Computers in Human Behavior*, *63*, 796–804. https://doi.org/10.1016/j.chb.2016.06.016

Parker, E., & Saker, M. (2020). Art museums and the incorporation of virtual reality: Examining the impact of VR on spatial and social norms. *Convergence: The International Journal of Research into New Media Technologies*, *26*(5–6), 1–15. https://doi.org/10.1177/1354856519897251

Pelowski, M., Leder, H., Mitschke, V., Specker, E., Gerger, G., Tinio, P. P. L., Vaporova, E., Bieg, T., & Husslein-Arco, A. (2018). Capturing aesthetic experiences with installation art: An empirical assessment of emotion, evaluations, and mobile eye tracking in Olafur Eliasson's "Baroque, Baroque!" *Frontiers in Psychology*, *9:1255*, 1–22. https://doi.org/10.3389/fpsyg.2018.01255

Plumhoff, J. E., & Schirillo, J. A. (2009). Mondrian, eye movements, and the oblique effect. *Perception*, *38*(5), 719–731. https://doi.org/10.1068/p6160

Puig, A., Rodríguez, I., Arcos, J. Ll., Rodríguez-Aguilar, J. A., Cebrián, S., Bogdanovych, A., Morera, N., Palomo, A., & Piqué, R. (2020). Lessons learned from supplementing archaeological museum exhibitions with virtual reality. *Virtual Reality*, *24*(2), 343–358. https://doi.org/10.1007/s10055-019-00391-z

Rauss, K., & Pourtois, G. (2013). What is bottom-up and what is top-down in predictive coding? *Frontiers in Psychology*, *4:276*, 1–8. https://doi.org/10.3389/fpsyg.2013.00276

Roettl, J., & Terlutter, R. (2018). The same video game in 2D, 3D or virtual reality – How does technology impact game evaluation and brand placements? *PLOS ONE*, *13*(7), e0200724. https://doi.org/10.1371/journal.pone.0200724

Rothacher, Y., Nguyen, A., Lenggenhager, B., Kunz, A., & Brugger, P. (2018). Visual capture of gait during redirected walking. *Scientific Reports*, *8*(1), 17974. https://doi.org/10.1038/s41598-018-36035-6

Scarfe, P., & Glennerster, A. (2015). Using high-fidelity virtual reality to study perception in freely moving observers. *Journal of Vision*, *15*(9:3), 1–11. https://doi.org/10.1167/15.9.3

Schindler, I., Hosoya, G., Menninghaus, W., Beermann, U., Wagner, V., Eid, M., & Scherer, K. R. (2017). Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PLOS ONE*, *12*(6), e0178899. https://doi.org/10.1371/journal.pone.0178899

Sharma, S., & Chakravarthy, B. K. (2013). How people view abstract art: An eye movement study to assess information processing and viewing strategy. In A. Chakrabarti & R. V. Prakash (Eds.), *ICoRD'13* (pp. 477–487). Springer. https://doi.org/10.1007/978-81-322-1050-4_38

Simola, J., Le Fevre, K., Torniainen, J., & Baccino, T. (2015). Affective processing in natural scene viewing: Valence and arousal interactions in eye-fixation-related potentials. *NeuroImage*, *106*, 21–33. https://doi.org/10.1016/j.neuroimage.2014.11.030

Smith, L. F., Smith, J. K., & Tinio, P. P. L. (2017). Time spent viewing art and reading labels. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(1), 77–85. https://doi.org/10.1037/aca0000049

Specker, E., Forster, M., Brinkmann, H., Boddy, J., Pelowski, M., Rosenberg, R., & Leder, H. (2020). The Vienna Art Interest and Art Knowledge Questionnaire (VAIAK): A unified and validated measure of art interest and art knowledge. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(2), 172–185. https://doi.org/10.1037/aca0000205

Steinicke, F. (2016). *Being really virtual*. Springer. https://doi.org/10.1007/978-3-319-43078-2

Stevanov, J., Zanker, J., & Holmes, T. (2019). Mobile eye tracking in the Royal Academy of Arts: Analysing scanpath sequences in Jackson Pollock's paintings. *Perception*, *48(S1)*, 69. https://doi.org/10.1177/0301006618824879

Stevanov, J., & Zanker, J. M. (2020). Exploring Mondrian compositions in three-dimensional space. *Leonardo*, *53*(1), 63–69. https://doi.org/10.1162/leon_a_01583

Sutherland, I. E. (1968). A head-mounted three-dimensional display. *AFIPS Conference Proceedings*, 757–764.

Swami, V., & Furnham, A. (2012). The effects of symmetry and personality on aesthetic preferences. *Imagination, Cognition and Personality*, *32*(1), 41–57. https://doi.org/10.2190/IC.32.1.d

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta Psychologica*, *135*(2), 77–99. https://doi.org/10.1016/j.actpsy.2010.02.006

Tricart, C. (2017). *Virtual reality filmmaking: Techniques & best practices for VR filmmakers* (1st ed.). Routledge.

Troy, N. J. (1980). Mondrian's designs for the Salon de Madame B..., à Dresden. *The Art Bulletin*, *62*(4), 640–647. https://doi.org/10.2307/3050061

Tymkiw, M., & Foulsham, T. (2019). Eye tracking, spatial biases, and normative spectatorship in museums. *Leonardo*, 1–10. https://doi.org/10.1162/leon_a_01746

Ventura, S., Brivio, E., Riva, G., & Baños, R. M. (2019). Immersive versus non-immersive experience: Exploring the feasibility of memory assessment through 360 degree technology. *Frontiers in Psychology*, *10:2509*, 1–7. https://doi.org/10.3389/fpsyg.2019.02509

Vessel, E. A., & Rubin, N. (2010). Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of Vision*, *10*(2:18), 1–14. https://doi.org/10.1167/10.2.18

Wagemans, J. (2011). Towards a new kind of experimental psycho-aesthetics? Reflections on the Parallellepipeda project. *I-Perception*, *2*(6), 648–678. https://doi.org/10.1068/i0464aap

Walker, F., Bucker, B., Anderson, N. C., Schreij, D., & Theeuwes, J. (2017). Looking at paintings in the Vincent Van Gogh Museum: Eye movement patterns of children and adults. *PLOS ONE*, *12*(6), e0178912. https://doi.org/10.1371/journal.pone.0178912

Wilson, C. J., & Soranzo, A. (2015). The use of virtual reality in psychology: A case study in visual perception. *Computational and Mathematical Methods in Medicine*, *2015*, 1–7. https://doi.org/10.1155/2015/151702

Wiseman, B., Carusi, A., Briggs, E., Poyntz, S., Pelowski, M., Alcock, L., & Mazzà, C. (2019). Embodied viewing and Degas's Little Dancer Aged Fourteen: A multi-disciplinary experiment in eye-tracking and motion capture. *The Senses and Society*, *14*(3), 284–296. https://doi.org/10.1080/17458927.2019.1652413

Wooding, D. S. (2002a). Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 518–528. https://doi.org/10.3758/BF03195481

Wooding, D. S. (2002b). Fixation maps: Quantifying eye-movement traces. *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '02*, 31. https://doi.org/10.1145/507072.507078

Wu, D. W.-L., Anderson, N. C., Bischof, W. F., & Kingstone, A. (2014). Temporal dynamics of eye movements are related to differences in scene complexity and clutter. *Journal of Vision*, *14*(9:8), 1–14. https://doi.org/10.1167/14.9.8

Yarbus, A. L. (1967). *Eye movements and vision*. Springer. https://doi.org/10.1007/978-1-4899-5379-7

Zanker, J., Stevanov, J., & Holmes, T. (2017). Mobile eye tracking in the Royal Academy: Analysing the interaction with abstract paintings. *40th European Conference on Visual Perception (ECVP 2017)*, 212. http://journals.sagepub.com/page/pec/collections/ecvp-abstracts/index/ecvp-2017

3.7. Figures

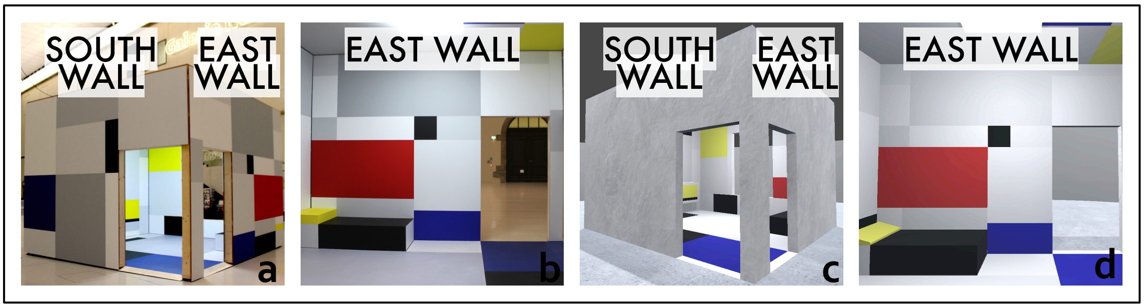


Figure 1. Physical and virtual versions of Mondrian's room design. (a) The exterior and (b) the interior photographs of the physical installation created by artist Heimo Zobernig. Similarly, (c) the exterior and (d) the interior views of VR reconstruction, developed by our research team. In the physical installation, one of the artistic decisions of Heimo Zobernig was to extrude the interior patterns onto exterior surfaces of the room, whereas our VR reconstruction had a homogeneous grey texture for the exterior surfaces. Also note that since counterbalancing of the conditions was implemented to minimize the temporal order effects in repeated measures design, the participants were randomly assigned to view either (a-b) the physical installation first, or (c-d) the VR reconstruction first.

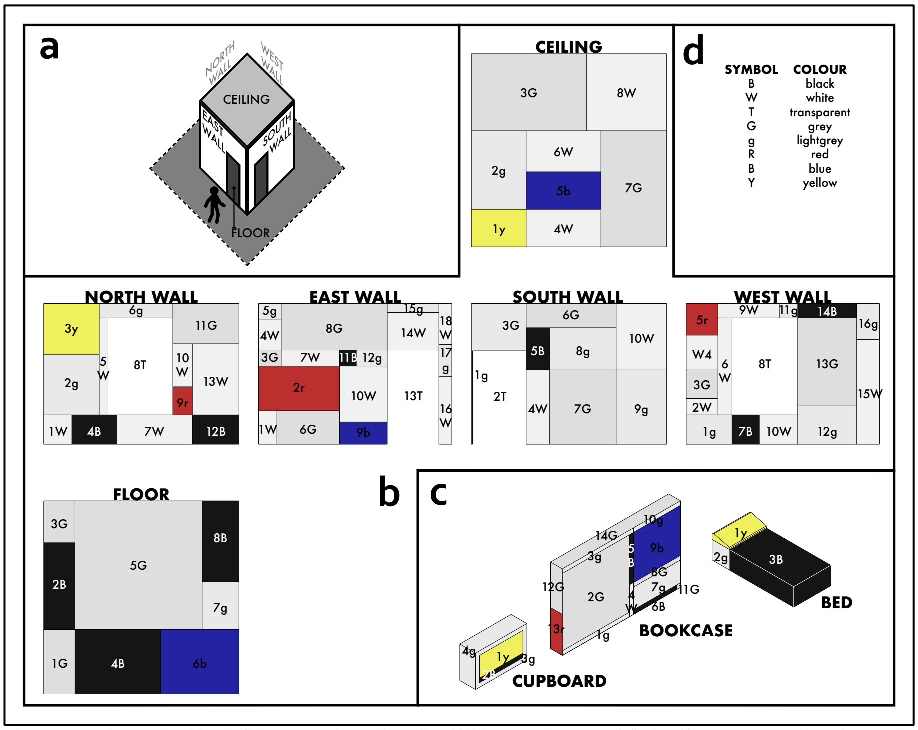


Figure 2. An overview of 3D AOI mapping for the VR condition. (a) A diagrammatic view of the cuboid room, in which (b) six surfaces of the room and (c) three pieces of furniture were present. Each individual coloured 2D panel was coded as an individual AOI during the development of the VR environment, and (d) had a unique colour value out of six possible colours. For example, the dwell time on red colour patches was calculated as the cumulative sum of fixation duration on four AOIs, namely North-9r, East-2r, West-5r, and Bookcase-13r. Along the same lines, the 3D AOI mapping was also formed for the physical condition, regarding the same sets of AOIs.

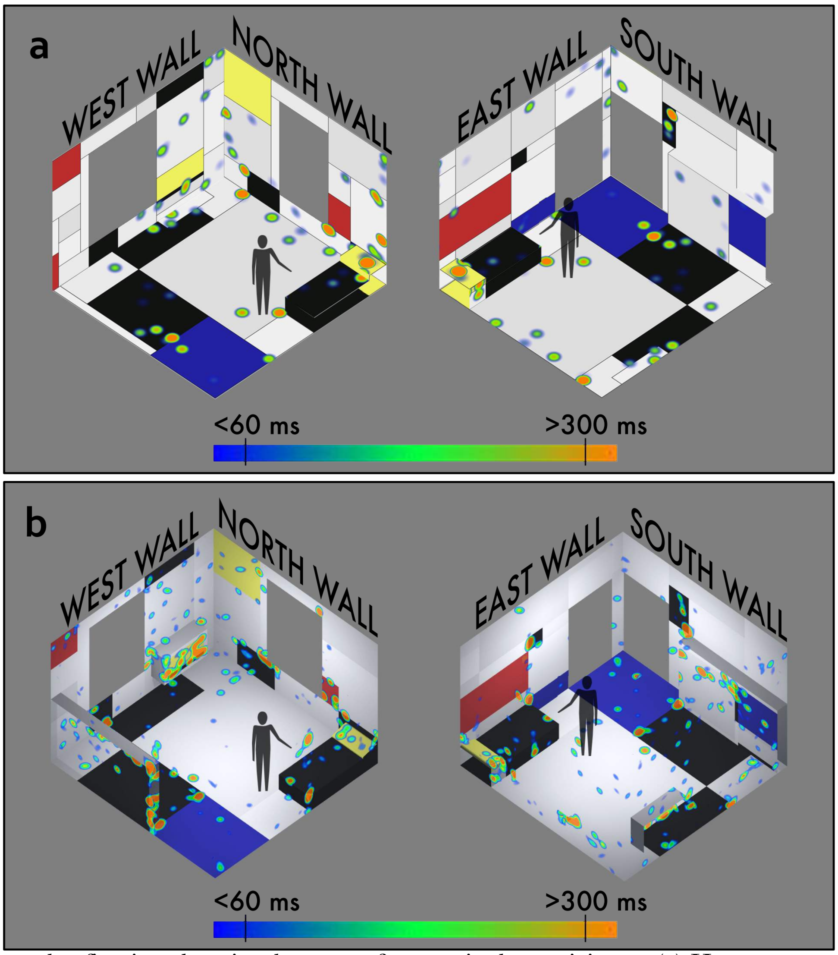


Figure 3. Exemplar fixation duration heatmap from a single participant. (a) Heatmap corresponding to the physical condition was formed after 89 seconds of interaction inside the physical installation, and (b) the heatmap corresponding to the VR condition was formed after 162 seconds of interaction inside the virtual environment.

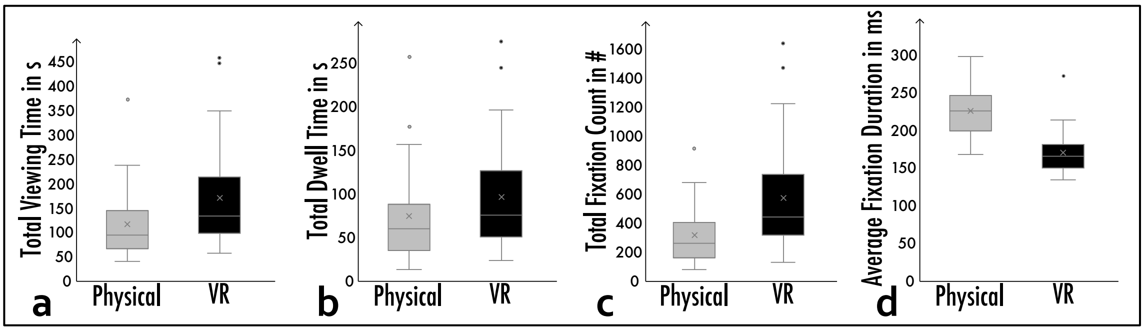


Figure 4. Main descriptive statistics comparing physical and VR environments: (a) total viewing time in seconds, (b) total dwell time in seconds, (c) total fixation count, and (d) average fixation duration in milliseconds illustrated as box-plots. Each box was drawn from first quartile (Q1) to third quartile (Q3) with a horizontal line denoting the median, and a cross denoting mean. Whiskers indicate minimum and maximum except outliers. Outliers were visualised as points ± 1.5 interquartile range. On average, participants spent about two minutes in physical installation and three minutes in virtual installation. Since viewing duration correlates with fixation duration and fixation count, the same trend of difference was observed both for dwell time and fixation count. However, the average fixation duration showed an opposite trend, where longer individual fixations were observed in physical installation compared to VR. The sample sizes were $N_{Physical} = 24$, $N_{VR} = 30$.

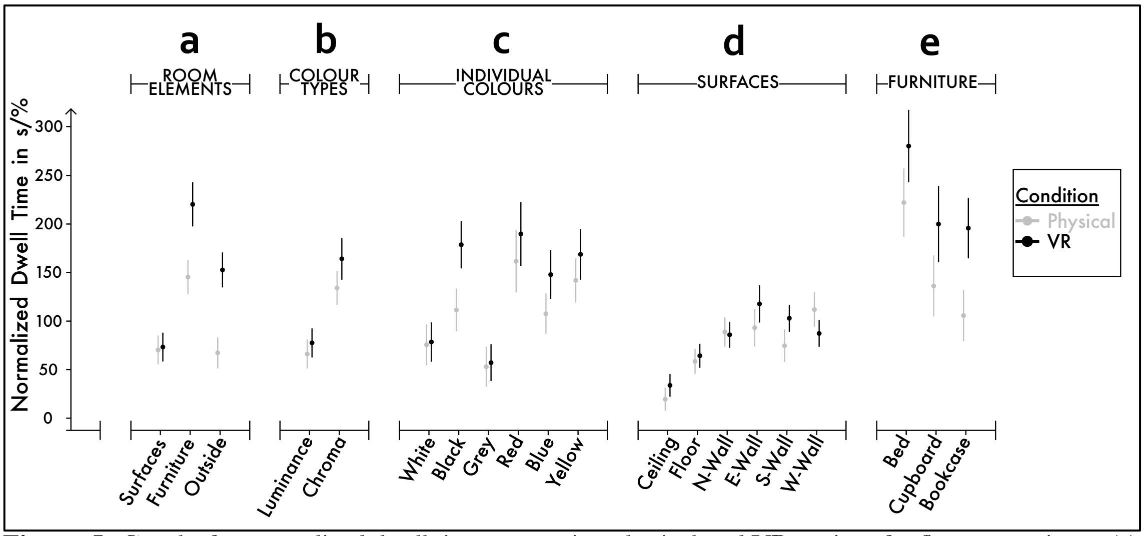


Figure 5. Graphs for normalized dwell time comparing physical and VR settings for five comparisons. (a) Room elements, (b) colour types, (c) individual colours, (d) surfaces, and (e) furniture. The x-axis shows the levels of AOIs, and the y-axis shows absolute dwell time in seconds. Physical and VR conditions were colour-coded as grey and black, respectively. The visualized data based on means, with whiskers indicating 95% of confidence intervals (CIs). The sample sizes were $N_{Physical} = 24$, $N_{VR} = 30$.

3.8. Supplementary figures

SCREENING FORM

DEMOGRAPHICS

Participant ID	Age	Sex	Education (Highest degree)	Occupation

VISION STATUS

Questions	Yes	No
Are you wearing corrections glasses or contact lenses?	o	o
Are you near-sighted (corrected for distant objects) or far-sighted (for reading)?	o	o
In a gallery, would you wear glasses or contacts?	o	o
Do you know your prescription?	o	o
Are you depending on bi-focal or multi-focal contact lenses?	o	o
Are you wearing corrections for ASTIGMATISM?	o	o
Did you have corrective laser surgery in the last 12 to 24 months (LASIK)?	o	o
Are you colour blind?	o	o
Do you know about any other visual disorder?	o	o
Do you know about any other neurological conditions that effect vision?	o	o

PRIOR KNOWLEDGE IN ART

Art expertise	Not at all	Slightly	Moderately	Very	Extremely
To what extend are you interested in modern art?	o	o	o	o	o
To what extend are you interested in Bauhaus?	o	o	o	o	o
Are you familiar with works of Piet Mondrian?	o	o	o	o	o
Are you familiar with works of Heimo Zobernig?	o	o	o	o	o

ADDITIONAL INFO

Do you think that is there any points to take into consideration prior to experiment?

EXIT QUESTIONNAIRE

VIEWS IN ART

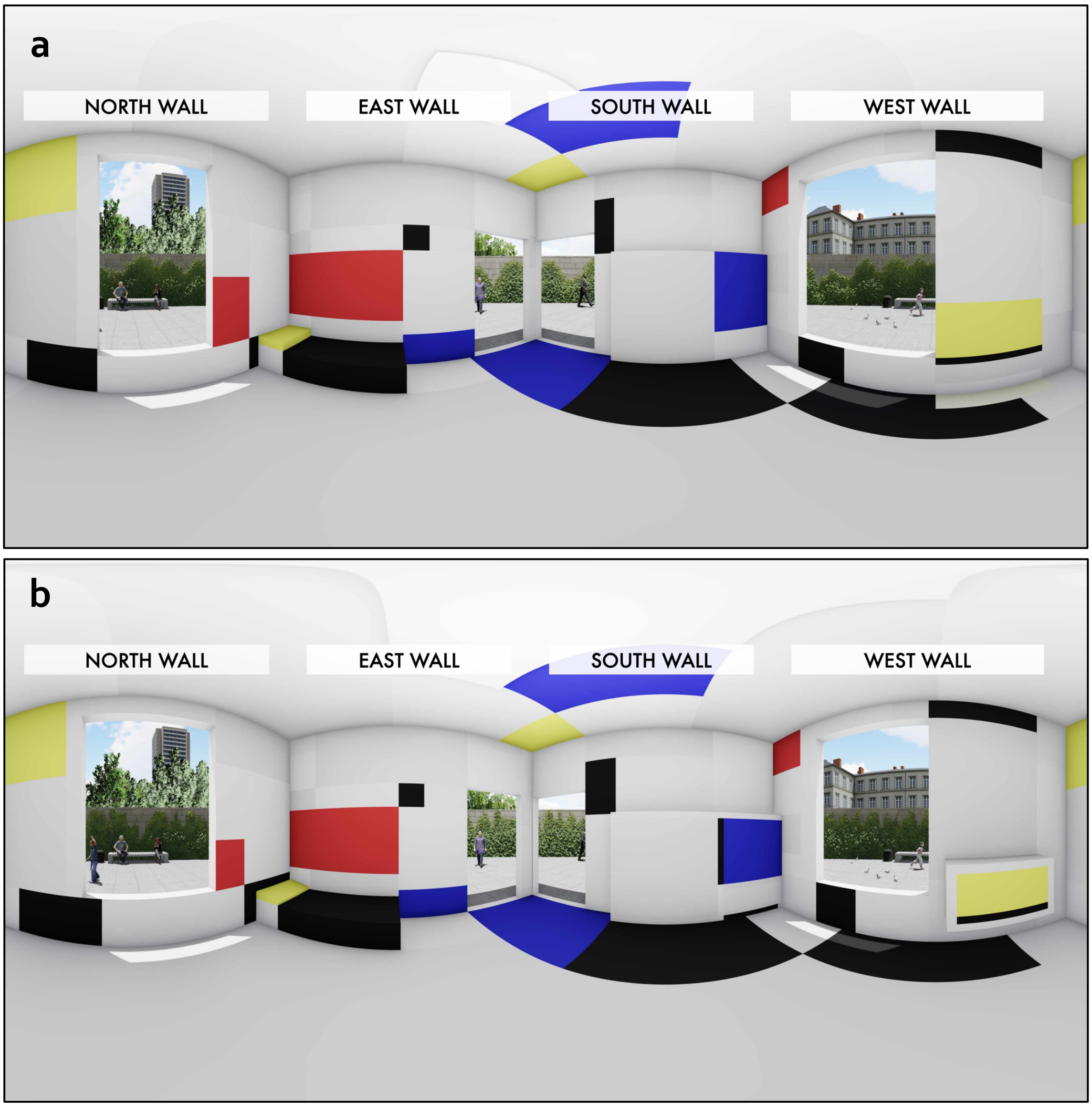
Opinions on art	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly agree
Viewing experience in installation was enjoyable.	o	o	o	o	o
Viewing experience in VR was boring.	o	o	o	o	o
The VR environment felt like a real room.	o	o	o	o	o
Art museums and galleries are losing their significance.	o	o	o	o	o
All public art objects/spaces should be digitized and available online.	o	o	o	o	o
The VR reconstruction was more exciting than the real installation.	o	o	o	o	o

Questions on art exposure	Never	Seldom	Sometimes	Often	Always
How often do you visit art museums, galleries, or events?	o	o	o	o	o
How often do you view art digitally?	o	o	o	o	o
How often do you use art-related sources (such as books, journals or websites)?	o	o	o	o	o
How often do you pursue an artistic activity or a hobby (e.g. painting, photography, workshops)?	o	o	o	o	o
How often do you play video games?	o	o	o	o	o
How often do you use VR?	o	o	o	o	o
How long would you comfortably view arts in a gallery or museum?	≤15 mins o	~30 mins o	~1 hour o	~2 hours o	≥4 hours o
How long would you comfortably view arts in a VR environment?	≤15 mins o	~30 mins o	~1 hour o	~2 hours o	≥4 hours o

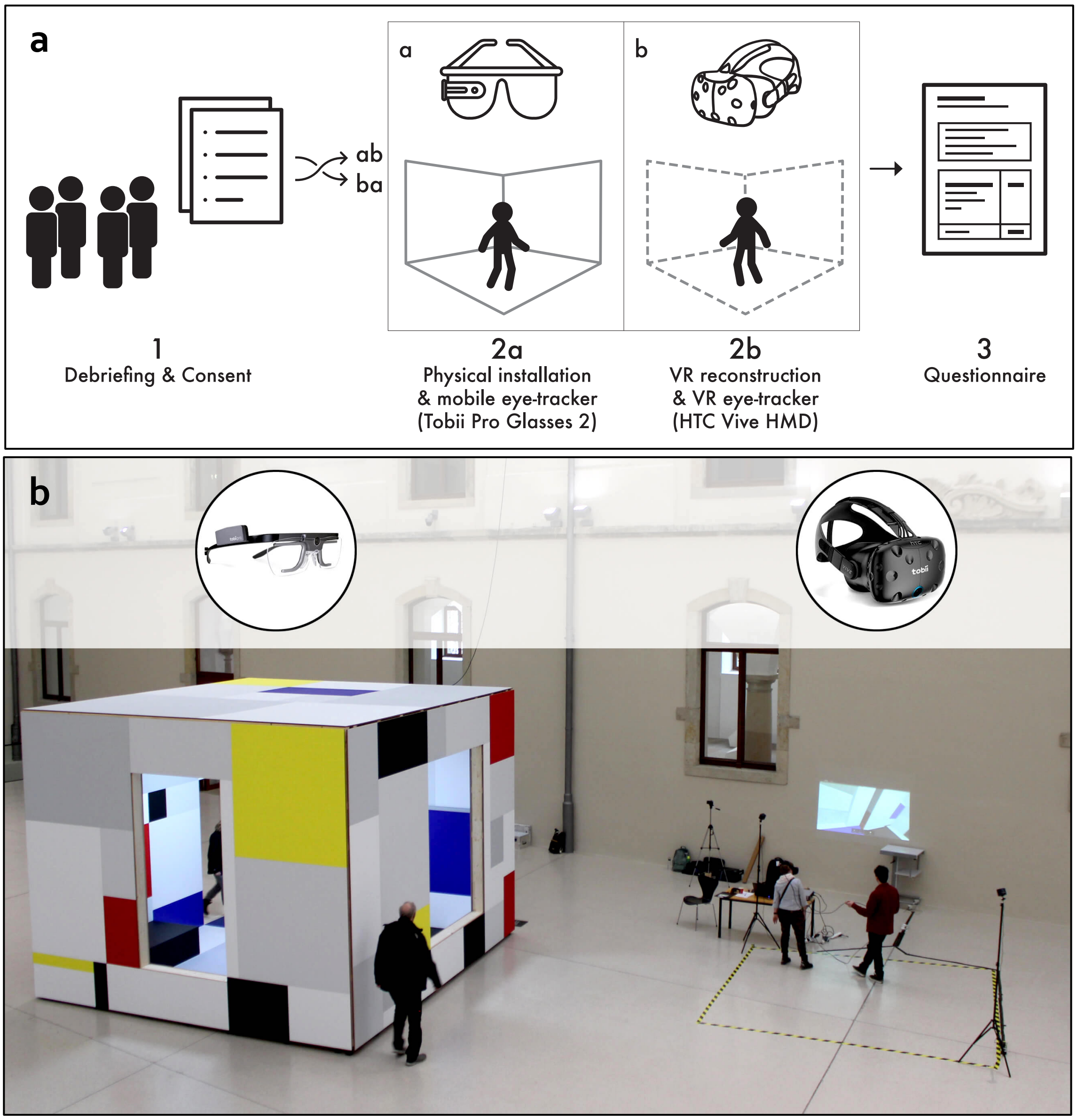
FEEDBACK

What did you like about the experiment?
Was there anything you did not like?
Can you please provide 3 to 6 words which best describe your experience in the VR environment?
Do you have any additional comments (e.g., a couple of keywords)?

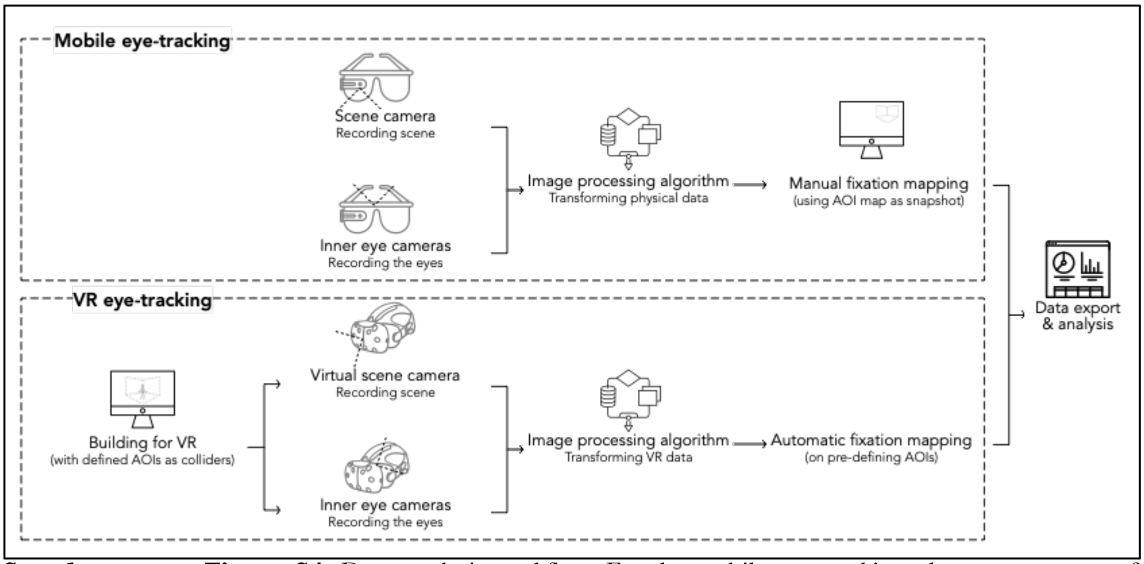
Supplementary Figure S1. The screening form along with the exit questionnaire (English version), presented here in a compact, alternative layout compared to the layout used during the experiment.



Supplementary Figure S2. Panoramic rendered images of (a) the physical and (b) the VR versions of Mondrian's room design. Note that, the exterior scene visible from the window and door openings are added to these renders to illustrate the scale of the installation. During the experiment, the visible exterior scene was the exhibition space of the Albertinum Museum in the physical installation, and a static grey scene in the VR.

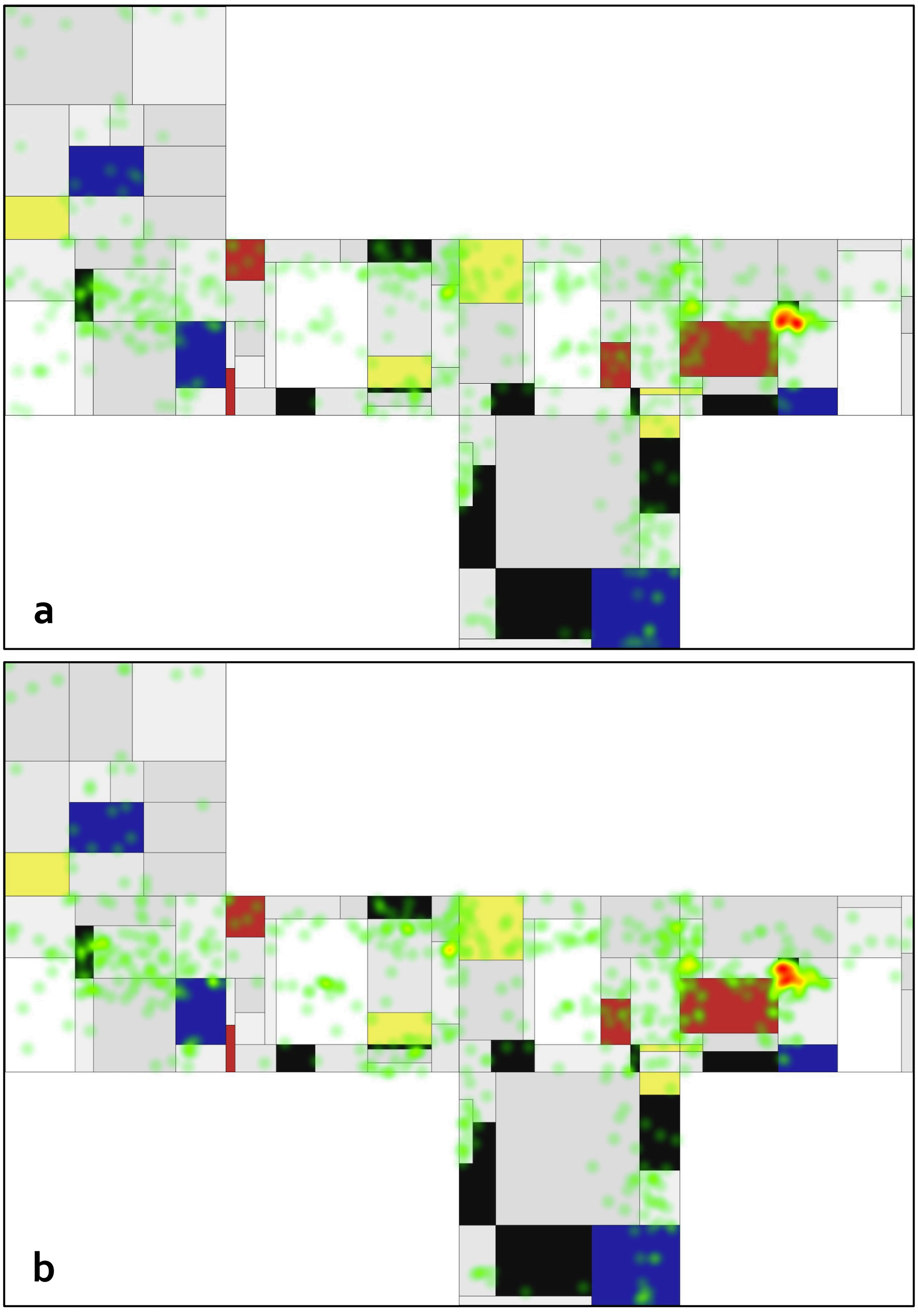


Supplementary Figure S3. (a) Overview of the experimental procedure. Following the debriefing and receiving consent, participants engaged with physical installation and digital reconstruction one by one, in a counter-balanced order. A wearable eye-tracker (Tobii Pro Glasses 2) wirelessly connected to a computer (Microsoft Surface) was used to collect data from the physical installation, whilst a VR eye-tracker (HTC Vive HMD with embedded eye-tracker) wired to another computer (Dell Alienware 15 R4) was used for the digital reconstruction. Note, the weight of the mobile eye-tracking glasses used for the physical installation was approximately 45 g, and the weight of the VR headset was approximately 470 g. Lastly, participants were asked to complete an exit questionnaire. (b) View from the gallery space. The physical installation can be seen on the left and the VR counterpart can be seen on the right, both situated side-by-side in the gallery. On top, the thumbnail images are showing the hardware used during the experiment: the mobile eye tracker (Tobii Pro Glasses 2) on the top-left, and the VR eye tracker (HTC Vive) on the top-right. Note that, the safe walkable area in the VR was limited to approximately 3.5 by 3.5 metres, regarding the optimum tracking area of the hardware. This inability of infinite walking in VR was mentioned to participants both before the experiment and during the experiment whenever needed, for example, when a participant has intended to exit from the doors in VR, or tried to get very close to the virtual walls.

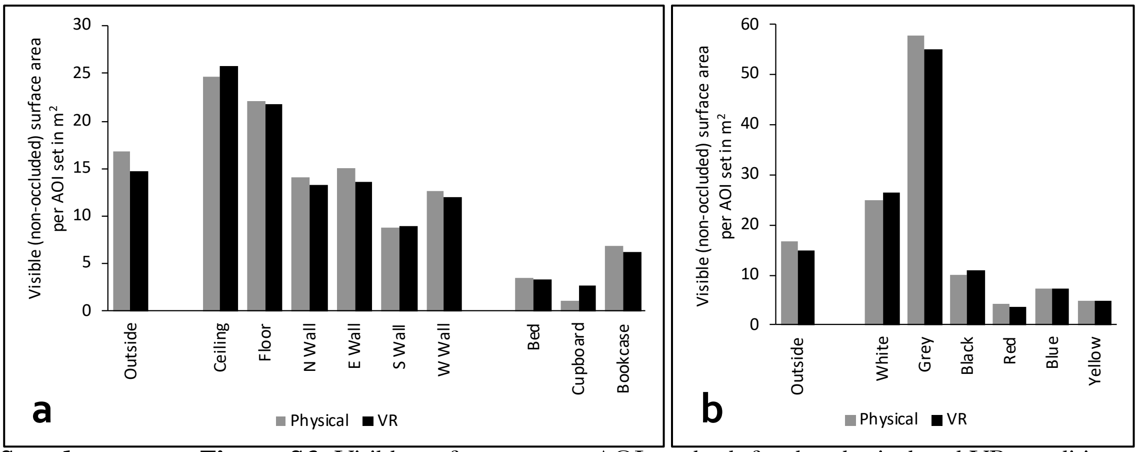


Supplementary Figure S4. Data analysis workflow. For the mobile-eye tracking, the scene camera of the eye tracker captured complete video footage from the participant's point-of-view, whilst eye-tracking sensors with infra-red illuminators in the glasses recorded eye orientation, providing the direction of eye gaze. Tobii Pro Lab software matched the retinal coordinates of gaze to individual frames of the scene video. The velocity-threshold identification (I-VT) filter (Salvucci & Goldberg, 2000) was applied to classify gaze positions as fixations or other events, with a dispersion threshold of 1.0°, a minimum fixation duration of 60 ms, allowing for a gap of a maximum of 75 ms between fixations (namely, maintaining gaze on a single location is only defined as a fixation if the event lasted at least for 60 ms, and two or more consecutive fixations were combined into a single fixation if they were temporally separated by less than 75 ms), as suggested by (Komogortsev et al., 2010), and by Tobii User's Manual for Glasses Pro 2). To localize the fixation positions, firstly we mapped manually the 3D interior of the installation to a 2D representation: this process can be defined as a backward texture mapping or an inverse UV mapping, since the 3D model's three-dimensional XYZ coordinates were projected back onto a 2D image as two-dimensional UV coordinates. Note that, U and V refer to the two axes of the 2D image; and X, Y, and Z refer to the three axes of the 3D model; and this notation simply aims to distinguish labelling of axes between 2D image and 3D model. As a single image file, this unwrapped version of the installation consisting of six faces of the installation stitched together was used as a reference image, which is often referred to as a snapshot in Tobii Pro Lab software. On top of the reference image, we added a second level of information called the AOI map (as illustrated in Fig. 2): the AOI map was a secondary image layer in which individual rectangular pieces of the installation were redrawn and tagged with unique identifiers. Each individually defined AOI had a set of features associated with it, such as colour and location, and whether it is a part of a specific wall, floor, ceiling, or furniture. Thus, we were able to manually code the fixation positions from the video footage to a single reference image. At this point, the real-world data output for each fixation had the duration, location, and corresponding AOI information. This interim manual coding was approximately 1/30X in real-time, e.g., for a 1-minute recording, the mapping took about half an hour. Being a time-intensive task, the coding was carried out by the lead author, but the reliability of coding (similar to the intercoder reliability) was initially checked with the additional independent coding by the last senior author for a single recording (see Supplementary Fig. S5 for a comparison): given the overwhelming locational overlap between two mappings, this interim mapping was assumed to be minimally error-prone, and a complete set of secondary mapping was not sought after.

