Keeping it real: the potential for accessible mixed-reality usability testing environments

Bethan Gordon BSc (Hons) MSc PGCE FHEA

Director of Studies	Prof Gareth Loudon Associate Dean Research Cardiff School of Design
Supervisor	Prof Steve Gill Director of Research
Supervisor	Prof Andy Walters Director of Research PDR

Submitted for the Degree of PhD by Research

Awarding body: Cardiff Metropolitan University

Date of Submission: 16/12/2020

This research was undertaken under the auspices of Cardiff Metropolitan University

Abstract

This body of research contributes to design praxeology: the study of process and methodology with the intention to enhance the designer's experience of the human-centred design process by means of validating an optimum usability testing environment for recreating the context of use early in the design process.

Usability testing will frequently make the difference between an excellent product and a poor one. Moreover, in certain fields such as medical device development or training, the defence field or the automotive industry, such testing can literally be the difference between life and death. Unfortunately, design teams rarely have the luxury of either time or budget to usability test every aspect of a design at every stage, and so knowing where and when to devote time to testing, and the fidelity required for accurate results, are all critical to delivering a good result.

This thesis introduces research aimed at defining the optimum fidelity of mixed-reality usability testing environments. It aims to develop knowledge enabling the optimization of usability testing environments by balancing effort vs reward and thus developing critical and accurate data early in the design process. This research also seeks to validate the findings of an optimum environment in a final study and highlight the significance of early usability testing in a simulated context of use.

Testing in a traditional laboratory setting brings advantages such as the ability to limit experimental variability, control confidentiality and measure performance in detail. Its disadvantages over 'in the wild' or field studies approaches tend to be related to ecological validity and the small but vitally important changes in user behaviour in real-life settings. Virtual reality and hybrid physical-virtual testing environments should theoretically give designers the best of both worlds, finding critical design flaws cheaply and early. However, many attempts have focused on high-fidelity, technology-rich approaches that make them simultaneously more expensive, less flexible and less accessible. Additionally, the design literature does not take due account of computer science and psychology when dealing with recreating environments. The final result is that they are less viable and hence somewhat counter-productive.

This thesis presents the results of testing at a variety of fidelity levels within a mixed-reality testing environment, and offers a contribution to new knowledge in determining how usability testing can be maximized to recreate context of use environments early in the design process. The findings recommend the use of recreated environments throughout the design process complemented by field/in-the-wild testing once the product is of high fidelity.

Acknowledgments

This has been one journey I will not forget in a hurry; however, survival on this journey has been predicated on a handful of people. I would like to say a heartfelt thank you to my supervisory team, my Director of Studies Professor Gareth Loudon and Supervisor Professor Steve Gill. Their time and dedication in supporting me has been incredible and without them and their knowledge of how I learn, and their extensive knowledge and expertise, this would not have been possible. I would also like to thank Professor Andrew Walters for his supervisory insight during this work.

I would also like to extend a thank you to Dr Joe Baldwin, who supported me on some of my studies as an extra set of hands and offered support and guidance when needed. I would like to extend a thank you to Window Cleaning Warehouse and to Peter Mundy, who developed the aircraft cleaning product. I would also like to thank Professor Olwen Moseley for giving me the space to write up and being patient with me whilst I tackled this body of work.

Next I would like to thank my parents and family for their continued support, faith, and for putting support mechanisms in place to help me focus on writing and getting me over the last hurdle (doing my ironing and cleaning once a week!). And thank you to all my friends who have been there for me, Leah Warman and Rhian Jenkins and Dr Pete Dorrington for encouraging me over the finishing line.

This is for my husband, Sean, and children, Noni and Lona; sorry girls, this was not meant to take this long to complete, but your patience and support has been amazing. Thanks, Alfred, for being my PhD Pug. Sean, you deserve a medal, but it is done and yes you can have your Bethan back.

Table of Contents

Contents

Glossary

Participant: An individual involved in the research process who represents the intended target market.

Simulated environment: An environment that includes mixed media and a mixture of either a physical or virtual or physical-virtual representation of an intended environment.
Simulations: How real-world context and scenarios can be recreated and imitated.

Simulator: A purpose-built device that provides realistic encounters of real-world scenarios or environments.

Fidelity: Concerns the level of likeness in a reproduced object.

Presence: Defined here as the degree to which a participant feels they are present within a simulation, without necessarily believing it to be real; *"a psychological, perceptual and cognitive consequence of immersion and involvement"* (Witmer & Singer, 1998).

Immersion: A psychological response to the content found in a simulated environment.

Face validity: *"the extent to which experts agree that the measures capture the intended construct"* (Reimer, D'Ambrosio, Coughlin, Kafrissen & Biederman, 2006).

Ecological validity: A means of determining to what extent a virtual environment induces behaviour found in the real world (Deniaud *et al.*, 2015), i.e. Did the study measure what it was designed to measure?

Relative validity: A means of determining to what extent a virtual environment induces a relative experience to that of an identified comparison.

Absolute validity: A means of determining to what extent a virtual environment induces an absolute experience to that of an identified comparison.

Augmented reality: A real-world experience of an environment with digitally augmented attributes.

Tangible user interfaces: The use and engagement in a digital interface of a system or product in a physical environment.

Virtual reality: The generation of digital content to recreate an environment for interaction purposes.

Prototypes: The manifestation of ideas to be tested and challenged, or to communicate intent; this can be in a physical or virtual format.

PANAS: Positive and Negative Affect Schedule, to determine emotional state.

SUS: System Usability Scale to determine the overall usability of a system or product.

ANOVA: Analysis of Variance is a statistical method used to establish the difference in means between two or more groups.

T-test: A statistical method deployed to analyse the means of two groups of people. An independent T-test is used for two different conditions and people and is used with a sample less than 30.

Independent study / between subjects: Population sample size is different in each condition.

Repeated measures / within subjects: Population sample size is exposed to each condition.

List of Figures

List of Tables

Chapter 1 Introduction

Chapter 1 Introduction and Overview

1.1 Introduction

Norman (1998) relates the frustration faced by users of under-developed products in everyday life and identifies them as 'devices that lead to error' and 'products that are misunderstood'. A good Product Development Process (PDP) requires products to be tested with users throughout the design process (Hare *et al*., 2014; Rubin 2008), but often the environment used to test the product is a laboratory setting, normally a plain enclosed space so tests can be regulated. Brehmer & Dorner (1993) recognized that field research is often too complex to arrive at conclusions and the laboratory environment does not offer enough complexity for more defined conclusions. Woolley *et al*. (2013) challenged the pitfalls of laboratory usability testing and developed studies exploring usability testing of products in their intended natural (or real) context of use. An example of testing a product in its natural context would be usability testing a football boot on a football pitch rather than in a laboratory environment. Woolley *et al*. (2013) found that for computer-embedded products, context of use changes the results of usability tests, particularly those relating to subtle but important design details. Their research concluded that in-context usability testing was a potentially valuable tool at the early stage of the design process. Doing so, however, is predicated on finding ways to produce and test prototypes very quickly and effectively in the correct environment. Unfortunately, prototypes that are robust enough for usability testing in context-of-use situations outside of a laboratory can be costly and take a long time to make. Likewise, Woolley *et al*. recognized the benefits of in-context testing and laboratory environment testing as both have benefits and drawbacks. Gill *et al*. (2008) demonstrated that the reality of employing sophisticated high-fidelity prototypes for usability testing means that companies either ignore the resulting design flaws or cancel the product, as the costs of making changes to the product late in the development process are too high. Gould & Lewis (1985) and Abra *et al*. (2005) highlighted that users need to be involved earlier in the development process of a product, so that behaviours and attitudes can be documented, analysed and fed back into the development of the product. Hare *et al*. (2014) also identified the need to test physical products as realistically as possible early in the PDP, supporting the need for increasing realism during product usability testing.

Therefore, a key challenge is to be able to test product prototypes in the early stages of the design process while trying to consider the importance of context of use during testing. This thesis takes on this challenge and explores the potential of using accessible mixed-reality usability testing environments during the ideation phase of the HCD process, to provide the benefits of laboratory settings (i.e. the ability to limit experimental variability, control confidentiality and measure performance in detail) while trying to simulate real contexts of use. Theoretically, mixed-reality usability testing environments should give designers the best of both worlds, finding critical design flaws cheaply and early. However, the challenge is also how to design such simulated environments. This leads to the main research questions. For the scope of this research, computer embedded products are used, and one is linked to a mobile application. The product is defined as having physical attributes that encourages interaction and will be recreated for the purposes of the usability studies in a prototype that has active physicality (Hare 2015). This work can also extend to any product requiring usability testing; however, the research to date does not include complex systems of services.

1.2 Research Questions

Are simulated environments useful in the usability testing of product prototypes early in the design process? What are the key design requirements for said simulated environments?

1.3 Aim and Objectives

The aim of this research is to explore the potential of using simulated environments early in the design process to enhance the effectiveness of prototype-based usability testing, and to uncover key requirements in the design of such simulated environments.

The objectives required to answer the research questions are:

1. *To understand the landscape and the literature concerning usability testing approaches.*
2. *To explore methods of recreating environments and how they can influence the requirements of the usability testing space.*
3. *To ascertain the fidelity of a simulated environment that is most effective in discovering product design flaws early in the product development process.*
4. *To validate the findings and detail the outcome of this research exploration.*

1.4 Why these two research questions?

Context allows individuals to ground themselves in what they are doing. If context is removed then scenarios become arbitrary, abstract and alien. Context builds a picture and familiarity. Kneebone (2013) describes himself as a capable surgeon able to stitch wounds together. However, when he was removed from his medical theatre and placed in a stitch room where he attempted to stitch a jacket—a far less life-threatening scenario—he displayed a lack of confidence, dithering and consequently making mistakes that he was unable to undo because he was in a different frame of mind and a different context. Jumisko-Pyykkö and Vainio (2010) reviewed extensive literature to frame the meaning of the term 'context of use' in the Human computer Interaction (HCI) community, and highlighted key definitions that are pertinent to this body of research noting that they use the term context of use *"to clarify that the perspective taken is that of usage and user, rather than the application or system context".* Jumisko-Pyykkö and Vainio propose that context of use captures the holistic experience (Roto 2006) and takes account of where an activity takes place and is a construct that considers time, use, social interaction and local influences (Greenberg's 2001). Therefore, during this thesis, context of use, will relate to the environment a product is used in, considering the holistic nature of an activity.

The research questions were arrived at as a result of a body of work that contributed to the author's Master's project. Although this work identified an opportunity to further explore context of use in early product development, it lacked the depth and understanding of what was currently being used to fill this void. Some observations made in that body of research were about the use of simulated environment in cognitive therapies and its positive impact.

The key components of the Master's work was the setup and development of an environment to test the principle of a low-cost, rapidly configurable physical virtual environment. At that stage, it used a wooden frame (figure 3), bed sheets and back projectors to form a hybrid physical-virtual testing space that could be compared to the real environment. While the setup was primitive (see figure 1), its results highlighted the potential of such an approach.

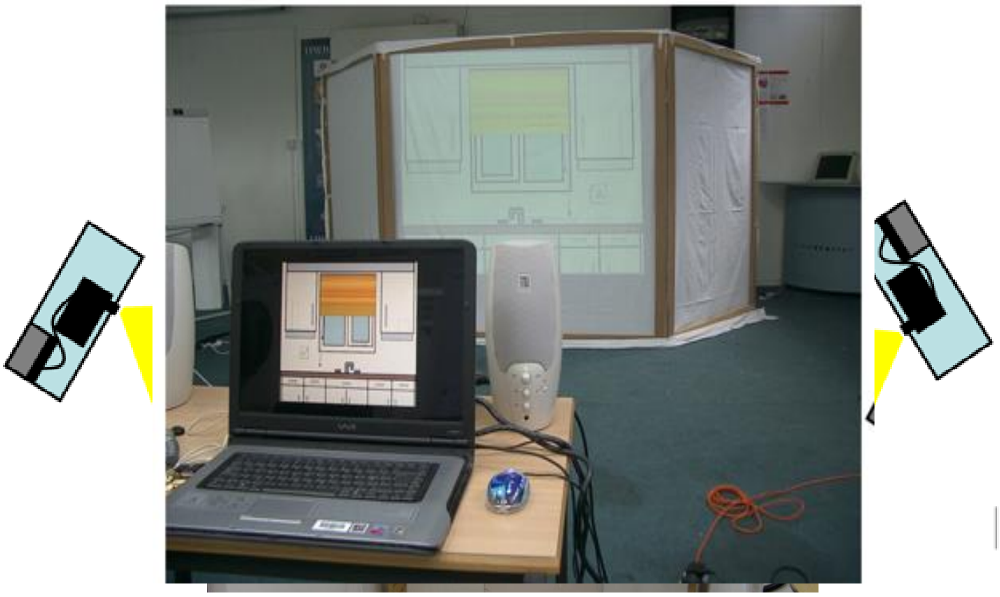


Figure 2 Initial plan view of the usability testing setup

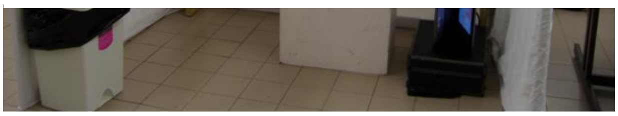


Figure 3 Wooden frame set up (Gordon 2007)

Two key findings from the Master's initial research were that (1) a replicated environment yielded better results vs a plain laboratory setting, but only when the fidelity of the model matched the fidelity of the environment; and (2) the usability testing of a real, high-fidelity product in a low-fidelity environment created a sense of jarring for participants (Gordon, 2007). A key limitation of the initial research was the lack of quantifiable research on achieving participant 'presence' in a re-created environment. Presence is defined as a user's subjective sensation of 'being there' in a simulated environment (Kneebone *et al.*, 2010; Steuer, 1992; Slater, 2013; Deniaud *et al.*, 2015) with Witmer & Singer (1998) defining presence as *"a psychological, perceptual and cognitive consequence of immersion and involvement".* There was also a lack of appreciation for ecological validity required to induce presence and its role when recreating real environments and testing their impact on product development. When conducting studies, ecological validity is a means of determining to what extent a virtual environment induces behaviour found in the real world (Deniaud *et al.*, 2015). Another definition of ecological validity is:

> *"The degree where results obtained from research are representative to conditions in the wider world. Ecological validity deals with the results from experiments being valid in the real world".*
>
> *Psychology Dictionary (2013).*

This body of research is not generative research, whereby the study of people or specific product development problem is explored, rather this work is situated as usability evaluation, with the intention of developing insights that can be applied in future usability studies.

The laboratory that facilitated the studies found in this thesis (The Perceptual Experience Laboratory described below) was based on the original prototype developed by the author.

1.5 The Perceptual Experience Laboratory (PEL)

The author was a member of the Programme for Advanced Interactive Prototype Research (PAIPR), founded by Professor Steve Gill and Professor Gareth Loudon in 2002. The group later evolved into the User Centred Design Research (UCD-R) group in collaboration with the National Centre for Product Design & Development Research (PDR). The research group's initial aim was to investigate methods for the rapid development of computer-embedded products and later diversified into broader user-centric methods that applied these approaches to all products. The group later pooled resources to develop a multi-disciplined laboratory, known as the Perceptual Experience Laboratory (PEL), with FovoLab, a multidisciplinary research team developing theories around experiential perspective displays. PEL is a mixed-reality simulated environment incorporating a range of recording equipment. Now in its 3rd generation, having been developed from the author's original Masters' work, PEL was built for two purposes: to allow FovoLab to further develop insights around how we see and for UCD-R to fill the void between a traditional laboratory and 'in the wild' usability testing research. It was purpose-built to afford the restricted and consistent testing conditions of a standard usability laboratory within simulated context-of-use environments.

With each iteration, PEL's sophistication and cost increased. In its third iteration, PEL uses perspective algorithms, developed by FovoLab, that allow PEL to display imagery in a way that is literally unique. Before developing PEL in its current form, a great deal of research into analogous environments and their capability was required. This included scoping activities to ascertain the needs of PEL in the guise of a workshop conducted with anticipated stakeholders, a visit to the Welsh National Opera sets at the Wales Millennium Centre and a two-day 'Train the Trainer' (medical insights) session in the Cochrane Simulation Laboratory at the University Hospital Wales. The information gleaned from these activities regarding recreating simulated environments is discussed in more detail in Chapter 3. With this additional insight, the 2nd generation of PEL (PEL2) was built using a cardboard frame with soundproofing foam to create an enclosed laboratory (see figure 5). PEL2 benefitted from the use of observational cameras and Observer XT software to analyse user behaviour along with the capability of using heartbeat variability data to gain detailed feedback on participants' psychophysiological state. Observer XT is a standard software package found in usability laboratories and is utilized in the analysis of data obtained from behavioural coding. Heartbeat variability allows biometric data collection, offering a study the addition of objective qualitative data on the impact of an induced environment on participants.