The workflow of eye-tracking data analysis from the VR eye tracker was similar, and comparatively simpler because there was already a direct allocation of gaze points to surface regions in the VR model, without a need for interim manual coding. During the design of the VR environment, individual surfaces in the 3D space of the digital installation were tagged in a similar way as when generating an AOI map for the physical setting. The oculomotor data were recorded by eye-tracking sensors with infra-red illuminators embedded within the VR headset, and the same fixation filter algorithm was applied, based on display screen coordinates in the headset. Localizing the fixation positions was an automatic process, since the positions of individual AOIs were already predefined (as illustrated in Fig. 2), and the fixation positions were converted to 3D XYZ coordinates inside the virtual environment. Therefore, the VR data output of fixations already contained the duration, location, and corresponding AOI information, without the need for an intermediary manual mapping procedure.



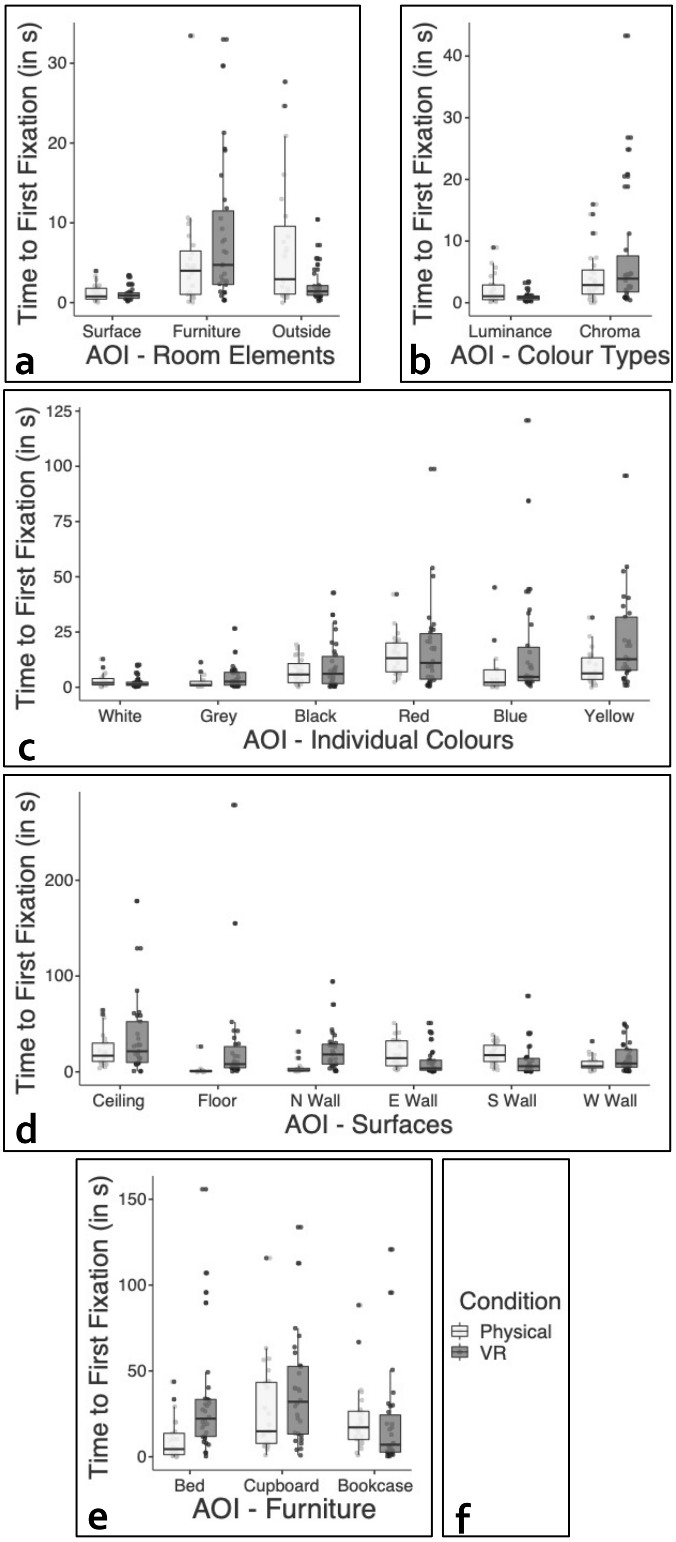
Supplementary Figure S5. Two independently coded fixation heatmaps for a single participant's data, (a) by the first author with a kernel sizes 25 pixels for the visualization and (b) by the last author with a kernel size of 30 pixels for the visualization, to visually inspect the reliability of the interim manual coding that was only required for the mobile eye-tracking data. Given the overwhelming locational overlap of localizing the fixations on particular AOIs, and given the time-consuming nature of this interim coding, all the remaining coding was carried out by the lead author without a secondary coder.



Supplementary Figure S6. Visible surface area per AOI set, both for the physical and VR conditions. The AOIs can be clustered together either (a) as a collection of outside (as window and door openings), six surfaces, and furniture elements; or (b) as a collection of outside and individual colours. In both instances, the total surface area of the physical installation was 125.727 m^2, and the VR counterpart was 122.648 m^2. Note that these minor discrepancies between two conditions have created a slight difference in terms of surface area, but the main analysis was based on area-normalized dwell time, which also aimed to normalize these minor discrepancies.

MEAN (± STANDARD ERROR OF MEAN)		AOI CONDITIONS																			
		ROOM ELEMENTS			COLOUR TYPES		INDIVIDUAL COLOURS						SURFACES						FURNITURE		
ENVIRONMENTS		Surfaces	Furniture	Outside	Luminance	Chroma	White	Grey	Black	Red	Blue	Yellow	Ceiling	Floor	N Wall	E Wall	S Wall	W Wall	Bed	Cupboard	Bookcase
PHYSICAL	**Dwell Time in s (Absolute)**	53.82 (±8.83)	13.09 (±1.66)	9.11 (±1.61)	49.00 (±8.05)	17.92 (±2.45)	15.45 (±2.90)	24.57 (±4.13)	8.98 (±1.34)	5.95 (±1.14)	6.19 (±0.72)	5.78 (±0.76)	3.82 (±0.63)	9.95 (±1.71)	10.37 (±1.67)	12.51 (±2.51)	5.50 (±1.26)	11.67 (±2.28)	6.01 (±0.75)	1.28 (±0.24)	5.80 (±0.86)
	Dwell Time in s/% (Area-normalized)	69.48 (±11.40)	142.90 (±18.11)	68.10 (±12.03)	66.50 (±10.92)	138.39 (±18.91)	77.97 (±14.62)	53.51 (±9.00)	112.88 (±16.82)	178.09 (±34.01)	108.62 (±12.63)	147.83 (±19.39)	19.52 (±3.19)	56.48 (±9.69)	92.56 (±14.95)	104.25 (±20.87)	78.31 (±17.94)	116.38 (±22.74)	218.12 (±27.22)	143.84 (±26.59)	105.20 (±15.57)
	Fixation Count in # (Absolute)	230.21 (±32.38)	51.33 (±5.72)	40.54 (±6.15)	209.67 (±29.58)	71.88 (±8.55)	66.92 (±10.28)	107.08 (±15.65)	35.67 (±4.89)	22.29 (±3.73)	27.00 (±2.63)	22.58 (±2.84)	20.29 (±3.00)	40.50 (±6.32)	44.54 (±6.24)	53.71 (±9.16)	23.25 (±4.57)	47.92 (±7.54)	21.50 (±2.19)	4.92 (±0.82)	24.92 (±3.31)
	Fixation Count in #/% (Area-normalized)	297.17 (±41.80)	560.36 (±62.46)	303.18 (±45.99)	284.56 (±40.14)	555.19 (±66.01)	337.79 (±51.87)	233.22 (±34.09)	448.29 (±61.42)	667.30 (±111.68)	474.14 (±46.16)	577.45 (±72.75)	103.57 (±15.31)	229.77 (±35.85)	397.73 (±55.76)	447.41 (±76.31)	331.22 (±65.12)	477.94 (±75.22)	780.69 (±79.44)	551.93 (±92.38)	451.72 (±59.94)
	Fixation Duration in ms (Absolute)	221.92 (±7.4)	247.79 (±8.7)	214.25 (±8.1)	221.83 (±7.2)	240.90 (±9.2)	213.99 (±7.0)	217.80 (±7.5)	247.26 (±12.7)	242.85 (±13.2)	221.43 (±9.6)	252.86 (±12.8)	183.51 (±11.0)	240.70 (±11.8)	222.80 (±8.5)	213.83 (±10.0)	222.25 (±10.7)	222.55 (±10.1)	274.42 (±13.1)	252.61 (±25.0)	222.53 (±9.2)
	Fixation Duration in ms/% (Area-normalized)	286.47 (±9.53)	2704.95 (±95.19)	1602.23 (±60.40)	301.06 (±9.82)	1860.84 (±71.10)	1080.18 (±35.39)	474.34 (±16.23)	3107.79 (±159.59)	7269.69 (±394.55)	3888.58 (±167.99)	6465.54 (±326.39)	936.65 (±56.11)	1365.60 (±66.97)	1989.46 (±76.17)	1781.31 (±83.24)	3166.21 (±152.82)	2219.82 (±100.81)	9964.47 (±474.75)	28356.93 (±2802.76)	4034.29 (±167.27)
	Area in %	77.47%	9.16%	13.37%	73.68%	12.95%	19.81%	45.92%	7.96%	3.34%	5.69%	3.91%	19.59%	17.63%	11.20%	12.00%	7.02%	10.03%	2.75%	0.89%	5.52%
VR	**Dwell Time in s (Absolute)**	57.06 (±7.80)	22.23 (±2.73)	18.39 (±2.29)	58.44 (±6.91)	20.85 (±3.22)	16.89 (±2.15)	25.70 (±3.19)	15.85 (±2.18)	5.65 (±1.36)	8.61 (±1.60)	6.59 (±1.21)	7.13 (±1.49)	11.45 (±1.99)	9.31 (±1.35)	13.07 (±2.69)	7.56 (±1.01)	8.55 (±1.22)	7.82 (±1.10)	4.43 (±1.08)	9.98 (±1.40)
	Dwell Time in s/% (Area-normalized)	73.30 (±10.02)	220.00 (±27.02)	152.60 (±19.05)	77.66 (±9.19)	164.23 (±25.39)	78.60 (±9.99)	57.25 (±7.10)	178.78 (±24.63)	189.96 (±45.68)	148.00 (±27.55)	168.84 (±31.01)	33.94 (±7.08)	64.39 (±11.18)	85.96 (±12.47)	117.66 (±24.25)	102.92 (±13.78)	87.32 (±12.48)	280.21 (±39.33)	199.97 (±48.68)	195.76 (±27.52)
	Fixation Count in # (Absolute)	348.90 (±47.52)	121.20 (±14.52)	109.03 (±12.74)	351.03 (±42.48)	119.07 (±18.34)	104.97 (±13.24)	158.13 (±20.11)	87.93 (±12.06)	33.37 (±7.45)	50.40 (±9.36)	35.30 (±6.10)	45.30 (±8.97)	67.23 (±12.05)	56.40 (±8.11)	82.73 (±15.74)	45.67 (±6.01)	51.57 (±7.07)	40.73 (±5.69)	23.77 (±5.25)	56.70 (±7.41)
	Fixation Count in #/% (Area-normalized)	448.19 (±61.04)	1199.44 (±143.67)	904.94 (±105.75)	466.45 (±56.44)	937.85 (±144.44)	488.35 (±61.59)	352.21 (±44.80)	992.04 (±136.06)	1122.43 (±250.58)	865.97 (±160.87)	904.46 (±156.28)	215.81 (±42.74)	378.11 (±67.78)	520.85 (±74.87)	744.68 (±141.64)	621.71 (±81.83)	526.70 (±72.18)	1459.89 (±203.77)	1071.64 (±236.89)	1112.46 (±145.31)
	Fixation Duration in s (Absolute)	165.34 (±4.10)	182.44 (±6.44)	169.46 (±7.08)	168.45 (±4.66)	174.67 (±5.79)	161.84 (±4.23)	163.60 (±4.93)	181.24 (±6.36)	163.95 (±5.36)	172.03 (±6.58)	179.66 (±7.91)	153.64 (±4.20)	166.37 (±5.64)	165.31 (±4.94)	153.42 (±5.71)	170.77 (±9.26)	164.63 (±5.56)	192.18 (±7.89)	171.18 (±8.91)	172.22 (±5.72)
	Fixation Duration in s/% (Area-normalized)	212.39 (±5.27)	1805.52 (±63.69)	1406.50 (±58.77)	223.84 (±6.20)	1375.83 (±45.60)	752.95 (±19.68)	364.38 (±10.99)	2044.67 (±71.70)	5515.19 (±180.46)	2955.86 (±113.11)	4449.77 (±248.73)	707.55 (±31.13)	935.63 (±31.74)	1475.72 (±67.32)	1380.89 (±51.37)	2324.91 (±126.11)	1681.53 (±56.79)	6887.60 (±282.72)	7204.16 (±517.73)	3379.01 (±112.25)
	Area in %	77.85%	10.10%	12.05%	75.26%	12.70%	21.49%	44.90%	8.86%	2.97%	5.82%	3.90%	20.99%	17.78%	10.83%	11.11%	7.35%	9.79%	2.79%	2.22%	5.10%

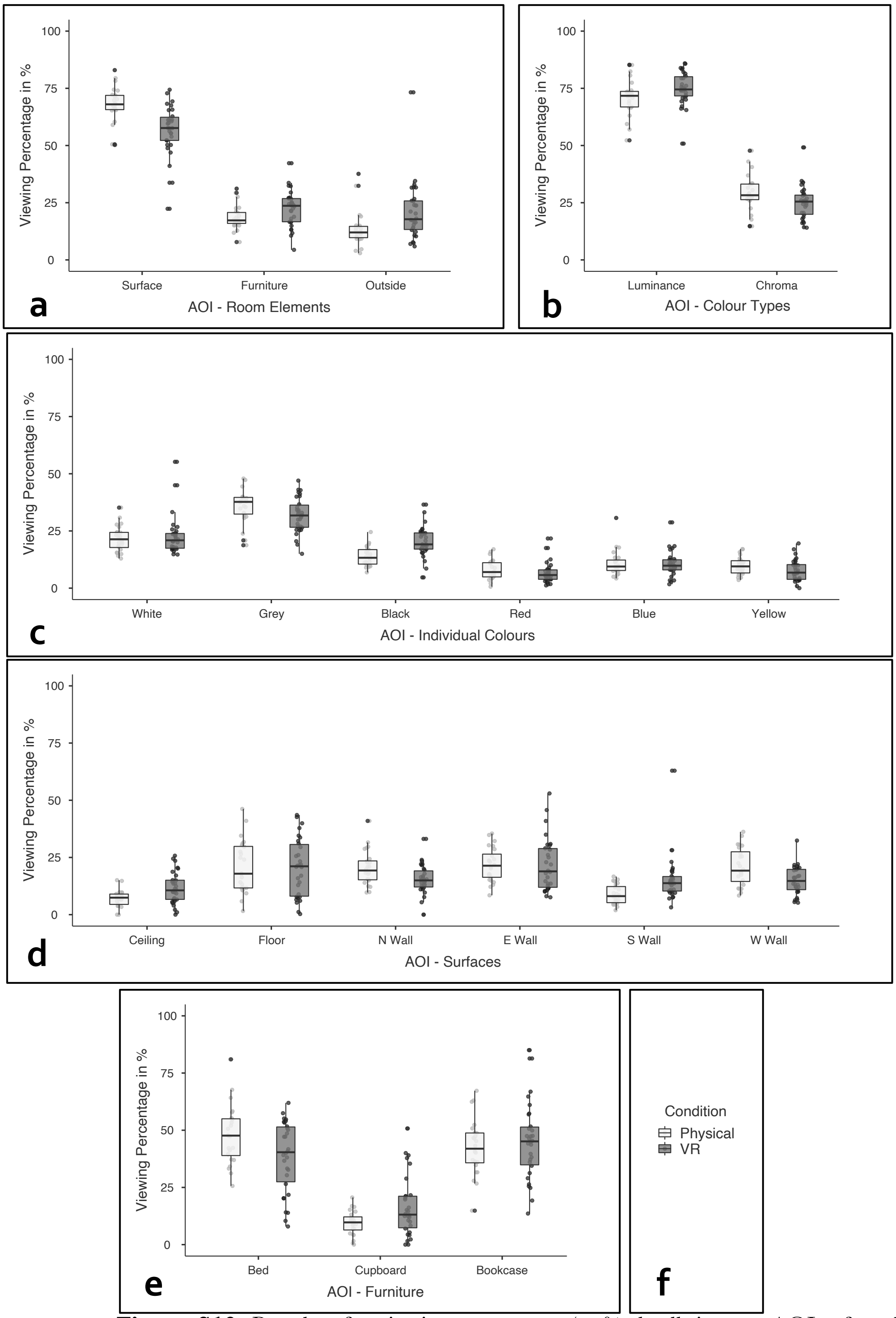
Supplementary Figure S7. Summary data table of eye-tracking metrics. The sample sizes were $N_{Physical} = 24$, $N_{VR} = 30$.



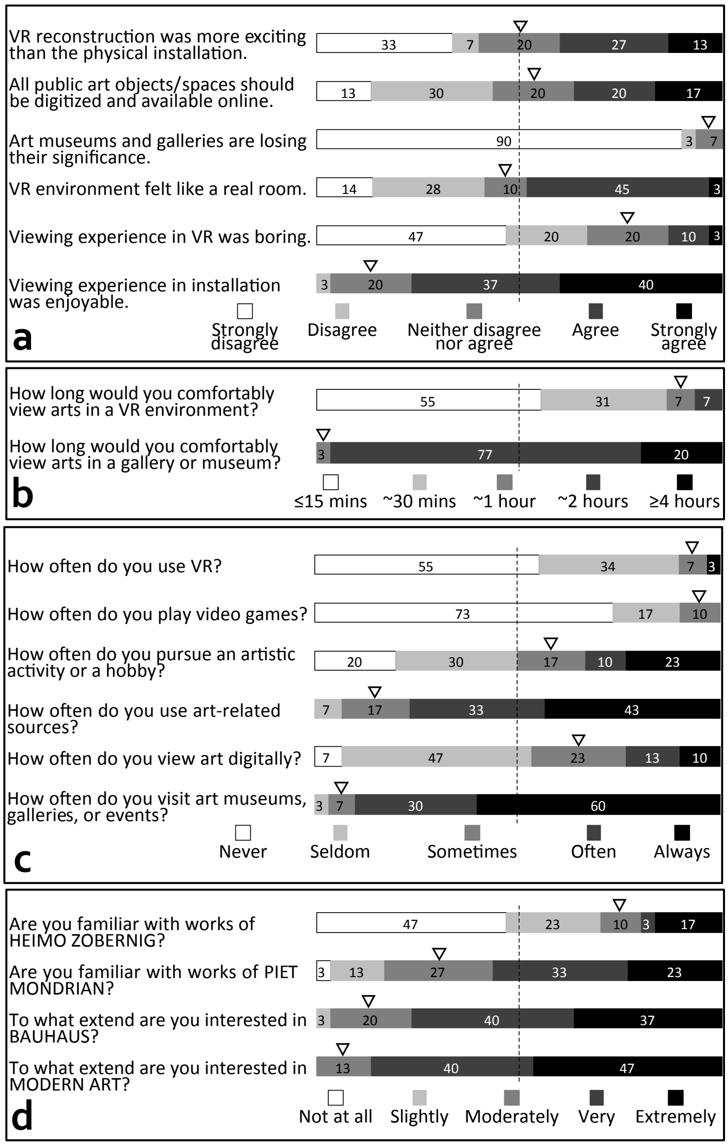
Supplementary Figure S8. Boxplots for time to first fixation comparing physical and VR settings for five comparisons: (a) room elements, (b) colour types, (c) individual colours, (d) surfaces, and (e) furniture. The x-axis shows the levels of AOIs, and the y-axis shows time to first fixation in seconds. Physical and VR conditions are shown as (f) light and dark grey, respectively. The sample sizes were $N_{Physical} = 24$, $N_{VR} = 30$.

CORRELATION BETWEEN FIXATION DURATION AND FIXATION COUNT								
			CORRELATION COEFFICIENTS (Physical Condition, N = 24)			CORRELATION COEFFICIENTS (VR Condition, N = 30)		
			Pearson's r_P	Spearman's r_S	Kendall's τ	Pearson's r_P	Spearman's r_S	Kendall's τ
AOIs	Surfaces	Ceiling	.961***	.923***	.830***	.994***	.982***	.917***
		Floor	.982***	.955***	.876***	.990***	.989***	.927***
		N Wall	.982***	.958***	.859***	.978***	.987***	.923***
		E Wall	.986***	.959***	.854***	.989***	.942***	.846***
		S Wall	.984***	.976***	.896***	.970***	.914***	.788***
		W Wall	.976***	.975***	.889***	.986***	.972***	.904***
	Furniture	Bed	.897***	.929***	.796***	.956***	.962***	.852***
		Cupboard	.960***	.934***	.832***	.982***	.980***	.908***
		Bookcase	.965***	.951***	.847***	.970***	.974***	.892***
	Colours	White	.985***	.990***	.940***	.986***	.956***	.857***
		Grey	.987***	.953***	.860***	.988***	.966***	.869***
		Black	.945***	.932***	.788***	.980***	.966***	.871***
		Red	.985***	.986***	.934***	.991***	.977***	.900***
		Blue	.942***	.971***	.874***	.988***	.934***	.833***
		Yellow	.948***	.837***	.684***	.974***	.982***	.914***

Supplementary Figure S9. Correlation table indicating the strength of the relation between dwell time (as total fixation duration) and fixation count for sets of AOIs, summarized separately for physical and VR conditions. Both dwell time and fixation count can be treated either as continuous variables (and analysed using Pearson's r_P) or ordinal variables (and analysed using Spearman's r_S or Kendall's τ). Irrespective of the statistical standpoint, the data showed significant, positive, linear relationship for every AOI set: the numbers on the table denote the self-correlation coefficients (such as the correlation between dwell time on the ceiling and fixation count on the ceiling, etc.) corresponding to three different statistical measures, and *** denotes $p < .001$. Note that, since the area-normalized metrics were directly derived from the regular metrics, the dwell time and area-normalized dwell time has a perfect linear positive relationship with a correlation coefficient of +1. Similarly, the fixation count and area-normalized fixation count has also a perfect linear positive relationship with a correlation coefficient of +1. Hence, the correlation coefficients between area-normalized dwell time and area-normalized fixation count are the same as this table. One interpretation of these highly strong correlations is that the results of a potential set of supplementary analysis using absolute and area-normalized fixation count should be very similar to our main results in this study where we have chosen to use absolute and area-normalized dwell time. The sample sizes were $N_{Physical} = 24$, $N_{VR} = 30$.



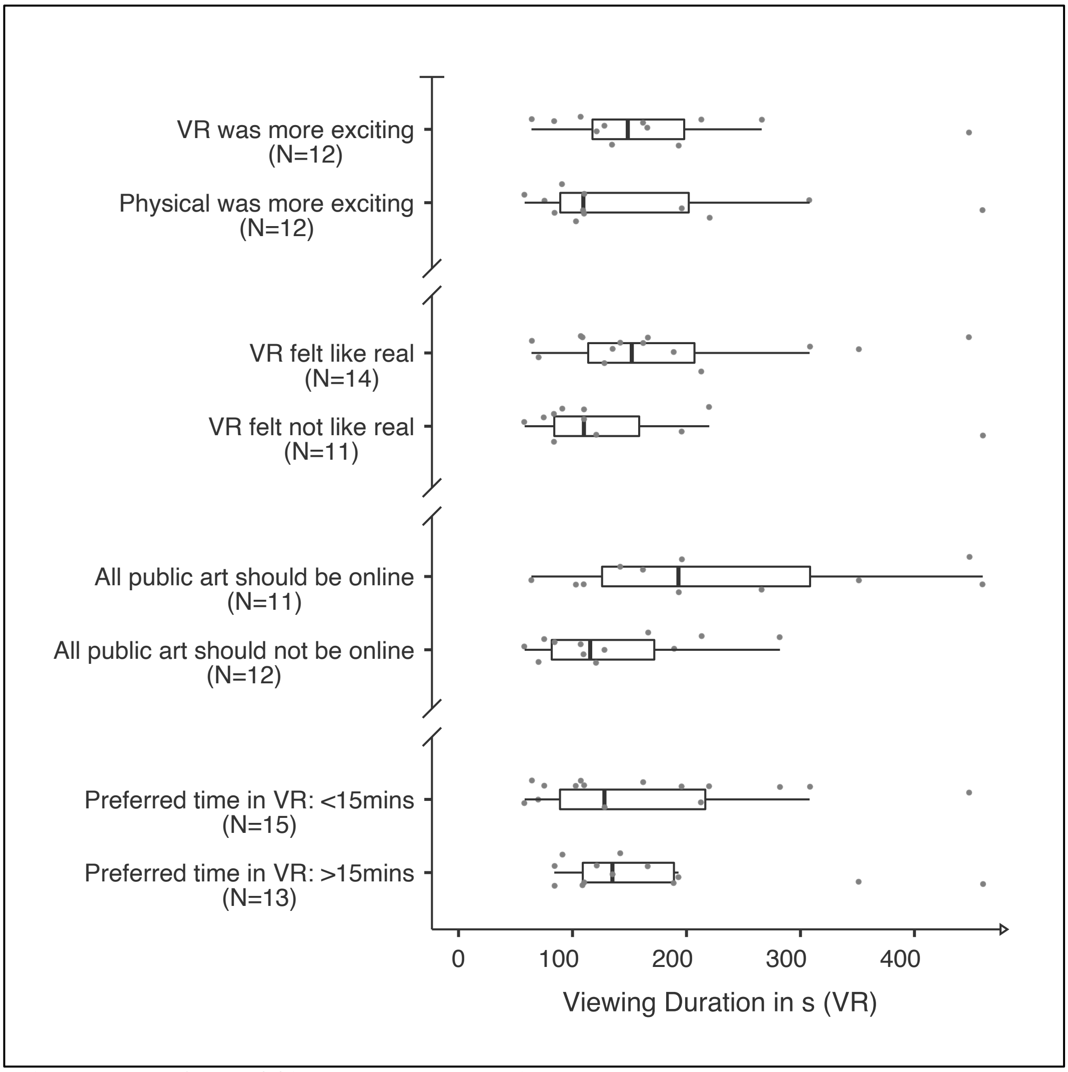
Supplementary Figure S10. Boxplots for viewing percentage (as % dwell time on AOIs of total dwell time), comparing physical and VR settings for five comparisons: (a) room elements, (b) colour types, (c) individual colours, (d) surfaces, and (e) furniture. The x-axis shows the levels of AOIs of a given grouping, and the y-axis shows the viewing proportions in %. Physical and VR conditions are shown as (f) light and dark grey, respectively. The sample sizes were $N_{Physical}$ = 24, N_{VR} = 30. Note, particularly to compare the physical and the VR condition, the normalized data show an overlapping trend of viewing percentage between the two conditions, and is similar to the results of both the absolute and area-normalized dwell time analysis in this sense (see Fig. 5 and Supplementary Fig. S14, respectively).



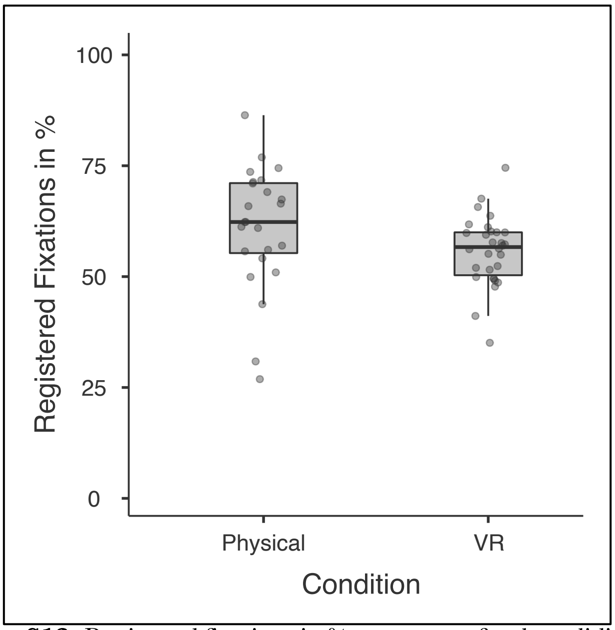
Supplementary Figure S11. Questionnaire results visualized as frequency plots, where numbers on the bars indicate the percentage of responses rounded to the nearest one, the dashed lines aim to correspond to the median, and the triangle above the bars indicates the mid-point of the middle response (i.e., the mid-point of the scale as the centre of the 3rd rating response value out of 5-point Likert-scale, to further illustrate the central tendency): (a) attitudes as agreement ratings, (b) comfort as ratings on preferred view duration, (c) exposure to art forms as frequency ratings, and (d) knowledge as familiarity ratings. The sample sizes were $N_{Physical} = 24$, $N_{VR} = 30$.

(a) Attitude ratings showed that visitors were split into two (40%:40%) on judging whether the VR reconstruction was more exciting than the physical installation. Similarly, the VR environment felt like a real room for around half of the participants (48%). A similar split was also visible on whether all public art should be digitized and available online (37%) or not (43%). At the same time, visitors strongly disagreed with the idea that the art museums and galleries are losing their significance (90%). The physical installation was enjoyable for most of the participants (77%), and the VR experience was not boring for most participants (67%). (b) Preferred viewing duration that participants would think to allocate to viewing artworks was radically different between two conditions: participants seemed to be comfortable to spend much more time in a gallery (77% for circa 2 hours) in comparison to VR (55% for less than 15 minutes). (c) Frequency responses on their exposure to art forms revealed that most of the participants were not VR users (55%) and did not play video games at all (73%). Visitors showed diverse responses on their pursuit of artistic activities and use of art-related sources. Around half of the visitors reported that they view art digitally either seldom or never (54%), whereas they were keen on visiting art museums, galleries and events very frequently (60%). (d) Level of interest in art periods and familiarity with artists was overall high: visitors

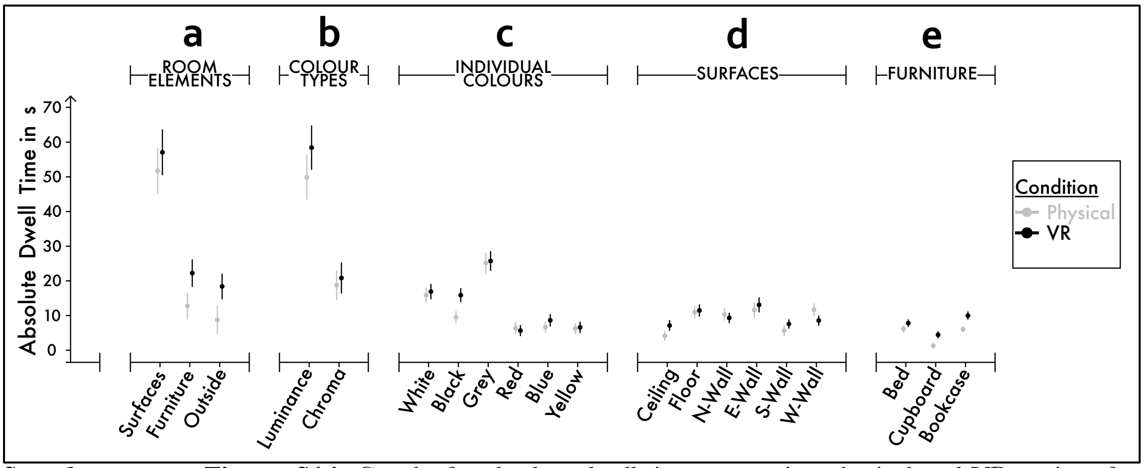
were extremely interested in modern art (47%), and very interested in Bauhaus (40%). Visitors were very familiar with works of Mondrian (33%), but not at all familiar with works of Heimo Zobernig (47%).
Open-ended feedback revealed an overall satisfaction from the experience, and visitors described it using keywords such as novel, surprising, stimulating, realistic, exciting, interesting, impressed, curious, enriching, calm, etc. According to visitors, a prominent drawback of the study was the technical properties of devices, some participants particularly remarked on the spatial resolution of VR screen, limited field of view, and discomfort due to screen. In regards to the additional question about formal educational level (not presented here), the participants were biased towards higher education, as qualifications showed a highly skewed distribution among participants compared to the population: PhD (23%), graduate 2 (33%), graduate 1 (17%), secondary 2 (23%), and secondary 1 (3%).



Supplementary Figure S12. The distribution of viewing duration in the VR, bisected into two based on four questionnaire item responses which resulted in close to bimodal splits. The x-axis shows the viewing duration in s for the VR condition, the categorically separated y-axis shows the response pairs of four items, with individual datapoints corresponding to individual participants.



Supplementary Figure S13. Registered fixations in %, as a proxy for the validity of the gaze data: here, the registered fixation in % was simply calculated by the total fixation duration divided by the total recording duration (i.e., total viewing duration). The remaining percentage accounts for all non-fixational metrics such as saccades, as well as any data loss. In both conditions, the registered fixations in percentage ($M_{Physical}$ = 61.1, $SEM_{Physical}$ = ±2.83; M_{VR} = 55.8, SEM_{VR} = ±1.43) showed normal distribution based on the Shapiro-Wilk tests ($W_{Physical}$ = .946, p = .221; W_{VR} = .973, p = .625), and a similarity between the two conditions was assumed based on a comparison using a simple GLM that resulted in a non-significant result ($F_{(1,52)} = 3.19$, $p > .05$, $\eta_p^2 = .058$). The boxplots are based on the distribution of the means from participants, and each individual data point shown as an overlay on top of the boxplots represents a single participant. The categorically separated x-axis denotes two conditions, and the y-axis denotes the registered fixations in percentage. The sample sizes were $N_{Physical}$ = 24, N_{VR} = 30.



Supplementary Figure S14. Graphs for absolute dwell time comparing physical and VR settings for five comparisons: (a) room elements, (b) colour types, (c) individual colours, (d) surfaces, and (e) furniture. The x-axis shows the levels of individual AOIs, and the y-axis shows absolute dwell time in seconds. Physical and VR conditions were colour-coded as grey and black, respectively. The visualized data were based on means, with whiskers indicating 95% of confidence intervals (CIs). The sample sizes were $N_{Physical} = 24$, $N_{VR} = 30$.

The analysis method for absolute dwell time (defined as time spent on looking on particular regions irrespective of the size of these regions, namely, cumulative fixation duration per AOI) was the same compared to area-normalized dwell time, using linear mixed-effects models for five comparisons.

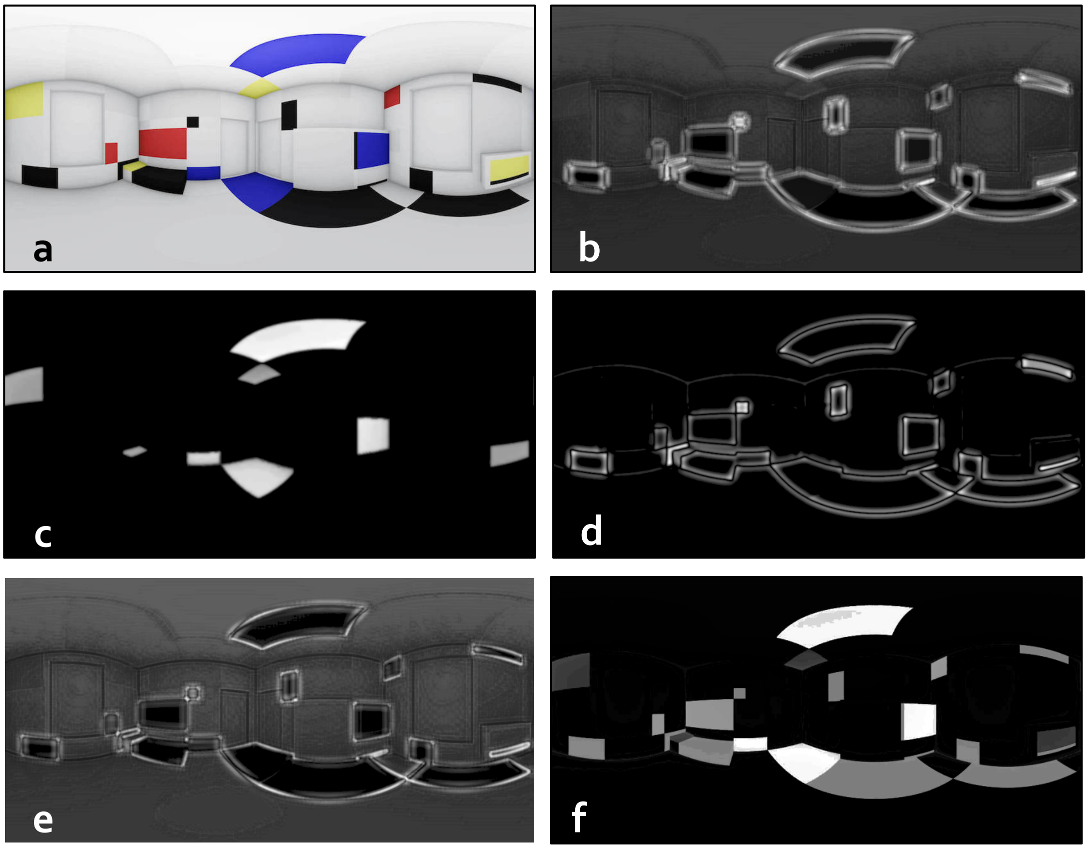
Comparison 1 on room elements (a): A significant difference between room elements was found, but no difference was observed between environments or in terms of an interaction: $F_{(2, 46.5)} = 30.605$, $p < .001$; $F_{(1, 30.4)} = 2.503$, $p > .05$; and $F_{(2, 104.3)} = 0.322$, $p > .05$ respectively. Dwell time on surfaces ($M_{Surfaces} = 55.62 \pm 5.79$s) was higher than both furniture ($M_{Furniture} = 18.17 \pm 1.79$s) and outside ($M_{Outside} = 14.26 \pm 1.58$s): $t_{(28.9)} = 7.06$, $p < .001$; and $t_{(28.9)} = 7.73$, $p < .001$, respectively. Overall, in both environments visitors spent most time looking at surfaces of the installation in comparison to furniture elements or outside space.

Comparison 2 on colour types (b): A significant difference between colour types was found, but no difference was observed between environments or in terms of an interaction: $F_{(1, 32.6)} = 72.091$, $p < .001$; $F_{(1, 30.5)} = 0.617$, $p > .05$; and $F_{(1, 53.3)} = 1.635$, $p > .05$, respectively. The dwell time on luminance-only colours ($M_{Luminance} = 54.25 \pm 5.24$s) was higher than chroma-containing colours ($M_{Chroma} = 19.55 \pm 2.09$s). Overall, in both environments visitors spent more time looking at black, grey and white colours in comparison to red, blue, and yellow.

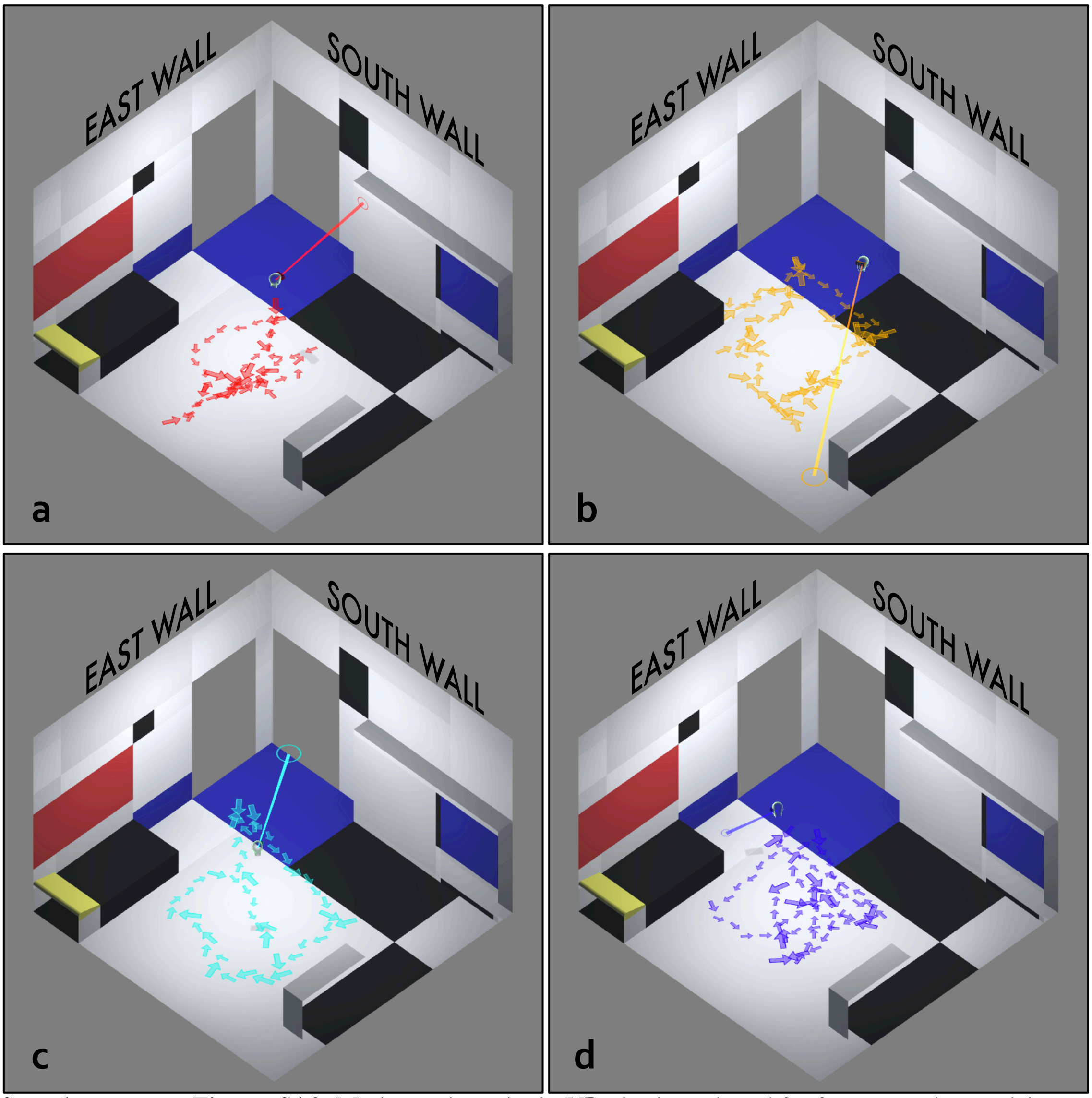
Comparison 3 on individual colours (c): Significant differences between individual colours was found, no difference was observed between environments, and a difference was present in terms of an interaction: $F_{(5, 56.7)} = 17.579$, $p < .001$; $F_{(1, 30.6)} = 0.524$, $p > .05$; and $F_{(5, 233.0)} = 2.451$, $p = .034$; respectively. Visitors spent most time on grey ($M_{Grey} = 25.20 \pm 2.53$s), and least on red ($M_{Red} = 5.78 \pm 3.63$s). When the interaction was broken down to check the relationship between two environments and six colour types, we found little differences for any level of colour: all six post hoc comparisons yielded non-significant results (where all $p > .05$). The interaction was only pronounced, when it was broken down by focusing on the individual colours to check how paired colour differences are affected differently for VR and physical environment: in this approach, some trend changes were visible, such as the dwell time difference between $black_{VR}$ and $grey_{VR}$ was significant ($p = .005$), but the dwell time difference between $black_{VR}$ and $grey_{Physical}$ was not significant as ($p > .05$), suggesting that for some colour pairs, the amount of dwell time difference could be dependent on the environment.