PEL2 also included a section for storage and a separate room for a moderator to observe the user in real time. Although the screens were large, they relied on front projection. This created a large space with good light levels. Unfortunately, outside of a small ‘sweet spot’, shadows were cast when a user was in the space.

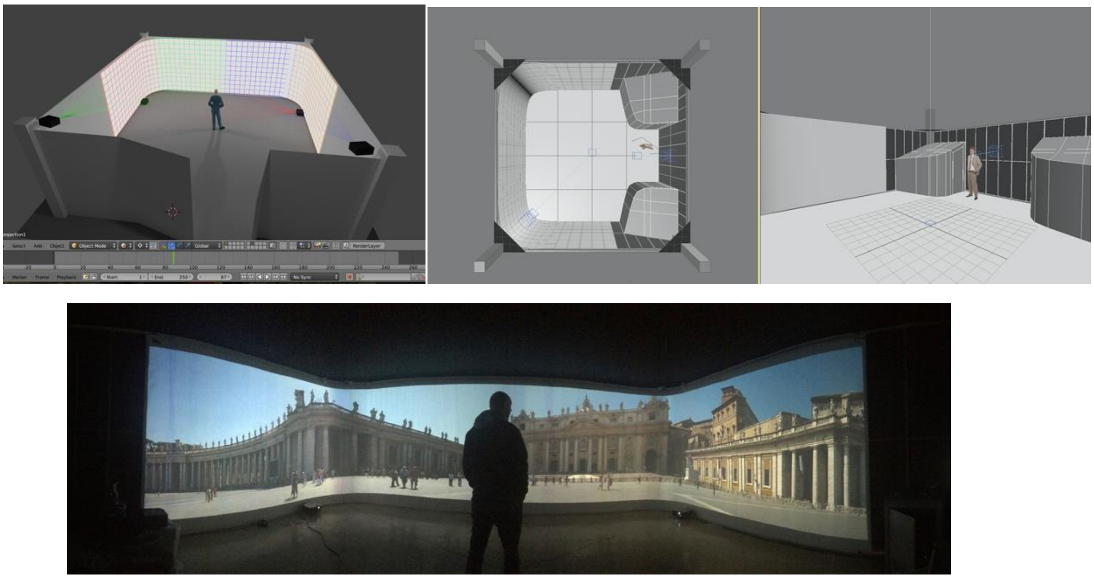


Figure 4 PEL 2 with limited sweet spot.

To mitigate against the limited usable area in the configuration (as shown in Figure 4), a new curved screen was developed with back projectors for an uninterrupted and shadow-free environment, as shown in figure 5 below.

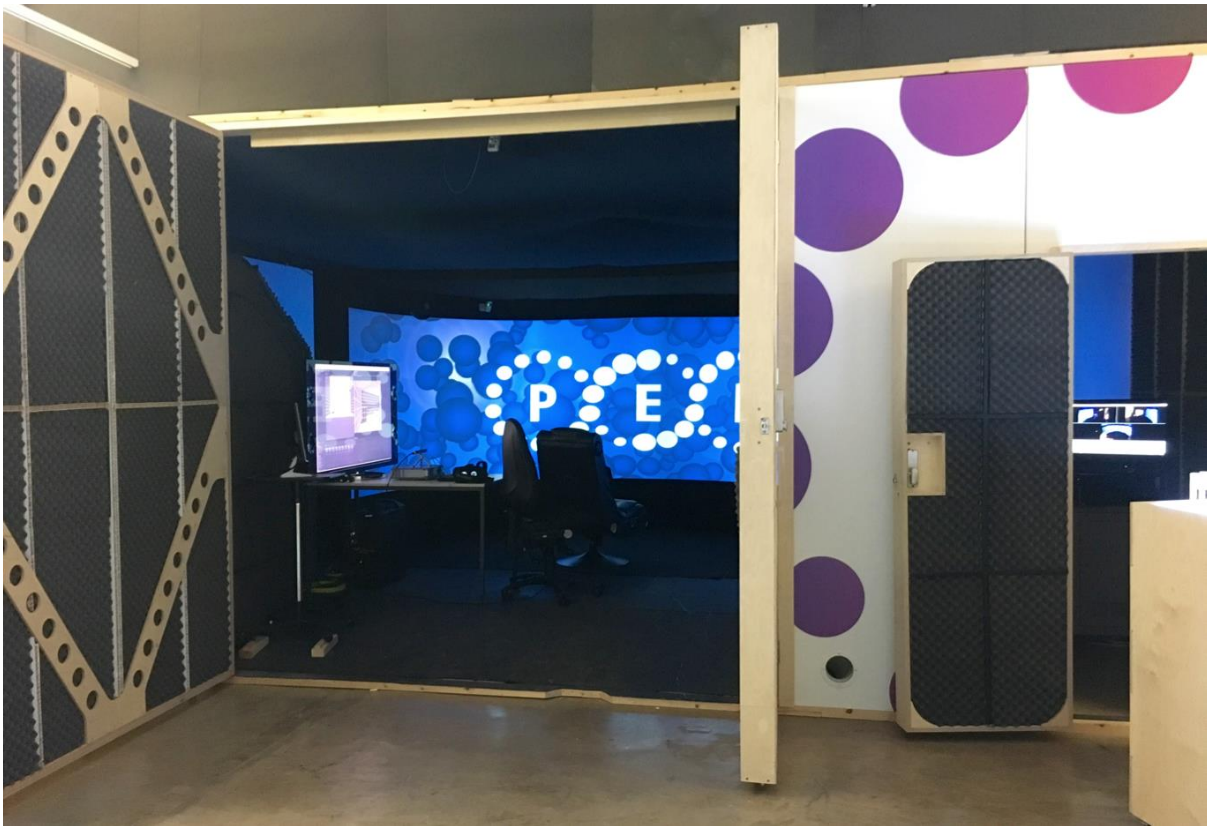


Figure 5 PEL 2 curved screen & back projection

The 3rd generation of PEL (PEL3) was built in a designated room. It uses rear projection to eliminate shadows and has an ambisonic sound capability facilitated by a 20.4 Dolby surround system in a 360° array and is capable of mimicking the sound characteristics of any space being simulated. The new configuration allows for a bigger, frameless curved screen (see figure 6 below) that reaches to the floor and displays a 5280px by 1980px (4K +) image on a 200° curved screen to accommodate participants' field of view (FoV) of 180- 200°. As in PEL2, PEL3 used observational cameras and Observer XT software for capturing research data. However, in PEL3, eye tracking equipment was also incorporated to enhance the detailed data gathering required when conducting usability testing, so areas of interest could be identified and tracked utilizing analysis software that allows for visual coding, again facilitating the research capabilities of the laboratory.

In total, PEL3 enables a mixed-reality method that couples physical objects with a 4K, 200° panoramic visual surround screen, artificially added smell and 3D ambisonics to engender an appropriate sense of immersion (see figures 6, 7 and 8 below).

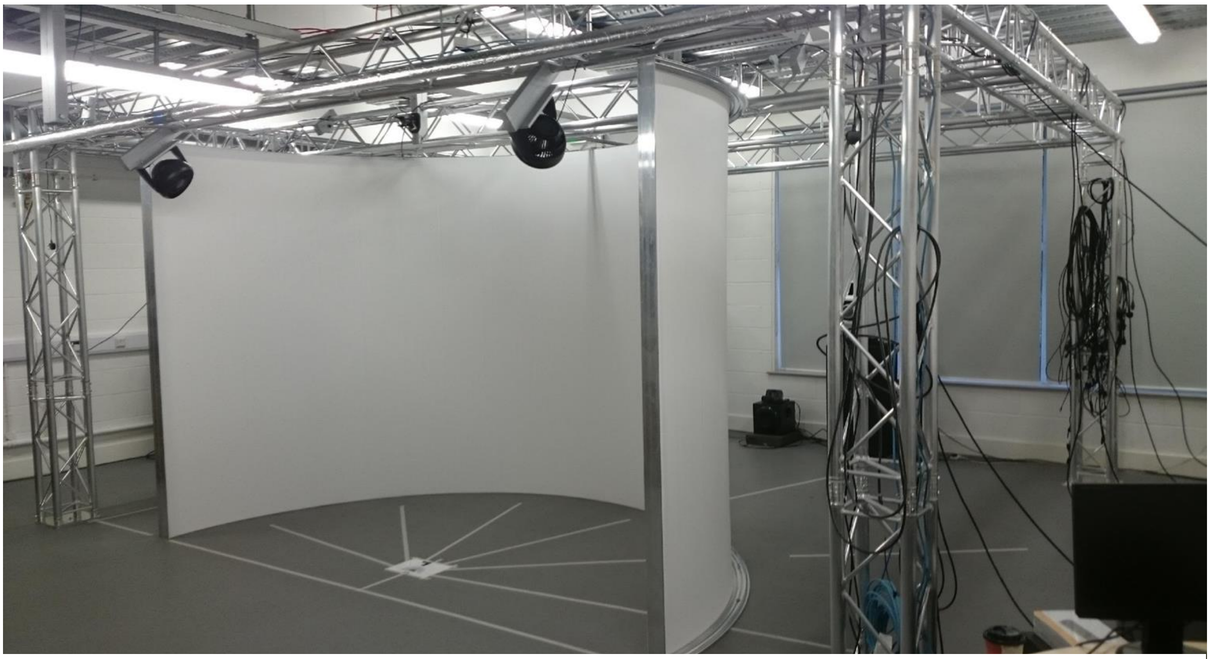


Figure 6 PEL 3 construction in designated room (Baldwin 2019)