Comparison 4 on cube surfaces (d): A significant difference between cube surfaces was found, no difference was observed between environments, and an effect was present in terms of an interaction: $F_{(5, 54.4)} = 9.819$, $p < .001$; $F_{(1, 29.9)} = 0.068$, $p > .05$; and $F_{(5, 231.5)} = 2.653$, $p = .024$, respectively. The dwell time on the ceiling ($M_{Ceiling} = 5.66 \pm 0.89$s) was lowest, and on the east-wall ($M_{East\text{-}Wall} = 12.80 \pm 1.85$s) was highest. Overall, visitors looked approximately five to thirteen seconds at each one out of six faces of the cube. When the interaction was broken down to check how the two types of environment interact with six cube surfaces, all six post hoc comparisons yielded non-significant results (where all $p > .05$). The interaction was only pronounced, when the interaction was broken down to check how levels of cube surfaces have a different

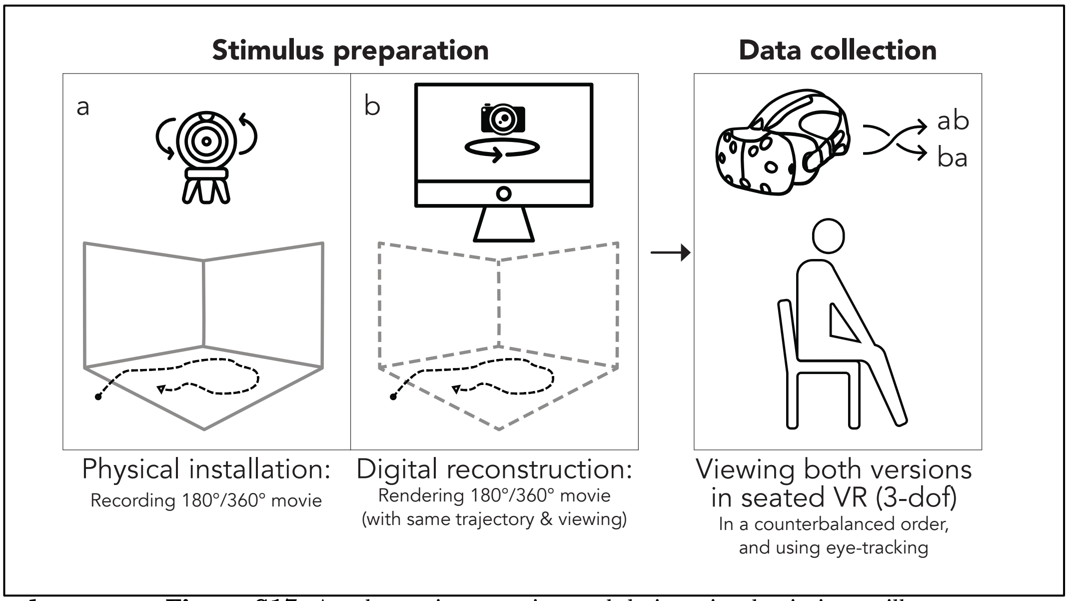
effect for VR and physical environment: in this approach, some trend changes were visible, such as the dwell time difference between south-wall$_{Physical}$ and west-wall$_{Physical}$ was significant ($p = .009$), but the dwell time difference between south-wall$_{Physical}$ and west-wall$_{VR}$ was not significant as ($p > .05$), suggesting that for some paired cube surfaces, the amount of dwell time difference was dependent on the environment.
Comparison 5 on furniture (e): A significant differences between types of furniture was found, a difference was observed between environments, but no effect was present in terms of an interaction: $F_{(2, 26.1)} = 20.21$, $p < .001$; $F_{(1, 30.2)} = 7.77$, $p = .009$; and $F_{(2, 63.3)} = 1.22$, $p > .05$, respectively. In terms of furniture, the dwell time on cupboard ($M_{Cupboard} = 3.03 \pm 0.64s$) was lower than both bookcase ($M_{Bookcase} = 8.12 \pm 0.91s$) and bed ($M_{Bed} = 7.01 \pm 0.70s$): $t_{(28.2)} = 5.88$, $p < .001$; and $t_{(28.1)} = 4.74$, $p < .001$, respectively. In terms of environments, the dwell time for furniture in VR ($M_{VR} = 7.41 \pm 0.73s$) was longer than physical ($M_{Physical} = 4.36 \pm 0.36s$): $t_{(28.4)} = 2.77$, $p = .010$. Overall, visitors spent more time looking at furniture in VR compared to in the physical environment, and in both conditions the cupboard was the least fixated furniture out of three.



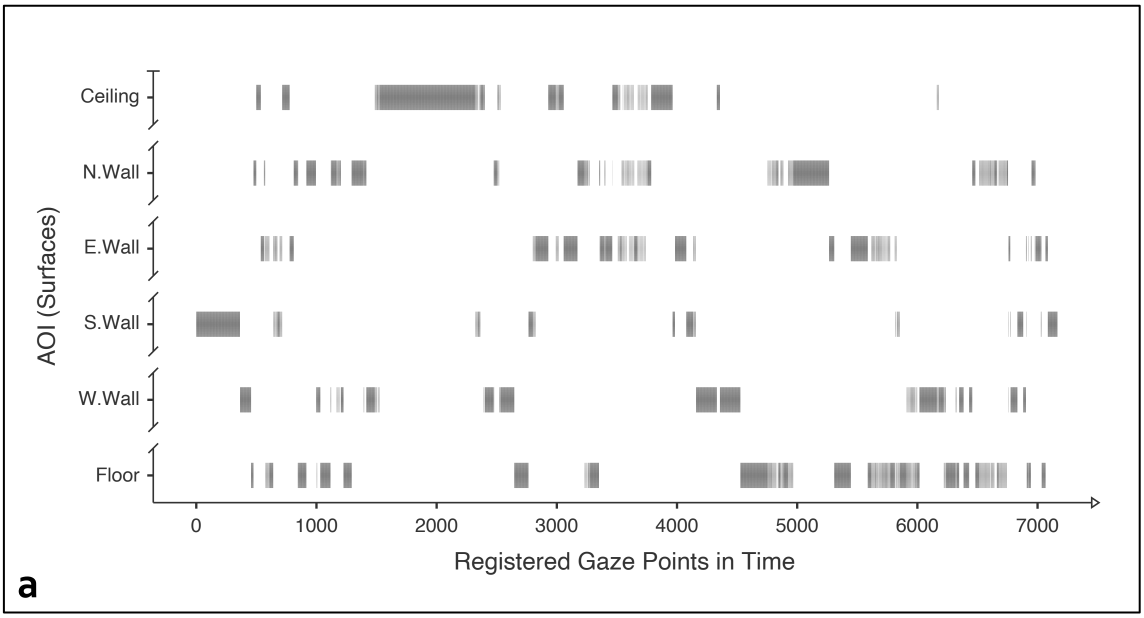
Supplementary Figure S15. (a) Panoramic render of the VR version, (b) a saliency map generated using the Itti algorithm and its (c) colour, (d) intensity, and (e) orientation components; and (f) another saliency map generated using histogram contrast. (b-e) For the Itti algorithm, following the default settings to calculate features using linear centre-surround operations in the RGB colour space, the centre is defined as a pixel at scale $c \in \{2, 3, 4\}$, and the surround is defined as the corresponding pixel at scale $s = c + \delta$ where $\delta \in \{3, 4\}$ (Itti et al., 1998). (f) The histogram contrast mapping is simply based on mean colour difference to image pixels. Note that, (c) the colour component of the Itti algorithm disregards the black surfaces, whereas (f) a simpler saliency map based on histogram contrast includes black surfaces, which becomes a more suitable saliency map since normalized dwell time was relatively high on both chroma-containing colours and blacks, and relatively low on white and grey surfaces.

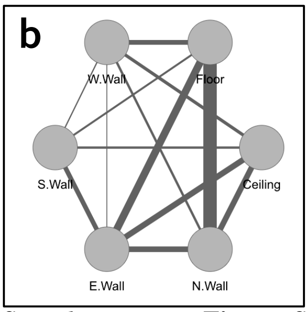


Supplementary Figure S16. Motion trajectories in VR viewing, plotted for four exemplar participants to show the individual differences in terms of general movement within the installation, illustrated by the arrows on the floor.



Supplementary Figure S17. An alternative experimental design, simply aiming to illustrate a potential way of equating the movement and viewing duration between physical and digital settings, whilst still providing partial immersiveness to some extent. Instead of allowing the participant to freely move in two settings without time constraints (as in the case of our presented study); alternatively, the 180° or 360° video footage can be recorded from the physical museum, and using the same trajectory and viewing angles, a virtual replication of a rendered movie can be created. After that, both versions of the videos, basically as a form of "a guided viewing", can be viewed by the participants in a seated-VR as full or semi-spherical movies in a counter-balanced order in the case of a within-subjects experiment, (or can be viewed by two groups of participants, if the design is a between-subjects experiment).





c

	Ceiling	**N.Wall**	**E.Wall**	**S.Wall**	**W.Wall**	**Floor**
Ceiling	0	.017	.060	.015	.027	0
N.Wall	.057	0	.018	0	.016	.146
E.Wall	.021	.057	0	.043	0	.077
S.Wall	.015	0	.044	0	.004	.011
W.Wall	.025	.016	.001	.002	0	.047
Floor	0	.149	.076	.012	.044	0

Supplementary Figure S18. (a) The plot showing the evolution of the registered gaze points (i.e., raw data to be used to classify fixations) evolved over time, here, for a single exemplar participant in the VR condition. The x-axis shows the order of the registered gaze points in time, the lanes on the y-axis show the corresponding AOI (here, out of six cardinal surfaces), and each datapoint displayed as a thin slice on the plot is a single registered gaze point in a temporal order. This temporal illustration, for example, shows that the participant stated by particularly looking on the south wall at start, and after some time, spend significant time viewing the ceiling, etc. (b) A graph-theoretical representation for registered gaze point changes for the same participant (as an approach for the scan path similarity): in the weighted undirected graph, the nodes denote set of AOIs, here illustrated only as six cardinal surfaces, and each edge with a stroke width denotes the amount of gaze point change from one AOI to another, disregarding loops, which can be formed if two or more consecutive gaze points fall on the same AOI. (c) The corresponding adjacency matrix, denoting the amount of registered gaze point change from one AOI to another, reported as a percentage of all fixation data. Note, the amount of change either equates or approximates zero for the surface pairs orthogonal to each other (i.e., ceiling-floor, north wall – south wall, east wall – west wall), since it is unlikely to observe such fixational jumps.

References

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Do Hyong Koh, & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, *57*(11), 2635–2645. https://doi.org/10.1109/TBME.2010.2057429

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '00*, 71–78. https://doi.org/10.1145/355017.355028

Chapter 4

Virtual Reality Gallery

Aesthetic judgments of 3D arts in virtual reality and online settings

Doga Gulhan* · Szonya Durant · Johannes M. Zanker

Department of Psychology, Royal Holloway, University of London, UK

*doga.gulhan@rhul.ac.uk

4.1. Abstract

Empirical aesthetics is beginning to branch off from conventional laboratory-based studies, leading to in-situ, immersive, often more accessible experiments. Here, we explored different types of aesthetic judgments of three-dimensional artworks in two contexts: virtual reality (VR), aiming for an immersive experience, and online settings aiming for an accessible setup for a remote audience. Following the pilot experiment conducted to select a set of 3D artworks, in the first experiment, participants freely engaged with virtual artworks via an eye-tracking-enabled VR headset and provided evaluations based on subjective measures of aesthetic experience such as ratings on liking, novelty, complexity, and viewing duration. Results showed positive, linear, and mostly moderate correlations between liking and the other perceived judgment attributes. Supplementary eye-tracking data showed a range of viewing strategies and variation of viewing durations between participants and artworks. Results of the second experiment, adapted as a short online follow-up, showed converging evidence on correlations between the different aspects contributing to aesthetic judgments and suggested similarity of judgment strategies across contexts. In both settings, participants provided further insights via exit-questionnaires. We speculate that both VR and online settings offer ecologically-valid experimental contexts, create immersive visual arts experience, and enhance accessibility to cultural heritage.

4.2. Introduction

Evaluating visual artworks can be described as a partially overlapping extension of aesthetic experience, and also as a complex cognitive-emotional process. Engaging with arts often involves both general emotions such as surprise, joy, or disgust, and emotions associated with arts such as sublime or aesthetic pleasure. On the other hand, when observers are asked to evaluate an artwork whilst they are interacting with art, they tend to assign and update a set of values towards the artwork. The assigned values can be related to any potential aspect of the artwork, for example, beauty, compositional properties, or monetary worth. These highly subjective assigned values are thought to depend on visual properties such as contrast and colour (Mallon et al., 2014), could be affected by contextual information (Grüner et al., 2019), artwork title (Turpin et al., 2019) or artists' names (Cleeremans et al., 2016) and can change over time (Isik & Vessel, 2019). Taking on board these complexities from empirical studies, a set of design guidelines for using artworks as stimuli has recently been proposed (Hayn-Leichsenring, 2017). The authors of these guidelines highlighted the vagueness of aspects of the research in this area. An additional conceptual challenge is that an observer can assign a value for an artwork either as an absolute judgment or a relative judgment. Nevertheless, previous work tends to propose that the evaluative aspect of aesthetic experience can be operationalized and thus at least partially measurable.

Evaluation of an artwork can be as simple as a single binary judgment of like or dislike, or a long interpretive narrative from an observer. On the theoretical level, aesthetic and non-aesthetic based values can be assigned to evaluate arts, and they have the potential to influence each other (Aumann, 2014). The conceptual richness of visual arts leads to the possibility of using many adjectives, adverbs, or metaphors to evaluate an artwork; and thus aesthetics becomes a challenging research topic for the philosophy of language and semantics as well (Young, 2017). On the empirical level, researchers have previously investigated themes related to aesthetic values (often by incorporating relative judgments) in varied contexts, such as (i) perceptual and representational attributes describing paintings as a basis to form an assessment tool (Chatterjee et al., 2010), (ii) use of highly specific modes of expression such as "feeling like crying" in relation to aesthetic experience (Pelowski, 2015), (iii) predicting aesthetic preference by other perceived attributes such as meaning(fulness), and whether these attributions are robust to image manipulation such as blurring (Moore & West, 2012), (iv) category-dependent generality and specificity of word usage describing subsets of artworks, and aiming to form a language of aesthetics for the visual modality (Augustin et al., 2012), (v) extent of choice reversal following a type of experimental biasing by pairing "average-beauty" paintings with either relatively more or relatively less beautiful paintings, where observer makes a binary preference choice between two abstract paintings (Belchev et al., 2018), among many others. As a common framework implied in many studies, a general form of positive aesthetic judgment (such as finding an artwork good) is often linked to either other positive emotional judgments (such as finding an artwork pleasurable) or forms of positive cognitive or moral judgments (such as finding an artwork beneficial). A general interpretation regarding such associations can be described such that most types of aesthetic judgments can be conceptually aligned along a single negative-positive judgment axis, although some research has explicitly investigated the counter-intuitive associations between judgments, as well as specifically negative aesthetic emotions and judgments (Landau et al., 2006; Silvia & Brown, 2007; Cooper & Silvia, 2009; Wagner et al., 2014).

Since empirical research on aesthetic judgments is often conducted in laboratory settings, a common limitation is using reduced artworks such as 2D snapshots of paintings or a manipulated visual stimulus as a substitute for artworks. The tendency to favour well-controlled stimulus presentation in laboratory settings often results in a diminished resemblance between experimental paradigm and genuine aesthetic experience. The generalizability of findings outside the lab settings to a real-world has been described as a common weakness of these studies (Locher et al., 1999; Brieber et al., 2015). Recent developments such as ease of implementing virtual reality (VR) environments, using 3D modelling software as an artistic tool, and more specifically photogrammetry methods to translate physical objects and environments into 3D models has come to offer, in some aspects, ecologically valid alternatives to real-world scenarios and useful tools for cultural heritage (Clini et al., 2018; Liarokapis et al., 2020). As immersive environments aim to enhance user experience in gallery and museum settings, many exploratory studies have started to investigate visitors' experience and the feasibility of these VR applications (Hoang & Cox, 2017; Petrelli, 2019; Parker & Saker, 2020). Experiments have mostly focussed so far on the general cognitive implications of using these environments, for example, crowd movement on navigation decisions in VR (Zhao et al., 2020), mental imagery and eye movements in VR (Chiquet et al., 2020), visual search in 3D scenes (Helbing et al., 2020), replication of findings from a lab-based inattentional blindness paradigm in VR (Schöne et al., 2021), or episodic memory in virtual museum rooms (van Helvoort et al., 2020). Experimental aesthetics research in VR remains to be explored. Apart from screen-based and VR-based studies often conducted in a lab setting, internet-mediated research can be seen as a distinct research setting, and its validity can be linked to the increased viewing of visual arts by a remote audience, away from physical museums and galleries. Additionally, online research in general offers the possibility of a diverse and large sample and minimizing some biases, such as the observer-expectancy effect where researchers unintentionally influence the behaviour of participants (Palan & Schitter, 2018; Peer et al., 2017; Woods et al., 2015). In this sense, immersive experiments utilizing VR can be framed as a proxy for the real-world art experience, whereas web experiments can be seen as a proxy for the online art experience.

The present research aimed to incorporate various commonly used judgment types for visual arts from previous studies including liking and novelty, as well as to include relatively unusual types of judgments such as liking from a third-person point-of-view (as a proxy for assessing normativity, i.e., whether participants' personal liking judgments align with their expected judgments from others' perspective) and perceived viewing duration (as a not directly aesthetic, artwork based judgment). The pilot experiment aimed to select a set of 3D models as artworks, the first experiment using VR was designed to measure observers' conscious choices towards a set of artworks as indexed by rating scales, and the second experiment as an online follow-up was a shortened version of the VR experiment. As a supplementary (and more implicit) measure, eye-tracking data were collected whilst participants were engaging with artworks in VR, to inspect visual exploration patterns of observers. In both the first and second experiments, exit-questionnaires were included to provide additional insights into participants' attitudes towards visual arts, art-related arguments, and the experiments. The experiments' main aim was to investigate the strength of correlations between liking ratings and all other rating types in VR and online settings, and we expected significant correlations between aesthetic ratings in both settings. In this sense, the main alternative hypothesis can be formulated such that there are statistically significant positive linear relationships between liking rating and other ratings, whereas the null hypothesis as the default state can be formulated that there are no relationships between liking and other ratings.

4.3. Pilot experiment: Selection of 3D artworks

4.3.1. Methods

4.3.1.1. Participants

All three authors participated in the pilot experiment.

4.3.1.2. Stimulus and material

The stimulus was a set of 2D snapshot images of 3D models from SketchFab (sketchfab.com), an online platform for publishing 3D content. The 3D models were a small subset of the collection, selected with the following criteria: the top hundred, most viewed, downloadable models (according to all-time website usage metrics provided by SketchFab) were listed, in line with the four suitable categories available on the website: architecture, art and abstract, cultural heritage, places and travel. Since a digital model might belong to more than one category, eliminating duplicate models in this subset resulted in a total selection of 336 models instead of 400. After downloading a batch of 2D snapshot images of these 3D models, all images were cropped proportionally and equated in size to 720 by 400 pixels using Adobe Photoshop (adobe.com). Stimuli were presented using MatLab (mathworks.com) with Psychtoolbox (psychtoolbox.org), viewed on a screen of personal monitors of varying display resolution. Responses were recorded via the keyboard of personal computers.

4.3.1.3. Design

A single, binary variable of judgment (i.e., "interesting" or "not interesting") was present in the 2-alternative-forced choice (2-AFC) design, for each of the individually displayed artworks.

4.3.1.4. Procedure

Participants performed a judgment task to decide whether the presented model is "interesting enough to be included in the upcoming VR-based experiment". Each trial consisted of displaying a single 2D snapshot image of a 3D model at the centre of the screen, and participants categorised them either as "interesting" or as "not interesting" by pressing rightward or leftward arrows on the keyboard. Each snapshot image, corresponding to a single model, was presented only once. A total of 336 artworks were presented in three blocks, with an estimated time of completion of 30 minutes. The order of presented images was randomized for each participant.

4.3.1.5. Data analysis

Since the aim of this experiment was to choose models by unanimous agreement of all researchers, agreement percentages per artwork were calculated, followed by the determination of artworks which all participants unanimously agreed upon.

4.3.2. Results

Data revealed that 78 models out of 336 were found to be interesting by all participants, which was a 23.21% unanimous agreement, where the mean duration of decision per artwork was 1631ms. Following on from that, the models were further categorised into two-by-two binary categories, based on "physicality" and "spatiality": physicality was operationalized as whether the 3D model was a recreation of a physical artwork (and labelled as "physical"), or completely created as a digital artwork (and labelled as "digital"). Spatiality was operationalized as whether the 3D model was a small-scale artwork (and labelled as "object"), or a large-scale artwork (and labelled as "space"). As a result, the two-by-two clustering resulted in four labels for four sets: physical object, physical space, digital object, and digital space. Note that this clustering was performed by the lead author, and might be prone to miscategorization: for example, it is debatable whether a 3D model of a set of trees created via photogrammetric reconstruction from a public park is a physical object or physical space. To increase the diversity of types of selected artworks for the next experiment, from 78 models, four models were randomly selected per category, resulting in a total of sixteen artworks, including, for example, a digitally created room model and a photogrammetric model of a sculpture. Whenever required, a selected model might be disregarded due to the lack of feasibility of implementing it into a walkable virtual gallery space, therefore a new random selection was performed (see Fig. 1 for snapshot images of the final set of selected 16 artworks).

4.4. Experiment 1: Aesthetic judgments in a virtual reality setting

4.4.1. Methods

4.4.1.1. Participants

Participants were students or members of staff from Royal Holloway, University of London, and they were recruited using convenience sampling. They were compensated monetarily (£5). A total of 31 participants (17 females, 11 males, M_{Age} = 22.74 years, SD_{Age} = 4.83 years, R_{Age} = 18-38 years) were recruited for the experiment, and all were naïve to the hypotheses of experiments. All participants reported having normal or corrected to normal vision. Participants could wear their glasses or contact lenses in the VR headset. All participants provided written informed consent prior to the experiment. All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki.

A sensitivity power analysis using G*Power (gpower.hhu.de) for a sample of 31 participants in a one-tailed correlational design (as the main analysis) with a significance level of a = .05, with an assumed power of 80% as a power level of 1-β = .80, and under the assumption about the null hypothesis that there is no correlation in the population distribution, such that the correlation coefficient $\rho_0(H_0) = 0$, resulted in estimates of r_{critical} = .30 and $\rho_1(H_1)$ = .43: the estimates entail the ability to detect positive relations with medium-to-large effect sizes (i.e. $r \geq .30$ or $r \geq .50$, following the conventional guideline values (Cohen, 1992)), in line with our range of interest. Nevertheless, as a general disclaimer, this research had relatively low power to detect true low-to-medium effects, which in turn might minimise the likelihood of reproducibility of results presented.

4.4.1.2. Stimulus and material

Stimuli of the first phase (referred to as VR pre-screening) were sixteen artworks in the form of 3D digital models, where participants were expected to engage with these artworks one by one in context, walking around in a 1:1 scale gallery space. All models used in the experiment were available to be used under Creative Commons licenses (refer from snapshot images illustrated in Fig. 1 to Supplementary Fig. S1 listing attributions of individual models including their titles, uploaders, and hyperlinks). Models were digitally revised whenever needed for inter-stimulus consistency, using a set of modelling software such as Trimble SketchUp Autodesk Maya, Mudbox, and Blender (sketchup.com, autodesk.com, blender.org). Using Unity game engine (unity.com), models were placed individually in virtual gallery spaces, where all environments had roughly equal illumination levels. Participants could freely move around by using physical space at the Psychology VR Lab, a walkable area of approximately 280x360 cm due to the limits of trackable area for the VR headset. This spatial limit on walking was explained to the participants, and participants were reminded if they got close to the boundaries of the trackable area. Participants could also use the teleport function via hand-held VR controllers to instantaneously shift to a further position (see Supplementary Fig. S2*a* for an exemplar 1st person point-of-view of the instructions, and Supplementary Fig. S2*b-f* for an exemplar 1st person point-of-view of the artwork viewing). Note that, to overcome any temporal order effect, conditions (as presentation order of gallery spaces) were randomized for each participant. To collect eye-tracking data, a software plug-in called Tobii Pro VR Analytics (tobiipro.com) was also implemented in the environment. Stimuli of the second phase (referred to as the judgment task) were static snapshot images of these digital models. All images of artworks were placed individually on a mid-grey, square background of 80x80 pixels (≈1.5x1.5° of visual angle), scalable to 240x240 pixels (≈4.5x4.5° of visual angle) by right mouse click (see Supplementary Fig. S2*g* for a view of the judgment task, performed on a regular 2D monitor). No visual stimulus was present in the exit questionnaire, which was designed to collect demographic data, five-point Likert-scale rating questions, and some open-ended questions as feedback. Note that, one set of questionnaire items was particularly aimed to ascertain the task difficulty in making each judgement, phrased to participants as "how challenging was each of your judgments during the experiment", and we refer to these items as the level of difficulty.

Stimuli were displayed using an HTC Vive VR Headset (vive.com) with an embedded eye tracker, wirelessly connected to a PC (Lenovo ThinkStation, with Xeon E5-1630 @ 3.70GHz CPU, Nvidia Quadro M5000 GPU, 40 GB of RAM, running on Windows 10 Pro). HTC Vive controllers were also used to teleport within the environment, and to proceed between gallery spaces. Software used to present VR stimulus and record gaze data was an executable file built using Unity. The judgement task was coded using MatLab with Psychtoolbox and was displayed via a separate PC (Dell Alienware 17, with Intel i7-8759H @ 2.20GHz CPU, Nvidia GeForce RTX 2060 GPU, 16 GB of RAM, running on Windows 10 Home, with a display resolution of 1920 by 1080 pixels at 60Hz refresh rate). The exit questionnaire was a simple form created using Google Forms (google.com/forms) and displayed again using the same PC.

4.4.1.3. Design

Designed mainly as a correlational experiment, nine variables from the task were the ratings, provided by the participants for each artwork: liking, liking from the third-person point of view (POV), emotional valence, meaningfulness, novelty, artfulness, complexity, colourfulness, and perceived viewing duration. Liking from third-person POV was explicitly described to the participant such that this judgment should be based on the expectancy on how other participants would rate the images. During the task, these ratings as binary labels indicated minimum and maximum ends of the rating scale, as in "most-*p*/least-*p*", where *p* denotes a given variable of a judgment type (see Supplementary Fig. S3 for an overview of all judgment questions, and their labels as presented to the participants). Viewing duration was the tenth variable. The eye-tracking data were only used for visual inspection of the exploration patterns of observers. The main hypothesis was solely based on correlations between participant ratings: it was expected that participants' assigned liking ratings are positively correlated with other ratings towards artworks. Additionally, using a linear regression model, liking rating was defined as an outcome variable, and all other ratings as predictor variables, to check whether participants' ratings on these dimensions can predict liking rating. Lastly, mean liking ratings of individual artworks per category were compared using a 2x2 RM-ANOVA in line with 2x2 stimuli categories (namely, physical vs. digital arts, and objects vs. spaces), and post hoc comparisons using t-tests were Bonferroni corrected. The first factor was defined as physicality (namely, whether the displayed artwork was a representation of a physical artwork, or completely created as a digital artwork), where two levels were physical and digital artworks. The second factor was defined as spatiality (whether the displayed artwork was small-scale and defined as an "artwork-object" or is large-scale and defined as an "artwork-space"), where two levels were object and space. The reasoning was to check whether category-specificity affected liking ratings. Here, the dependent variable (DV) was the liking rating in percentage, and the two independent variables (IVs) were two artwork categories (labelled as physicality and spatiality).

4.4.1.4. Procedure

Following the briefing and receiving consent from participants, the experiment consisted of three phases. In the first phase, referred to as the VR pre-screening, information on hardware, software, and the user interface of the VR headset were provided to the participant. Participants were in standing position in the Psychology VR Lab. After putting the head-mounted display (HMD) on, an eye-tracking calibration was executed to ensure the reliability of gaze data to be collected using the 5-point default in-built calibration. Following the calibration, participants firstly visited three gallery spaces. Using VR controllers, participants had a chance to practice travelling between these gallery spaces by the trigger-button press on the VR controllers, as well as teleporting themselves around the artworks by pressing the trackpad-button. Participants then engaged with sixteen artworks one by one in randomized order, without any time constraints. Participants could revisit a specific artwork if they wished to do so, but none of the participants revisited a previously seen artwork again during the experiment. 3D objects and spaces were presented in an otherwise empty digital room, similar to a white-box gallery space. During this period, gaze data were recorded.

In the second phase, referred to as the judgment task, participants sat in front of a computer screen, approximately 57 cm away. For each question, participants could see and judge all artworks at once. Participants thus could drag and drop thumbnail images of artworks on screen, in relation to a "most-*p*/least-*p*" scale for the given property *p* such as most liked and least liked, visualized by a background gradient from black to white (see Supplementary Fig. S2*g* for exemplar view from the judgment task). This way of sorting the stimuli allowed for more precise and relative judgments from participants. Responses were recorded using a keyboard and mouse. Lastly, a brief exit questionnaire was also presented on-screen. During the VR phase and in-between sessions, participants were reminded that they may pause or stop the experiment whenever they feel discomfort or motion sickness (although none of the participants reported such issues, paused or ended the experiment). The experiment was conducted without a time limitation, but on average the time of completion was around 30 minutes.

4.4.1.5. Data analysis

Three main data streams were formed: data from the VR eye-tracking, the rating judgments, and the exit questionnaire. Software used for the data analysis and visualisation were R and jamovi (r-project.org, jamovi.org). Apart from descriptive statistics, the eye-tracking data were only analysed in terms of fixation duration visualization (referred to as heatmaps, associated with the amount of visual attention). In terms of the rating tasks, as a participant dragged and dropped images relative to background gradient, the image coordinates in pixels corresponded to the rating scores from 0 (on the lowest end of the scale, referring to the "least-*p*") to 100 (on the highest end of the scale, referring to the "most-*p*"). Linear correlations between liking rating and all other ratings were indexed using the Pearson correlation coefficient, since the rating scores were treated as continuous variables. Additionally, mean liking ratings of individual artworks per category were compared using a 2x2 RM-ANOVA. A complementary set of analyses for the exit questionnaire included graphs indicating frequency distribution of rating responses on a Likert-scale. Main reported descriptive values were mean (M) and standard error of the mean (±SEM) for any given analysis, unless stated otherwise

4.4.2. Results

Participants spent more than half a minute viewing each artwork on average (M_{Duration} = 44.35 ±4.12 s). A 2x2 repeated-measures analysis of variance (RM-ANOVA) to check whether spatiality and physicality (as artwork categories) affected viewing duration yielded a significant difference in mean viewing duration for spatiality ($F_{(1,30)} = 29.574$, $p < .001$, $\eta_p^2 = .496$), but not for physicality ($F_{(1,30)} = .485$, $p > .05$, $\eta_p^2 = .016$) or interaction ($F_{(1,30)} = .617$, $p > .05$, $\eta_p^2 = .020$). The post hoc comparison ($t_{(20)} = 5.437$, $p < .001$) showed that participant spent more time engaging with artwork-spaces (M_{Space} = 55.50 s ±6.24 s) compared to artwork-objects (M_{Object} = 33.19 ±2.95 s), and illustrated in Fig. 2*a* and in Supplementary Fig. s6*a*.

To measure the relation between liking rating and other ratings, a set of correlational analyses were calculated using Pearson correlation. For the analysis, continuous rating scores from thirty-one participants on each individual artwork out of sixteen were treated as a single data point, resulting in $N = 496$. Results showed that liking positively and significantly correlated with all measures, namely with liking from the third-person POV ($r_p = .538$), emotional valence ($r_p = .571$), meaningfulness ($r_p = .381$), novelty ($r_p = .360$), artfulness ($r_p = .441$), complexity ($r_p = .502$), colourfulness ($r_p = .443$), perceived viewing duration ($r_p = .680$), and real viewing duration ($r_p = .183$), where all $p < .001$. Perceived viewing duration also positively correlated with real viewing duration ($r_p = .214$, $p < .001$). Cross-correlation between all ratings showed mostly positive and moderate significant correlations as indexed by $0.3 < r_p < 0.7$, but some weak correlations as indexed by $0 < |r_p| < 0.3$ were present, especially with real viewing duration (see Fig. 3 for the overall cross-correlation matrix, Supplementary Fig. S4 for individual correlation plots, and Supplementary Fig. S5 for rating scores illustrated as boxplots and density curves drawn from the individual data points).

A multiple linear regression model was built to test whether liking can be predicted by other measured variables, following the previous findings on significant correlations. The analysis consisting of the liking rating as the dependent variable and all nine other metrics as the covariates resulted in a significant model: $F_{(9, 486)} = 77.11$, $p < .001$. The overall model explained 58.1% of the variance in liking ratings as indexed by adjusted

R^2. Testing significance of individual predictors yielded four significant predictors as liking from the third-person POV ($\beta = .169$, $t = 4.467$, $p < .001$), positive emotional valence ($\beta = .237$, $t = 5.732$, $p < .001$), meaningfulness ($\beta = .123$, $t = 3.757$, $p < .001$), perceived viewing duration ($\beta = .389$, $t = 10.494$, $p < .001$); and five non-significant predictors as novelty ($\beta = .047$, $t = 1.504$, $p > .05$), artfulness ($\beta = -.028$, $t = -.726$, $p > .05$), complexity ($\beta = .034$, $t = .912$, $p > .05$), colourfulness ($\beta = .016$, $t = .508$, $p > .05$), real viewing duration ($\beta = .033$, $t = 1.543$, $p > .05$). To form the predictive model, all five non-significant predictors were removed from the model generation, and new constants were calculated based only on the significant predictors, therefore the parameter estimates became: $\hat{y} = 2.530 + 0.174x_1 + 0.261x_2 + 0.127x_3 + 0.416x_4$, where $\hat{y}$ = liking, , x_1 = liking from the third-person POV, x_2 = positive emotional valence, x_3 = meaningfulness, x_4 = perceived viewing duration.

To check whether category-specificity affected liking ratings using a 2x2 RM-ANOVA, four ratings from a single participant corresponding to four artworks from the same category level was averaged (as a common practice to average data over participants,), thus resulting in $N_{\text{Participant}} = N_{\text{Observation}} = 31$, and $df = 30$. Physicality did not alter observer's liking ratings ($F_{(1, 30)} = 2.95$, $p > .05$, $\eta_p^2 = .090$), but spatiality significantly altered the liking ratings ($F_{(1, 30)} = 20.31$, $p < .001$, $\eta_p^2 = .404$), and observers liked spaces more compared to objects ($M_{\text{Space}} = 59.54 \pm 1.61\%$; $M_{\text{Object}} = 47.32 \pm 1.75\%$). A significant interaction effect was also present ($F_{(1, 30)} = 16.61$, $p < .001$, $\eta_p^2 = .356$), implying that liking ratings of two levels of physicality differed across the two levels of spatiality. The difference between $\text{Object}_{\text{Physical}}$ ($M = 40.70 \pm 2.11\%$) and $\text{Space}_{\text{Physical}}$ ($M = 61.57 \pm 2.09\%$) was pronounced in the post-hoc comparisons ($t_{(56.8)} = -6.06$, $p < .001$), whereas no significant difference between $\text{Object}_{\text{Digital}}$ ($M = 53.98 \pm 2.67\%$) and $\text{Space}_{\text{Digital}}$ ($M = 57.52 \pm 2.43\%$) was observed ($t_{(56.8)} = - 1.03$, $p > .05$). See Fig. 2*b* for average liking scores per artwork category, and Supplementary Fig. S6*b* for average liking scores per individual artwork.

Ratings from the exit-questionnaire on the five-point Likert were used to calculate the most frequent responses, and reported as a percentage for the highest frequency choice in brackets alongside the given questionnaire items. In terms of the amount of difficulty for each rating judgment (Supplementary Fig. S7*a*) based on the most frequent responses, overall, participants found that (i) liking (45%), complexity (48%), and colourfulness (74%) were not at all difficult; (ii) positive emotional valence (39%), meaningfulness (45%), novelty (32%), and perceived viewing duration (48%) were slightly difficult; (iii) artfulness (35%) and liking from the third-person point of view (39%) was moderately difficult. In terms of general attitudes (Supplementary Fig. S7*b*), overall, participants (i) strongly agreed that viewing experience in VR was enjoyable (81%); (ii) strongly disagreed that the experiment was boring (48%); (iii) disagreed that judgment tasks were challenging (32%); (iv) strongly disagreed that video games are not art (45%); (v) agreed that anything can be art (39%); (vi) disagreed that all public art objects/spaces should be digitized and available online (29%); (vii) disagreed that art museums and galleries are losing their significance (52%); (viii) disagreed that aesthetic experience cannot be investigated empirically (35%). Note that, although part of these results indicated an overall positive response on the VR experience and the experiment, this finding might be prone to novelty effects of VR or experimenter bias. In terms of items related to the frequency of exposure to arts (Supplementary Fig. S7*c*), overall, participants (i) sometimes visit art museum, art galleries, or art events (45%); (ii) never view art digitally (32%); (iii) seldom pursue an artistic activity or a hobby, such as painting or participating workshops (48%); (iv) never (26%) or sometimes (26%) play video games; and (v) never use VR (55%). Open-ended feedback yielded an overall liking of the experiment and diversity of artworks in particular as commented by multiple participants. Various keywords from the feedback were extremely interesting, very enjoyable, interactive, amazed by the level of immersion, etc. Minor drawbacks such as rare connection issues and a potentially better UI for the on-screen interactive questionnaire were also noted by some participants. Lastly, individual fixation duration plots as heatmaps showed viewing strategy differences among participants for each artwork. To briefly demonstrate the individual differences, exemplar heatmaps as a visualisation of fixation duration for two artworks from two randomly selected participants were plotted (Fig. 4).

In summary, the main findings of Experiment 1 indicated that (i) participants spent more time viewing spatial artworks compared to objects, (ii) liking rating showed a linear and positive relation with all other judgment types, (iii) a linear model to predict liking ratings can be based on four judgments (namely, liking from the third-person POV, emotional valence, meaningfulness, perceived viewing duration), (iv) although a relatively high variance was present for liking ratings per artwork subsets, an interaction between physicality and spatiality was observed, where the liking rating difference between spatiality depends on whether the artwork is physical or digital (such that spatial artworks were preferred more compared to artwork objects if they were physical artworks, but not if they were digital artworks), (v) the level of difficulty was reasonably low for judgment types, and (vi) participants showed diverse opinions about presented arguments related to arts.

4.5. Experiment 2: Aesthetic judgments in an online setting

4.5.1. Methods

4.5.1.1. Participants

120 people (60 females, 60 males, M_{Age} = 33.28 years, SD_{Age} = 11.51 years, R_{Age} = 18-65 years) were recruited for the online experiment via Prolific (prolific.co), an online participant recruitment tool. All participants were naïve to the hypotheses of experiments, and they were compensated monetarily (equivalent to £5/hour for a 5-to10-minute-long experiment). All were from the UK, as selected via Prolific pre-screening. All participants provided informed consent prior to the experiment. All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki.

The sample size for the online study is based on the assumption to equate the total number of observations between the VR and the online experiments (i.e., N_{Stimulus} X $N_{\text{Participant}}$): previously, 16 artworks were presented to 31 participants in the VR setting, here, for 4 artworks were presented to 120 participants in the online setting.

4.5.1.2. Stimulus and material

One artwork from each category, aimed to represent each of the 2x2 categories, was selected from the sixteen artworks of the lab-based experiment, resulting in four artworks (number 1, 8, 12, and 14 from the lab experiment, see Fig. 1). Due to the nature of the online experiment, artworks were only displayed as 2D static images. Stimuli and the questionnaire were created on a simple online form using Google Forms and viewed on a screen of personal monitors. Responses were recorded via the keyboard of personal computers.

4.5.1.3. Design

In a correlational experimental design, as before, eight variables were ratings on aesthetic judgments: liking, liking from the third-person point of view, positive emotional valence, meaningfulness, novelty, artfulness, complexity, and colourfulness per artwork. Viewing duration could not be measured as the rating scales were simultaneously presented with the snapshot images of the artworks, to minimize the online experimental duration, thus the viewing time and decision time could not be separated.

4.5.1.4. Procedure

Following a written briefing and receiving a consent form from the participants online, the experimental workflow consisted of two brief phases: in the first phase, each artwork was presented as a 2D snapshot image of the 3D model, and participants viewed the snapshots without any time constraints. Rating questions per artwork were displayed beneath the images, and participants were asked to rate on eight aesthetic judgments for each of the four artworks (see Supplementary Fig. S9 for a diagrammatic view of the judgment task). For the sake of simplicity, the ratings were on 5-point Likert scales, instead of a continuous interactive interface presented in Experiment 1. We aimed to test whether the correlations between judgment ratings in a simplified online experiment were comparable to the results in the more nuanced lab-based experiment. The second phase was a brief exit questionnaire, where participants were asked to rate the level of difficulty for each type of judgment, again on a 5-point Likert scale, and the same as in Experiment 1. This phase also contained two open-ended questions asking for (i) any other terms, including adjectives or metaphors which might be useful in describing and judging artworks; and (ii) any liked or disliked aspects of the experiment. The experiment was conducted without a time limitation, but on average the time of completion was expected to be approximately five minutes, based on previous piloting.

4.5.1.5. Data analysis

Apart from descriptive statistics, for each judgment task, linear correlations between liking rating and all other ratings were indexed using the Spearman correlation coefficient (as ratings were discrete variables). Software used for the data analysis was R and jamovi.

4.5.2. Results

The overall duration of the experiment was around five minutes with large variance (M_{Duration} = 268.32s, SD_{Duration} = 137.82s), where duration minima and maxima were approximately two and seventeen minutes.

For the correlation analysis, ratings of each individual artwork were treated as a single data point, resulting in $N = 480$ (see Fig. 5 for the overall cross-correlation matrix, and Supplementary Fig. S10 for individual correlation plots). Similar to the VR based experiment, results showed that liking positively and significantly correlated with all seven measures, namely with liking from the third-person POV ($r_s = .599$), positive emotional valence ($r_s = .442$), meaningfulness ($r_s = .503$), novelty ($r_s = .410$), artfulness ($r_s = .647$), complexity ($r_s = .382$), and colourfulness ($r_s = .262$), where all $p < .001$. Note that, although the correlations were significant both in Experiment 1 and here in Experiment 2, the relation seemed to be relatively less pronounced here mainly due to higher variance present in the ordinal data (compare Supplementary Fig. S4 of Experiment 1 and Supplementary Fig. S10 of Experiment 2). Overall, the correlational patterns were similar (also see Supplementary Fig. S11 for a similarity assessment of the correlations between VR and online settings).

Building a prediction model was slightly different compared to the first experiment in VR, since here the ratings were on a 5-point Likert scale, the ratings cannot be treated as continuous variables for a linear regression. Instead, following the findings on significant correlations, an ordinal logistic regression can be used. A model can be built to test the probability of a liking rating occurring given the known values of the other ratings. The analysis consisting of the liking rating as the dependent variable and all seven other metrics as covariates resulted in a significant model: $\chi^2_{(7)} = 366.006$, $p < .001$, $R^2_{McF} = .245$, where R^2_{McF} refers to McFadden R^2 and is not analogous to the R^2 in multiple linear regression. Testing significance of individual predictors yielded four significant predictors: liking from the third-person POV ($\beta = .122$, $z = 7.244$, $p < .001$, meaningfulness ($\beta = .121$, $z = 3.318$, $p < .001$), novelty ($\beta = .095$, $z = 2.300$, $p = .021$), artfulness ($\beta = .844$, $z = 6.724$, $p < .001$); and three non-significant predictors as positive emotional valence ($\beta = .084$, $z = .761$, $p > .05$), complexity ($\beta = -.075$, $z = -.845$, $p > .05$), colourfulness ($\beta = .060$, $z = .996$, $p > .05$). Note that, the two significant predictors as liking from the third-person POV and meaningfulness were the same as the previous model from the VR-based experiment, but here, the two other predictors were novelty and artfulness instead of emotional valence and perceived viewing duration (which was not a measured rating here).

In terms of the level of difficulty in making the judgements, based on the most frequently provided responses (where the percentage of responses reported in brackets), participants found (i) colourfulness (62%), liking (49%), positive emotional valence (32%), and artfulness (30%) were not at all difficult; (ii) complexity was slightly to not-at-all difficult (30%:30%), (iii) novelty was slightly difficult (34%), and (iv) liking from the third-person point of view (32%) and meaningfulness were moderately difficult (35%), (see Supplementary Fig. S12). Open-ended feedback yielded an overall liking of the experiment and the simplicity of the design. Various keywords from feedback were enjoyable, fun, easy, etc. A minor suggestion was a potential addition of other viewing angles per artworks, and another participant suggested to have more "traditional art" in the stimulus set. When asked to provide other terms for describing and judging the artworks, an extensive list of suggestions was produced, some of which were: provocative, inspirational, absorbing, deep, soothing, nostalgic, strange, etc. Some particular suggestions were disturbing, sad, deceptive, confusing, and dark, which may be inspirational for a relatively understudied research direction on aesthetic judgments associated with negative connotations (see Supplementary Fig. S8 for all suggested keywords visualised as a word cloud).

In summary, the main findings of Experiment 2 indicated that (i) liking rating again showed a linear and positive relation with all other judgment types, similar to the lab-based VR experiment, (ii) a logistic model to predict liking ratings can be based on four judgments (namely, liking from the third-person, meaningfulness, novelty, artfulness), (iii) the level of difficulty was again reasonably low for judgment types, and (iv) participants were able to suggest a diverse set of keywords which might be useful in aesthetic judgments.

4.6. Discussion

We began this work by demonstrating in the pilot experiment that adapting a 2AFC task on the judgment of interest could provide an alternative way of generating a stimulus set, whilst aiming to minimize stimulus selection bias. The first lab-based experiment in VR showed that the liking ratings significantly and moderately correlated with various other judgments with positive denotations such as positive emotional valence, meaningfulness, or novelty. A multiple linear regression model suggested that liking ratings can be predicted by some of those judgments (in this case, by liking from the third-person POV, emotional valence, meaningfulness, and perceived viewing duration ratings). Although the experimental design did not include an in-depth analysis of eye-tracking data, fixation duration visualised as 3D heatmaps, showed diverse viewing strategies of the artworks. Supplementary questionnaire results provided insights into the diverse opinions of participants towards arts, and also an overall positive attitude towards the experiment itself, such that for example, the level of difficulty for the judgments was reasonable for all tasks, which implies the feasibility of the presented experimental method for future studies. The second experiment, in the form of an online follow-up, resulted in similar and comparable correlational trends between ratings, suggesting that reducing the immersiveness of the artwork presentation medium (from the VR environment to the 2D image) did not radically change the relation between measured aesthetic ratings. An ordinal logistic regression model suggested that liking ratings can be predicted again by some judgments (in this case, by liking from the third-person POV, meaningfulness, novelty, artfulness ratings); a partially overlapping finding compared to the first VR-based experiment. Similar to the lab-based experiment, the level of difficulty for the judgments was again reasonable for all tasks in the online experiment. Additionally, participants were able to provide a diverse set of terms including adjectives or metaphors to be used in describing and judging artworks.

In terms of using artworks as stimulus, our approach had a relatively unconventional methodology compared to psychophysical tradition, where generally well-controlled or categorizable stimuli and high internal validity were often sought: for example, generating random dot textures with varying visual complexity levels (Friedenberg & Liby, 2016) or comparing representational and abstract artworks (Schepman et al., 2015). Since the physical properties activating the senses can be generally described as an initial state of the aesthetic experience, these properties are often linked to the experience and the judgment of art. Following on from that, one common rationale behind well-controlled stimuli sometimes relies on the assumption that some intrinsic (physical) properties of artworks are the main (or only) factors that shape both the experience and the judgment of art. However, similar bold assumptions have recently started to be criticized, and for example, a conceptual dissociation between the evaluation of artworks (as a specific research case) and the aesthetic experience (as a more broad research area) is proposed (Skov & Nadal, 2020). In this study, the stimuli selection process resulted in the inclusion of, for example, the bust of Nefertiti, a photogrammetric model of a graffiti wall, and an abstract digital sculpture resembling a trefoil knot. These items do not share much similarity or have an obvious common property. This intentional divergence across properties makes it even more interesting that we found relationships between aesthetic judgments themselves, irrespective of artwork properties or categories. Put differently, instead of parametrically modifying physical properties to see how judgements depend on these changes, we relied on the existing or "natural" variation of such properties to see how the different judgements are linked together. More broadly, investigating higher-level associations of aesthetic judgements separate from physical properties of visual arts can be indeed a meaningful empirical research context.

Following viewing artworks in VR or online, participants were able to provide elaborative judgments on rating scales for all given judgment items using a rating task, which aimed to eliminate serial dependence bias present in aesthetic judgments (Kim et al., 2019). In both experiments, participants were able to provide their judgments without much effort, as indexed by their responses indicating a low level of difficulty on the questionnaire. Since we mainly compared perceptual judgements across a wide range of artworks, relying only on the correlational analysis between judgment ratings (some of which can be defined as "perceived" properties) without measuring any "physical" properties of artworks avoided having to ask some of the questions raised by previous research, for example, whether authenticity and presentation context of artworks effects participant ratings (Brieber et al., 2015), whether participant assumed the 3D objects as original or reproduction (Locher, 2016), or indeed whether participants classified 3D objects as art (Pelowski et al., 2017).

Regarding the conceptual link between physical properties of artworks and aesthetic judgments, one exception in line with this link in our analysis for the VR experiment was following the 2x2 categorization of presented artworks as spatiality (whether the artwork is large-scale or small-scale) and physicality (whether the artwork is a modelled version of a real-world object/space or a digital-only): the observed liking rating difference between spatiality depends on whether the artwork is physical or digital, and the difference between artwork-objects and artwork-spaces was pronounced only for physical artworks but not for digital artworks, but arguably, participants might not easily distinguish between physical and digital artworks, which were not introduced to participants. More specifically, irrespective of those categories formed by the researchers, in the VR experiment, all presented artworks might just be labelled as virtual by participants, whereas in the online experiment a common label might be digital. Additionally, the finding of a significant positive relationship between personal liking ratings and liking ratings from the third-person point of view suggest that people assume that other people think like them, and more speculatively, imply that a type of aesthetic normativity was present in the experiment. From this perspective, explicit measures as a part of the experimental design can capture commonalities between types of judgments, and may also contribute to evaluating conceptual frameworks related to aesthetic judgments.

On the other hand, in both experiments, although the regression models aimed to predict liking ratings did not depend on all the rating scores for a best fit, the existence of a significant, positive, and linear relationship between liking and all other ratings (albeit explaining only a medium amount of variance) does not entirely tally with existing research. For example, an inverted U-curve relation between complexity and preference has often been suggested (Berlyne, 1958; Vitz, 1966; Güçlütürk et al., 2016) and was not present in our results. Some potential explanations might be (i) the stimuli set did not cover the full range of (measurable) complexity levels, since the stimulus selection procedure did not specifically aim for it, (ii) participants might have treated the concept of complexity not only as visual complexity, and therefore assigned varied meanings to it regarding other associated words such as complicated or hard to understand, among others.

Additionally, two main types of judgments can be formulated, either related to artworks' properties as perceived by the observer (such as pleasantness and interestingness) or related to properties as measurable features (such as colour and form). Some of those properties related to either of the judgment types can be treated as more elemental concepts and might be merged into relatively canonical factors such as arousal or regularity, respectively, for example using factor analysis on judgments (Marković & Radonjić, 2008). Although building an overarching model is beyond the scope of this research, we can speculate that a potential inclusion of additional judgments with a positive connotation (such as interesting, successful, engaging, impressive; some of which were already suggested by the participants, see Supplementary Fig. S8) would still give similar results (namely, positive correlation with liking ratings). From this point of view, participants might have been using a common judgment strategy across all (or most of the) ratings, such as assigning an aesthetic value to an artwork as a common factor followed by providing isolated ratings. In this sense, targeting such potential judgment strategies, instead of isolated judgments, might be a promising research theme for future studies.

On considering viewing duration, a variable only present in the VR experiment, firstly, participants spent more time engaging with large-scale artworks (referred to as artwork-spaces) compared to small-scale artworks (referred to as artwork-objects), and the viewing durations were around 55 s and 33 s, respectively. Especially for the small-scale artworks, viewing duration in VR was similar to the real-world scenario, for example, a museum-based research with a large sample size of 456 visitors found the average viewing duration as 28.63 s (Smith et al., 2017), where the researchers also underline the large variance between participants and between different artworks, and attributed these arguably brief viewing durations to visitors' potential need for rapid art consumption. Here, one common-sense interpretation of duration increase for large-scale artworks might be that large-scale artworks simply provided more area to explore, in line with a previous finding from another museum-based study where for larger viewing angles (as a derivative metric from painting size and viewing distance), a trend of longer viewing times were observed (Carbon, 2017). Additionally, although the duration judgment (as the perceived time spent viewing artworks) was one of the moderately difficult ratings according to the exit questionnaire, it was a strong predictor of liking in the regression model. However, the weak correlation between real duration and perceived duration suggests that there were some misestimations of time and interestingly this was related to how much the artwork was liked. A related study involving a temporal reproduction paradigm in a between-groups design found a trend of duration underestimation for one group where the visual stimuli were described as "artworks" compared to another group, where the same stimuli were described as "photographs used in psychological experiments" (Arai & Kawabata, 2016). Other previous research found expert-novice differences, such that the trained participants underestimated and naive participants overestimated the viewing duration of paintings, and conceptually linked this to perceptual and cognitive effort (Cupchik & Gebotys, 1988). Although previous research underlined the importance of temporal dynamics during aesthetic judgments (Cupchik & Gebotys, 1988; Smith et al., 2006; Muth et al., 2015), types of temporal distortion during viewing artworks, and more specifically, their relations to aesthetic judgments seem to call for further controlled experiments.

In terms of limitations and future directions, firstly, although the pilot experiment aimed to minimize the selection bias, implementing auto-generative algorithms into the experimental procedure might eliminate a potential selection bias. Emerging machine learning methods specific for media, visual arts, and cultural heritage might be relevant in terms of a stimulus generation, such as creating 3D objects from a single image (Chen et al., 2019) and constructing complex real-world scenes in 3D from a photo sample (Mildenhall et al., 2020). Secondly, the immersive experience using VR (and the ability to collect eye-tracking data) was not easily feasible for online experiments. However, the recent developments on online VR tools such as WebXR API, in line with the adoption of personal VR-HMDs with built-in eye trackers might provide more compatible online experiments soon. An immersive online experiment to show artworks as 360-degree images/videos, or as 3D environments can be created. Following on from that, a potential use of eye-tracking comparing online-2D and online-VR setups might provide a more direct measure of similarity between the two contexts. In both cases, the effects of the novelty of using VR might be minimized with longer experiments involving training sessions. Lastly, although this research was primarily concerned with the descriptive aspects of aesthetic judgments, predictive aspects beyond regression models can be further explored, especially using machine-learning-based tools: if a research direction is to build for example personalized predictive models (with a potential implication for online consumerism in art), many forms of artificial neural networks can be adapted to make better predictions about the liking judgment from other judgment types, or from supplementary measures including eye-tracking. Following the observed benefits of the multisensory interactions in VR museums (Koutsabasis & Vosinakis, 2018), immersive experimental paradigms and novel approaches to behavioural data analysis call to be extended to the multisensory research context, such as implementing immersive art pieces, video games or gamified experiences, to test for example whether aesthetic judgments hold their relationships across the senses.

Acknowledgements

The authors would like to thank Royal Holloway University of London (RHUL) for the doctoral scholarship of Doga Gulhan, and access to the Psychology VR lab.

Ethical statement

All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki. Research Ethics Committee (REC) Project ID: 1396, Year: 2018, User ID: PDJT007.

Funding statement

We received no funding for this study.

Data availability

All anonymised data are accessible via the Open Science Framework (osf.io/ec46q).

Declaration of conflicting interests

The authors declare no conflicting interests.

Author contributions

Conceptualization, DG, JMZ; Methodology, DG, JMZ; Software, DG; Formal Analysis, DG, SD; Investigation, DG, SD, JMZ; Writing – original draft, DG; Writing – review & editing, DG, SD, JMZ; Visualization, DG, JMZ; Supervision, SD, JMZ; Project management, JMZ.

Timeline

The lab-based VR experiment was conducted in May-June 2019, and the online experiment was conducted in June 2019. Please also refer to the last paragraph of the introduction (p. 31), 1st thesis chapter, for a detailed overall timeline and the effects of the pandemic on the experiments.

4.7. References

Arai, S., & Kawabata, H. (2016). Appreciation contexts modulate aesthetic evaluation and perceived duration of pictures. *Art and Perception*, *4*(3), 225–239. https://doi.org/10.1163/22134913-00002052

Augustin, M. D., Wagemans, J., & Carbon, C.-C. (2012). All is beautiful? Generality vs. specificity of word usage in visual aesthetics. *Acta Psychologica*, *139*(1), 187–201. https://doi.org/10.1016/j.actpsy.2011.10.004

Aumann, A. (2014). The relationship between aesthetic value and cognitive value: The relationship between aesthetic value and cognitive value. *The Journal of Aesthetics and Art Criticism*, *72*(2), 117–127. https://doi.org/10.1111/jaac.12073

Belchev, Z., Bodner, G. E., & Fawcett, J. M. (2018). Inducing preference reversals in aesthetic choices for paintings: Introducing the contrast paradigm. *PLOS ONE*, *13*(4), e0196246. https://doi.org/10.1371/journal.pone.0196246

Berlyne, D. E. (1958). The influence of complexity and novelty in visual figures on orienting responses. *Journal of Experimental Psychology*, *55*(3), 289–296. https://doi.org/10.1037/h0043555

Brieber, D., Leder, H., & Nadal, M. (2015). The experience of art in museums: An attempt to dissociate the role of physical context and genuineness. *Empirical Studies of the Arts*, *33*(1), 95–105. https://doi.org/10.1177/0276237415570000

Carbon, C.-C. (2017). Art perception in the museum: How we spend time and space in art exhibitions. *I-Perception*, *8*(1), 1–15. https://doi.org/10.1177/2041669517694184

Chatterjee, A., Widick, P., Sternschein, R., Smith, W. B., & Bromberger, B. (2010). The assessment of art attributes. *Empirical Studies of the Arts*, *28*(2), 207–222. https://doi.org/10.2190/EM.28.2.f

Chen, W., Ling, H., Gao, J., Smith, E., Lehtinen, J., Jacobson, A., & Fidler, S. (2019). Learning to predict 3D objects with an interpolation-based differentiable renderer. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 9609–9619). Curran Associates, Inc.

Chiquet, S., Martarelli, C. S., & Mast, F. W. (2020). Eye movements to absent objects during mental imagery and visual memory in immersive virtual reality. *Virtual Reality*. https://doi.org/10.1007/s10055-020-00478-y

Cleeremans, A., Ginsburgh, V., Klein, O., & Noury, A. (2016). What's in a name? The effect of an artist's name on aesthetic judgments. *Empirical Studies of the Arts*, *34*(1), 126–139. https://doi.org/10.1177/0276237415621197

Clini, P., Ruggeri, L., Angeloni, R., & Sasso, M. (2018). Interactive immersive virtual museum: Digital documentation for virtual interaction. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII–2*, 251–257. https://doi.org/10.5194/isprs-archives-XLII-2-251-2018

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cooper, J. M., & Silvia, P. J. (2009). Opposing art: Rejection as an action tendency of hostile aesthetic emotions. *Empirical Studies of the Arts*, *27*(1), 109–126. https://doi.org/10.2190/EM.27.1.f

Cupchik, G. C., & Gebotys, R. (1988). The experience of time, pleasure, and interest during aesthetic episodes. *Empirical Studies of the Arts*, *6*(1), 1–12. https://doi.org/10.2190/5YN3-J3P8-FWHY-UDB3

Friedenberg, J., & Liby, B. (2016). Perceived beauty of random texture patterns: A preference for complexity. *Acta Psychologica*, *168*, 41–49. https://doi.org/10.1016/j.actpsy.2016.04.007

Grüner, S., Specker, E., & Leder, H. (2019). Effects of context and genuineness in the experience of art. *Empirical Studies of the Arts*, *37*(2), 138–152. https://doi.org/10.1177/0276237418822896

Güçlütürk, Y., Jacobs, R. H. A. H., & Lier, R. van. (2016). Liking versus complexity: Decomposing the inverted U-curve. *Frontiers in Human Neuroscience*, *10*, 112. https://doi.org/10.3389/fnhum.2016.00112

Hayn-Leichsenring, G. U. (2017). The ambiguity of artworks: A guideline for empirical aesthetics research with artworks as stimuli. *Frontiers in Psychology*, *8*, 1857. https://doi.org/10.3389/fpsyg.2017.01857

Helbing, J., Draschkow, D., & Võ, M. L.-H. (2020). Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. *Cognition*, *196*, 104147. https://doi.org/10.1016/j.cognition.2019.104147

Hoang, T. N., & Cox, T. N. (2017). Alternating reality: An interweaving narrative of physical and virtual cultural exhibitions. *Presence: Teleoperators and Virtual Environments*, *26*(4), 402–419. https://doi.org/10.1162/PRES_a_00307

Isik, A. I., & Vessel, E. A. (2019). Continuous ratings of movie watching reveal idiosyncratic dynamics of aesthetic enjoyment. *PLOS ONE*, *14*(10), e0223896. https://doi.org/10.1371/journal.pone.0223896

Kim, S., Burr, D., & Alais, D. (2019). Attraction to the recent past in aesthetic judgments: A positive serial dependence for rating artwork. *Journal of Vision*, *19*(12), 19. https://doi.org/10.1167/19.12.19

Koutsabasis, P., & Vosinakis, S. (2018). Kinesthetic interactions in museums: Conveying cultural heritage by making use of ancient tools and (re-) constructing artworks. *Virtual Reality*, *22*(2), 103–118. https://doi.org/10.1007/s10055-017-0325-0

Landau, M. J., Greenberg, J., Solomon, S., Pyszczynski, T., & Martens, A. (2006). Windows into nothingness: Terror management, meaninglessness, and negative reactions to modern art. *Journal of Personality and Social Psychology*, *90*(6), 879–892. https://doi.org/10.1037/0022-3514.90.6.879

Liarokapis, F., Voulodimos, A., Doulamis, N., & Doulamis (Eds.). (2020). *Visual computing for cultural heritage*. Springer. https://doi.org/10.1007/978-3-030-37191-3

Locher, P. (2016). Comparison of the perception and evaluation of paintings viewed in the original versus reproduction formats. In S. Mastandrea & F. Maricchiolo (Eds.), *The role of the museum in the education of young adults. Motivation, emotion and learning* (pp. 27–34). Roma Tre-Press.

Locher, P., Smith, L., & Smith, J. (1999). Original paintings versus slide and computer reproductions: A comparison of viewer responses. *Empirical Studies of the Arts*, *17*(2), 121–129. https://doi.org/10.2190/R1WN-TAF2-376D-EFUH

Mallon, B., Redies, C., & Hayn-Leichsenring, G. U. (2014). Beauty in abstract paintings: Perceptual contrast and statistical properties. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00161

Marković, S., & Radonjić, A. (2008). Implicit and explicit features of paintings. *Spatial Vision*, *21*(3–5), 229–259. https://doi.org/10.1163/156856808784532563

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. *ArXiv E-Prints*, arXiv:2003.08934.

Moore, K. M., & West, A. N. (2012). Global perception, meaning, and aesthetic appreciation. *Empirical Studies of the Arts*, *30*(1), 23–38. https://doi.org/10.2190/EM.30.1.d

Muth, C., Raab, M. H., & Carbon, C.-C. (2015). The stream of experience when watching artistic movies: Dynamic aesthetic effects revealed by the Continuous Evaluation Procedure (CEP). *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00365

Palan, S., & Schitter, C. (2018). Prolific.ac: A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Parker, E., & Saker, M. (2020). Art museums and the incorporation of virtual reality: Examining the impact of VR on spatial and social norms. *Convergence: The International Journal of Research into New Media Technologies*, *26*(5–6), 1–15. https://doi.org/10.1177/1354856519897251

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pelowski, M. (2015). Tears and transformation: Feeling like crying as an indicator of insightful or “aesthetic” experience with art. *Frontiers in Psychology*, *6*, 1006. https://doi.org/10.3389/fpsyg.2015.01006

Pelowski, M., Gerger, G., Chetouani, Y., Markey, P. S., & Leder, H. (2017). But is it really art? The classification of images as “art”/“not art” and correlation with appraisal and viewer interpersonal differences. *Frontiers in Psychology*, *8*, 1729. https://doi.org/10.3389/fpsyg.2017.01729

Petrelli, D. (2019). Making virtual reconstructions part of the visit: An exploratory study. *Digital Applications in Archaeology and Cultural Heritage*, *15*, e00123. https://doi.org/10.1016/j.daach.2019.e00123

Schepman, A., Rodway, P., Pullen, S. J., & Kirkham, J. (2015). Shared liking and association valence for representational art but not abstract art. *Journal of Vision*, *15*(5), 11. https://doi.org/10.1167/15.5.11

Schöne, B., Sylvester, R. S., Radtke, E. L., & Gruber, T. (2021). Sustained inattentional blindness in virtual reality and under conventional laboratory conditions. *Virtual Reality*, *25*(1), 209–216. https://doi.org/10.1007/s10055-020-00450-w

Silvia, P. J., & Brown, E. M. (2007). Anger, disgust, and the negative aesthetic emotions: Expanding an appraisal model of aesthetic experience. *Psychology of Aesthetics, Creativity, and the Arts*, *1*(2), 100–106. https://doi.org/10.1037/1931-3896.1.2.100

Skov, M., & Nadal, M. (2020). A farewell to art: Aesthetics as a topic in psychology and neuroscience. *Perspectives on Psychological Science*, *15*(3), 1–13. https://doi.org/10.1177/1745691619897963

Smith, L. F., Bousquet, S. G., Chang, G., & Smith, J. K. (2006). Effects of time and information on perception of art. *Empirical Studies of the Arts*, *24*(2), 229–242. https://doi.org/10.2190/DJM0-QBDW-03V7-BLRM

Smith, L. F., Smith, J. K., & Tinio, P. P. L. (2017). Time spent viewing art and reading labels. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(1), 77–85. https://doi.org/10.1037/aca0000049

Turpin, M. H., Kara-Yakoubian, M., Walker, A., Gabert, N. N., Fugelsang, J., & Stolz, J. A. (2019). Bullshit makes the art grow profounder. *Judgment and Decision Making*, *14*(6), 658–670. https://doi.org/10.2139/ssrn.3410674

van Helvoort, D., Stobbe, E., Benning, R., Otgaar, H., & van de Ven, V. (2020). Physical exploration of a virtual reality environment: Effects on spatiotemporal associative recognition of episodic memory. *Memory & Cognition*, *48*, 691–703. https://doi.org/10.3758/s13421-020-01024-6

Vitz, P. C. (1966). Preference for different amounts of visual complexity. *Behavioral Science*, *11*(2), 105–114. https://doi.org/10.1002/bs.3830110204

Wagner, V., Menninghaus, W., Hanich, J., & Jacobsen, T. (2014). Art schema effects on affective experience: The case of disgusting images. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(2), 120–129. https://doi.org/10.1037/a0036126

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: A tutorial review. *PeerJ*, *3*, e1058. https://doi.org/10.7717/peerj.1058

Young, J. O. (Ed.). (2017). *Semantics of aesthetic judgements* (First edition). Oxford University Press.

Zhao, H., Thrash, T., Grossrieder, A., Kapadia, M., Moussaïd, M., Hölscher, C., & Schinazi, V. R. (2020). The interaction between map complexity and crowd movement on navigation decisions in virtual reality. *Royal Society Open Science*, *7*(3), 191523. https://doi.org/10.1098/rsos.191523

4.8. Figures

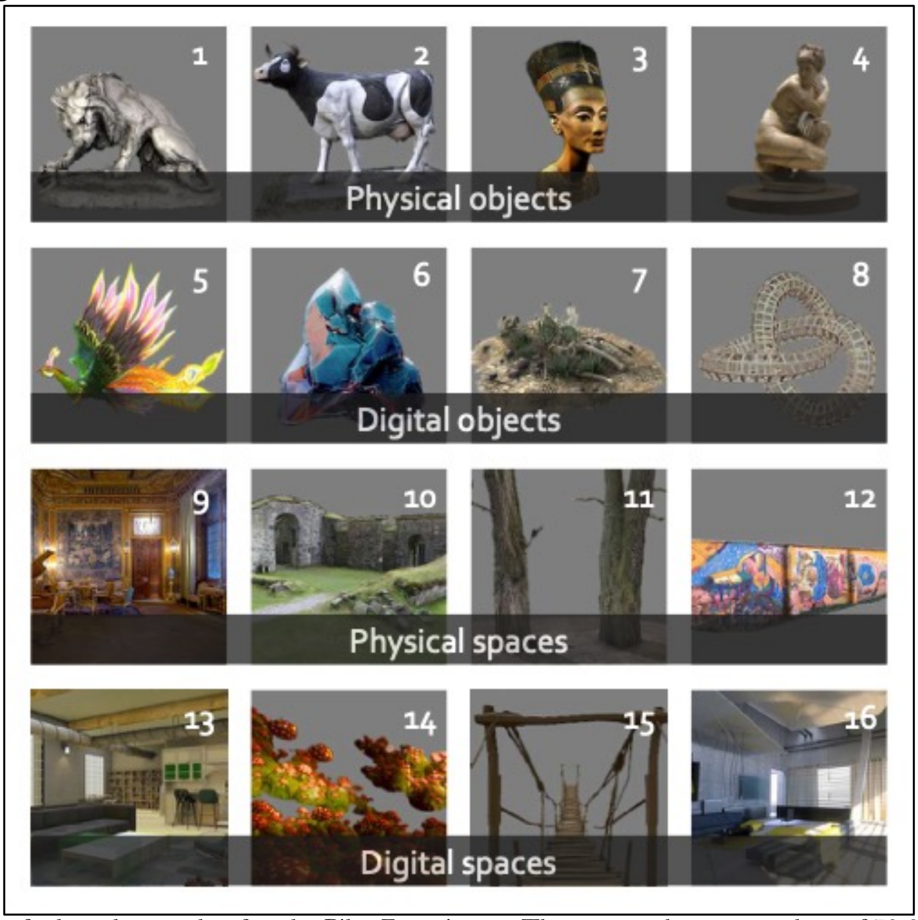


Fig. 1 Set of selected artworks after the Pilot Experiment. These artworks were a subset of 78 3D models which were found to be interesting by participants unanimously. For each of the 78 models, two descriptive tags were attached: whether the model was physical or digital (generated from an existing physical object/space using photogrammetry or not), and whether the model was an object or a space (depending on the model's relative size). Four models were randomly assigned to each category, resulting in a total set of 16 artworks to be used in Experiment 1 and 2. Also see Supplementary Fig. S1 listing attributions of individual 3D models including their titles, uploaders, and hyperlinks.

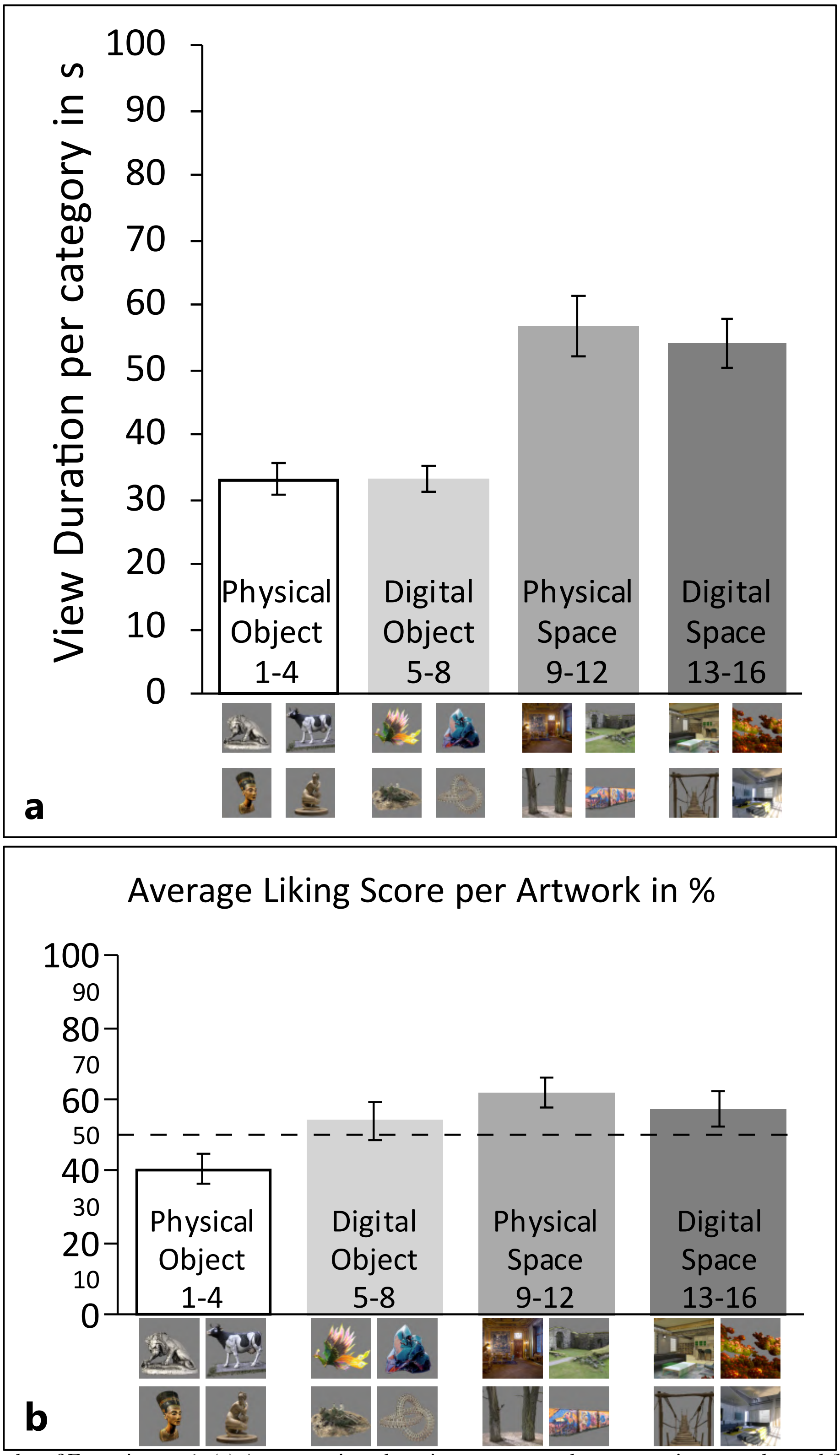


Fig. 2 Results of Experiment 1. (a) Average view duration per artwork category in seconds, and (b) average liking score binned into artwork categories were illustrated as bar graphs, indicating means with error bars showing ±1 standard error of the mean (SEM). In the second panel, a dashed line parallel to the x-axis denotes the mid-value of liking ratings. If the mid-value can be defined as a threshold point of a "neither liked nor disliked" artwork, then an artwork with a mean rating score (including variance) above the threshold can be defined as an overall liked artwork, and vice versa. The sample size was $N_{\text{Participant}} = 31$, and the number of observations for each category (out of four) was $N_{\text{Observation}} = 4$.

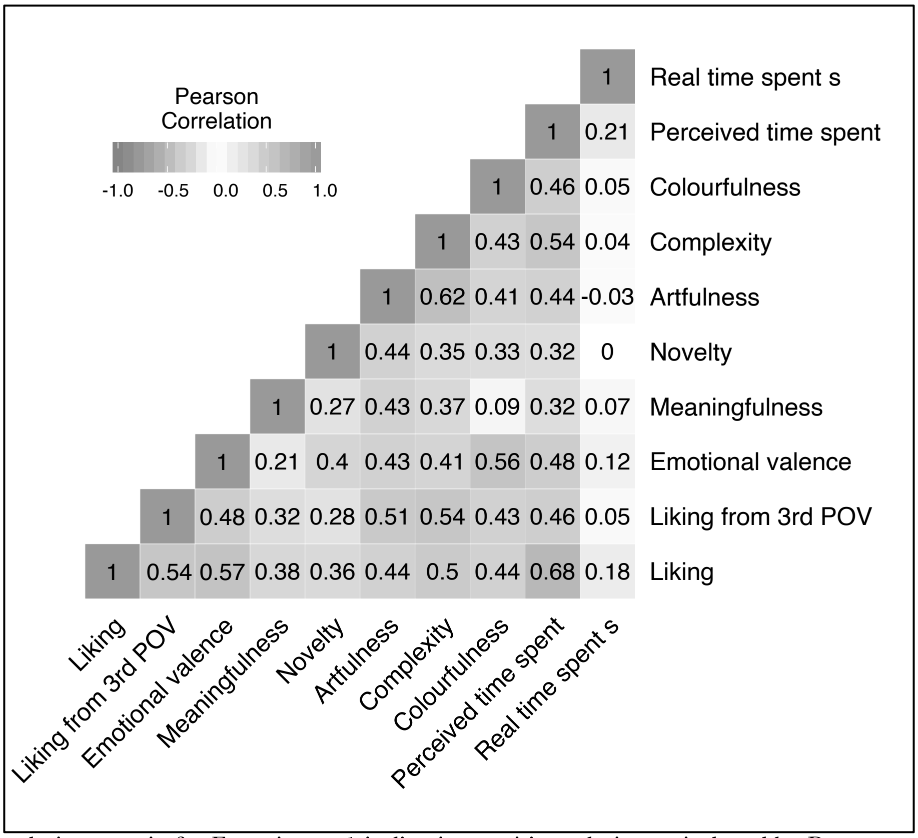


Fig. 3 Correlation matrix for Experiment 1 indicating positive relation as indexed by Pearson correlation, between all nine ratings that participants provided and one measured variable as real view time. Most of the individual responses showed positive and moderate cross-correlation to one another as indexed by $0.3 < r_p < 0.7$, except for real viewing time, which only showed either weak or no correlation to other variables as indexed by $0 < |r_p| < 0.3$. The sample size was $N_{\text{Participant}} = 31$, and the number of observations was $N_{\text{Observation}} = 16$.