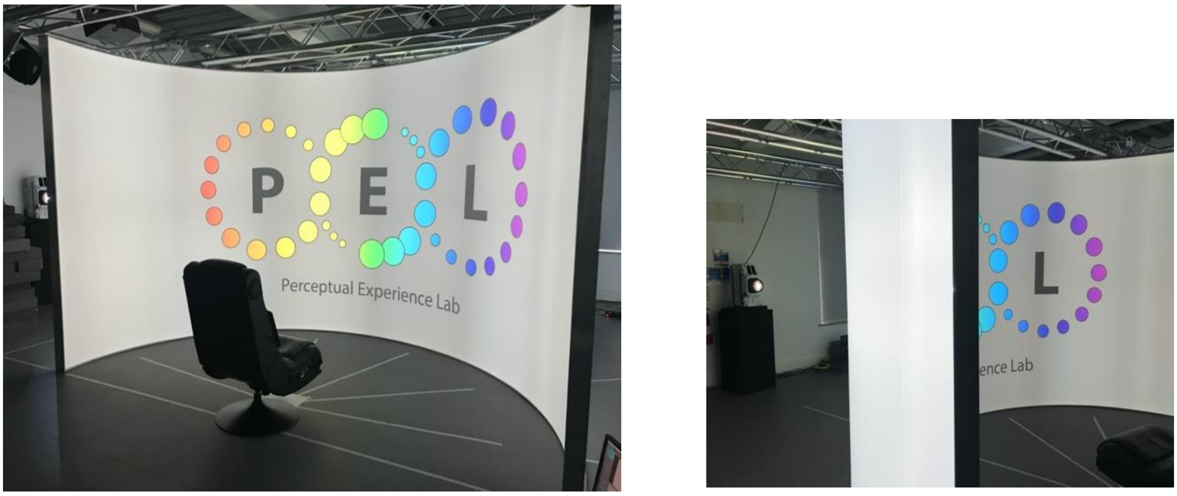


Figure 7 PEL 3, front view and behind the scenes (Baldwin 2019)

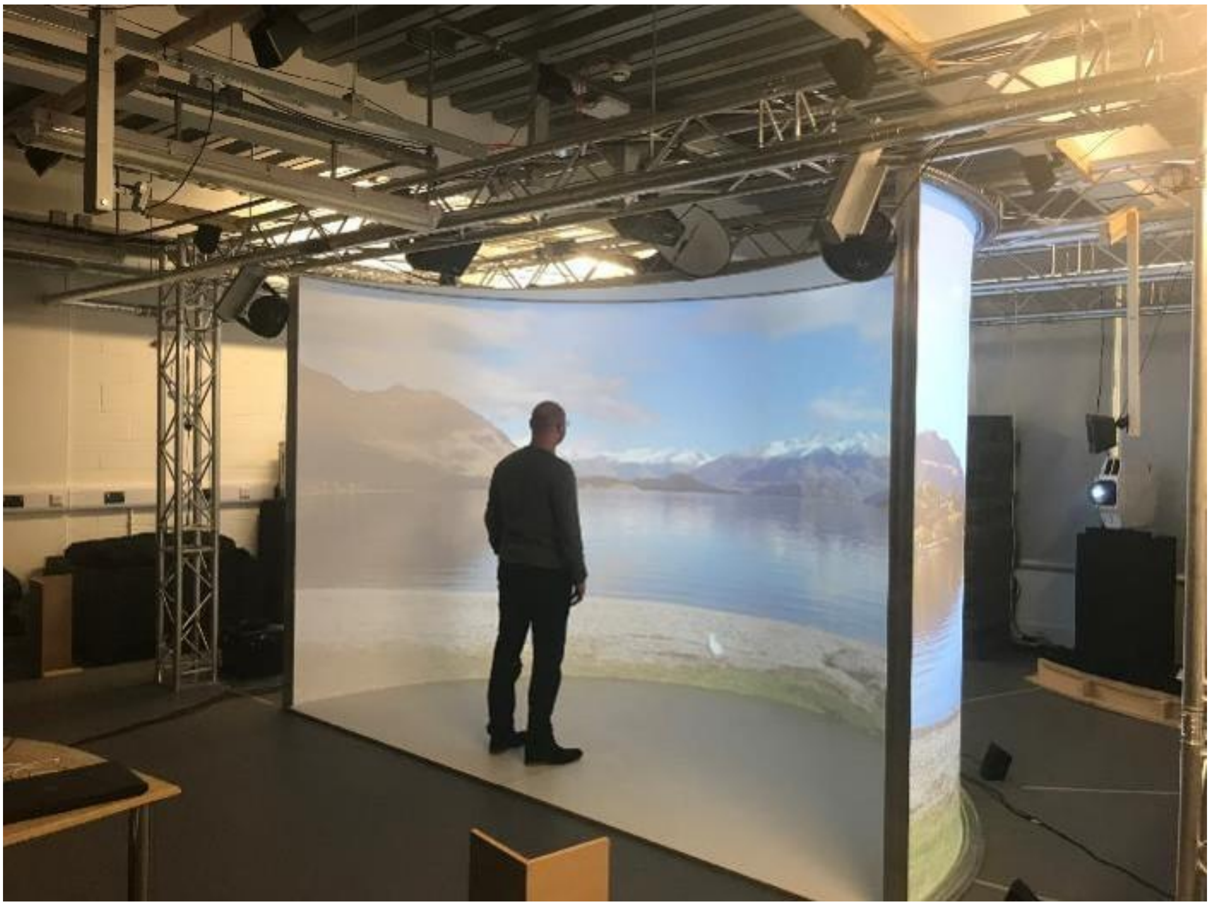


Figure 8 Perceptual Experience Lab. PEL3 (Baldwin 2019)