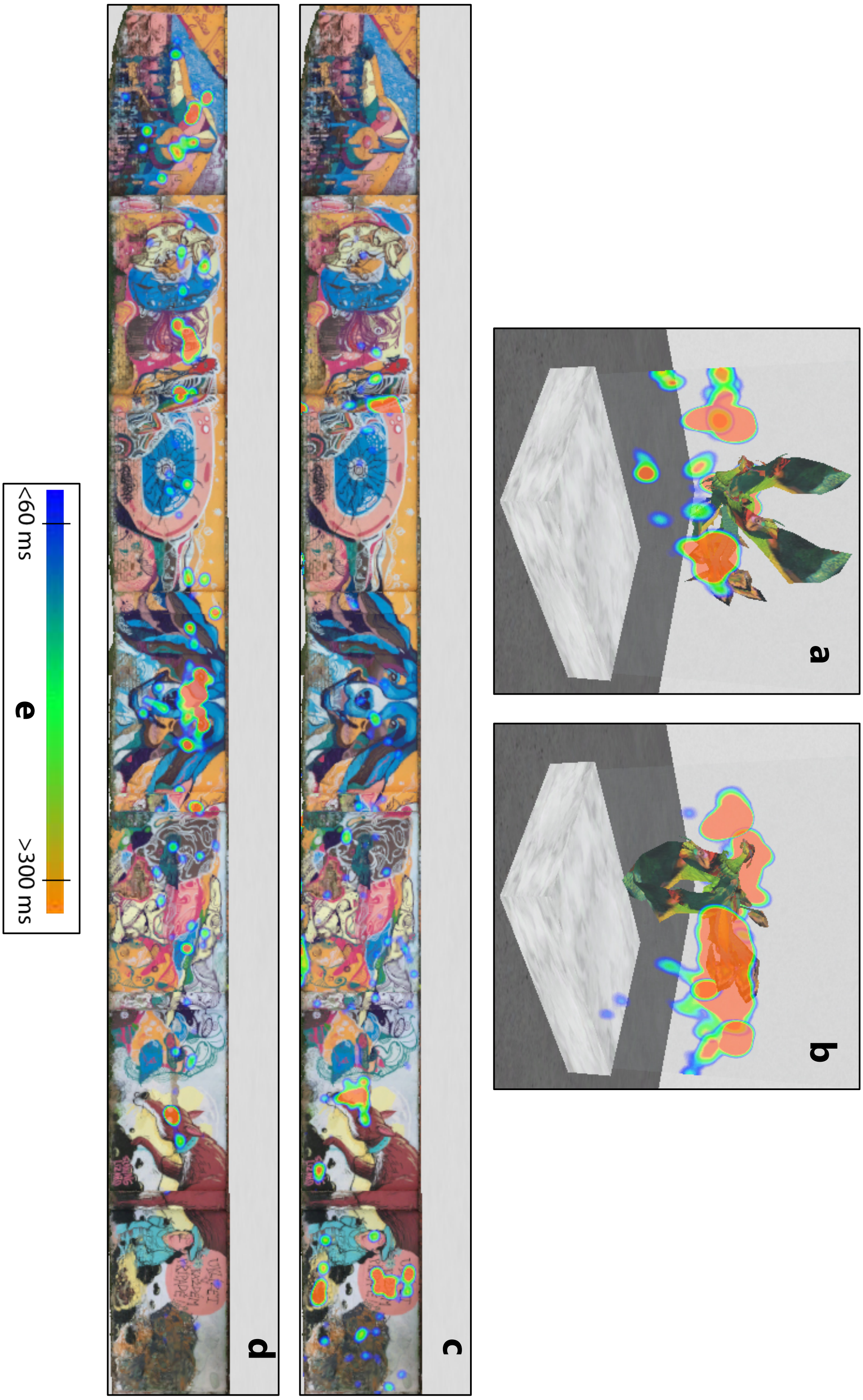


Fig. 4 Exemplar heatmaps from Experiment 1. (a-b and c-d) The visualisation of fixation duration for two artworks (as artwork #5 and #12) from two randomly selected participants with (e) a fixation duration scale ranging between 60 ms and 300 ms was generated to demonstrate various viewing strategies employed by participants. The heatmaps were created for participant's entire viewing time (here, 55 s and 105 s for a-b; 70 s and 71 s for c-d). Note that, artwork #5 was a moving digital artwork (with an animated wing motion) and displayed inside a glass box in VR, therefore the corresponding heatmaps were on the surface of this cuboid box, and not on top of the artwork surface.

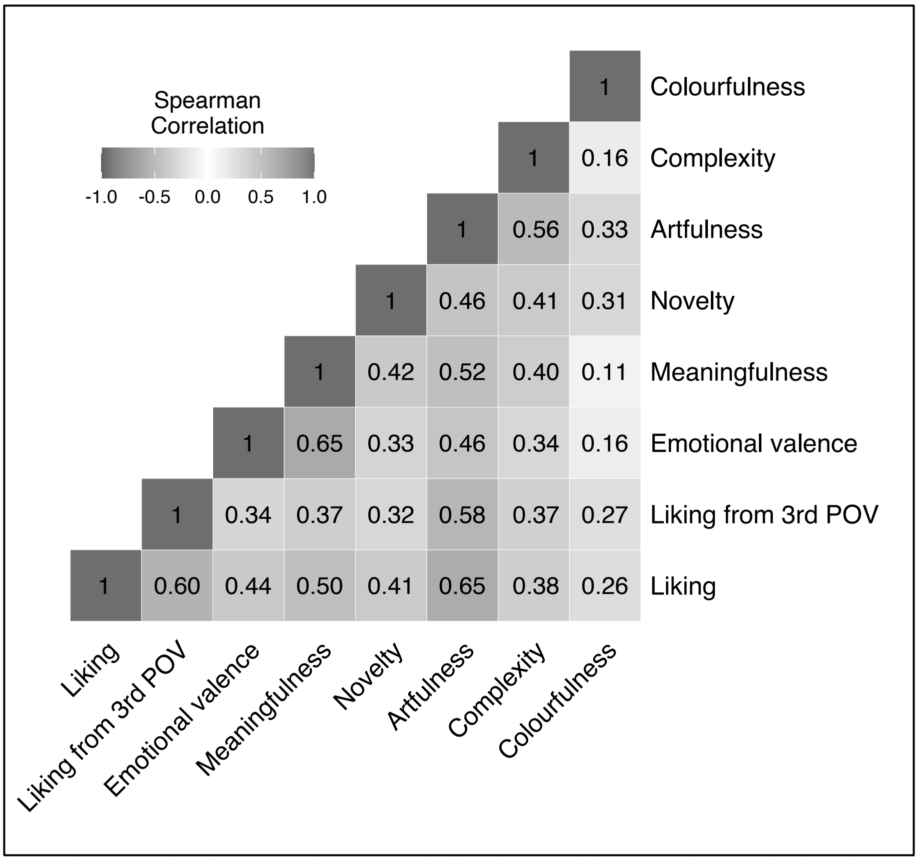
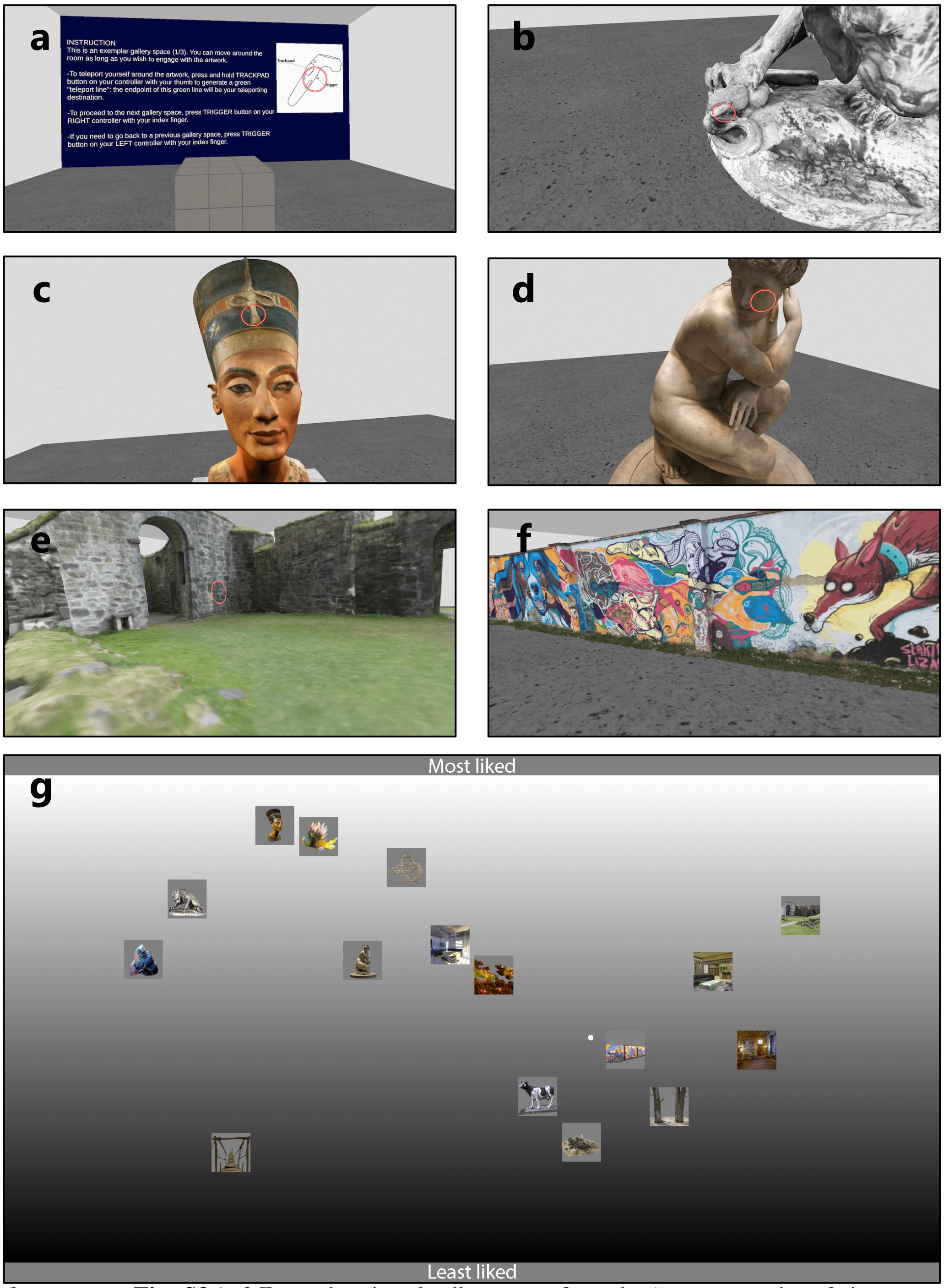


Fig. 5 Correlation matrix for Experiment 2 indicating positive relation as indexed by Spearman correlation, between all eight ratings that participants provided. Most of the individual responses showed positive and moderate cross-correlation to one another as indexed by $0.3 < r_s < 0.7$, with some exception of weak or no correlation as indexed by $0 < r_s < 0.3$. The sample size was $N_{\text{Participant}} = 120$, and the number of observations was $N_{\text{Observation}} = 4$.

4.8. Supplementary figures

No	Name of 3D Model	Username of Uploader	Hyperlink
1	Lion Crushing a Serpent	Rigsters	https://skfb.ly/68s9T
2	Plastic cow - statue	3dhdscan	https://skfb.ly/68YoO
3	Nefertiti's bust	C. Yamahata	https://skfb.ly/Mn7L
4	Aphrodite Crouching, British Museum	Thomas Flynn	https://skfb.ly/CVTr
5	phoenix bird	NORBERTO-3D	https://skfb.ly/6vLBp
6	Crystal stone	GenEugene	https://skfb.ly/6CsxD
7	Remains	seenoise	https://skfb.ly/6yoGs
8	Abstract 2- Torus Knot	Mike Rowley	https://skfb.ly/LOvy
9	The Great Drawing Room	The Hallwyl Museum (Hallwylska museet)	https://skfb.ly/6ypJL
10	Olavskirken ruin	Kulturarv i Vestfold og Telemark fylkeskommune	https://skfb.ly/68MPH
11	Trees in the park Anthropos	3dhdscan	https://skfb.ly/6uUwD
12	Graffiti Wall Mendlak Brno	3dhdscan	https://skfb.ly/6sSBx
13	Big Room	Francesco Coldesina	https://skfb.ly/6toBv
14	Mushroom Fields	d880	https://skfb.ly/KzFK
15	Bridge	filipeb2011	https://sketchfab.com/3d-models/bridge-ca35b5ecb93140a4b02a46f2e320093d
16	Mirror's Edge Apartment - Interior Scene	Aurélien Martel	https://skfb.ly/YZoC

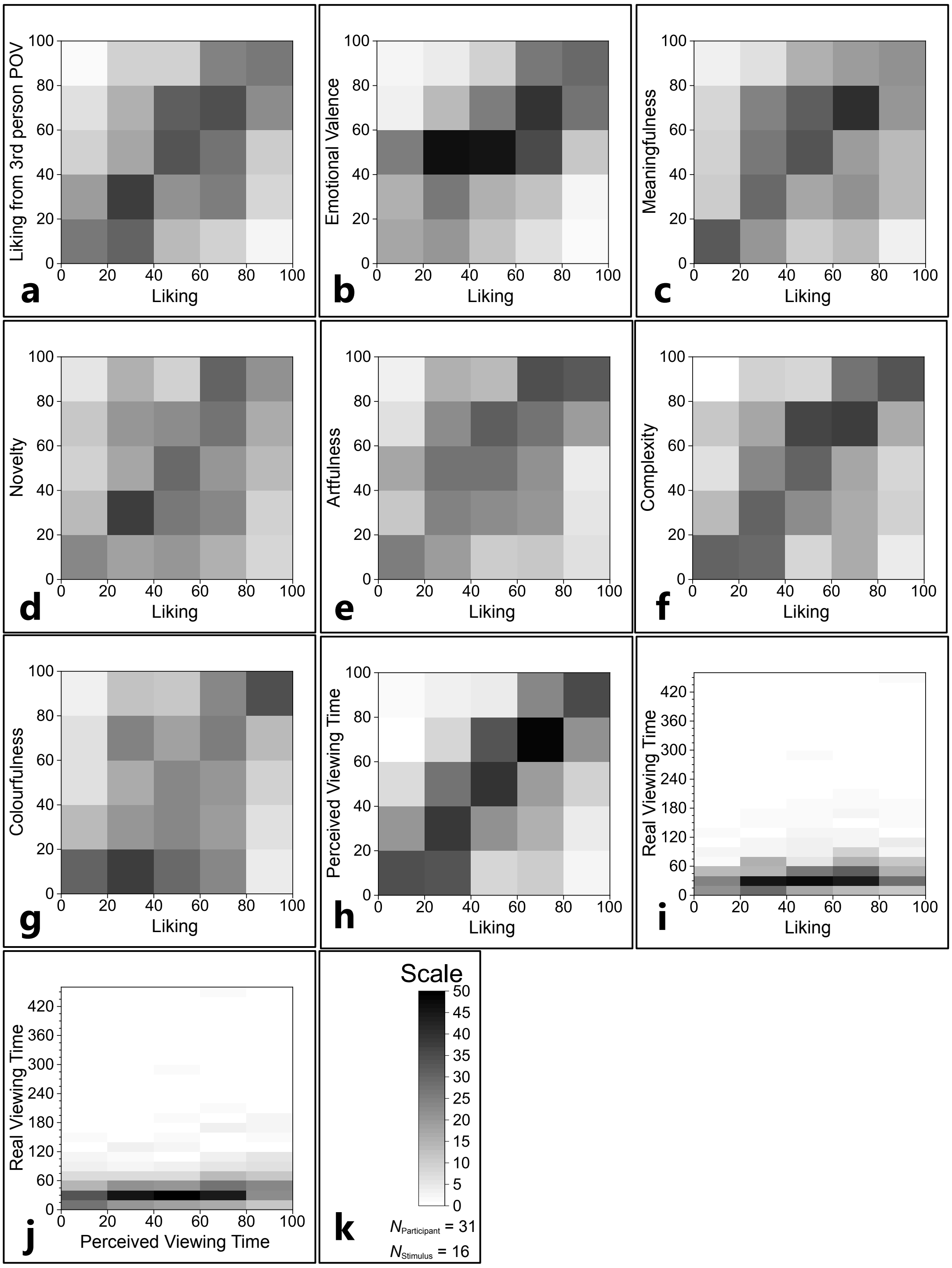
Supplementary Fig. S1 List of all selected 3D models, indicating 3D model names, usernames of uploaders, and direct hyperlinks to the webpages hosting the models.



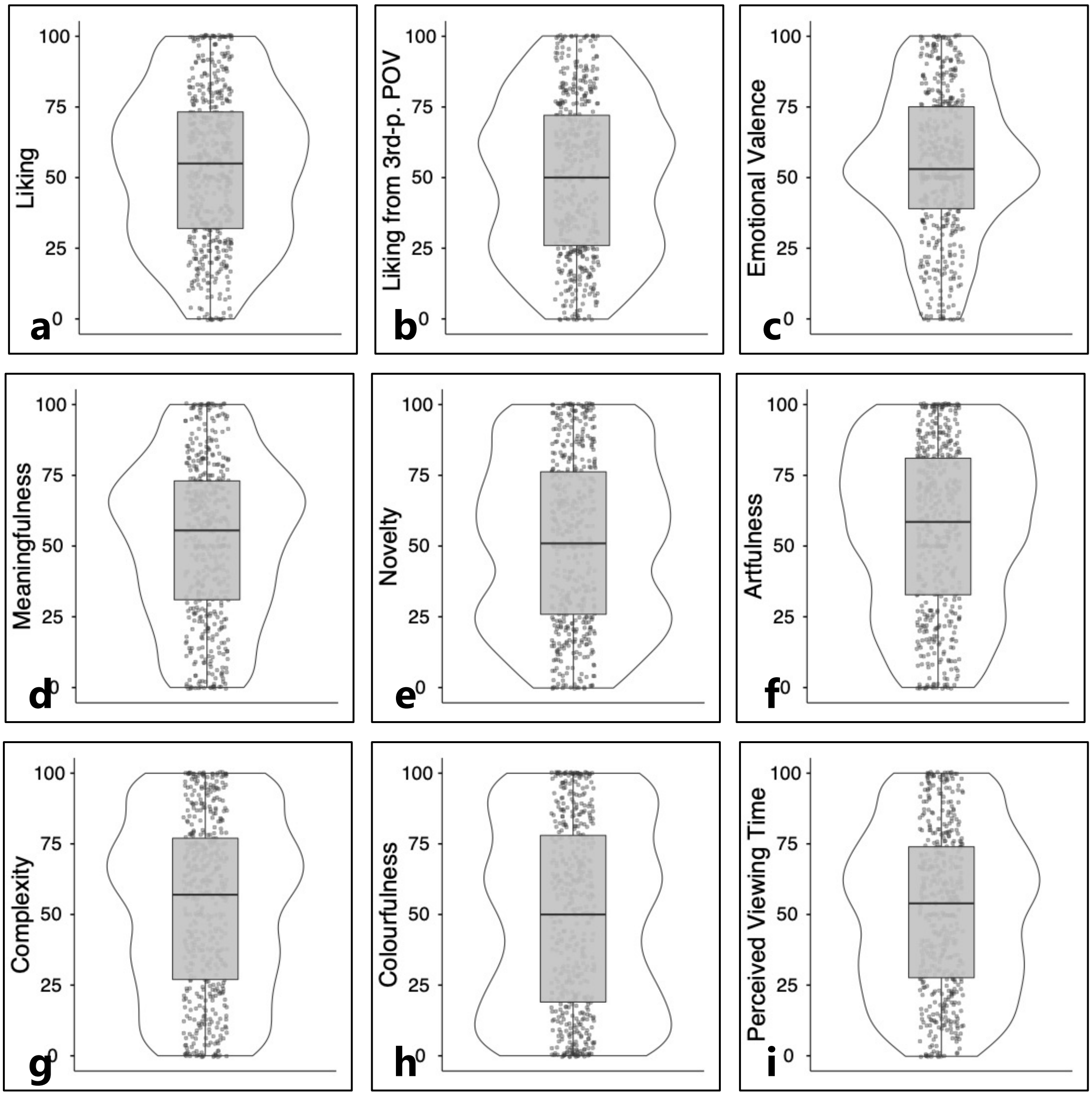
Supplementary Fig. S2 (a-f) Exemplar virtual gallery space from the 1st person point-of-view, snapshots were taken from instructional gallery space (a), from artwork #1 (b), #3 (c), #4 (d), #10 (e), and #12 (f) during the VR pre-screening phase of Experiment 1, where participants engaged with individual artworks one by one in a virtual white-box gallery space. Here, the red circles on the artworks indicate fixations to illustrate the heatmaps' input as 3D fixation coordinates, and these fixational circles were not visible to the participants during the experiment. (g). Exemplar view from the judgment task phase of Experiment 1. Here, participants were able to see and enlarge all the artworks simultaneously and they provided ratings on a "most *p* / least *p*" scale (such as most liked vs. least liked) by dragging and dropping the snapshot images of the artworks. Also see supplementary figure s1 listing attributions of individual 3D models including their titles, uploaders, and hyperlinks.

Instruction: This experiment is about aesthetic perception. You will be asked to provide various judgments on objects and spaces which you saw earlier in VR. There will be no time limit, but an estimated time of completion is 15 minutes. Each experimental trial will be an interactive questionnaire having an intuitive procedure: In each trial, small images of all objects and spaces will be displayed at the middle-centre of the screen. A background of a black-to-white gradient will also be present. Two rating keywords will also be presented on the top and bottom of the screen, such as most liked and least liked. Using your left mouse button, you can slide these images up and down, indicating your rating. You can fine-tune your choices as long as you want. If you wish, you can right-click on an image which will expand the thumbnail for further inspection. Please press a key to start the experiment.
Judgment 1/9: Please arrange the following images on a scale between MOST LIKED to LEAST LIKED.
Judgment 2/9: Please arrange the following images on a scale between MOST LIKED BY OTHERS to LEAST LIKED BY OTHERS. Only for this question, your judgment will be based on your expectancy on how other participants would rate the images.
Judgment 3/9: Please arrange the following images on a scale between EMOTIONALLY POSITIVE to EMOTIONALLY NEGATIVE.
Judgment 4/9: Please arrange the following images on a scale between MOST MEANINGFUL to LEAST MEANINGFUL.
Judgment 5/9: Please arrange the following images on a scale between MOST NOVEL to LEAST NOVEL.
Judgment 6/9: Please arrange the following images on a scale between MOST ARTFUL to LEAST ARTFUL.
Judgment 7/9: Please arrange the following images on a scale between MOST COMPLEX to LEAST COMPLEX.
Judgment 8/9: Please arrange the following images on a scale between MOST COLOURFUL to LEAST COLOURFUL.
Judgment 9/9: Please arrange the following images on a scale between MOST TIME SPENT during viewing to LEAST TIME SPENT during viewing, by guessing your viewing duration in VR.

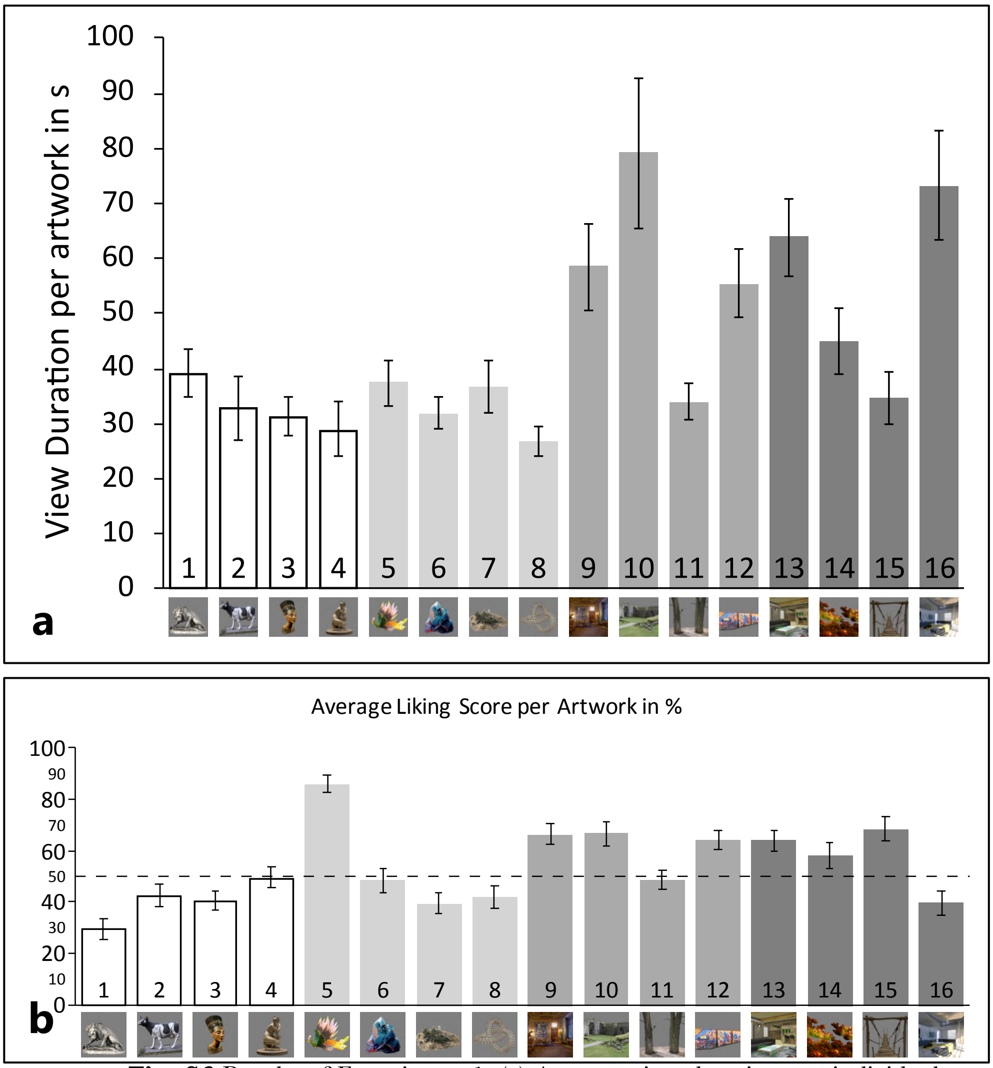
Supplementary Fig. S3 An overview of judgment questions in Experiment 1. Note that, except from the clarification on the liking from the third-person point-of-view (referred to as liking by others in Judgment 2/9), types of judgments were not explicitly defined or described to the participants, to minimize a form of instructional bias.



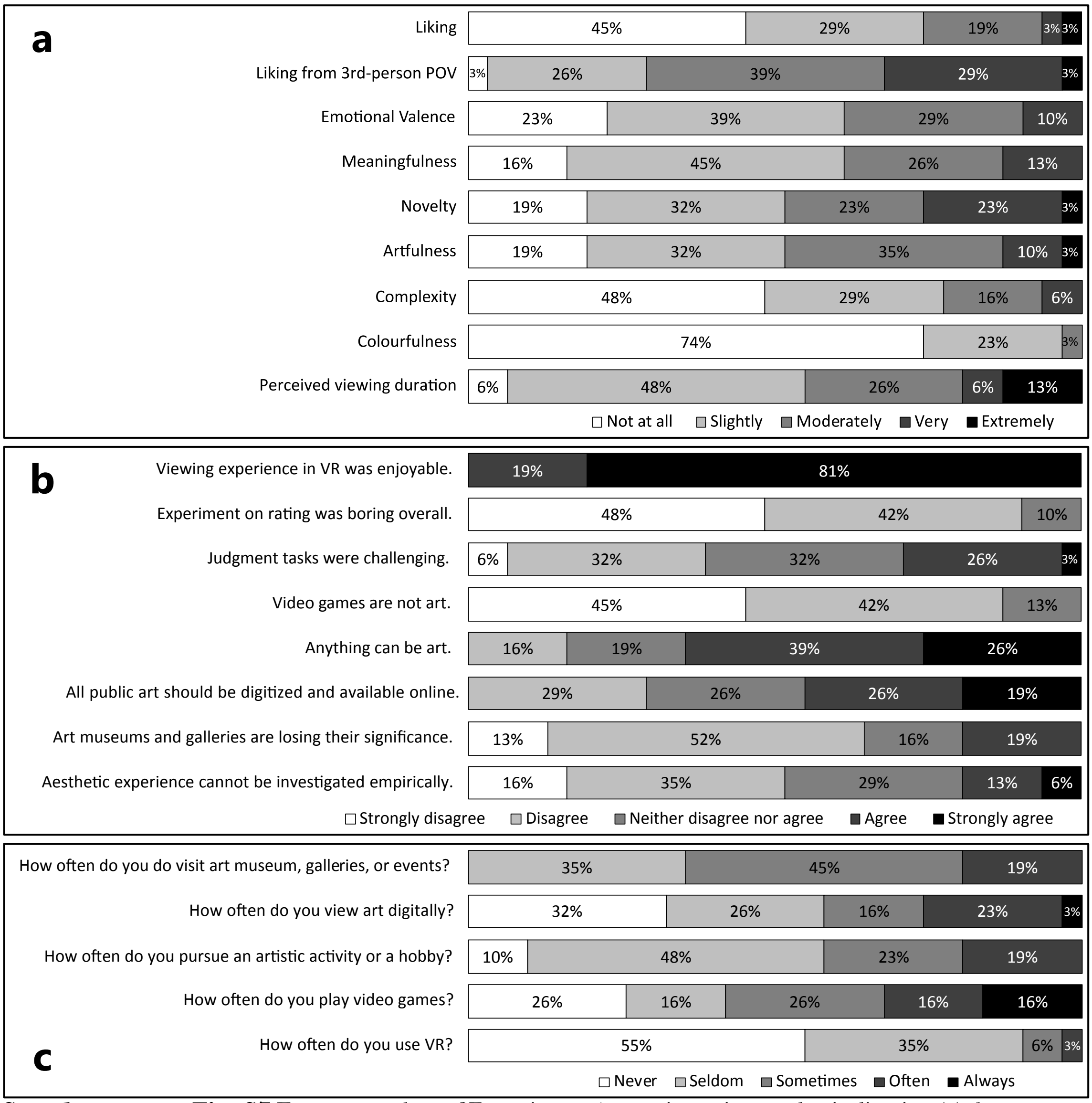
Supplementary Fig. S4 Correlations found in Experiment 1. Plots show correlations between liking rating and other types of ratings (a-i), and the last plot shows the correlation between perceived and real view time (j). For plots A-H, x and y axes denote rating judgements ranging between 0 and 100, binned into 20-points as a way of data visualization. For plots i-j, the x-axis again denotes the rating judgements, but the y-axis denotes real view time spent in s per artwork during the VR screening phase, ranging from 0 s to 480 s, binned into 20-points. All plots were illustrated with a shared scale ranging from zero to fifty number of data points (k), where an individual data point refers to a single judgment rating, by a single participant for a single artwork. Note that the cumulative sum of data points was 496 per correlation matrix, regarding 31 participants multiplied by 16 artworks (i.e. $\mathcal{N}_{\text{Participant}} = 31$, $\mathcal{N}_{\text{Observation}} = 16$).



Supplementary Fig. S5 Rating results of Experiment 1. All rating scores were visualized as boxplots and density curves drawn from the individual data points. An individual dot (as an individual data point) refers to a single judgment score from a single participant for a single artwork, with a cumulative sum of 496 data points per plot, regarding 31 participants multiplied by 16 artworks (i.e. $N_{\text{Participant}} = 31$, $N_{\text{Observation}} = 16$). The vertical two-sided density curves represent the distribution of those cumulative data points, covering the range of the rating scale ranging from 0 to 100.

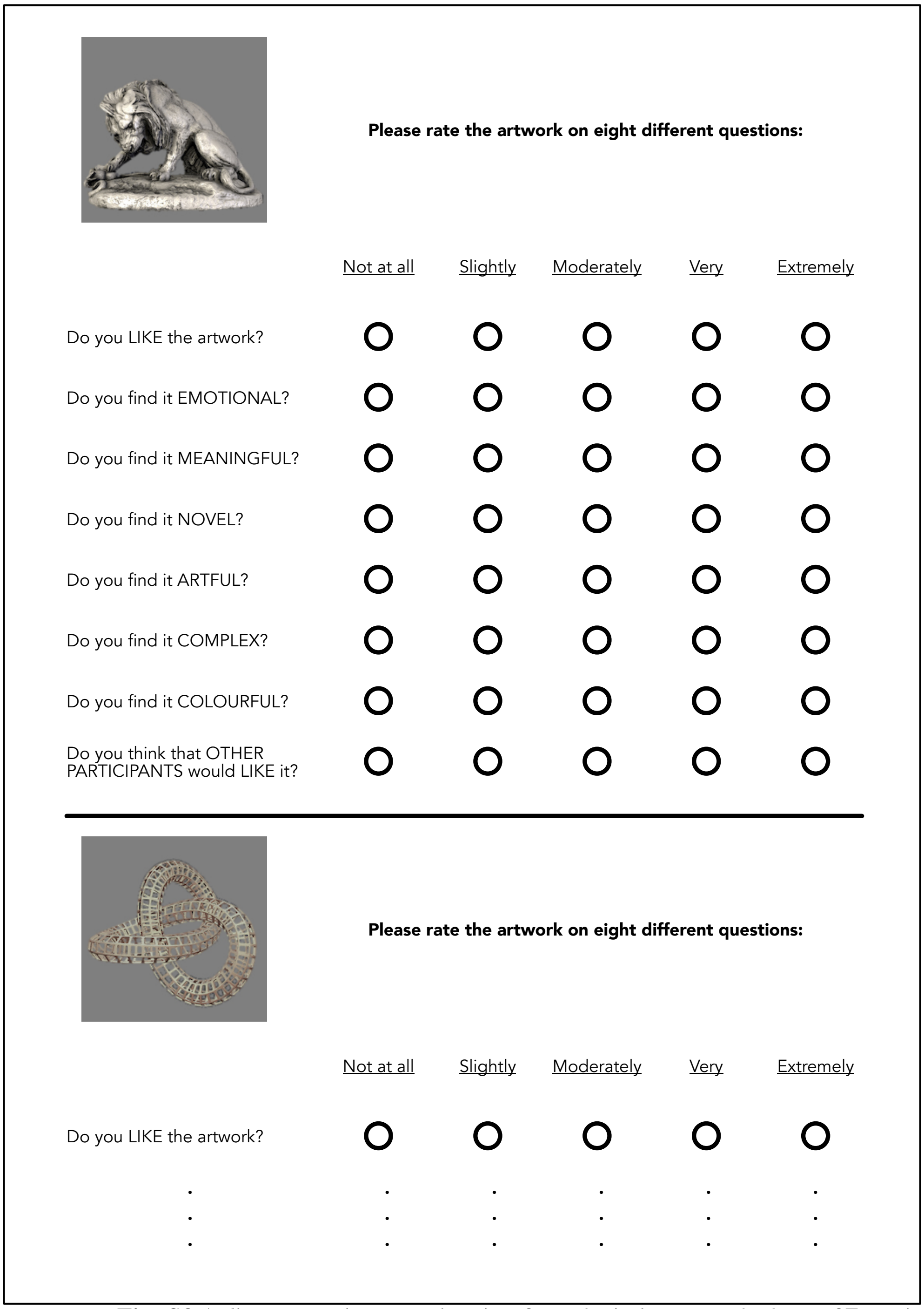


Supplementary Fig. S6 Results of Experiment 1. (a) Average view duration per individual artwork in seconds, and (b) average liking score binned into individual artworks were illustrated as bar graphs, indicating means with error bars showing ±1 standard error of the mean (SEM). In the second figure, a dashed line parallel to the x-axis denotes the mid-value of liking ratings. If the mid-value can be defined as a threshold point of a "neither liked nor disliked" artwork, then an artwork with a mean rating score (including variance) above the threshold can be defined as an overall liked artwork, and vice versa. Note that artwork #5 (as a colourful bird) had the highest liking rating score, and also it was the only non-static artwork: the basic wing movement present was part of the 3D model and preserved in the experiment, therefore this singularity might have created a type of an oddball effect, compared to all other static artworks. The sample size was $N_{\mathrm{Participant}} = 31$.

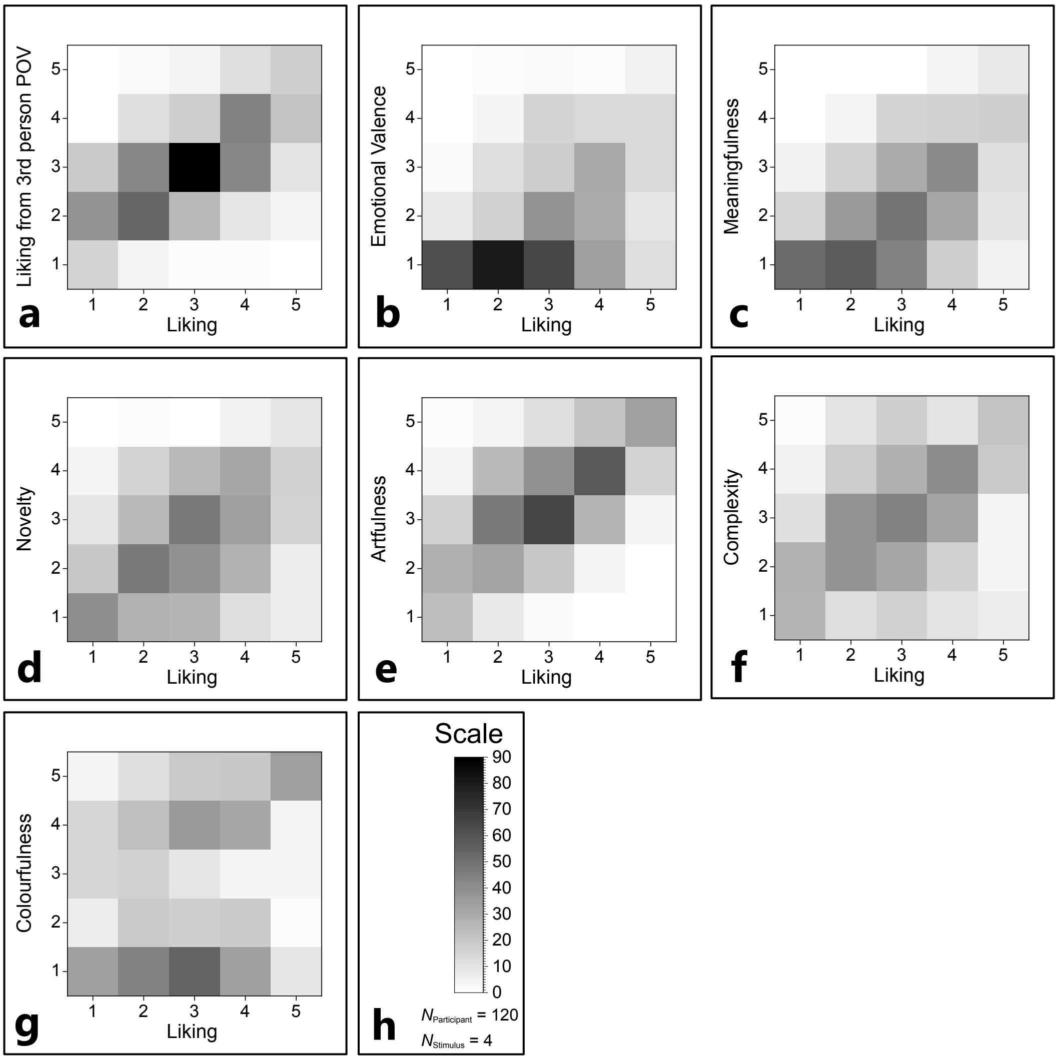


Supplementary Fig. S7 Frequency plots of Experiment 1 questionnaire results, indicating (a) the amount of difficulty on judgments, (b) general attitudes as agreement items to provide additional insight about the participants view, and (c) frequency of exposure to arts. The response frequencies corresponding to the five possible levels for each item were reported as frequencies, rounded to the nearest integer. The sample size was $N_{\text{Participant}} = 31$

Supplementary Fig. S8 A word bubble was generated from open-ended responses asking participants to provide any other term(s) including adjectives or metaphors which might be useful in describing and judging the given artworks in Experiment 2. To illustrate the frequency of responses, font sizes were linked to the number of responses, such that the terms with larger font size refer to frequently provided responses, whereas terms with the smallest font size refer to the responses provided only once. The sample size was $\mathcal{N}_{\text{Participant}} = 31$



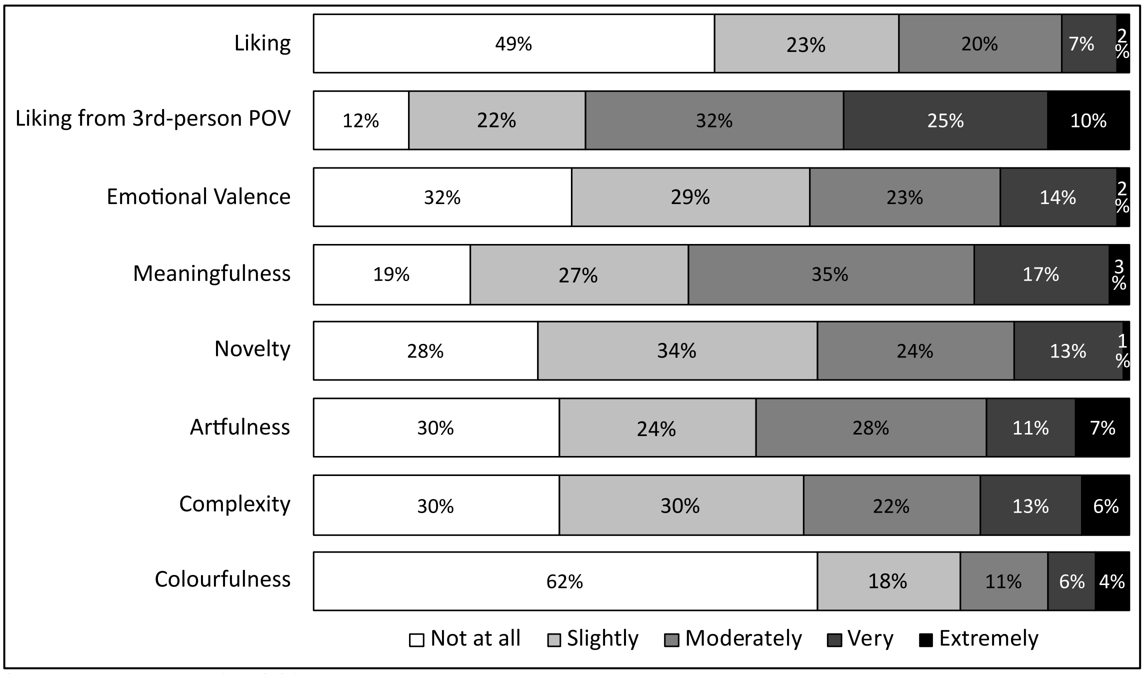
Supplementary Fig. S9 A diagrammatic, exemplar view from the judgment task phase of Experiment 2. Here, participants were able to see all the artworks and rating judgments simultaneously on the same webpage, and they provided ratings on a 5-point Likert scale (such as not at all liked, slightly liked, moderately liked, very liked, and extremely liked) by selecting the corresponding level of their responses.



Supplementary Fig. S10 Correlational matrices of Experiment 2. All plots showing the correlations between liking rating and other ratings (a-g), where x and y axes denote rating judgements ranging between 1 and 5. All plots were illustrated with a shared scale ranging from zero to ninety number of data points (h), where an individual data point refers to a single judgment rating, by a single participant for a single artwork. Note that the cumulative sum of data points was 480 per correlation matrix, regarding 120 participants multiplied by 4 artworks (i.e. $N_{\text{Participant}} = 120$, $N_{\text{Observation}} = 4$).

Covariates		Experiment 1		Experiment 2		*z*	*p*
#1	#2	*r(P)*	*p*	*r(S)*	*p*		
liking	liking from 3rd person POV	0.538	< .001	0.599	< .001	-0.43	> .05
liking	emotional valence	0.571	< .001	0.442	< .001	0.83	> .05
liking	meaningfulness	0.381	< .001	0.503	< .001	-0.72	> .05
liking	novelty	0.360	< .001	0.410	< .001	-0.28	> .05
liking	artfulness	0.441	< .001	0.647	< .001	-1.41	> .05
liking	complexity	0.502	< .001	0.382	< .001	0.71	> .05
liking	colourfulness	0.443	< .001	0.262	< .001	0.99	> .05

Supplementary Fig. S11 To assess the significance of the correlational difference between VR and online settings, the Fisher *r*-to-*z* transformation was used, although the online experiment was not designed to be a direct replication of the VR-based experiment. In this analysis, the correlation coefficients were Pearson correlation coefficient (denoted as r_p) for the first, lab-based experiment, and Spearman correlation coefficient (denoted as ρ or r_s) for the second, online experiment, and sample sizes were $N_{\mathrm{VR}} = 31$ and $N_{\mathrm{Online}} = 120$. For all correlation pairs where one variable was liking, no difference was observed for liking from the third-person POV ($z = -0.43$), positive emotional valence ($z = 0.83$), meaningfulness ($z = -0.72$), novelty ($z = -0.28$), artfulness ($z = -1.41$), complexity ($z = 0.71$), colourfulness ($z = 0.99$), where all $p > .05$.



Supplementary Fig. S12 Frequency plots of Experiment 2 questionnaire results, indicating the level of difficulty on judgments. The response frequencies corresponding to the five possible levels for each item were reported as frequencies, rounded to the nearest integer. The sample size was $N_{\text{Participant}} = 120$.

Chapter 5

Mural in Compton Verney Art Gallery

Visual experience of installation art in the gallery and aesthetic preference towards artwork variations

Doga Gulhan* · Szonya Durant · Johannes M. Zanker

Department of Psychology, Royal Holloway, University of London, UK

*doga.gulhan@rhul.ac.uk

5.1. Abstract

In empirical aesthetics, mobile eye-tracking allows researchers to observe aesthetic experience with genuine artworks outside the laboratory settings, and online experiments can provide a supplementary research context. Here, we investigated eye movements of gallery visitors whilst they were engaging with a room-scale installation by Lothar Goetz, consisting of diagonal colour patches of parallelograms, entitled Salon Diagonale, exhibited in Compton Verney Art Gallery. Initial gaze data visualisation showed substantial individual differences, and analysis on aggregated gaze data using absolute and area-normalized dwell time yielded significant trends, such as preference on edges and vertices between colour patterns, and a horizontal central tendency. In a second, online, experiment, we investigated participants' aesthetic judgments on six modified variations of the artwork presented as rendered videos and their reasoning behind the judgments. Although ratings showed high variance, most participants depreciated the version with the Gaussian-blur introduced edges, slightly more preferred a monochrome version over the original artwork unbeknownst to them, and provided vastly different justifications. In a follow-up and mainly an explorative experiment, using webcam-based eye-tracking, we identified a promising potential of online eye-tracking research, and further reported a lack of linear relation between liking ratings and gaze count.

5.2. Introduction

The aesthetic experience is a complex and an inherently subjective phenomenon, and its subjectivity creates ever-lasting conceptual challenges for researchers. Historically, for example, Koffka (1935) remarked the qualitative change of aesthetic appreciation of a given art object, exemplified by van Gogh paintings, depends not only on physical properties but also on the complementary factors such as time and (cultural) geography (pp. 347-348): he speculated that given a painting (P) and two art critics (A and B), two behavioural art-objects (P_A and P_B) emerge, which can account for the discrepancy between judgments (Koffka, 1935). This brief remark can be treated as a useful reminder echoed by many researchers that the same artwork can be perceived in vastly different and distinct ways. More recently, elaborate models of aesthetic experience have been proposed (Leder & Nadal, 2014; Pelowski, Markey, et al., 2017), aiming to capture the complexity of aesthetic experience and judgments, and to try establishing a common conceptual ground for researchers. On the empirical level, and generally in line with the psychophysical tradition, some behavioural aspects of aesthetic experience can be treated as an objectively measurable phenomenon, demonstrated by earlier seminal research (Fechner, 1876; Helmholtz, 1863), and expanded into neuroimaging studies (Vartanian & Skov, 2014), eye-tracking based approaches (Nayak & Karmakar, 2019), or combinations of these (Guo et al., 2019; Kesner et al., 2018), museum-based studies (Pelowski, Forster, et al., 2017) with additional incorporated measures such as heart rate and skin conductance (Tschacher et al., 2012), and virtual reality (Parker & Saker, 2020; Lee et al., 2020). Arguably less objective measures such as survey methods (Drought, 1929; Hager et al., 2012; Wanzer et al., 2020) and other qualitative-dominant methods such as content and discourse analysis (Lagerspetz, 2016) or hermeneutic phenomenological analysis (Ashrafi & Garbutt, 2017) have also been used extensively and provided valuable knowledge about aesthetic appreciation. Currently, the standpoint of many researchers seems to be that the aesthetic experience as a whole is highly context-dependent and not something reducible to a set of observational data, but meaningful insights can be derived from empirical investigations.

Eye-tracking in empirical aesthetics can be described as an objective measure, which can be targeted to unconscious choices or (automatic) cognitive-emotional processes during the aesthetic experience. Historically, employing eye-tracking methods on paintings have allowed researchers to investigate relations between gaze-patterns, painting properties, individual differences, and instructions given to observers (Buswell, 1935), backdating some seminal works on task-dependent eye movements during scene viewing (Yarbus, 1967). Often linking gaze metrics to concepts of visual attention (Rosenberg & Klein, 2015), studies investigated for example (i) the influence of bottom-up and top-down cognitive processes during artwork viewing (Massaro et al., 2012); (ii) the relationship between observers' fixation locations on abstract painting and emotional valence distribution of paintings derived using machine learning (Yanulevskaya et al., 2012); (iii) the relationship between gaze information, derived from viewing Van Gogh's paintings, and observer-independent computational aesthetics measures (Hao et al., 2020); (iv) perceived complexity and aesthetical pleasantness of Western paintings from various art periods, in conjunction with eye-tracking measures, and use of saliency modelling as a prediction tool (Wallraven et al., 2009); (v) exploring viewing trends and aesthetic exploration for installation art in-situ (Pelowski et al., 2018) and in virtual reality (Mu et al., 2020). As underlined by many researchers, if engaging with a visual artwork is a constituent of the aesthetic experience and can be described as an analogous behaviour to free viewing without an explicit task, then the eye-tracking metrics can provide both the individual visual exploration patterns and remark the overall salient regions or hotspots of visual attention of the artwork.

Many studies are necessarily bound to the laboratory conditions, mainly due to the experimental designs requiring precise stimuli control or due to the lack of accessibility of the required hardware and software to conduct research outside the laboratory. The lab settings offer precise stimuli control so that the contents of the presented stimuli such as colour, luminance, or other compositional properties can be easily manipulated according to the research questions, which is often followed by forms of aesthetic judgments in experimental designs. In this sense, researchers can investigate, for instance, (i) preference towards Munsell colour patches (McManus et al., 1981); (ii) preferred chromatic composition of paintings (Nascimento et al., 2017); (iii) effects of colour processing on the (automaticity of) the aesthetic judgment task (Mullennix et al., 2016); (iv) relation between statistical image properties and preference towards visual artworks and architectural photographs (Braun et al., 2013); (v) relation between edge-orientation entropy of images and preference (Grebenkina et al., 2018); (vi) a visual analogue of the auditory cocktail party by superimposing luminance and colour components of two paintings (Anstis et al., 2012).

It is important to note that, on the conceptual level, most researchers seem to be arguing for some form of aesthetic relativism, which roughly refers to the standpoint that aesthetic experience and judgment are highly context-dependent, contrasting Kantian ideas of aesthetic universals (Guyer, 1993; Wenzel, 2005). In terms of the experimental designs, asking seemingly the same research question, for example, whether an artwork would be appreciated more in its original form in comparison to in its manipulated version still holds validity, since the answer often at least depends on the specific artworks in question and the type of particular manipulation. Additionally, if an aesthetic judgment is conceptualized either as part of the aesthetic experience or as a subsequent elaborative state following the aesthetic experience, then, for example, the distributions of judgment ratings can be regarded as indicators of an overall attitude towards the artworks.

Diverging from the lab conditions, at least three additional contexts for empirical aesthetics can be specified: in-situ, online, and extended reality (XR) such as virtual and augmented reality (VR/AR). These emerging settings may tackle some drawbacks of the lab-based studies, such as restricted observer experience and dependence on the use of art reproductions, and offer, for example, experience of original artwork in natural context for in-situ experiments (Heidenreich & Turano, 2011; Walker et al., 2017; Yi et al., 2020), large sample size for online studies (Redi et al., 2013; Saiz et al., 2018), and otherwise impossible experimental configurations for XR research (Shehade & Stylianou-Lambert, 2020; Tennent et al., 2020). These settings are not analogues of the lab conditions, simply because the viewing behaviour might drastically differ between laboratory and real-world settings (Estrada-Gonzalez et al., 2020). Additionally, the use of virtual reality (Ivancic et al., 2016) and augmented reality (Kyriakou & Hermon, 2019) might enhance visitors' overall experience, and even relatively simple online tools to view arts promise accessibility and well-being (Tyack et al., 2017). Overall, capturing observers' behaviour in realistic in-situ, or in immersive VR settings, and exploring online art viewing seem to be relevant contexts for the current experimental aesthetics research.

In the present study, we aimed to investigate the aesthetic experience towards a room-scale abstract art installation in a gallery context and two online settings. In the first experiment, using mobile eye-tracking whilst the gallery visitors engage with the installation, we mainly investigated viewing trends such as the distribution of dwell time as cumulative fixation duration and fixation count on compositional features, such as edges and colours. In the second experiment, using a set of variations of the same artwork presented as rendered videos in an online setting, we investigated participants' aesthetic judgments, the reasoning behind them, and their insights about digital art viewing. As a follow-up experiment, we also utilized state-of-the-art online, webcam-based eye-tracking on those variations. Given the abstract nature of the installation artwork and an expected high influence of artwork's topological properties on gaze patterns, for the first in-situ experiment, the main non-directional alternative hypothesis can be formulated such that the visual exploration patterns as indexed by absolute and area-normalized dwell time are different regarding sets of areas of interests (AOIs). Similarly, assuming that the aesthetic liking as a type of aesthetic judgment is prone to change depending on the compositional features of the installation, for the second online experiment, the main non-directional alternative hypothesis can be formulated such that the liking ratings are different regarding the variations of the artwork.

5.3. Experiment 1: Eye-tracking in the gallery

5.3.1. Methods

5.3.1.1. Participants

Participants who were visitors of the Compton Verney Art Gallery during data collection were recruited on-site by convenience sampling. A total of 15 visitors participated in the first experiment (M_{Age} = 52.60 years, SD_{Age} = 15.04 years, R_{Age} = 23-76 years, 9 females, 6 males). The analysis included thirteen participants due to a data-saving error corrupting two recordings. All were naïve to the hypotheses of experiments. All participants had a normal or corrected-to-normal vision. They took part in the study voluntarily. All participants provided written informed consent before the experiment. The experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki.

In total, we collected approximately 300 fixations per participant, each of which were spatially corresponding to a particular AOI on the installation. Note that, although the standard power or sample size estimations are not well suited for this particular eye-tracking experiment, and the pre-determining of sample size was not possible due to the unpredictable availability of gallery visitor during the day in the gallery, to illustrate the context of all the reported results with three different degrees of freedom (i.e. 2, 3, and 15), sensitivity analyses on generic $\chi 2$ tests with a significance level of a = .05 and a power level of 1-β = .80, critical $\chi 2$ values are estimated as 5.99, 7.81, and 25.00, respectively.

5.3.1.2. Stimulus and Material

The stimulus of this case study was an artwork entitled Salon Diagonale by Lothar Goetz, and temporarily exhibited in Compton Verney Art Gallery in Warwickshire, England. The room-scale installation covered the walls of the approximately 25 m^2 room with colourful oblique parallelograms (see Figure 1).

5.3.1.3. Design

The study was designed as a within-subjects experiment. Main dependent variables (DVs) were absolute and area-normalised dwell times as the cumulative fixation durations in seconds, as measured by the mobile eye-tracker. Main independent variables (IVs) were sets of AOIs as indexed by corresponding 3D geometry of the installation. As a supplementary descriptive analysis, the number of fixations in absolute and area-normalised forms were defined as DVs for the same IV sets.

5.3.1.4. Procedure

Before the start of the experiment, participants were greeted, researchers provided verbal and written information about the experiment and asked participants for their written informed consent using the consent forms. After that, the mobile eye-tracker (Tobii Glasses 2) was mounted on the participant's head and firmly adjusted, and a brief calibration process was initiated. Then participants engaged with the artwork by moving freely in the gallery space as long as they wanted and were free to verbally describe what they think during the viewing experience if they wish to do so. Lastly, they were asked to complete a brief exit-questionnaire. A typical experiment lasted from ten to fifteen minutes.

5.3.1.5. Data Analysis

The raw data output was obtained from the mobile eye tracker and was analysed using MatLab, R, and jamovi, and further visualizations such as heatmaps were generated in a combination of software including Tobii Pro Lab, Affinity Suite (mathworks.com, r-project.org, jamovi.org, tobiipro.com, affinity.serif.com). The main analysis was conducted using the Friedman test as a non-parametric alternative of the RM-ANOVA, due to the non-normality of the data, where Durbin-Conover pairwise comparisons were Bonferroni-corrected.

Fixation datasets were segmented according to the AOI-maps to be compared for the absolute and area-normalized dwell time analysis: the absolute dwell time was defined as the cumulative fixation duration per AOI and aimed to measure the overall visual attention, whereas the area-normalized dwell time aimed to measure fixation density or attentional density, after accounting for sampling at chance (i.e., to account for an expected bias towards larger fixation counts and longer cumulative fixation duration on larger AOIs), to correct for the relative size of areas. It was calculated as cumulative fixation duration multiplied by the given AOI area in percentage, such that any given AOI was expressed as a fraction of the total area of all the AOIs.

To analyse the overall dwell times and fixations, one main AOI map was defined with three levels: installation (as all the painted areas on the wall), windows (as two windows of the gallery space), and other (as all other surfaces of the gallery, such as the fireplace, door, baseboard, etc.). To analyse the spatial distribution of absolute dwell time, five AOI-maps were defined to reflect the topological and compositional structure of the installation, excluding the windows and all other surfaces. Five conditions for AOI sets were created as the IVs of the statistical analyses (see Figure 2 for an overview).

These five AOI maps were the following: Map #1 on “horizontal levels” divides the installation into three portions, as the upper, middle, and lower levels (see Figure 2a). Map #2 on “cardinal walls” divides the installation into four walls, referred to as north, east, south, and west; approximately based on the physical location of the gallery space (see Figure 2b). Map #3 on “vertical edges” divides the installation into four parts, as convex and concave edges of the gallery room, illusory edges formed by the layout of the installation itself, and the remaining painted areas on the walls (see Figure 2c). Here, illusory edges refer to the two vertical edges on each wall, formed by the artwork layout. The four real vertical edges of the room between each cardinal wall are split into two as convex and concave edges: two convex edges are basically the borders between E-S and W-N walls, and two concave edges are the borders between N-E and S-W walls. Map #4 on “vertices” (i.e., the points where the edges of the installation meet) divides the installation into three portions, as all edges around the individual colour patches, vertex/intersection points between those edges, and the remaining painted areas on the walls) (see Figure 2d). Map #5 on “individual colours” divides the installation into sixteen parts, as sixteen individual colours are assigned to the installation (see Figure 2e). Note, the analysis workflow was the same for the spatial distribution of normalized dwell time, using values normalised by the surface area of given AOIs, instead of absolute values.

The basic flow of quantitative data analysis was the following: inner sensors on the mobile eye tracker recorded raw eye movement data as eye orientation providing the direction of eye gaze, whilst a regular scene camera captured video footage from the participant’s point-of-view. Using Tobii Pro Lab, the retinal coordinates of gaze were matched to individual frames of the scene video. To detect fixations, an I-VT filter (Salvucci & Goldberg, 2000) was used, with a dispersion threshold of 1.0 degrees, a minimum fixation duration of 60 ms, allowing for a gap of a maximum of 75 ms between fixations (namely, maintaining gaze on a single location is only defined as a fixation if the event lasted at least for 60 ms, and two or more consecutive fixations were combined into a single fixation if they were temporally separated by less than 75 ms, as suggested by (Komogortsev et al., 2010)), and used as default by Tobii User’s Manual for Glasses Pro 2). Individual fixations were mapped on a scaled reference image, previously created from the photographs of the gallery space. Fixation duration, number of fixations, and their locations on the artwork as XY-coordinates were exported.

5.3.2. Results

The types of eye-tracking variables were analysed and reported either as the main results or summarized as supplementary material. To allow a qualitative evaluation and to illustrate the distribution of visual attention, initial visualizations were performed to generate heatmaps derived from dwell times (see Figure 3 for a cumulative heatmap aggregated across 13 participants, and Supplementary Figure s1 for fixation heatmaps from exemplar participants): here, hotspots of attention as indicated by densely fixated regions seemed to be preferentially located on edges and intersection areas between the individual-coloured patches. This preference trend was overlapping especially with the intensity channel of the artworks' basic saliency map (Supplementary Figure s2) generated using the Itti algorithm (Itti et al., 1998), although no further analysis was performed. Additionally, open-ended feedback from the exit-questionnaire showed that participants described the experiment using various keywords and phrases such as 'interesting', 'different', 'being involved', 'more focused', 'artist led me to look out of the window', 'more aware of windows', etc.

5.3.2.1. Total dwell times and fixations in the room

A general analysis was conducted, following the separation of the data by three basic regions (i.e., the installation itself, windows, and all other surfaces). In terms of absolute dwell time, a significant difference was observed ($\chi2(2) = 19.8$, $p < .001$) where dwell time on installation ($M = 103 \pm 10.87$s) was significantly higher compared to both windows and all other surfaces such as the fireplace, baseboard, etc. ($Z = 6.97$, $p < .001$; $Z = 8.13$, $p < .001$, respectively), and shown in Figure 4a. Similarly, a significant difference was also observed for fixation count ($\chi2(2) = 19.5$, $p < .001$), where the number of fixations was highest for the installation ($M = 288 \pm 36.19$) compared to both the windows and all other surfaces ($Z = 7.18$, $p < .001$; $Z = 7.56$, $p < .001$, respectively), and shown in Figure 4b. Following the area-normalization, the normalized dwell time reached significance, ($\chi2(2) = 12.2$, $p = .002$) but here, no difference was observed between installation and windows ($Z = 1.03$. $p > .05$), and the normalized dwell time on the AOIs of other areas was shortest compared to installation and windows ($Z = 4.39$, $p < .001$; $Z = 3.36$, $p = .003$) and shown in Figure 4c. Similarly, normalized fixation count reached significance, ($\chi2(2) = 12.9$, $p = .002$) but here, no difference was observed between installation and windows ($Z = 1.59$. $p > .05$), and normalized fixation count on the AOIs of other areas was smallest compared to installation and windows ($Z = 4.78$, $p < .001$; $Z = 3.19$, $p = .004$), and shown in Figure 4d. Lastly, average fixation durations (without area normalization) were similar across three levels, without reaching a statistically significant difference ($\chi2(2) = 0.154$, $p > .05$), and shown in Figure 4e.

In summary, participant spent most of their time viewing the installation (as indexed by absolute dwell time and fixation count), their average dwell time (i.e., cumulative fixation duration) was around two minutes, where approximately 300 fixations were captured per participant, and their attentional focus was distributed similarly between the installation and windows (as indexed by normalized dwell time and fixation count).

5.3.2.2. Spatial distribution of absolute dwell time on artwork's painted areas

Here, a more detailed analysis was conducted for the absolute dwell time: The painted geometry of the artwork was scrutinized into five sets of AOIs, mentioned in the previous Data Analysis subsection.

For comparison #1 of horizontal levels, a significant difference was observed ($\chi 2(2) = 19.8$, $p < .001$), where the absolute dwell time on the middle level was higher than both the lower level and upper level ($Z = 3.88$, $p < .001$; $Z = 2.67$, $p = .013$, respectively). For comparison #2 of the cardinal directions, no difference was observed between the four walls ($\chi 2(3) = 5.40$, $p > .05$). For comparison #3 of edges, a significant difference was observed ($\chi 2(3) = 32.9$, $p < .001$) where all pairwise comparisons reached statistical significance: the dwell time was highest on wall compared to illusory edge, convex edge, and concave edge ($Z = 5.54$, $p < .001$; $Z = 10.34$, $p < .001$; $Z = 12.92$, $p < .001$), lowest on concave edge compared to convex edge and illusory edge ($Z = 2.58$, $p = .014$; $Z = 7.39$, $p < .001$), and dwell time on illusory edge was also higher compared to convex edge ($Z = 4.80$, $p < .001$). For comparison #4 of vertices, a significant difference was observed ($\chi 2(2) = 12.9$, $p = .002$), where dwell time on intersection was lower than both edge and wall ($Z = 4.78$, $p < .001$; $Z = 3.19$, $p = .004$). For comparison #5 of colours, a significant difference was observed ($\chi 2(15) = 101$, $p < .001$), mainly due to singularities such as black or brown, not individually reported further. All results were shown in Figure 5a, and corresponding supplementary data on fixation count was shown in Supplementary Figure s3a.

5.3.2.3. Spatial distribution of area-normalized dwell time on artwork's painted areas

Similar to the previous section, a similarly detailed analysis was conducted for the area-normalized dwell time. For comparison #1 of horizontal levels, a significant difference was present ($\chi 2(2) = 7.38$, $p = .025$), where area-normalized dwell time on middle level was highest compared to upper and lower levels ($Z = 2.67$, $p = .013$; $Z = 2.67$, $p = .013$). For comparison #2 of cardinal directions, no difference was observed between four walls ($\chi 2(3) = 1.34$, $p > .05$), similar to the results of absolute dwell time. For comparison #3 of edges, a significant difference was present ($\chi 2(3) = 15.6$, $p = .001$), but the order of levels was changed compared to absolute dwell time: here, the normalized dwell time on convex edge was highest compared to concave edge and wall ($Z = 2.26$, $p = .030$; $Z = 4.15$, $p <$

.001) and was lowest on wall compared to illusory edge (Z = 4.15, p < .001), also a difference was present between concave edge and illusory edge (Z = 2.26, p = .030). For comparison #4 of vertices, a significant difference was present ($\chi 2(2)$ = 14.9, p < .001), but the order of levels was changed compared to absolute dwell time: here, the normalized dwell time on intersection was highest compared to edge and wall (Z = 4.04, p < .001, Z = 5.48, p < .001). Lastly, for comparison #5 of colours, a significant but diminished difference was present ($\chi 2(15)$ = 26.9, p = .030). All results were shown in Figure 5b, corresponding supplementary data on area-normalized fixation count was shown in Supplementary Figure s3b.

In brief, the results of both absolute and normalized dwell times were obviously similar when the area percentages of surface levels were similar (i.e., particularly for comparison #1 of horizontal levels and #2 of cardinal directions): these results indicated a horizontal central tendency and lack of a strong cardinal directional preference. However, when the area percentages of surface levels were dissimilar (i.e., particularly for comparison #3 of edges and #4 of vertices), the trend of longer absolute dwell time observed mainly on the walls were reversely translated into smaller normalized dwell time. Lastly, although some trends were visible for the individual colours for absolute dwell time, the area normalization had an equating effect across colours.

5.4. Experiment 2: (A) Online judgments and (B) online eye-tracking

5.4.1. Methods

5.4.1.1. Participants

For Experiment 2A, 120 individuals (60 females, 59 males; M_{Age} = 26.85 years, SD_{Age} = 8.17 years, R_{Age} = 18-63 years, four participants did not report their age), and for Experiment 2B, 48 individuals (24 females, 24 males; M_{Age} = 38.40 years, SD_{Age} = 14.47 years, R_{Age} = 19-79 years) were recruited via Prolific (prolific.co), an online participant recruitment tool. All participants were naïve to the hypotheses of experiments, and they were compensated monetarily (equivalent to £6/hour for an expected experimental duration of six minutes for Experiment 2A, and twelve minutes for Experiment 2B). For both online experiments, all participants provided informed consent before the experiment. All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki.

Prior to experiment 2A, a priori power analysis was carried out, indicating that a sample size of 102 or 119 would be sufficient for an RM-ANOVA (within factors), with a significance level of $a = .05$, a power level of $1\text{-}\beta = .80$, an assumed effect size of $f(U) = .2$ (corresponding to the partial $\eta 2 \approx .04$), depending on whether the six variations of artworks are treated as six levels of the IV (i.e. as six measurements) or the original artwork is defined as a baseline and the differences between the baseline and all other five variations are treated as five levels of the IV (i.e. as five measurements), respectively. For Experiment 2B, a priori power analysis indicated that a sample size of 29 would be sufficient for a non-directional correlation with a significance level of $a = .05$, a power level of $1\text{-}\beta = .80$, a projected population correlation under the alternative hypothesis of .5, and under the null of 0. In the case of a partial data exclusion (such as some participants might either spend exceptionally short or long time viewing artworks, or provide extreme liking ratings by treating the continuous liking rating scale as a binary scale), a sample size above the estimated value to some extent was sought to allow flexibility for potential supplementary analysis.

5.4.1.2. Stimulus and Material

For Experiment 2A, firstly, a digital model of the artwork was built. The 3D model's set of surfaces can be manipulated, for example, in colour, shape, boundary properties. Following the modifications, visual stimuli were six digitally rendered videos of the artwork, recorded within the one of the six manipulated 3D models, including the original artwork and five variations (see Supplementary Figure s4a for an overview). The iterations were referred to as (i) "diagonal" as the original artwork, (ii) "monochrome" generated as a black-and-white version, (iii) "blurred" generated as a version with introducing Gaussian blur on edges, (iv) "polygonal" generated by segmenting parallelograms into further polygons, (v) "horizontal" generated by transforming parallelogram into horizontally-aligned rectangles, and (vi) chevron generated by vertically flipping the middle set of parallelograms on each wall, thus creating an additional, large zigzag pattern. Note that the participants were not aware that the diagonal version was the original artwork. All videos were created using a rendering software called Lumion (lumion.com) and were based on 3D models of the artwork created using a 3D modelling software called SketchUp (sketchup.com). The duration of each video was 30 seconds, and all videos were rendered using the exact camera pathway and rendering settings (such as interior luminance and field-of-view) to preserve inter-stimulus consistency since the only manipulated aspect of the stimuli was the artwork itself. The renders were recorded from a first-person point-of-view in its true relative-sized gallery space to mimic a viewing experience of the installation. The presentation order of the videos during the viewing period and static snapshots of them displayed afterwards during the rating task were randomized for each participant. The experiment was created using Qualtrics (qualtrics.com), an online experiment management software. Stimuli and the whole experiment were displayed only on participants' personal computers and not on tablets or smartphones, using a pre-screening option on device type via Prolific.

For Experiment 2B, four cardinal views of the gallery room were presented for each variation, instead of presenting rendered videos as in Experiment 2A. To minimize the primacy effect, an empty view of the gallery room was also presented, allowing participants to familiarize themselves with the room layout, before viewing variations (see Supplementary Figure s5b for an overview of the stimuli). Stimulus presentation order was randomized for each participant, and the stimuli were again rendered using Lumion from a first-person point-of-view with a virtual camera-height at 160 cm, to create a relatively realistic looking stimulus, i.e., as if the gallery was viewed by a visitor at standing position.

The experiment was created using Labvanced (labvanced.com), which allowed us to record eye movements using participants' webcams. Stimuli and the whole experiment were displayed only on participants' personal computers and not on tablets or smartphones, using a pre-screening option on device type via Prolific. Note that, presenting static 2D renders instead of videos allowed us to more accurately record the gaze data and map it onto the layout of each variation.

5.4.1.3. Design

Designed mainly as within-subject experiments, the main measured dependent variable was liking rating ranging from 0 (labelled as most disliked) to 100 (labelled as most liked) provided for each variation of the artwork, and the main independent variable was the type of the artwork variation. For experiment 2B, gaze count (as a proxy for fixation count) and liking rating were covariates for the correlational analysis. All other ratings and questions in the experiment were treated as descriptive items.

5.4.1.4. Procedure

Both experiments started with the briefing, receiving consent from the participants, and providing experimental instructions, explaining the experimental procedure in detail.
For experiment 2A, the experimental workflow followed three brief parts: in the first part, participants viewed six artwork variations presented as 30 s long rendered videos. Following on from that, they were asked to provide liking ratings for each variation using continuous sliders, ranging from 0 (labelled as most disliked) to 100 (labelled as most liked).

Lastly, a brief exit-questionnaire was presented (see Supplementary Figure s5a for an overview of the procedure).

For Experiment 2B, the experimental phase started with webcam eye-tracking calibration using a proprietary algorithm based on a deep learning pipeline offered by the Labvanced platform, lasting about three minutes. In brief, during the calibration, participants were asked to fixate the presented points on the screen one by one, and repeated this procedure while viewing the monitor from particular angles and positions, as instructed. After that, participants were asked to view the empty gallery space (i.e., all the walls normally occupied by the mural were rendered as neutral grey, along with avatars as black silhouettes), followed by six installation variations inside the same gallery space one by one in a randomized order. They were first presented by four walls of the gallery space at the same time, located on the quadrants of the screen, without any time limitations, and were instructed to move into the test phase. After that, they viewed four individual walls one by one, with a fixed duration of 4 s. In this way, participants firstly engaged with the artworks as long as they wanted, followed by a 16 s long fixed viewing. In this way, the duration of the first free-viewing part is solely based on the participants' decisions, and the second part of fixed viewing was aimed to equate the duration of collected eye-tracking data across all participants. The eye-tracking, following the calibration, was only enabled during artwork viewing, to maximize the recording accuracy by minimizing the instances of gaze drift correction, and also to avoid redundant recording of eye tracking data, for example, during responding to the questionnaire. Lastly, participants were asked to provide liking ratings for each variation using continuous sliders ranging from 0 to 100 (see Supplementary Figure s5b for an overview of the procedure). Note, both before and after artwork viewing, a five-point calibration check was recorded to measure the reliability of the data as directly as possible.

5.4.1.5. Data Analysis

The liking ratings were treated as continuous variables as they were ranging from 0 to 100, and descriptive statistics for questionnaire items were provided. For the first analysis using the Friedman test as a non-parametric alternative of the RM-ANOVA, due to the non-normality of the data, where pairwise comparisons were Bonferroni-corrected, the liking ratings (as the dependent variable) was treated as a continuous variable, and the artwork variation consisting of six levels (as the independent variable) was treated as a nominal variable. For the second analysis to check whether a significant relationship, either a linear or monotonic one, exists between liking ratings and dwell time, irrespective of artwork variation, both measures were defined as covariates, and both Pearson correlation coefficient and Spearman correlation coefficient were calculated. Software used for the data analysis were R and jamovi (r-project.org, jamovi.org).

5.4.2. Results

For Experiment 2A, firstly, rating scores were visualized as boxplots and density curves to show the overall judgment trends across conditions (see Figure 6). Friedman test as a non-parametric alternative of RM-ANOVA was conducted to compare the liking ratings for six versions of the artwork, and results yielded a significant difference between the variations ($\chi 2(5) = 131.468$, $p < .001$, $W_{\text{Kendall}} = .206$). As post hoc tests, Durbin-Conover pairwise comparisons with a Bonferroni adjustments were carried out: (i) blurred-version was liked least ($M = 23.500$, $Mdn = 12$, $SEM = 2.421$) compared to all other versions (where all five $p < .001$); (ii) monochrome-version as the most liked artwork ($M = 65.783$, $Mdn = 72$, $SEM = 2.905$) showed a significant difference only compared to horizontal- ($p = .022$), polygonal- ($p = .006$), and blurred-versions ($p < .001$); and (iii) all other pairwise comparisons were non-significant, were all ps > .05.

Note that, before the analysis of Friedman test, the Shapiro–Wilk test to check the normality with a significance value set at $a = .05$ yielded that ratings were not normally distributed ($W_{\text{Diagonal}} = .957$, $p < .001$; $W_{\text{Monochrome}} = .970$, $p < .001$; $W_{\text{Blurred}} = .832$, $p < .001$; $W_{\text{Polygonal}} = .937$, $p < .001$; $W_{\text{Horizontal}} = .970$, $p = .009$; $W_{\text{Chevron}} = 964$, $p = .003$). Additionally, the test of sphericity yielded a significant result as well ($W_{\text{Mauchly}} = .594$, $p < .001$). Due to the violations of repeated measures analysis of variance (rmANOVA), a non-parametric alternative was selected. If an RM-ANOVA could have been carried out from a less strict standpoint (for example, the skewness and kurtosis levels were between -2 and 2 for all conditions), the result would have still showed a significant difference between the variations ($F(5,595) = 38.171$, $p < .001$, $\eta^2_{\text{p}} = .243$). In that case, post-hoc comparisons using t-tests with Bonferroni corrections resulted in almost the same significant (and non-significant) pairwise statistics, with one exception: here, an additional significant difference was observed between the diagonal and the polygonal conditions ($t = 2.295$, $p = .022$), which was not present in the previous Durbin-Conover pairwise comparisons tests.

Ratings from the exit-questionnaire on five-point Likert scale were treated as ordinal data, and as a common practice of descriptive analysis, modes as most frequent responses were calculated. The liking judgment was not at all challenging (42%), and participants were moderately confident about their judgments (45%). The rendered videos were perceived as moderately realistic (39%). Participants thought that they would find both the ability to wander around (52%), hearing and ambient sound (39%), and having information about the artwork and artist (26%) very helpful. Although most participants stated that the additional information about the artist and artwork would have not changed the liking ratings (42% compared to 26%), a considerable number of participants were uncertain about it (27%). Most participants (42%) preferred to have an empty virtual gallery space without any digital avatars (see Supplementary Figure s6 for all items). Although explicitly stated as optional open-ended items, around three-fourths of participants provided elaborate feedback for the reasoning behind the most liked (74%) and least liked (78%) versions (see Supplementary Figure s7), and more than half of the participants (60%) suggested alternative variations of the artwork: to underline some of them, participants proposed use of triangles, circles, smooth curves, thinner stripes, only blue nuances, a pastel variation, etc.

For Experiment 2B, firstly, rating scores were visualized as boxplots and density curves to show the overall judgment trends across conditions (see Figure 7a). The Friedman test as a non-parametric alternative of RM-ANOVA was conducted to compare the liking ratings for six versions of the artwork, and results yielded a significant difference between the variations ($\chi 2(5) = 49.879$, $p < .001$, $W_{\text{Kendall}} = .069$). As post hoc tests, Durbin-Conover pairwise comparisons with a Bonferroni adjustments were carried out: (i) blurred-version ($M = 24.500$, $Mdn = 18$, $SEM = 3.57$) was liked least compared to all other versions (where $p = .003$ between blurred and polygonal versions, and $p < .001$ for all other four pairwise comparisons); (ii) chevron-version version as the most liked artwork ($M = 67.7$, $Mdn = 67$, $SEM = 3.57$) showed a significant difference only compared to blurred-version ($p = .003$); and (iii) all other pairwise comparisons were non-significant, were all ps > .05. Overall, the results were partially overlapping with the Experiment 2A. As an initial data quality check, participants' raw gaze count data coming from the five-point calibration check (displayed to the participant both before and after the eye tracking recording, both for 10 seconds) were plotted. Note, the metric of gaze count, as a proxy for fixation count, refers to the number of valid detections of gaze on estimated x-y pixel coordinates without applying an additional fixation filter, since the data quality does not easily allow the use of frequently used filtering. Although some promising participants were present (see Supplementary Figure s8), most of the calibration visualization unfortunately showed poor data quality. Therefore, compared to the mobile eye-tracking, no in-depth analysis was executed for the online eye-tracking data.

Nevertheless, to measure the relation between liking ratings and gaze count (as a proxy for fixation count) irrespective of artwork variation, a correlational analysis was calculated. The results showed that no correlation was present, either as $r_{\text{Pearson}} = .026$, p

> .05 indicating lack of linear relationship, or as $\rho_{Spearman} = .012$, $p > .05$ indicating lack of monotonic relationship (see Figure 7b).

Online eye-tracking data were treated as a supplementary feasibility check without an in-depth analysis, due to the expected lower data quality compared to the mobile eye-tracker. The initial check on data quality showed huge variance among participants (as indexed by the dispersion amount of registered gaze data during calibration, not reported further), indicating that the online web-cam based eye-tracking systems are not yet suitable for any detailed AOI-based analysis. Albeit the noise present in the data, a selection of fixation maps aggregated across all participants were illustrated in Supplementary Figure s9. Lastly, some participants provided optional feedback, underlining that they enjoyed the experiment and found it very interesting.

5.5. Discussion

The main research objectives were to investigate gallery visitors' exploration of compositional features of an original artwork in-situ, in this case a contemporary installation artwork as a room-scale mural, using mobile eye-tracker in Experiment 1, participants' aesthetic judgments on six variations of the same artwork online in Experiment 2A, and feasibility of implementing online eye-tracking in Experiment 2B. Analysing the set of fixation density on AOI maps from our in-situ eye tracking experiment we found that the dwell times were higher on coloured patches of the artwork, as compared to the edges and intersection areas between them. However, analysis using area-normalization showed an opposite trend, implying that the visitors' visual attention was densely directed towards the edges and intersections, as indexed by normalized dwell times. Additionally, for this case study, we further report that (i) a form of a central tendency during viewing was present, roughly corresponding to the viewing preference at eye-level on the horizontal axis, (ii) no strong difference was observed between four cardinal walls, and (iii) some singularities of preference were present between sixteen individual colours. Making use of the opportunity to create variations of the artwork for testing in online experiments, we found that the overwhelming majority of participants disliked when a Gaussian was used to blur edges over, as compared to the original version. The differences among other variations were not significant, due to the high variance on liking judgments, which would suggest that the liking judgments were more prone to substantially change depending on the edge information, in case of the installation used here. Open-ended feedback behind the liking judgments showed varied and insightful responses (for example, one participant's justification for the monochrome version to be the most liked was that "… colours didn't seem to work together to me", whereas another participant's justification again for the same monochrome version to be the least liked was that "… being much less striking"). Many of the liking ratings (per artwork variation) showed bimodal distributions and data points were scarce on central rating values (as neither liked nor disliked portion of the rating scale), implying the presence of participants' preference on binary judgments (either as liking or disliking) besides the more relative judgments using a continuous rating (either as liked or disliked, relatively more or less compared to another artwork variation). Additionally, irrespective of artwork variation, no linear or monotonic relationship was observed between gaze count and liking ratings, suggesting that the allocated viewing duration cannot be treated as a predicting factor for aesthetic liking in the online setting. Lastly, we report that although the online eye-tracking is a promising way forward for a more accessible experimental context, and a small subsample of participant data showed relatively high accuracy similar to the mobile eye-tracker data, currently available tools seem to be only partially suitable for more crude experimental designs.

The results in our experiment, which was regarded as a test case for dealing with some empirical aesthetics questions outside of the common practice, can be firstly discussed in terms of attention. Visual attention is often conceptualized as bottom-up and top-down, roughly referring to whether the attentional guidance is based on stimulus properties, such as information on colour or motion, or observer properties, such prior knowledge and action planning (Katsuki & Constantinidis, 2014). The main results suggest that the viewing patterns as indexed by mobile eye-tracking data and subsequent aesthetic judgments as indexed by questionnaire data towards a room-scale installation artwork and its variations were mostly driven by the bottom-up attentional mechanisms. More specifically, observers' fixations were densely located on the edges and vertices on the abstract room-scale installation, and parametric manipulations on stimulus resulted in changes in liking ratings. Previous research findings underline some effects of bottom-up processes present in art viewing, such as low-level visual features of nature-content images may drive gaze behaviour compared to figurative representational paintings (Massaro et al., 2012) or visual exploration of abstract paintings by artistically untrained participants might be guided by basic salience maps compared to artists (Koide et al., 2015).