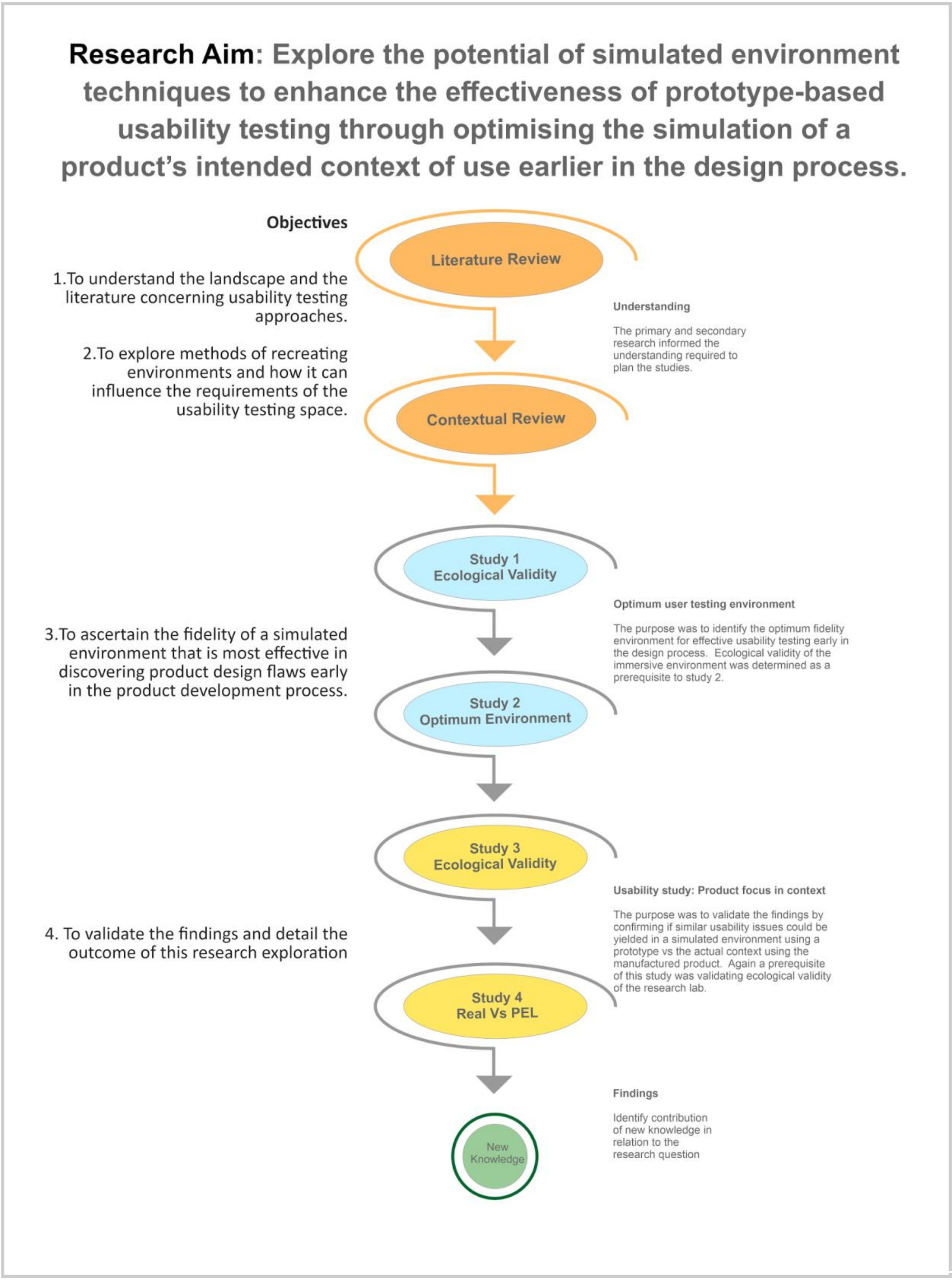
On the experimental monitoring side, PEL3 has three video cameras to capture activity from various angles, omni-directional mics, heartbeat variability, galvanic skin monitors and state-of-the-art wireless eye tracking. All of these are linked to software systems that allow all data to be captured on a single timeline for detailed analysis.

1.6 Personal Motivation

Having thought about this opportunity for many years, it has been a personal motivation to conduct meaningful research that can inform the design discipline. This need, coupled with a love of learning and passion for personal development, has led to establishing new findings and working with statistics. It has been an enriching process, with the thinking and knowledge learnt being applied along the way in a range of ways, from Knowledge Transfer Partnerships, leading to a 2017 Insider Business and Education Partnership Award, to the development of the new Welsh national curriculum for ages 4-18. It has also contributed to my day-to-day work in Higher Education where developing new processes is almost a daily occurrence. The journey has been tough but invaluable.

1.7 Project Structure

The diagram below visualizes the contents of this thesis and process (figure 9).



Research Aim: Explore the potential of simulated environment techniques to enhance the effectiveness of prototype-based usability testing through optimising the simulation of a product's intended context of use earlier in the design process.

Objectives

1.To understand the landscape and the literature concerning usability testing approaches.

2.To explore methods of recreating environments and how it can influence the requirements of the usability testing space.

3.To ascertain the fidelity of a simulated environment that is most effective in discovering product design flaws early in the product development process.

4. To validate the findings and detail the outcome of this research exploration

Understanding

The primary and secondary research informed the understanding required to plan the studies.

Optimum user testing environment

The purpose was to identify the optimum fidelity environment for effective usability testing early in the design process. Ecological validity of the immersive environment was determined as a prerequisite to study 2.

Usability study: Product focus in context

The purpose was to validate the findings by confirming if similar usability issues could be yielded in a simulated environment using a prototype vs the actual context using the manufactured product. Again a prerequisite of this study was validating ecological validity of the research lab.

Findings

Identify contribution of new knowledge in relation to the research question

Figure 9 Research Diagram

Chapter 2 describes the Literature Review to understand the landscape and the literature concerning usability testing approaches. Chapter 2 also explores ways of recreating environments and how they can influence the requirements for designing a simulated environment for usability testing. Chapter 3 further explores these issues through primary research. Chapter 4 describes the methodology and methods adopted for the empirical studies undertaken, with Chapter 5 providing details of four studies undertaken that help in the design and validation of the simulated environment for usability testing. Chapter 6 discusses the key findings and implications discovered from the research, with Chapter 7 providing a summary of the research, the new contributions to knowledge and areas for future research.

Chapter 2 Literature Review

Chapter 2 Literature Review

2.1 Introduction

The first two objectives focus on understanding the landscape of literature concerning usability testing and exploring ways of recreating environments to influence the requirements of a usability testing space. These two objectives are explored in this chapter, addressing key areas including HCD; physicality in design; usability testing; usability testing environments and analogous environments. The consequence of this review is to inform the methodology and user studies described in the later chapters.

Boothe, Strawderman & Hosea (2013) and Spicer *et al*. (2015) highlight the barriers to developing and implementing systems and products, while acknowledging the associated cost when products are underdeveloped and consequently fail. One of the main reasons cited for failure is a lack of usability testing prior to the systems and products being launched into the consumer market. In some fields, such approaches are mandatory. For example, the usability testing of prototypes is essential for computer-embedded medical products as enshrined in BS EN ISO 9241-210:2010. Legal issues notwithstanding, a good Product Development Process (PDP) requires usability testing throughout the design process (Hare, Gill, Loudon & Lewis, 2014; Rubin & Chisnell, 2008) including the testing of low-fidelity models at the earliest stages. The usability testing of low-fidelity models usually takes place in a controlled laboratory setting (Kaikkonen, Kekalainen, Canker, Kallo & Kankainen, 2005) both for ease of access and so that tests can be regulated. Unfortunately, such settings are unlikely to match the cultural, social and physical environment where the resultant product will be used, which could limit the validity of any resultant usability data.

Literature on testing environments goes back a long way. Dahl, Andreas & Svanaes (2010), for example, found that the social and physical attributes of an environment are often ignored or given low priority when testing a prototype, and highlight the need to consider their implications in design testing. Even further back, Brehmer & Dorner (1993) found that while field research is often too complex to be practical, laboratory testing does not offer enough complexity for the often-critical fine grain conclusions. Later literature shows that testing prototypes 'in the wild' can be achieved early in the design process, e.g. Woolley, Loudon, Gill & Hare (2013). That research found that context of use had a marked effect on the results of usability tests, particularly when exploring the effects of subtle but important design details. The study concluded that in-context and laboratory testing each have benefits and drawbacks. For example, prototypes of computer-embedded products—the subject of that particular study–that are robust enough for testing in the wild, are often costly in time and money. More recently, Kjeldskov & Skov's (2014) review, 'Was it Worth the Hassle?' concluded that while a definitive answer has not been reached in the laboratory vs field debate, it is not whether one or the other is better, but 'when and how', that is significant. Their research concluded that field studies offered little value apart from the 'ecological validity'—the degree to which a recreated environment emulates the real environment. Other literature explores the potential role of virtual environments in this regard. For example, Deniaud, Honnet, Jeanne and Mestre (2015) recognized the need to evaluate virtual environments, describing two types of validity: absolute and relative. Absolute validity concerns achieving the exact same data from real and virtual testing environments, while relative validity describes the achievement of different results that are "*in the same direction and have a similar magnitude".*

What is the optimum fidelity for a usability-testing environment in order to meaningfully inform design decisions early in the design process, before a design team has committed to a particular design path? The answer is not yet known and is one of the purposes of this research, but a number of studies have found that surprisingly strong results can be achieved with low-fidelity environments. IDEO coined the phrase 'Experience Prototyping' around 20 years ago (Buchenau & Suri, 2000) and their work is also captured in Benz' (2014) collection of case studies as a contribution of experience design and its role in constructing purposeful experiences. Buchenau and Suri's work included the mocking up of real-world environments at the concept generation phase and has shown that creative thinking can enable the prototyping of relatively complex real-world scenarios without a lot of either time or technology. In a completely different field, work conducted by Kassab *et al.* (2011) established that a low-fidelity mixed-reality environment could be effective in training medical students in surgery techniques. In this case, the authors made recommendations on how to decide what should be physical and what should be virtual based on where the users' attention needed to be, with physical objects being used in the areas of greatest focus. The authors report that, taken together, the physical-virtual environment created an overall sense of immersion without intruding on the task at hand. Dahl *et al.*'s (2009) work contributes additional underpinning theory here, describing environment fidelity requirements as being predicated on the behaviour needs of the participant in that environment: Where is the attention? What influences decision-making? The theory is further supported by Lessiter, Freeman, Keogh and Davidoff's (2001) ethnographic studies of a hospital training environment that focused on how real visual cues are used to heighten participants' 'presence' and confidence in the simulated environment, with 'presence' defined as *"a user's subjective sensation of 'being there'"* (Deniaud *et al.,* 2015; Slater, 2013; Steuer, 1992).