Similarly, we might associate both the visual exploration in results of the real-world experiment results and liking judgments in results of the online experiment for this case study with saliency maps, and especially the edge information compared to other information channels such as colour or orientation. Assuming that the distribution of fixations derived from eye-tracking data is an approximation for the (bottom-up or overt) visual attention, and saliency maps calculated from an image represent particular qualities of visual scenes, visual saliency models can be useful tools in empirical aesthetics (Leder, 2013), especially to make predictions on where the observers look (Le Meur et al., 2020). Although the computational modelling was beyond the scope of this behavioural and mostly descriptive research, saliency-based approaches as a part of generative algorithms for creating visual stimuli can target more controlled modifications for the abstract artworks, for example, parametrically changing a particular filtering channel (such as orientation) whilst preserving the information entropy on all filtering channels. Similar to our methodological decision, treating genuine artworks as modifiable stimuli, altered to target a particular research question (Francuz et al., 2018), can benefit from even more precise controlled manipulations.

Unsurprisingly, this approach has some limitations. Firstly, introducing topological manipulations is almost never feasible for artworks in an in-situ context, temporarily exhibited artworks such as in our case limits access for researchers, and often only a small selection of participants can be targeted. Therefore, real-world experiments can be complemented by online follow-ups, in particular to reach out to wide audiences remotely. Introducing specifically designed versions of the artwork (or "artoids") in the online experimental design also further expand the potential research questions, which cannot be asked in the physical setting alone. Additionally, online context arguably reduces the resemblance of genuine aesthetic experience to visual arts in real-world settings, and the findings from an online experiment alone may not be easily generalized to the physical encounter. Lastly, online eye-tracking is not yet sophisticated enough for detailed experimental designs, due to the quality of online eye tracking at the current state. Although the main long-term solution is better detection algorithms, for future work, one solution would be designing a crude paradigm where, for example, parsing the display into four quadrants is sufficient for the research questions, instead of aiming for high spatial accuracy. Albeit a resource-demanding option, another potential solution would be a type of an initial pre-screening experiment for the participant, followed by conduction the second experiment only with participants having the most accurate pre-screening data.

Online platforms provide adequate precision in terms of stimuli presentation duration and response time (Anwyl-Irvine et al., 2020), however, online eye movement research is currently limited, whilst some researchers underline the huge potential of eye-tracking in VR (Clay et al., 2019). Following on from that, and as some of our initial works and hands-on experience suggest (Gulhan et al., 2019a, 2019b, 2019c), a future direction is utilizing eye-tracking in VR, where immersive experiments and gamified experiences for experimental aesthetics research can be easily created and some major limitations of both in-situ and online settings can be minimized. In VR, both bottom-up effects (such as topological manipulation on the virtual artwork) and top-down effects (such as types of information provided to the observers) on viewing patterns and aesthetic judgments can be further investigated. To illustrate a particular direction, specific to the case study presented here, for example, we can investigate whether introducing Gaussian blur also shifts average fixation duration or saccade amplitude compared to the original version of the artwork. Furthermore, whilst targeting basic research questions, VR offers an enhanced experience for cultural heritage (Hürst et al., 2016; Škola et al., 2020), and the potential use of eye-tracking might provide context-aware designs for the museum experience (Mokatren et al., 2018). Lastly, digital replicas using parametric stimulus manipulation and utilizing VR, eye-tracking, or online crowdsourcing can be seen extending beyond the scope of empirical aesthetics and into participatory design processes, as digital twins of existing or planned environments, such as in the context of interior architectural spaces (Franz et al., 2005), architectural design features (Ergan et al., 2018), urban public spaces (Kim & Kim, 2019) or cityscape planning and protection (Zhang et al., 2019). In a way, VR seems to offer both realistic viewing conditions similar to physical museum and gallery spaces, and allow parametric stimuli manipulation, therefore has the potential to provide more comprehensive answers to the research questions, which are not feasible in a real-world or an online setting alone.

Acknowledgements

The authors would like to thank Royal Holloway University of London (RHUL) for the doctoral scholarship of Doga Gulhan, the Compton Verney Gallery for providing generous hosting and advisors including staff support, Sabine Zanker and Guzin Oztok for logistics and participant management.

Ethical statement

All experimental protocols were approved by the Royal Holloway, University of London Research Ethics Committee. All methods were performed in accordance with the ethical guidelines and regulations of the Declaration of Helsinki. Research Ethics Committee (REC) Project ID: 123, Year: 2017, User ID: UJJT152 (for the in-situ experiment). Research Ethics Committee (REC) Project ID: 2167, Year: 2020, User ID: PDJT007 (for the online experiments).

Funding statement

We received no funding for this study.

Data availability

All anonymised data are accessible via the Open Science Framework (osf.io/ec46q).

Declaration of conflicting interests

The authors declare no conflicting interests.

Author contributions

Conceptualization, DG, JMZ; Methodology, DG, JMZ; Software, DG, JMZ; Formal Analysis, DG, SD; Investigation, DG, SD, JMZ; Writing – original draft, DG; Writing – review & editing, DG, SD, JMZ; Visualization, DG, JMZ; Supervision, SD, JMZ; Project administration, JMZ.

Timeline

The in-situ experiment was conducted in September 2017, the online experiment was conducted in July 2020, and the online eye-tracking experiment was conducted in December 2020. Please also refer to the last paragraph of the introduction (p. 31), 1st thesis chapter, for a detailed overall timeline and the effects of the pandemic on the experiments.

5.6. References

Anstis, S., Vergeer, M., & Van Lier, R. (2012). Looking at two paintings at once: Luminance edges can gate colors. I-Perception, 3(8), 515–519. https://doi.org/10.1068/i0537sas

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. Behavior Research Methods. https://doi.org/10.3758/s13428-020-01501-5

Ashrafi, S., & Garbutt, M. (2017). Looking at looking: Two applications of a hermeneutic phenomenological analysis of head-mounted eye-tracking data in gallery visitation. Australian and New Zealand Journal of Art, 17(2), 221–237. https://doi.org/10.1080/14434318.2017.1451203

Braun, J., Amirshahi, S. A., Denzler, J., & Redies, C. (2013). Statistical image properties of print advertisements, visual artworks and images of architecture. Frontiers in Psychology, 4. https://doi.org/10.3389/fpsyg.2013.00808

Buswell, G. T. (1935). How people look at pictures: A study of the psychology and perception in art. University of Chicago Press.

Clay, V., König, P., & König, S. (2019). Eye tracking in virtual reality. Journal of Eye Movement Research, 12(1). https://doi.org/10.16910/jemr.12.1.3

Drought, A. (1929). A survey of studies in experimental aesthetics. The Journal of Educational Research, 20(2), 97–102. https://doi.org/10.1080/00220671.1929.10879972

Ergan, S., Shi, Z., & Yu, X. (2018). Towards quantifying human experience in the built environment: A crowdsourcing based experiment to identify influential architectural design features. Journal of Building Engineering, 20, 51–59. https://doi.org/10.1016/j.jobe.2018.07.004

Estrada-Gonzalez, V., East, S., Garbutt, M., & Spehar, B. (2020). Viewing art in different contexts. Frontiers in Psychology, 11, 569. https://doi.org/10.3389/fpsyg.2020.00569

Fechner, G. T. (1876). Vorschule der Aesthetik. Breitkopf & Härtel Verlag.

Francuz, P., Zaniewski, I., Augustynowicz, P., Kopiś, N., & Jankowski, T. (2018). Eye movement correlates of expertise in visual arts. Frontiers in Human Neuroscience, 12:87, 1–13. https://doi.org/10.3389/fnhum.2018.00087

Franz, G., von der Heyde, M., & Bülthoff, H. H. (2005). An empirical approach to the experience of architectural space in virtual reality—Exploring relations between features and affective appraisals of rectangular indoor spaces. Automation in Construction, 14(2), 165–172. https://doi.org/10.1016/j.autcon.2004.07.009

Grebenkina, M., Brachmann, A., Bertamini, M., Kaduhm, A., & Redies, C. (2018). Edge-orientation entropy predicts preference for diverse types of man-made images. Frontiers in Neuroscience, 12, 678. https://doi.org/10.3389/fnins.2018.00678

Gulhan, D., Durant, S., & Zanker, J. (2019a, August). A methodological evaluation on simultaneous use of wearable and VR eye tracking in museum: A comparative case study on Piet Mondrian's room design. Proceedings of the European Conference on Visual Perception (ECVP). Proceedings of the European Conference on Visual Perception (ECVP), Leuven, Belgium. https://doi.org/10.1177/0301006619863862

Gulhan, D., Durant, S., & Zanker, J. (2019b, August). Aesthetic judgments of physical and digital three-dimensional arts in virtual reality and online settings. Proceedings of the Visual Science of Art Conference (VSAC). Proceedings of the Visual Science of Art Conference (VSAC), Leuven, Belgium.

Gulhan, D., Durant, S., & Zanker, J. (2019c, December). Experimental aesthetics in virtual reality: Mondrians, unicorns, and squared circles. Proceedings of the Applied Vision Association (AVA) Christmas Meeting. Proceedings of the Applied Vision Association (AVA) Christmas Meeting, Cardiff, Wales.

Guo, F., Li, M., Hu, M., Li, F., & Lin, B. (2019). Distinguishing and quantifying the visual aesthetics of a product: An integrated approach of eye-tracking and EEG. International Journal of Industrial Ergonomics, 71, 47–56. https://doi.org/10.1016/j.ergon.2019.02.006

Guyer, P. (1993). Kant and the experience of freedom: Essays on aesthetics and morality. Cambridge University Press.

Hager, M., Hagemann, D., Danner, D., & Schankin, A. (2012). Assessing aesthetic appreciation of visual artworks—The construction of the Art Reception Survey (ARS). Psychology of Aesthetics, Creativity, and the Arts, 6(4), 320–333. https://doi.org/10.1037/a0028776

Hao, Q., Ma, L., Sbert, M., Feixas, M., & Zhang, J. (2020). Gaze information channel in Van Gogh's paintings. Entropy, 22(5), 540. https://doi.org/10.3390/e22050540

Heidenreich, S. M., & Turano, K. A. (2011). Where does one look when viewing artwork in a museum? Empirical Studies of the Arts, 29(1), 51–72. https://doi.org/10.2190/EM.29.1.d

Helmholtz, H. von. (1863). Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik. Friedrich Vieweg & Sohn Verlag.

Hürst, W., Tan, X. J., & de Coninck, F. (2016). Using digital extensions to create new VR museum experiences. Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology - ACE2016, 1–6. https://doi.org/10.1145/3001773.3001822

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254–1259. https://doi.org/10.1109/34.730558

Ivancic, D., Schofield, D., & Dethridge, L. (2016). A virtual perspective: Measuring engagement and perspective in virtual art galleries. International Journal of Arts and Technology, 9(3), 273. https://doi.org/10.1504/IJART.2016.10000209

Katsuki, F., & Constantinidis, C. (2014). Bottom-up and top-down attention: Different processes and overlapping neural systems. The Neuroscientist, 20(5), 509–521. https://doi.org/10.1177/1073858413514136

Kesner, L., Grygarová, D., Fajnerová, I., Lukavský, J., Nekovářová, T., Tintěra, J., Zaytseva, Y., & Horáček, J. (2018). Perception of direct vs. averted gaze in portrait paintings: An fMRI and eye-tracking study. Brain and Cognition, 125, 88–99. https://doi.org/10.1016/j.bandc.2018.06.004

Kim, J., & Kim, S. (2019). Finding the optimal d/h ratio for an enclosed urban square: Testing an urban design principle using immersive virtual reality simulation techniques. International Journal of Environmental Research and Public Health, 16(5), 865. https://doi.org/10.3390/ijerph16050865

Koffka, K. (1935). Principles of Gestalt Psychology. Harcourt, Brace and Company.

Koide, N., Kubo, T., Nishida, S., Shibata, T., & Ikeda, K. (2015). Art expertise reduces influence of visual salience on fixation in viewing abstract-paintings. PLOS ONE, 10(2), e0117696. https://doi.org/10.1371/journal.pone.0117696

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Do Hyong Koh, & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. IEEE Transactions on Biomedical Engineering, 57(11), 2635–2645. https://doi.org/10.1109/TBME.2010.2057429

Kyriakou, P., & Hermon, S. (2019). Can I touch this? Using natural interaction in a museum augmented reality system. Digital Applications in Archaeology and Cultural Heritage, 12, e00088. https://doi.org/10.1016/j.daach.2018.e00088

Lagerspetz, M. (2016). Lay perceptions of two modern artworks. Art and Perception, 4(1–2), 107–125. https://doi.org/10.1163/22134913-00002047

Le Meur, O., Le Pen, T., & Cozot, R. (2020). Can we accurately predict where we look at paintings? PLOS ONE, 15(10), e0239980. https://doi.org/10.1371/journal.pone.0239980

Leder, H. (2013). Beyond perception – information processing approaches to art appreciation. In P. P. L. Tinio & J. K. Smith (Eds.), The Cambridge Handbook of the Psychology of Aesthetics and the Arts (pp. 115–138). Cambridge University Press. https://doi.org/10.1017/CBO9781139207058.008

Leder, H., & Nadal, M. (2014). Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode - Developments and challenges in empirical aesthetics. British Journal of Psychology, 105(4), 443–464. https://doi.org/10.1111/bjop.12084

Lee, H., Jung, T. H., tom Dieck, M. C., & Chung, N. (2020). Experiencing immersive virtual reality in museums. Information & Management, 57(5), 103229. https://doi.org/10.1016/j.im.2019.103229

Massaro, D., Savazzi, F., Di Dio, C., Freedberg, D., Gallese, V., Gilli, G., & Marchetti, A. (2012). When art moves the eyes: A behavioral and eye-tracking study. PLOS ONE, 7(5), e37285. https://doi.org/10.1371/journal.pone.0037285

McManus, I. C., Jones, A. L., & Cottrell, J. (1981). The aesthetics of colour. Perception, 10(6), 651–666. https://doi.org/10.1068/p100651

Mokatren, M., Kuflik, T., & Shimshoni, I. (2018). Exploring the potential of a mobile eye tracker as an intuitive indoor pointing device: A case study in cultural heritage. Future Generation Computer Systems, 81, 528–541. https://doi.org/10.1016/j.future.2017.07.007

Mu, M., Dohan, M., Goodyear, A., Hill, G., Johns, C., & Mauthe, A. (2020). User attention and behaviour in virtual reality art encounter. ArXiv. https://arxiv.org/abs/2005.10161

Mullennix, J. W., Varmecky, A., Chan, C. H., Korenoski, S., Mickey, Z., & Polaski-Hoffman, L. (2016). The effect of color on automaticity of aesthetic judgments. Empirical Studies of the Arts, 34(1), 8–34. https://doi.org/10.1177/0276237415621183

Nascimento, S. M. C., Linhares, J. M. M., Montagner, C., João, C. A. R., Amano, K., Alfaro, C., & Bailão, A. (2017). The colors of paintings and viewers' preferences. Vision Research, 130, 76–84. https://doi.org/10.1016/j.visres.2016.11.006

Nayak, B. K., & Karmakar, S. (2019). A review of eye tracking studies related to visual aesthetic experience: A bottom-up approach. In A. Chakrabarti (Ed.), Research into Design for a Connected World (Vol. 135, pp. 391–403). Springer Singapore. https://doi.org/10.1007/978-981-13-5977-4_33

Parker, E., & Saker, M. (2020). Art museums and the incorporation of virtual reality: Examining the impact of VR on spatial and social norms. Convergence: The International Journal of Research into New Media Technologies, 26(5–6), 1–15. https://doi.org/10.1177/1354856519897251

Pelowski, M., Forster, M., Tinio, P. P. L., Scholl, M., & Leder, H. (2017). Beyond the lab: An examination of key factors influencing interaction with 'real' and museum-based art. Psychology of Aesthetics, Creativity, and the Arts, 11(3), 245–264. https://doi.org/10.1037/aca0000141

Pelowski, M., Leder, H., Mitschke, V., Specker, E., Gerger, G., Tinio, P. P. L., Vaporova, E., Bieg, T., & Husslein-Arco, A. (2018). Capturing aesthetic experiences with installation art: An empirical assessment of emotion, evaluations, and mobile eye tracking in Olafur Eliasson's "Baroque, Baroque!" Frontiers in Psychology, 9:1255, 1–22. https://doi.org/10.3389/fpsyg.2018.01255

Pelowski, M., Markey, P. S., Forster, M., Gerger, G., & Leder, H. (2017). Move me, astonish me… delight my eyes and brain: The Vienna Integrated Model of top-down and bottom-up processes in Art Perception (VIMAP) and corresponding affective, evaluative, and neurophysiological correlates. Physics of Life Reviews, 21, 80–125. https://doi.org/10.1016/j.plrev.2017.02.003

Redi, J. A., Hoßfeld, T., Korshunov, P., Mazza, F., Povoa, I., & Keimel, C. (2013). Crowdsourcing-based multimedia subjective evaluations: A case study on image recognizability and aesthetic appeal. Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia - CrowdMM '13, 29–34. https://doi.org/10.1145/2506364.2506368

Rosenberg, R., & Klein, C. (2015). The moving eye of the beholder: Eye tracking and the perception of paintings. In J. P. Huston, M. Nadal, F. Mora, L. F. Agnati, & C. J. C. Conde (Eds.), Art, Aesthetics, and the Brain (pp. 79–108). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199670000.003.0005

Saiz, A., Salazar, A., & Bernard, J. (2018). Crowdsourcing architectural beauty: Online photo frequency predicts building aesthetic ratings. PLOS ONE, 13(7), e0194369. https://doi.org/10.1371/journal.pone.0194369

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '00, 71–78. https://doi.org/10.1145/355017.355028

Shehade, M., & Stylianou-Lambert, T. (2020). Virtual reality in museums: Exploring the experiences of museum professionals. Applied Sciences, 10(11), 4031. https://doi.org/10.3390/app10114031

Škola, F., Rizvić, S., Cozza, M., Barbieri, L., Bruno, F., Skarlatos, D., & Liarokapis, F. (2020). Virtual reality with 360-video storytelling in cultural heritage: Study of presence, engagement, and immersion. Sensors, 20(20), 5851. https://doi.org/10.3390/s20205851

Tennent, P., Martindale, S., Benford, S., Darzentas, D., Brundell, P., & Collishaw, M. (2020). Thresholds: Embedding virtual reality in the museum. Journal on Computing and Cultural Heritage, 13(2), 1–35. https://doi.org/10.1145/3369394

Tschacher, W., Greenwood, S., Kirchberg, V., Wintzerith, S., van den Berg, K., & Tröndle, M. (2012). Physiological correlates of aesthetic perception of artworks in a museum. Psychology of Aesthetics, Creativity, and the Arts, 6(1), 96–103. https://doi.org/10.1037/a0023845

Tyack, C., Camic, P. M., Heron, M. J., & Hulbert, S. (2017). Viewing art on a tablet computer: A well-being intervention for people with dementia and their caregivers. Journal of Applied Gerontology, 36(7), 864–894. https://doi.org/10.1177/0733464815617287

Vartanian, O., & Skov, M. (2014). Neural correlates of viewing paintings: Evidence from a quantitative meta-analysis of functional magnetic resonance imaging data. Brain and Cognition, 87, 52–56. https://doi.org/10.1016/j.bandc.2014.03.004

Walker, F., Bucker, B., Anderson, N. C., Schreij, D., & Theeuwes, J. (2017). Looking at paintings in the Vincent Van Gogh Museum: Eye movement patterns of children and adults. PLOS ONE, 12(6), e0178912. https://doi.org/10.1371/journal.pone.0178912

Wallraven, C., Cunningham, D. W., Rigau, J., Feixas, M., & Sbert, M. (2009). Aesthetic appraisal of art—From eye movements to computers. Computational Aesthetics in Graphics, Visualization, 8 pages. https://doi.org/10.2312/COMPAESTH/COMPAESTH09/137-144

Wanzer, D. L., Finley, K. P., Zarian, S., & Cortez, N. (2020). Experiencing flow while viewing art: Development of the Aesthetic Experience Questionnaire. Psychology of Aesthetics, Creativity, and the Arts, 14(1), 113–124. https://doi.org/10.1037/aca0000203

Wenzel, C. H. (2005). An introduction to Kant's aesthetics: Core concepts and problems. Blackwell Pub.

Yanulevskaya, V., Uijlings, J., Bruni, E., Sartori, A., Zamboni, E., Bacci, F., Melcher, D., & Sebe, N. (2012). In the eye of the beholder: Employing statistical analysis and eye tracking for analyzing abstract paintings. Proceedings of the 20th ACM International Conference on Multimedia - MM '12, 349. https://doi.org/10.1145/2393347.2393399

Yarbus, A. L. (1967). Eye movements and vision. Springer. https://doi.org/10.1007/978-1-4899-5379-7

Yi, T., Chang, M., Hong, S., Kim, M., & Lee, J.-H. (2020). A study on understanding of visitor needs in art museum: Based on analysis of visual perception through eye-tracking. In T. Ahram, W. Karwowski, A. Vergnano, F. Leali, & R. Taiar (Eds.), Intelligent Human Systems Integration 2020 (Vol. 1131, pp. 1132–1137). Springer International Publishing. https://doi.org/10.1007/978-3-030-39512-4_172

Zhang, L.-M., Zhang, R.-X., Jeng, T.-S., & Zeng, Z.-Y. (2019). Cityscape protection using VR and eye tracking technology. Journal of Visual Communication and Image Representation, 64, 102639. https://doi.org/10.1016/j.jvcir.2019.102639

5.7. Figures

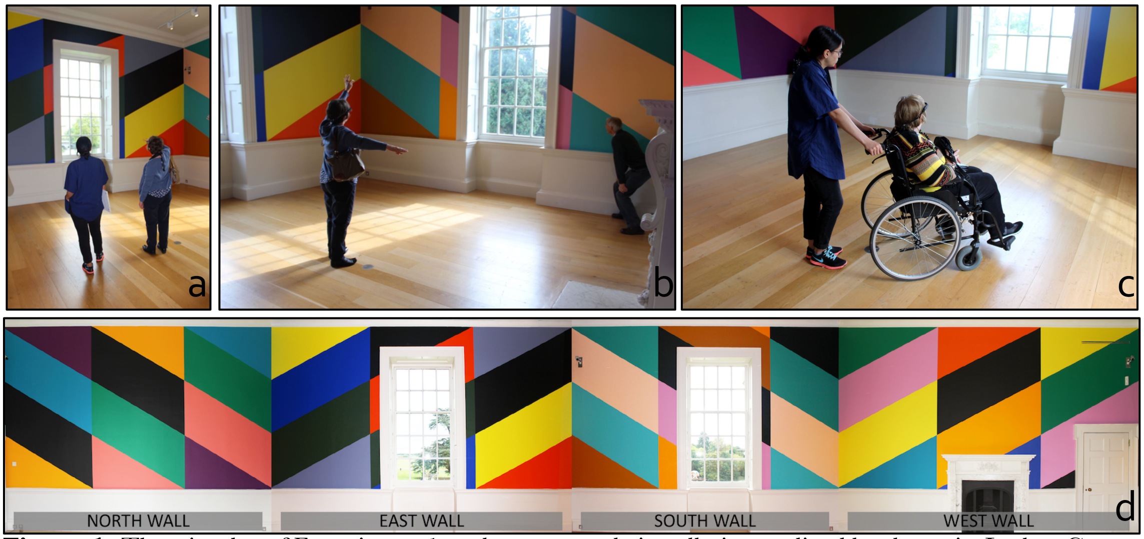


Figure 1. The stimulus of Experiment 1 as the room-scale installation realized by the artist Lothar Goetz, exhibited in the Compton Verney Art Gallery. (a-b-c) Three exemplar photographs from the installation; (d) and a composite image generated by combining photographs of four walls in the cardinal directions side by side, to illustrate the overall layout of the installation.

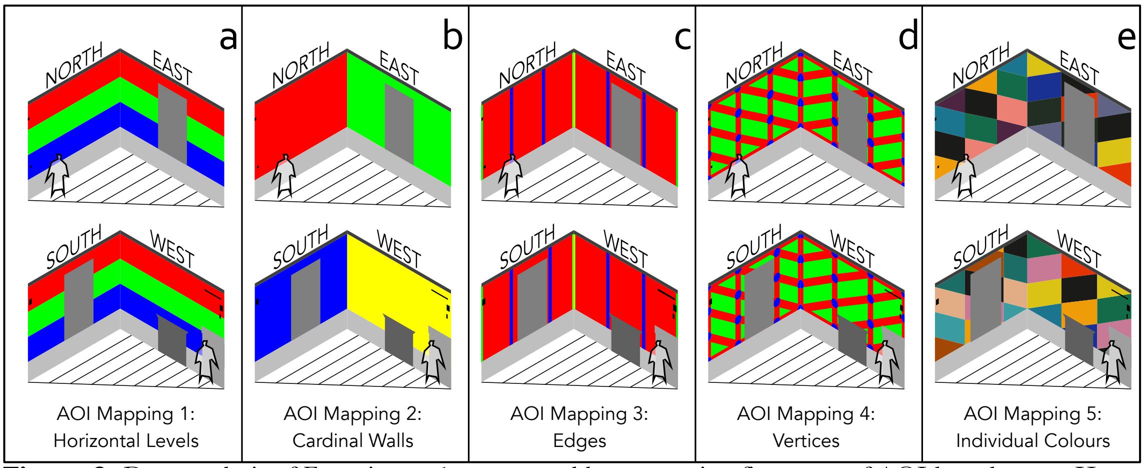


Figure 2. Data analysis of Experiment 1 was started by generating five types of AOI-based maps. Here, the coloured areas denote individual AOIs, and grey areas denote the areas outside of artwork such as the fireplace, door, windows, etc., which were excluded from the analysis. (a) Horizontal levels divided into three parts: upper (red), middle (green), and lower (blue) levels. (b) Cardinal walls divided into four parts: north (red), east (green), south (blue), and west (yellow). (c) Vertical edges divided into four parts: walls (red), convex edges (green), concave edges (yellow), illusory edges (blue). (d) Vertices divided into three parts: edges (red), intersections (blue) walls (green). (e) Individual colours divided into sixteen parts: coloured individually.

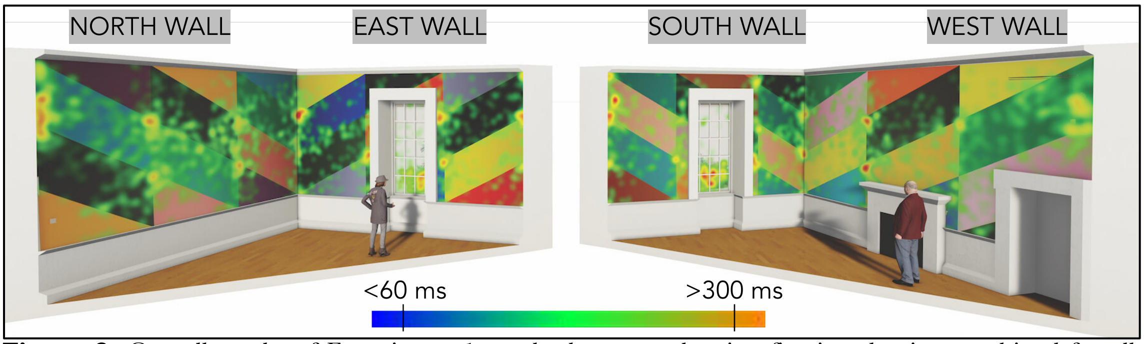


Figure 3. Overall results of Experiment 1, as the heatmap showing fixation density, combined for all participants, and layered on top of the rendered view of the gallery space derived from the 3D digital model, 13 participants. The heatmap scale illustrated below is colour-coded as a gradient ranging from blue-green to orange-red, and corresponding to the shorter to longer fixation duration, respectively.

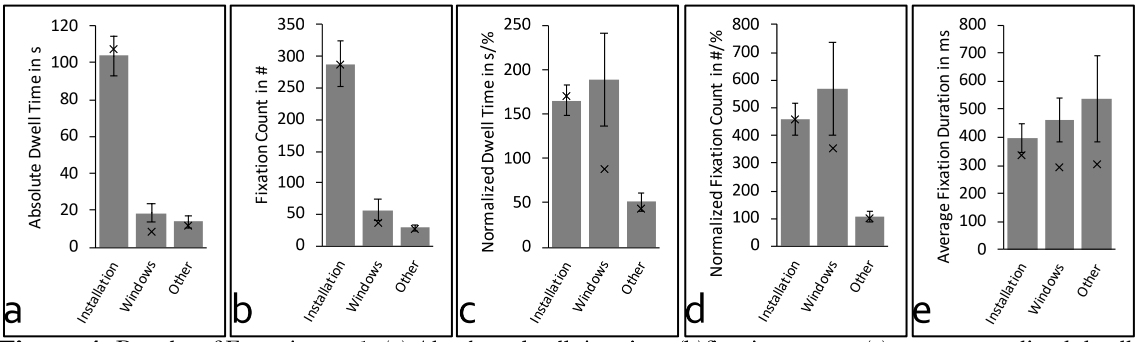


Figure 4. Results of Experiment 1. (a) Absolute dwell time in s, (b)fixation count, (c) area-normalized dwell time in s/%, (d)area-normalized fixation count, and (e) average fixation duration in ms. The categorical x-axis denotes three sets of main AOIs: the installation itself, windows of the gallery room, and all other surface areas including the door, fireplace, crown moulding, baseboard, etc. Bars indicate means with error bars showing ±1SEM, and crosses denote median, $N_{\text{Participant}} = 13$.

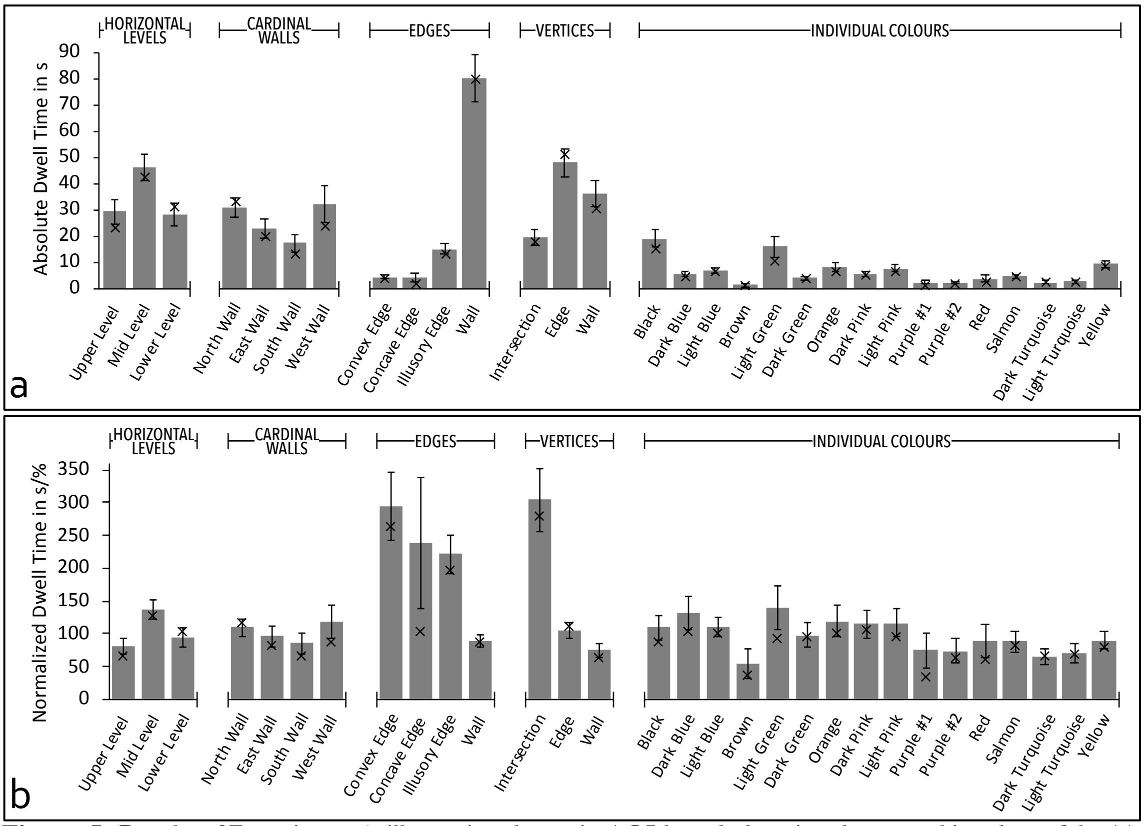


Figure 5. Results of Experiment 1, illustrating the main AOI-based, durational eye-tracking data of the (a) absolute dwell time in s, i.e. cumulative fixation duration in s, and (b) the area-normalized dwell time in s/%, i.e. cumulative fixation duration in s relative to the AOI size in percentage. The categorically separated x-axis denotes five sets of AOIs: horizontal levels, cardinal walls, edges, vertices, and individual colours. Bars indicate means with error bars showing ±1SEM, and crosses denote median, $N_{\text{Participant}} = 13$.

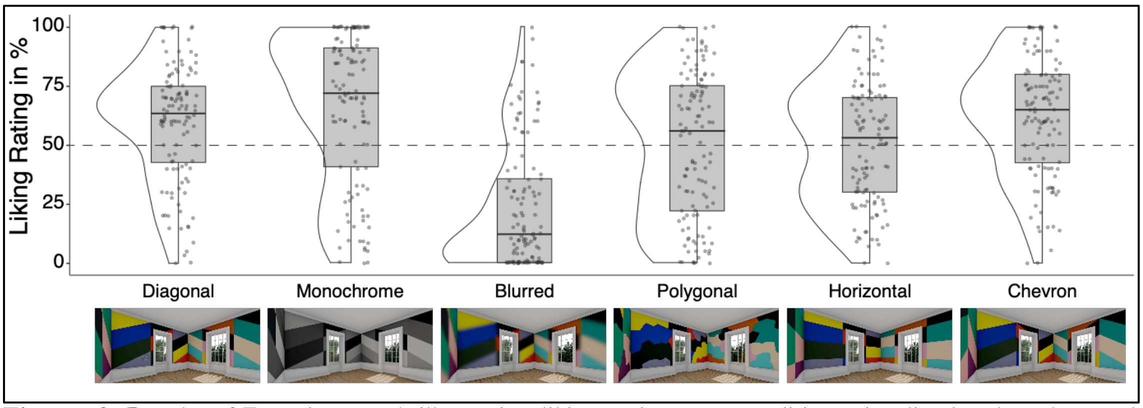


Figure 6. Results of Experiment 2A illustrating liking ratings per condition, visualized as boxplots and density curves drawn from the distribution of the individual data points, where $N_{\text{Participant}}$ = 120 for each condition. The density estimations as density curves drawn on the left-side of the plots are showing the data distribution succinctly, compared to a histogram. Dots as individual ratings are slightly and randomly jittered along the x-axis to illustrate all data on the plots. Six versions of the installation were illustrated on the x-axis, and the y-axis denotes the rating scores in percentage. 0% and 100% on the y-axis refer to the minimum and maximum possible rating a participant could provide, labelled as the least liked and the most liked during the experiment. If the mid-value of 50% can be defined as a threshold point of a neither liked nor disliked artwork (i.e., a bisection threshold), then a rating score above the threshold line can be defined as a liked artwork, and a rating score below the line as a disliked artwork.

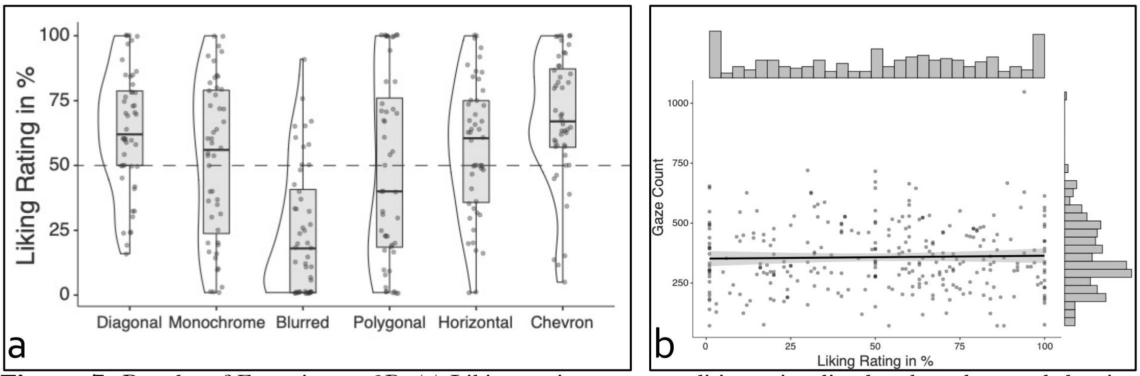
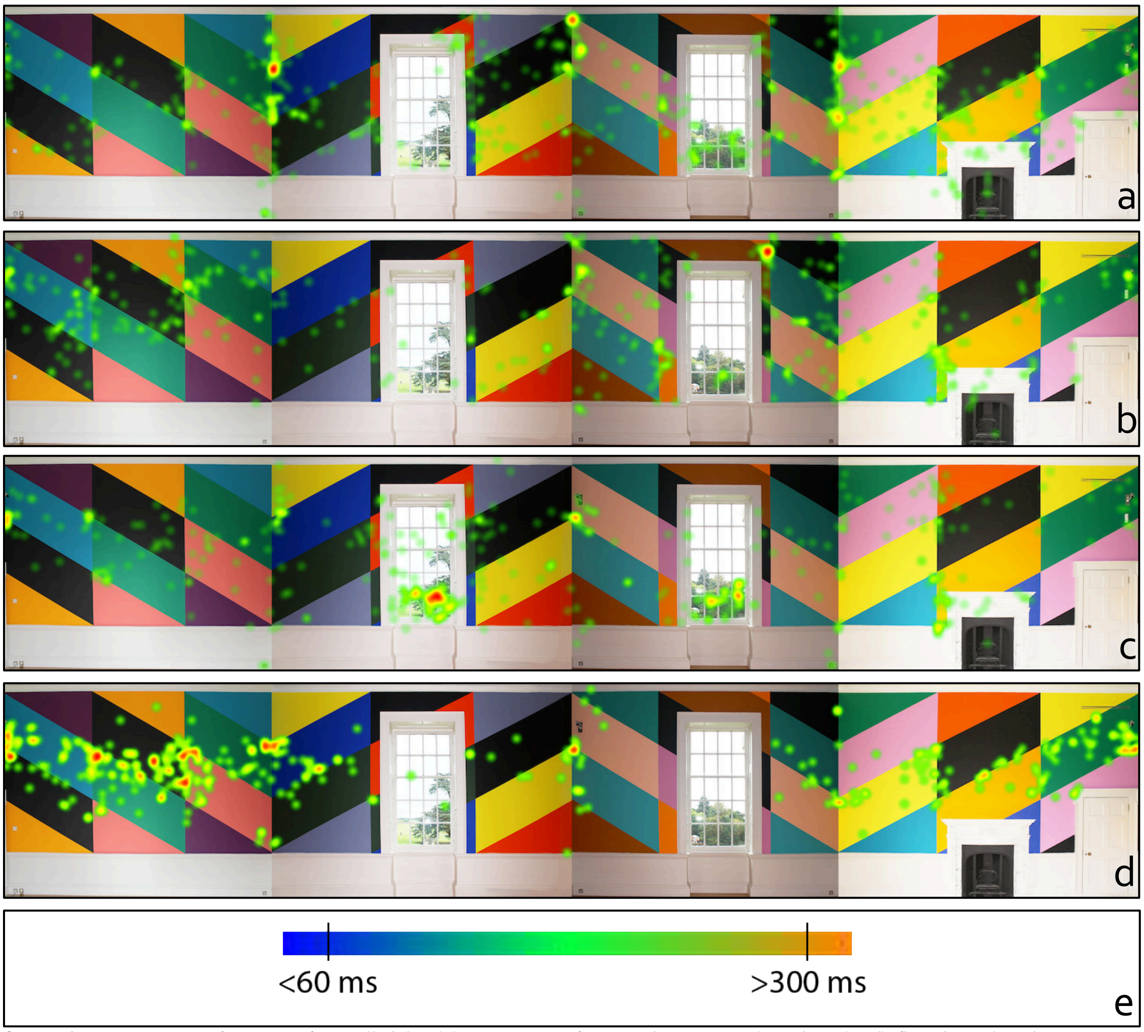
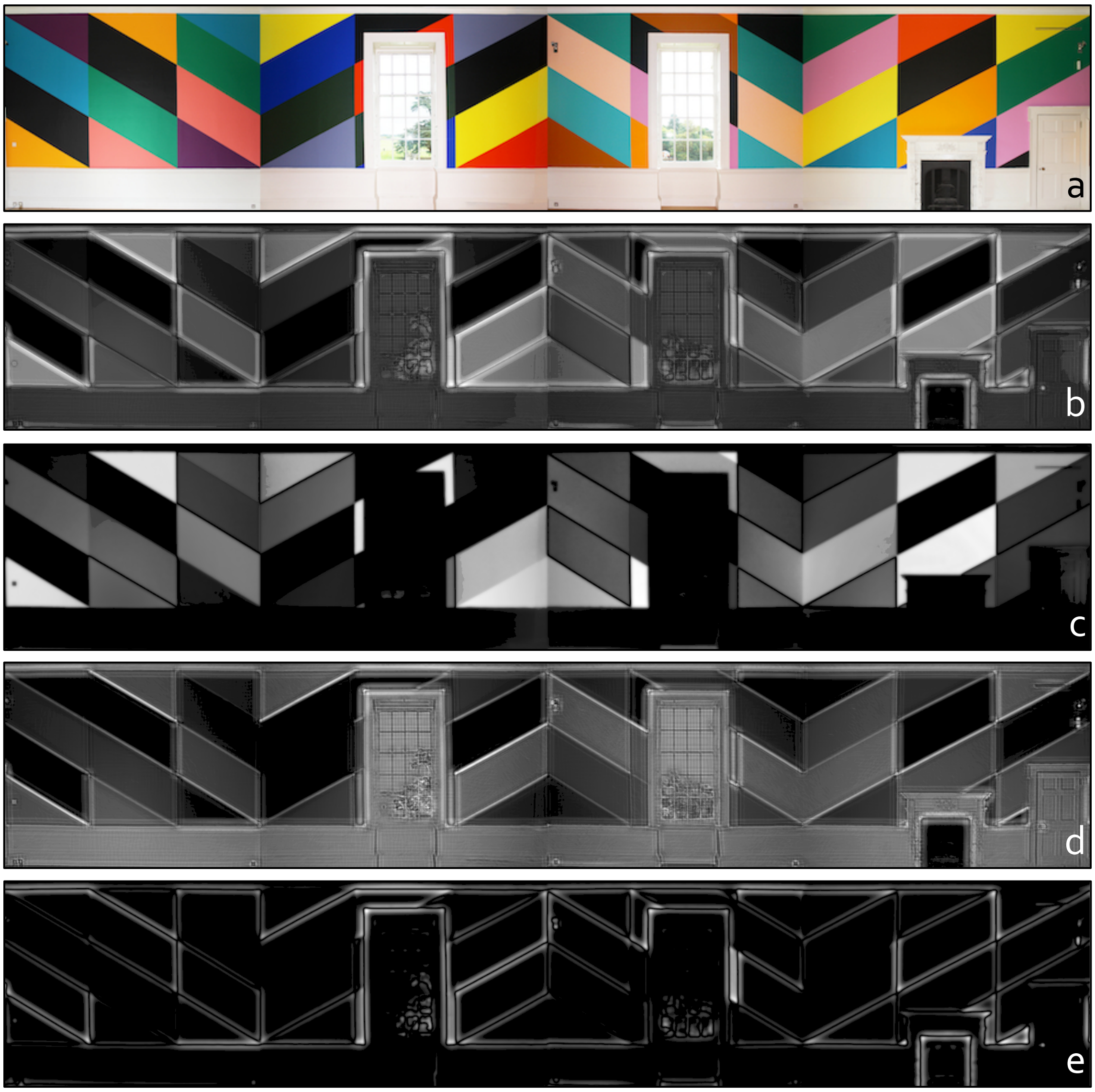


Figure 7. Results of Experiment 2B: (a) Liking ratings per condition, visualized as boxplots and density curves drawn from the distribution of the individual data points, where $N_{\text{Participant}}$ = 48 for each condition. The density estimations as density curves drawn on the left-side of the plots are showing the data distribution succinctly, compared to a histogram. Dots as individual ratings are slightly and randomly jittered along the x-axis to illustrate all data on the plots (b) Lack of relation between liking rating and gaze count irrespective of the artwork variation, where each individual data point came from the gaze count during free viewing for a single artwork variation. The sample size was $N_{\text{Participant}}$ = 48, and the number of observations was $N_{\text{Observation}}$ = 6 per participant.

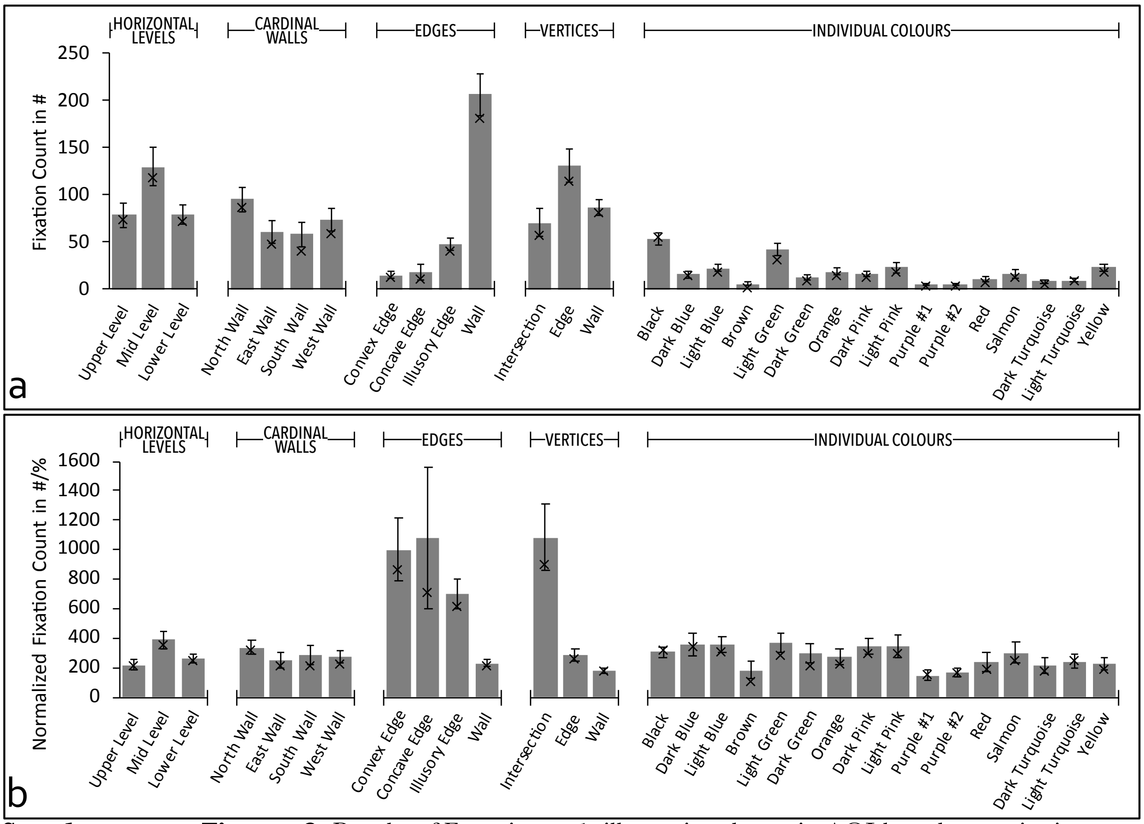
5.8. Supplementary figures



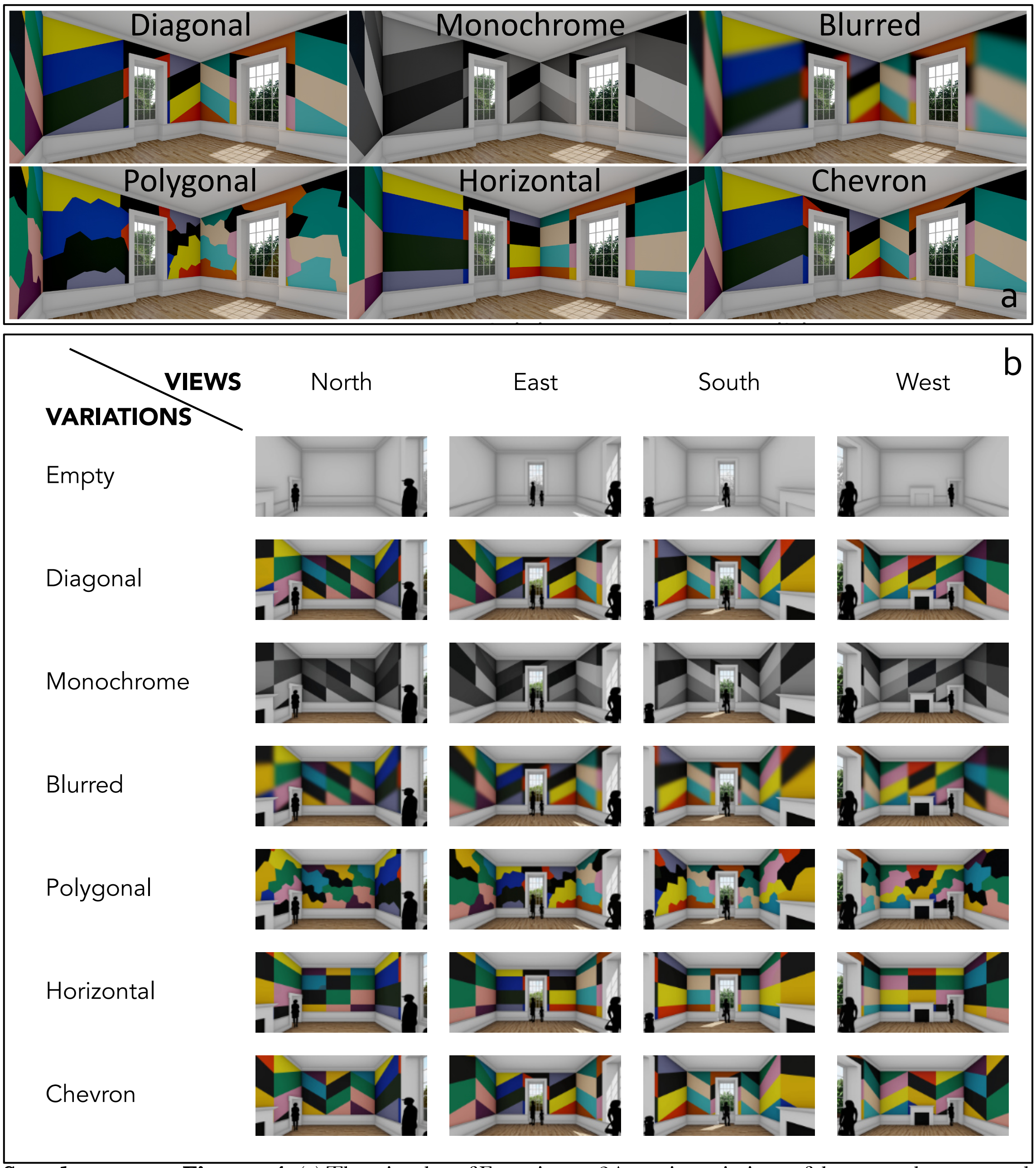
Supplementary Figure s1. Individual heatmaps of Experiment 1, showing (a-d) fixation density on top of the projected view of the room, from four exemplar participants, (e) with a shared scale.



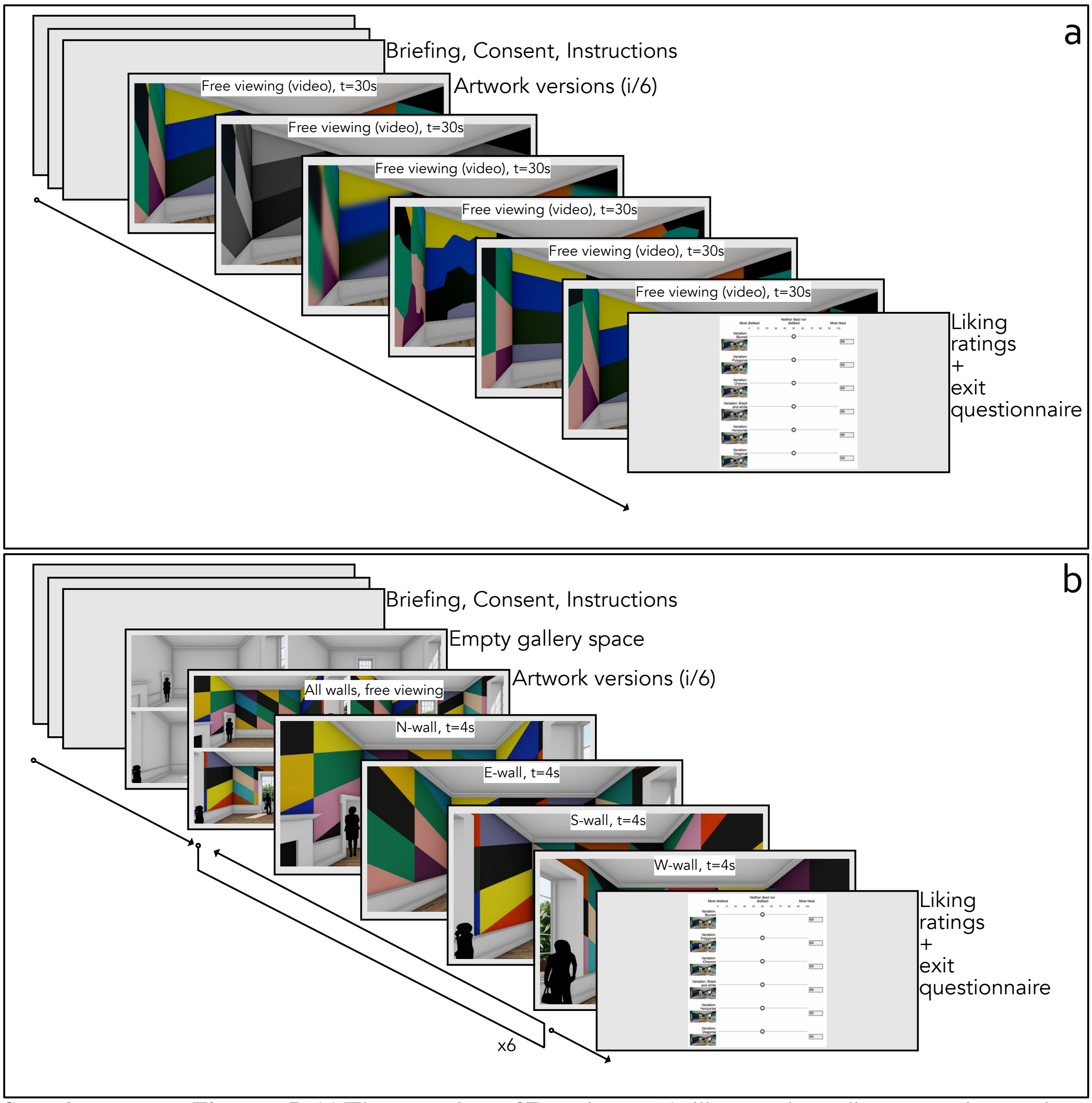
Supplementary Figure s2. (a) The projected view of the room-scale installation, (b) a saliency map generated using the Itti algorithm (Itti et al., 1998) and the saliency map's (c) colour, (d) orientation, and intensity (e) components. The gist of the saliency map is to combine various types of image features (i.e., in this example, three) into a single map. In brief, the intensity component is roughly derived from the local brightness changes on the image, the colour component is roughly derived from the hue information of the image, and the orientation component is roughly derived from local directions and taking into account both the local brightness and hue changes. For the algorithm, following the default settings to calculate features using linear centre-surround operations in the RGB colour space, the centre is defined as a pixel at scale $c \in \{2, 3, 4\}$, and the surround is defined as the corresponding pixel at scale $s = c + \delta$ where $\delta \in \{3, 4\}$. Note that, the intensity layer of the saliency map partially overlaps with the heatmap derived from fixations, as well as with the normalized dwell time analysis, since participants' visual attention were relatively more focused on the edges compared to the other areas.



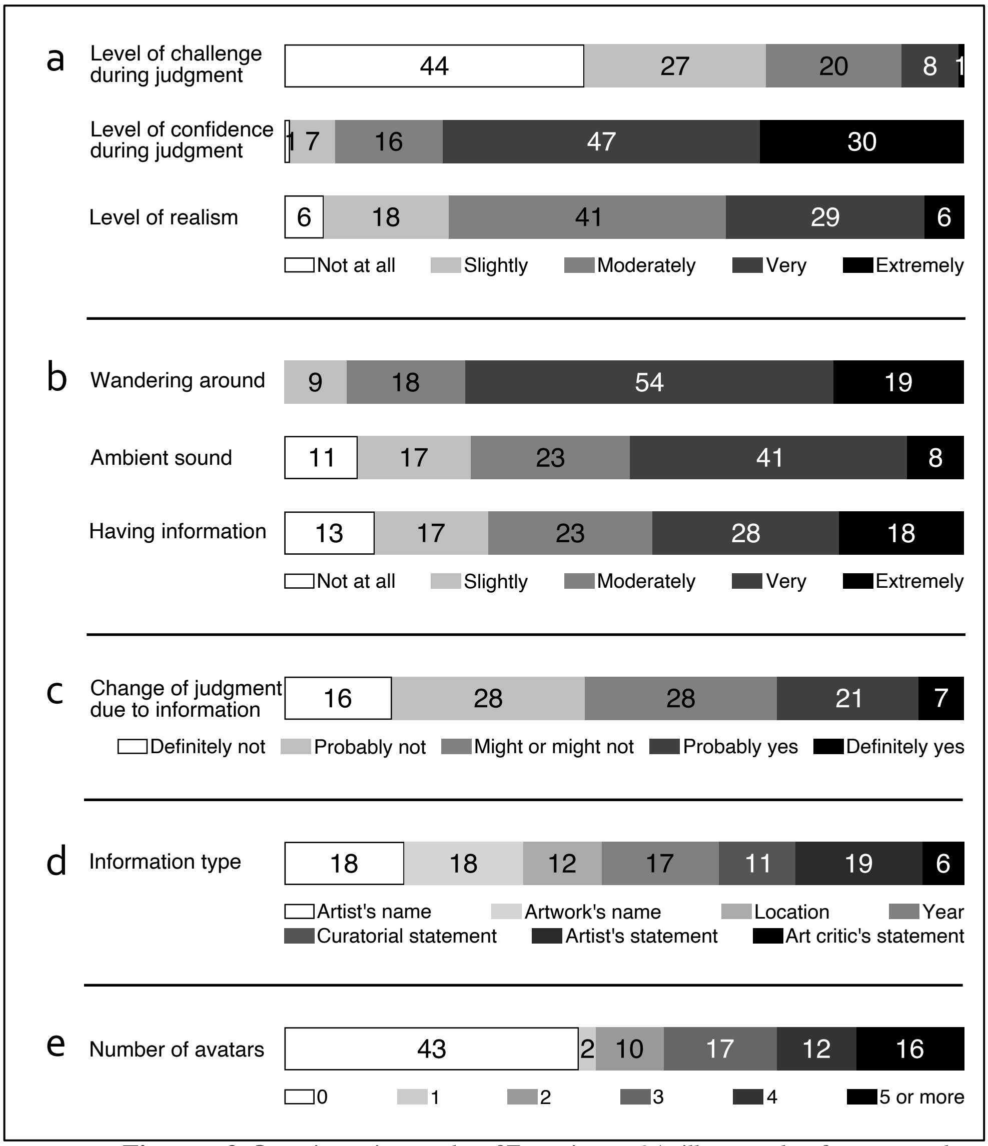
Supplementary Figure s3. Results of Experiment 1, illustrating the main AOI-based, quantitative eye-tracking data of the (a) fixation count, and (b) area-normalized fixation count. The categorically separated x-axis denotes five sets of AOIs: horizontal levels, cardinal walls, edges, vertices, and individual colours. Bars indicate means with error bars showing ± 1SEM, and crosses denote median, $N_{\text{Participant}}$ = 13.



Supplementary Figure s4. (a) The stimulus of Experiment 2A, as six variations of the artwork, presented as 30 s long video renders to the participants, and illustrated here as single 2D snapshot images from those videos, showing the east wall on the left side and the west wall on the right side of the snapshots. The versions of the installation were labelled as diagonal, monochrome, blurred, polygonal, horizontal, and chevron during the experiment. (b) The stimulus of Experiment 2B, as a set of artwork variations, illustrated as a matrix overview. Each row of the matrix shows a single variation of the artwork, including the first row illustrating the default gallery room view (labelled as empty) and each of the following rows illustrating one of six different variations of the installation (labelled as diagonal, monochrome, blurred, polygonal, horizontal, and chevron). Each column shows one of four cardinal views of the gallery space (labelled as north, east, south, and west, which were approximately based on the cardinal directions of the original gallery room). Note that, the human silhouettes in the rendered images were included to better illustrate the scale of the installation to the participants, and the installation itself was never occluded by any of those human figures in any of the views.



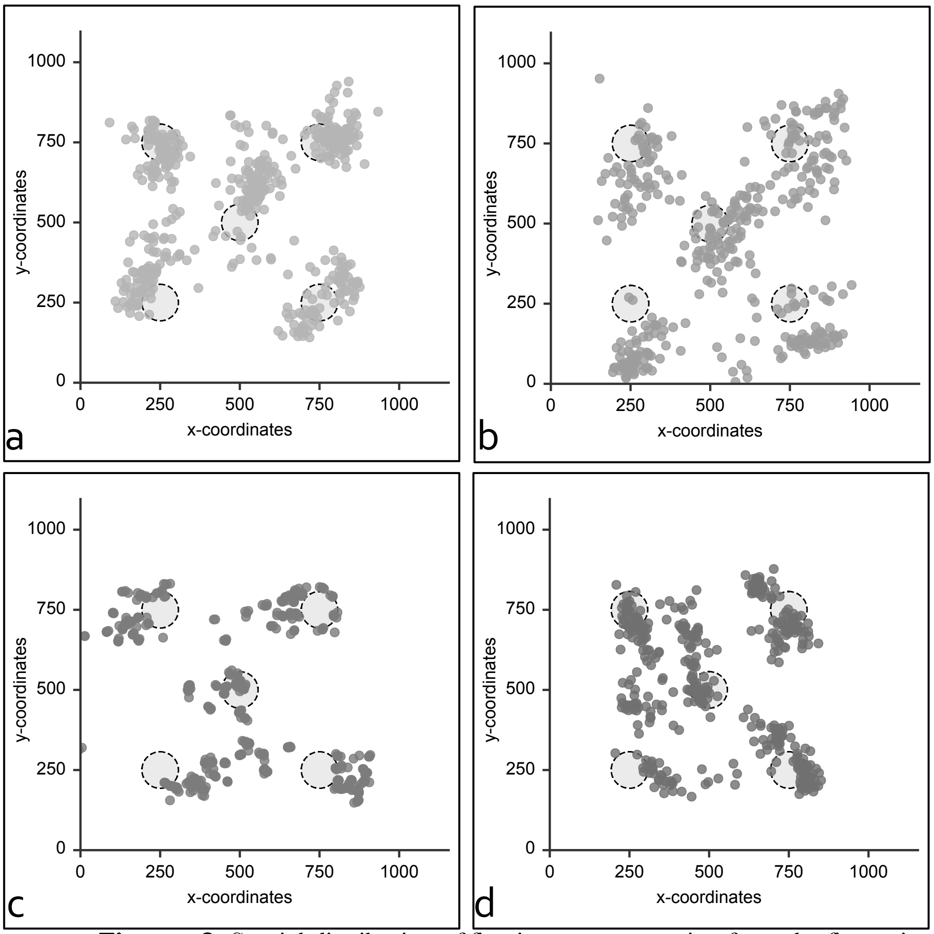
Supplementary Figure s5. (a) The procedure of Experiment 2A, illustrated as a diagrammatic overview. Following the briefing, receiving consent, and providing instructions; participants viewed each of the six variations of the artwork, rendered as 30 s long videos, one by one in a randomized order. After that they provided liking ratings on a continuous scale ranging from 0 (labelled as least liked) to 100 (labelled as most liked), also accompanied with 2D snapshot renders referring to the artwork variation. Lastly, they were asked to provide responses to the exit questionnaire items. (b) The procedure of Experiment 2B, illustrated as a diagrammatic overview. Following the briefing, receiving consent, and providing instructions; participants firstly viewed the empty gallery space on a single webpage, where four cardinal views of the room (namely, views of north, east, south, and west walls) were presented in a two-by-two layout. After that, participants viewed each artwork variation in the same layout without a time limit, followed by viewing each of the four cardinal directions, each lasting 4 s. The viewing order of the variations was randomized. After that, they provided liking ratings on a continuous scale ranging from 0 (labelled as least liked) to 100 (labelled as most liked), also accompanied with 2D snapshot renders referring to the artwork variation. Lastly, they were asked to provide responses to the exit questionnaire items.



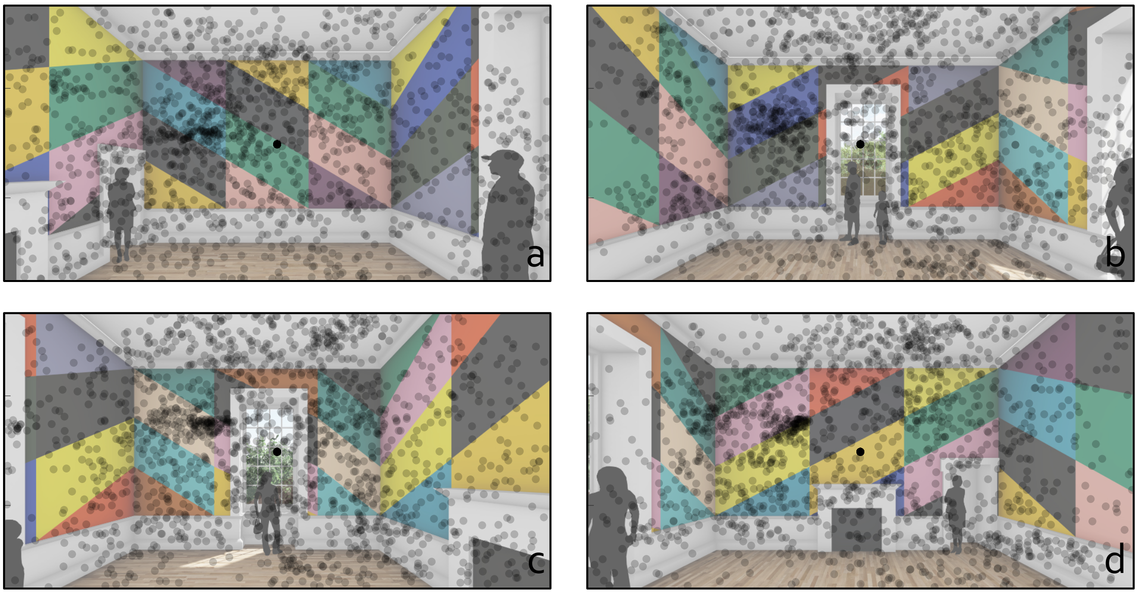
Supplementary Figure s6: Questionnaire results of Experiment 2A, illustrated as frequency plots, where the response frequencies corresponding to the possible levels for each item were reported as frequencies, rounded to the nearest integer: (a) the perceived level of challenge and confidence on judgments, and level of realism of view renders, (b) the amount of helpfulness of potential features in VR during viewing experience, (c) participant's view whether having information about the artist and artwork would change their judgments, such as liking ratings, (d) type of information that participants would like to have, and (e) the preferred number of digital human-like figures or avatars, to make the digital viewing experience more realistic. Note, all items were on a five-point Likert scale, resulting in $N_{participant} = N_{observation} = 120$, with the exception of information type (d), which was a multiple-choice item and resulted in $N_{observation} = 424$.

The reasoning behind the judgment for the most liked artwork	The reasoning behind the judgment for the least liked artwork
I thought the B&W scheme would go with more forms of furniture and wasn't so 'in your face'. It also seemed a bit more 'classy'.	Would hurt the eyes and just look bad.
I preferred black and white to the many different colours that didn't seem to work together to me	I just didn't like looking at the blurred lines
It's looking good, the colors are matching, in the same tone	It just looks boring and incoherent
I really appreciate the way a 2d decoration can define the 3d perception of a room	The moving wall effect wasn't strong enough
I liked the polygonal video because of unstricted randomized shapes. It's much more creative than strict lines on most of other videos.	From my opinion horizontal bricks is most boring painting. I may associate vertical bricks with books, but lying bricks is just bricks.
I liked the horizontal lines - it made me feel less 'dizzy' when viewing the artwork!	I did not like the version without colour. It was much less striking than the others.
The Chevron pattern looked nice on all of the walls, whereas for some of the other options the pattern might have looked stunning from a few perspectives, but not throughout the entire room (which when only viewing one wall at a time was relatively important). It was also the most interesting without being too busy - if that makes sense.	A mess. Just chaos but not in a beautiful way. Got no pleasure out of seeing this art installation.
It makes the room seem bigger and deeper and the variation colours doesn't seem so abrupt.	It makes the room seem way smaller and makes me feel sad in some ways.
I really like the Polygonal look-- it seems like what I would most like to actually see in real life, or even have in my own home.	For some reason, the Horizontal one felt a bit uneasy to look at.
I could imagine spending the most time in that room. I liked the bold colours and the simplicity of the horizontal areas. It probably played with my eyes less that some of the others!	This reminded me of interior decor during a period of the late 80s/early 90s when there was a black, grey and red colour scheme. The block shapes of the colours remind me of this. I vividly remember a friends room with the wallpaper and curtains in shapes of those colours and black furniture.

Supplementary Figure s7: Partial open-ended feedback of the questionnaire items from Experiment 2A, as the reasoning behind the judgment for the most and least liked artwork variations from 10 exemplar participants. Note that the two types of reasoning on each row belong to the same participant, and the typographical errors of these listed exemplar responses were not corrected.



Supplementary Figure s8. Spatial distribution of fixation counts, coming from the five-point accuracy check phase. This exemplar selection was based on the data of four participants (i.e., a-b-c-d for participant #1, #8, #19, and #23, respectively). The plots particularly aim to show some of the most accurate participant data, following the visual inspection of all participant data. The x- and y-axis correspond to the screen positions units, where 1000 units were defined as the height of the participants' monitor in a landscape full-screen (and therefore, 1 unit roughly equals to 1 pixel). During the experiment, the square calibration screen was presented with a resolution of 1000 X 1000 units (e.g., for a participant with a monitor resolution of 1920 X 1080 pixels, 1 unit = 1.08 pixels). Individual datapoints were the registered and estimated gaze points. Five circular areas with dashed strokes represent the positions of the calibration dots, and each calibration dot was presented one by one for 2000 ms, both before and after the experiment. Note, in total, 17,396 gaze points were recorded for all participants (N = 48), and the duration of this accuracy check phase was 20 s per participant. Thus, on average, 362.41 gaze points were recorded per participant, which equates the mean frequency of webcam recording to 18.12 Hz.



Supplementary Figure s9. Part of the aggregated eye-tracking data from online Experiment 2B, illustrating the registered gaze data (as a proxy for fixation points) on the (a) north, (b) east, (c) south, and (d) west walls of only the original artwork version, where each fixation is at 10-unit radius, and each artwork view is at 500 by 1000 units, $N_{\text{Participant}} = 48$, $t_{\text{Recording}} = 4$ s per view and per participant.

Chapter 6

Overall discussion

In this section, I revisit the overarching research aims, and how these aims were addressed in specific experiments, followed by a brief summary and evaluation of the most relevant findings. Following on from that, I briefly highlight strengths and limitations of this research, future directions, and conclude with general remarks. Since the findings of individual empirical chapters were independently discussed in-depth in their corresponding discussion sections, this general discussion can be read as a brief reflection of the overall approach to empirical aesthetics, and not as a detailed summary or a concluding review.

6.1. Summary of findings

The Mondrian Room in the Albertinum Museum (thesis chapter 3) was one particular case study, aiming to compare gaze patterns of museum visitors as partial constituents of the aesthetic experience across physical and virtual versions of an installation artwork, based on Piet Mondrian's room design. Using mobile and VR eye-trackers and primarily analysing fixation-related metrics based on a set of AOI-maps, we found an overwhelming similarity of eye movements between the two settings. The results demonstrated the ability to make a relatively objective comparison between the two instantiations of the artwork, without relying on the subjective responses. Lastly, assuming the observed, partially measurable, physical contents of visual experience were similar across two settings, at least based on particular types of eye-tracking analysis, then speculatively, the similarity of the phenomenal content of experience could be suggested. One key implication of this comparative experiment was that a virtual installation art could be a suitable enough proxy for a physical installation art. Following more rigorous testing in the future, alongside the enhanced realism in virtual environments and accessibility to such settings, the ecological validity of simulated environments to engage with visual arts can be implied.

The Virtual Reality Gallery (thesis chapter 4) started with a pilot experiment aiming to select a set of a variety of 3D models as artworks, leading to the inclusion of digital replicas of the existing physical objects and spaces, as well as purely digital objects and spaces that do not have correspondence in the real world. Following the model selection, the main VR-based study aimed to explore the relationship between different types of aesthetic judgments of three-dimensional artworks in an immersive experience. Lastly, the study concluded with an online follow-up experiment aiming to create an accessible setup for a remote audience. In the main VR experiment, besides the supplementary eye-tracking and questionnaire data, the core data were the ratings measured in a judgment task. The main results showed mostly positive, linear, and moderate correlations between liking and the other perceived judgment attributes, such as perceived novelty, perceived complexity, and perceived viewing duration. Supplementary VR eye-tracking analysis showed a variety of viewing strategies, both across participants and across artworks; and the online follow-up also showed converging evidence on correlations. In the methodological sense, the original contribution to the area of the research was demonstrating the validity of a relatively novel research setting: the 3D artworks as stimuli were consisting of a variety of objects and environments, without sharing any common or obvious property deliberatively, and even in this fringe case, the relations between aesthetic judgments were investigated. It was speculated that both VR-only or online-only settings offer ecologically-valid behavioural experimental contexts, create immersive visual arts experience, and have the potential to enhance accessibility to cultural heritage.

The Compton Verney Mural (thesis chapter 5) aimed to make use of eye-tracking data from an in-situ installation, where gallery visitors' visual attention was mainly on the edges and intersections of the room-scale abstract mural installation. Based on the captured viewing trends from a mobile eye-tracker in the in-situ experiment, as well as the artwork's topological layout, five additional "variations" of artwork were created as 3D models. These variations were then presented in two consecutive online experiments, the first one aiming to investigate observers' liking ratings, judgment strategies, and reasoning behind their judgments on these variations; and the second one aiming to purely implement an online-eye tracking method as a promising and emerging research tool. Liking ratings on those variations showed high variance, participants showed a variety of nuanced reasoning behind their judgments, but online eye-tracking data quality was very variable between participants. The variability could be attributed to two related aspects as the eye-tracking algorithm's detection quality given the variable testing conditions as well as the behavioural differences between participants, for example, varied hardware, lighting conditions and the amount of head movement during the experiment. A potential implementation of this research workflow was not necessarily about "art production" but design optimization, for example, as a data-driven participatory design process.

6.2. Novelty and strength

The main strength of the research is its ability to create transferable knowledge beyond the domain of empirical aesthetics, particularly in the methodological aspects. For example, in the most general sense, this research showed the overlooked potential of making use of the novel behavioural data measured in the real-world or in close proxies of it. In one particular sense, the emerging eye-tracking analysis using 3D area-of-interests (or more like "volume-of-interests") in-situ and in VR settings can go beyond the basic research questions about aesthetics. Therefore, the methods employed in the empirical chapters can be easily adapted to answer research questions, for example, in cognitive psychology and game studies. The methods can be easily adapted into curatorial practice as a tool for decision making, design practice as a tool for participatory design, and immersive A-B testing (i.e., as a two-sample hypothesis testing) in the creative industry as a tool for assessing potential consumer behaviour.

Additionally, throughout the empirical research, unusual questions were asked to participants, either as additions in a judgment task or in a questionnaire. This aspect was aiming to create further specific insight into the aesthetic experience and judgment. This approach is often disregarded in psychological research compared to the bold and concrete results (but sometimes praised, for example, in the emerging field of experimental philosophy). It is my belief that at least for some researchers, these minor results should not be overshadowed by the main reported findings, and these minor findings could be equally stimulating, for example, compared to in-depth AOI-based eye-tracking analysis.

6.3. Challenges and limitations

All studies can be defined as behavioural experiments, in line with the cognitive studies, and mainly designed to be set in relatively realistic environments. This decision inevitably translates into relatively limited control over the stimulus and the environment, particularly from vision science's point of view about strict experimental requirements and rigorous stimulus control. Similarly, neither the data from VR nor from mobile nor from online eye-tracking are matched to the quality of the research-grade desktop eye-trackers, and similarly, the data from a freely moving participant are also not matched to the data coming from a restricted lab participant. Moreover, the visual quality of VR-HMDs is relatively low fidelity compared to the calibrated research-grade 2D desktop monitors. These limitations are expected to be resolved over time, especially in line with current and future hardware and software developments.

All the analyses were based on the cumulative, atemporal data, for example, in the form of the dwell time as the cumulative fixation duration or the average of a particular judgment rating. The time-related aspects were partially addressed as part of supplementary research questions, and for example, the discrepancy between real viewing duration and perceived viewing duration estimation was found during the VR artwork viewing, and the importance of the viewing duration on liking and other judgments were discussed. Therefore, the temporal dynamics is a missing element. In an alternative experimental design (i.e., one not as an exploratory research), the temporal domain could have been the focus of the research questions, for example, particularly investigating how an aesthetic episode evolves over time using time series analysis.

The studies at their core were exploratory research, and arguably their research questions were not strongly based on theories. Therefore, the studies could not be easily described as rigorous hypothesis testing, and this issue can be seen as a potential weakness. This weakness arguably aligns with the general characteristics of the field: partially mentioned in the introduction section, the inclusivity of empirical aesthetics does often result in the lack of solid research expectations or an overarching research agenda, compared to the other specialized areas of vision sciences. This in turn minimizes the cumulative collection potential of knowledge in a single particular direction and makes theory/model building a challenging task. Speculatively, steps towards a theory/model of visual aesthetic experience in a traditional scientific sense might not necessarily be the ultimate, or at least current aim for many researchers.

Additionally, sample size and stimulus set size were limited in the traditional sense, which is also often a shared weakness in empirical aesthetics. This reduces the generalizability of findings, even though the empirical aesthetics research often does not assume universal expectations.

6.4. Future directions

The research presented here relied on some basic and common-sense descriptions about the concepts of the artwork and the spectator. Obviously, in the current state of the art-world, an artwork can be more than just a static installation, and a spectator can be more than just a museum visitor. As one particular example to illustrate, in participatory art practice or interactive art or relational art, traditional definitions of the concepts radically change depending on the artwork, and the boundaries between the artist, agent, and art get blurred: as for example, the agent may become part of the artwork, the artist might have a diminished or non-existing role, the artwork may be defined as an event compared to the traditional physical object, or not be defined at all. Moreover, AI-assisted or AI-generated art as peculiar “movements” compared to the precursor works in generative and algorithm art have gained high interest in the last years. Such instances might offer more up-to-date and interesting research settings. Similarly, since valuable insights were gained from participants, particularly in online experiments in the form of open-ended responses, during the initial study design process, developing the research alongside working with the artist and spectators might provide beneficial, eye-opening perspectives.

Simulated environments using virtual, augmented, or mixed reality tools offer a way of recreating the existing physical settings, enhancing those settings, as well as generating completely novel experiences. Some of these experiences were drafted during the thesis process as technical demos and learning examples. To briefly illustrate the current and upcoming research potential and use-cases, further research can particularly incorporate, for example, photogrammetric workflows, augmented reality, thematic VR exhibitions, physically impossible spaces such as fractal environments and 3D visual illusions, multi-user environments, interactivity and gamified environments, design on public spaces. In terms of user measurements, these concepts can be coupled with additional tools, such as phone-camera based eye-tracking, or other behavioural and electrophysiological measures using mobile EEG or electrodermal activity recording. Particularly with the increased accessibility in tools and workflows, each of these concepts can provide a rich and novel research setting.

The presented research was not based on predictive aspects, with minor exceptions of simple regression models built following the correlational analysis, in the analysis section of the VR Gallery (thesis chapter 4). Future directions aligning with predictive sciences, particularly implementing machine learning into the analysis, could be one way forward. This might be as simple as, for example, predicting the outcome of aesthetic judgments to novel artworks based on the previous data; or predicting the level of preference towards an artwork based solely on the eye-movement metrics. Unfortunately, such predictive directions are often investigated beyond the basic research but instead applied mainly in consumer-based research and product development, often with direct implementations in technological applications in mind. It should be noted that these directions could conflict with some core principles of research ethics. For example, the excessive use of personal and sometimes sensitive use in the case of big data might not be well justified, and is a core discussion topic in the emerging field of data ethics.

6.5. Concluding remarks

Conducting three distinct and parallelly progressing studies was an invaluable opportunity. Although each experimental set had specific conclusions, some concluding remarks might be helpful to integrate these aspects.

Acknowledging the subjectivity of experiences towards visual arts, this thesis was driven by a single research ambition in empirical aesthetics. The thesis rejects both ends of the conceptual spectrum, i.e., neither states that the aesthetic experience is not at all measurable, nor comes to the conclusion that the aesthetic experience is a fully measurable phenomenon. From this standpoint of common-sense, this research was not attempting to, for example, fully translate a philosophical question about aesthetics into an experiment. Similarly, this research was not aiming to test aesthetics models/theories at all. Remarks on philosophical and psychological models were provided whenever suitable, particularly to provide more context about the theoretical framework, without seeking bolder research aims to derive hypotheses solely based on those. All the empirical work conducted was, in a way, based on one fundamental question of the measurement: how to extend the ways of partially measuring aesthetic experience and judgment, in novel research settings using some emerging tools. This seemingly vague question provided the primary drive for three studies, each having a partially overlapping reference frame. An overarching achievement in this sense can be described as a demonstration of emerging or overlooked investigation methods on the elusive questions of scientifically describing aesthetic experience and judgment.

The richness of the aesthetic experience and various types of aesthetic encounters has the potential to generate endless questions to be speculated upon, potentially resulting in many exemplar experimental designs. Parts of the experience in various emerging settings can be measured with an increasingly high number of potential tools, and these observations can always be analysed using more and more sophisticated methods, for example, by using stochastic models and chaotic dynamics. Without a particular endgame in sight, experimental aesthetics as a strange amalgamation seems to continue being an idiosyncratic yet stimulating research field.