The following sections explore the literature in more detail, starting with an overview and funnelling into more specific areas of interest.

The Design Process is a well-established method of identifying needs and opportunities and producing a product or service as a consequence. However, individual organizations/ designers and *"teams should adapt and modify the approaches to meet their own needs and to reflect the unique character of their in-situation environment"* (Ulrich & Eppinger, 2003 & 2019). This was further evidenced when the author led on a Knowledge Transfer Partnership (KTP), working with the KTP Associate to develop a bespoke design process for cycle storage manufacturer Odoni Elwell. Regardless of the type of process or the iteration used, the process is a fundamental design philosophy (Norman, 2013). The emergence of user-centred design has allowed for improved empathy during the design process and, as a consequence, one more informed by user insights and user needs (Van der Bijl-Brouwer & Dorst, 2017).

Design process exists in many forms, but there are common key steps in each approach documented by leading authors in the field. Key phases of the design process, as recognized by Ulrich & Eppinger (2016), are:

1. Planning.
2. Concept development.
3. System level design.
4. Detail design.
5. Testing and refinement.
6. Production ramp-up.

The characteristics in the above example are common traits found in the Product Development Process (PDP). Their primary function, when adapted to an individual company's needs, is to mitigate risk when investing in the development of a new product, process or service (Unger & Eppinger, 2009). The steps above are intended to be iterative and, although it is not explicit when reading the PDP steps, the user should be considered. Unger and Eppinger note that testing should be conducted during the system level design stage; nevertheless, it was still deemed that prototyping earlier was not feasible or practical (Unger & Eppinger, 2009). However, Walters & Evans (2011) highlight that when the predominant driver is 'market needs' this can sometimes result in the lack of human involvement. It is interesting to note that emotions, empathy and behaviour do not feature clearly in the classic product development process of old, and Utterback & Vedin (2006) support this sentiment by recognizing the need for an emotional connection between person and product. The fundamental difference between the PDP and a Human-centred Design (HCD) approach is identified by Gordon *et al.* (2017) (not the author) as having two key differences: the first being the distance between the design team and the intended users, and the second recognizing that design teams often have little to no experience in the system, services or products they are designing. Therefore if the user/persons who interact with the new product, service or system are not considered from the beginning (and also in terms of their behaviour, emotions and thinking), then the process is not taking due account of them as humans at the centre of the process. The consequence is more likely to be an ill fit or poorly designed product, system or service, as noted by Norman (2013) in his book *'The Design of Everyday Things'.* A known consequence of not engaging people and validating ideas early in the design process is a subsequent commitment to high-fidelity prototypes. The result is often a reduction in the likelihood of changes being made when further end-user observations are sought because of the level of commitment to an already detailed design solution (Gill *et al*., 2008). *"Design teams in companies like Virgin Atlantic Airways and BSkyB* *conduct user research at a stage where a prototype is well developed, rather than involving users at the concept development stage."* British Design Council (2007).

2.2 Human-centred Design (HCD)

BS EN ISO 9241-210-2010 outlines the principles of HCD as *"complementary to existing design methodologies and provides a human-centred perspective that can be integrated into different design development processes in a way that is appropriate to the particular context".* The standard goes on to list six principles:

a) The design is based upon an explicit understanding of users, task and environments.

b) Users are involved throughout the design and development process.

c) The design is driven and refined by user-centred evaluation.

d) The process is iterative.

e) The design addresses the whole user experience.

f) The design team includes multidisciplinary skills and perspectives.

ISO 9241-210:2010

This level of commitment to HCD, driven by policy, should and, to some extent, does ensure compliance; but the level of compliance that allows for cost-effective implementation in a design organization is still an area that requires work to make HCD accessible and affordable, as noted in Walters & Evans' (2011) study on developing a user-centric accessible framework.

One tool that is available to designers as part of the HCD process is usability testing. It is a tool that enables designers to challenge assumptions of a design proposal and embed the human in the design process by engaging participants in usability tests. The aim is to continually enhance a design proposal by teasing out the measures of usability, as defined by Barnum (2011):

- effectiveness
- efficiency
- satisfaction

Although he is not a designer, Donald Norman is one of the leading authors and advocates for the term User-centred Design (UCD) and one of the first to use the term 'user experience': "*In the 1990s, the group I headed at Apple called itself 'the User Experience Architect's Office'."* Norman (2013). UCD (also known as Human-centred Design—HCD) is recognized as an approach that is empirical in nature. It is an iterative approach that requires participation and has to be realistic in what can be achieved when considering the bigger picture. At its core, UCD is a pragmatic systems-orientated approach designed to take due account of human diversity (Pheasant, 2005).

"Innovation begins and ends with people. It calls for keen and caring observation. The discipline of Human-centered Design involves careful investigation. It requires curiosity, objectivity, and empathy. You need to engage all of your senses (looking, listening, and so forth) in pursuit of meaningful findings."

Luma Institute (2012, p1).

In a bid to enhance design research and inform the development of new products, Norman's work challenged the notion of intuitive use over aesthetics, with a conscious eye on the human at the centre of the process whilst acknowledging the role of appropriate aesthetics that was informative and contributed to intuitive use of a product. Psychology plays a significant role in Norman's work, including how the human senses inform cognition and perception. He often refers to door handle design as an example of when the inherent language of the design might tell one to do one thing—like pull—but the door has been designed to be pushed. There is a design grammar that must be applied when designing products and Norman unpicks this notion: "*the most important characteristics of good design is discoverability and understanding*" (Norman, 2013). Design grammar is also known as affordances in design, a term dating back to James Gibson's work in 1979, where he recognized visual cues as affordances, and later Norman expands on the term to note 'perceived affordances' (Weinschenk, 2011). The significance of the term affordance is premised around usability, and how a product is designed to invite a user to interact with it; for example, a chair with a seat and back tells the user to sit on the flat section—it is a perceived affordance. Norman goes on to describe UCD as putting human needs first and then observing actual behaviour, since humans do not necessarily know how or why they interact with the designed world—what we say we do and what we actually do can often be different. This too is noted by Margaret Mead, a leading anthropologist (Luma, 2012), therefore gathering design insights from close observations is critical, especially context. Gill (2009) goes one step further to explore the impact on users' behaviour in the social context. He notes that products are designed in studio environments but recognizes there is limited 'good' testing conducted in context that captures social influences on a product in development. Social context enables the notion of discoverability that is also echoed in Norman's work, as well as the Luma Institute, and supported by Gordon *et al.* (2014) who highlight that *"people are creative and resourceful in their own contexts".* The challenges that are offered by the introduction of the social context into the UCD process relates directly back to the first HCD principle of developing an understanding of task and environment. To differentiate from participatory design, whereby the people are seen as the expert, HCD retains the designer as the expert (Walters & Evans, 2011). The notion of people playing a role in informing product use and subsequent development is noted in Gill's work (2009) where he cites research conducted by Loudon *et al*. (2002) whereby the original intention of a camera on a phone was for video calling, but when observing the users in their intended context, it was noted that what users actually wanted to capture were images of the outer world, just as they would behave when using a normal camera. This is discoverability of the users' intended social context.

You will have noted that the above section interchanges between the terms UCD and HCD. Norman's 2013 edition of '*The Design of Everyday Things'* specifically identifies the change in industry terminology, and he too changed his terminology from 'user' to 'human' (Norman, 2013). User-centred design is a widely used term and shares the same values as the human-centred design process, but there are potentially negative connotations to the term 'user', as opposed to the all-encompassing term 'human'. For example, during the development of this research, the author worked within a design and manufacture organization. One of the projects involved drug users, where a connotation of the word 'user' proved difficult and prone to misinterpretation by a designer immersed in a drug-using / needle safe project. Therefore, from this point forward in this thesis, Human-centred Design (HCD) is used. HCD is also the recognized term in BS EN ISO 9241-210:2010, the BSI standard publication for Ergonomics of human-systems interaction Part 2 10: Human-centred design for interactive systems, the standard that replaced ISO 13407:1999.

2.2.1 Key phases of the HCD process

HCD includes the fundamentals of a classic development process and overlays on it key characteristics, as identified in ISO 9241-210:2010. IDEO, an international design consultancy with a base in Silicon Valley, is one of the leading organizations in the successful implementation of the HCD process. IDEO have clearly illustrated a process that recognizes HCD as its own redefined process. Their design philosophy includes three key stages: Inspiration, Ideation and Implementation (figure 10).

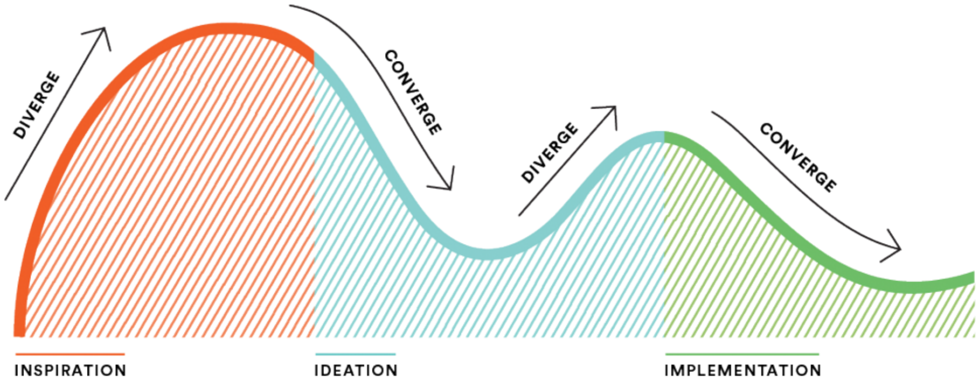


Figure 10 IDEO toolkit (2015)

IDEO's design philosophy is supported by a detailed tool kit that is open source and available to ensure the human is at the centre of the work. Although this echoes Cross's early works in the 1980s and 1990s, Cross's inflection was on designers as people rather than people as participants of the process. The fundamental differences between a classic design process,

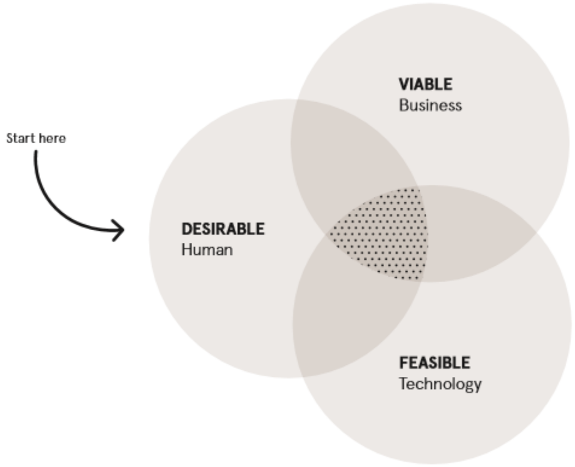


Figure 11 IDEO HCD Process (2015)

even ones published as recently as 2015, and the IDEO process is the consideration of the human or user throughout the process, rather than at the later refinement and modelling phase or recognizing people as a collective marketing group (Walters & Evans, 2011). HCD has empathy at its core; requires a multidisciplinary team (Smaradottir *et al.*, 2015; Walters & Evans, 2011); and advocates failing early and often in its approach to testing and validating ideas, to help deliver feasible solutions (IDEO, 2015). To help ensure success and impact, IDEO developed an approach that considered three areas throughout the HCD process (figure 11), and enabled a design team to view their design through the lens of desirability, viability and feasibility (IDEO, 2015). To validate these three lenses, usability testing should be implemented early and often, as supported by the Nielsen Norman Group (2012).

IDEO's HCD process is in line with the identified five phases in the International Encyclopaedia of Ergonomics and Human Factors (2006). Gordon *et al.* (2014) also note the five stages as *"project commitment; user and technology research; innovation sprint; concept creation and validation; and project assessment"*. Much of IDEO's approach has integrated these five stages, as identified in figures 12 and 13. This divergent and convergent approach has also been adopted in the UK Design Council's 2005 Double Diamond process and later reviewed in the Framework for Innovation (2019) process. It highlights at what point a designer should be diverging and converging their thinking.

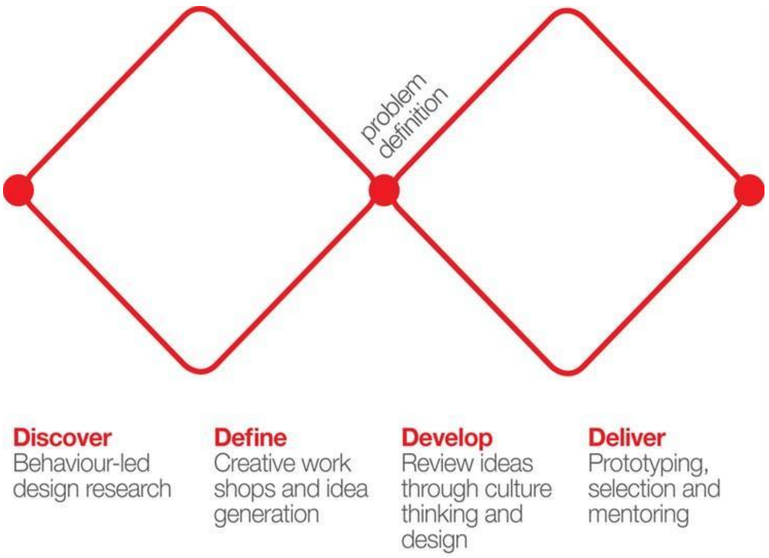


Figure 12 British Council Double Diamond (2007)

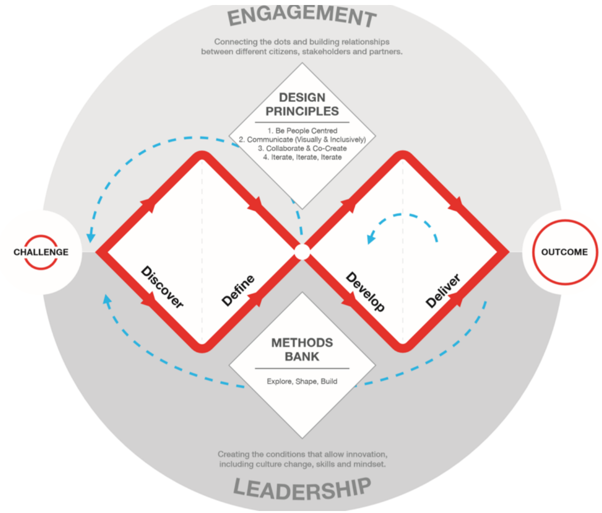

ENGAGEMENT
Connecting the dots and building relationships
between different citizens, stakeholders and partners.
DESIGN
PRINCIPLES
1. Be People Centred
2. Communicate (Visually & Inclusively)
3. Collaborate & Co-Create
4. Iterate, Iterate, Iterate
CHALLENGE
Discover
Define
Develop
Deliver
OUTCOME
METHODS
BANK
Explore, Shape, Build
Creating the conditions that allow innovation,
including culture change, skills and mindset.
LEADERSHIP

Figure 13 British Design Council Framework for innovation (2019)

IDEO and Design Council have offered a brief explanation as to what is involved in each of the phases. The Inspiration phase is premised around information gathering, including the utilization of user research tools to study the behaviour and motivations of people/humans by watching and learning (Marsh, 2018). The lens of desirability, viability and feasibility (figure 11) would be used to ascertain humans' thoughts and emotions towards the current products or services, to help inform the behaviour research. This is an opportunity to watch, ask and explore potential problems, opportunities, new markets and disruptive technologies without presumptions, in order to shape an appropriate and meaningful design proposal.

IDEO's intention is that observations are conducted, and users engaged from the very beginning. With a focus on behaviour and experiences, IDEO have made observations explicit in their philosophy, whereas it is implicit in Design Council's 'Discover' phase. This phase is about the triangulation of findings to inform the next phase.

The Ideation phase is about applying what was learnt in the first phase and generating multiple ideas. This phase involves validation and consultations throughout, via the generation of physical low-fidelity models, and includes empathy tools, such as body storming and usability testing to determine how to refine the ideas. Convergent thinking at this phase includes the utilization of the lens of desirability, viability and feasibility to further refine and validate ideas. Design Council's 'Define' stage is described as an opportunity to generate and illuminate ideas and to introduce physicality via prototyping. The IDEO process highlights ideating, rapid prototyping and user feedback and interaction with the aim of preparing ideas for the intended users to interact with and feedback on.

Finally, the Implementation phase is about production, launch and impact. This stage is seen as key to evidencing the impact HCD can have on the world. Implementation, or 'Deliver' as it is known in the Double Diamond, is when a product is taken to market and formally implemented. Fidelity levels of prototype testing are significant in both IDEO's and Design Council's design processes: IDEO put the emphasis on starting with low-fidelity, cardboard models while the UK Design Council's Double Diamond hints at testing as a 'development' role and eludes to an activity that happens much later with higher fidelity models. Fidelity is defined in Section 2.4.

Figure 12 and now the newly refined model in figure 13 are the two common processes in the industry. In 2007, Design Council investigated the use of the design process in 11 companies and mapped it back to their process and highlighted the prevalence of design process and thinking in leading design companies. It concluded that all 11 companies utilized a design process with a similar core to the HCD process but using different terminology. However, when comparing usability testing in the companies, it was evident that usability testing was confined to concepts near completion and implementation with only one company recognizing the role of *in situ* testing over a prolonged period of time (Design Council, 2007). When referring to the HCD process in this research, the Inspiration, Ideation and Implementation phases will be applied due to their simplicity, accessibility, and reference in the sector literature. Early validation of design proposals via usability testing is still an area that needs refining in terms of accessibility of setup—that includes context and social cues early in the Ideation phase of the HCD Process.

2.2.2 User Insights

User insights are gleaned via tools that are used during the HCD process. The role of these tools is obtaining meaningful observations and insights, ideally in the intended context of use. Primarily utilized, although not exclusively, in the Inspiration phase of the HCD process, they allow design teams to collect insights that can be applied to the Ideation phase. Marsh (2016) is an advocate for engaging in user research on a cyclical basis, so errors can be found, new opportunities identified and changing behaviours can be ascertained. A current global pandemic of COVID 19 is a casing example of how behaviours can change significantly to adapt to new scenarios. For example, in the way people communicate when physical proximity is restricted by legislation (Gov, 2020).

In the next few paragraphs, a selection of user insights tools are explored, as a plethora of tools are now available to understand the links between 'design and behaviour' (Hanington & Martin, 2019, p823). The relationship between user insights and usability testing is that insights are normally applied to design proposals and usability testing is normally applied during the Ideation phase of the HCD process to validate these applications of insights.

Cultural Probes are used as a means *"of obtaining inspirational response from people"* (Muratovski, 2016: p67) and a *"kit [that] typically includes items for gathering a variety of information in a creative manner" (*Milton & Rodgers, 2019, p47). This method can be used as a memory aid for participants and offers context in the form of images or sketches, either scenarios, ideas or storyboards. Cultural Probes can help ground participants in the scenarios that are being designed for and can offer participants clues without solutions. Cultural Probes can therefore help participants offer more meaningful or inspired insights (Gaver *et al*., 2004). Cultural Probes are predominantly used at the initial Inspirational phase of the HCD process to gather insights into human behaviour and emotions and serve to probe deeper than abstract questions. Storyboarding can also be used as a probe to solicit feedback on ideas under development (Buxton, 2007). Probes can also be used as data capture; for example, if participants are struggling to articulate their feelings, they can sketch them (Muratovski, 2016). Cultural probes can be used in individual settings but also as part of a focus group. The focus group seeks to gain the opinions of a group of the targeted demographic and is an opportunity to gain insights of the participants' opinions; however, ultimately, participants are there to contribute to a discussion premised around a particular theme (Dawson, 2009). Focus groups can also form part of an interview process, be it in a formal or semi-structured way (British Design Council, 2007; Muratovski, 2016; Ulrich & Eppinger, 2004). Focus groups can be used to inform the development of an idea (Barnum, 2011). Composed of stakeholders, they offer thoughts and feelings on a given topic (Muratovski, 2016). Sessions are guided by a moderator who keeps the group on topic, drawing out less dominant voices and working to reduce 'group think' as far as possible. The task may be predicated on memory of past experiences, but when others talk it should act as an *aide-mémoire* for other group members (Dawson, 2019).

Ethnography is more akin to the IDEO model and is a fundamental component of HCD. Ethnography has grown in prominence in design because, when coupled with interviews or cultural probes, it becomes a powerful tool (Norman, 2013). Ethnography, stemming from anthropology (the study of people), relies on design researchers being immersed in the relevant environment (Dawson, 2019). What people say they do and what they do, as noted by Margret Mead (Luma, 2012), can be rather different, and the ethnographic approach allows researchers to understand participants' values and beliefs. Ethnography is a visual research aid; the designer is able to observe individuals in context and capture their behaviour as well as interjecting with semi-structured questions. The visual aid is the environment the participant is in, grounding the participants in their context and allowing for the observer to engage in the discoverability, as noted by Norman (2013). To begin with, it informs the brief and gives an insight into participants' expectations and issues. It also allows for the designer to see the activities surrounding the participant and product and offers an overview of social context (Muratovski, 2016). Walters & Evans (2011) explored how ethnography is used in a very structured laboratory environment so as to gain data on user insights via eye tracking anthropology and touch cues. Although their paper notes it added value to the Ideation phase of the design process, each designer noted that observations in the product's natural environment (*in situ*) were missing and would have been of benefit. Redstorm (2006) and Norman (2013) both air caution that structured research may predetermine every aspect of a design and, as a consequence, lose the sense of play and discovery; for example, would the Stark juicer have ended up as it did if informed by structured research?

Structured ethnographic approaches have also been developed through ‘Living Labs’ which are explored in more detail in Section 2.5.2.

Surveys are widely used in the design industry. Used correctly, they can survey for customer opinion on existing products, opinions, values and beliefs. However, used incorrectly and in isolation they can be problematic. As mentioned above, what people say they do and want, and what they actually do and want, is not necessarily the same; and asking participants to recall past behaviours does not necessarily capture true behaviour (Marsh 2016). This is reinforced by McAdams *et al.*’s (2006) longitudinal studies of life stories and how they developed to meet the narrative premised around cultural and expectations and identities. The studies highlighted that, over time, adults *“constructed more emotionally positive stories”*, hence the importance of ethnographic research approaches, so data can be correlated. Surveys, however, are a means of obtaining big data on a given topic. Where more detailed insights are required, semi-structured interviews are preferred, allowing as they do for flexibility in obtaining feedback through follow-up questions and further probing in areas of interest as they occur; however, as a consequence, they can be more complex to analyse (Dawson, 2019). Surveys can play two roles when gaining insights: they can be used as a recruitment tool for usability studies and they can be used to collate information from the market about existing products or for identifying gaps for new product development. In many instances, product managers and/or marketing departments are more likely to conduct them and feed back the findings to the design team (Marsh, 2018). This was also echoed in semi-structured interviews conducted by the author in four manufacturing and design companies, and in Carulli’s (2012) work on developing early virtual reality prototypes to capture the voice of customer (VOC) to inform customized product development, Carulli notes that the consequences of a product-manager-led activity, when gaining insights, is that it can generate *"distortions in translating the VOC into functional requirements at the present moment".*

Unlike usability design research, the tools discussed in the user insights section are predominantly qualitative data gathering methods that allow designers to build a picture that will inform the Ideation phase of the HCD process. When convergence of ideas is needed, more qualitative data gathering methods are used to rule out a design or parts of a design through continual iterations. The significance of gaining stakeholder insights is detailed in the latest HCD ISO standard. *"Constructing systems based on an inappropriate or incomplete understanding of user needs is one of the major sources of systems failure."* (BS EN ISO 9241-210:2010.) Tools utilized to gain user insights during the Inspiration phase of the HCD process can also be used during the Ideation phase to test early design iterations. This is evident in Milton & Rogers' (2019, p16-17) book *'Research Methods for Product Designers',* whereby a summary of research methods normally associated with gaining insights are also used when developing designs.

2.2.3 Physicality and Fidelity in the Design Process

The advantage of HCD as an applied process is the ability to gain a more holistic and deeper understanding of psychological, social, economic, ergonomic and behavioural factors, which means that products and or systems or services are produced to match the intended expectation of the users, or even surpass it, and users can make sense of the product and consequently its intuitive interaction (Norman, 1998). However, the disadvantages are the time and cost associated with gathering insights into the intended context in which the product will be used (Abras *et al.,* 2004). It is this disadvantage that leads the author to explore physicality in HCD. The role of physicality is particularly important in this body of work, because to explore context and usability testing early in the HCD process, physicality of a prototype and the physical attributes of the intended context of use need to be considered in terms of accessibility. For example, how much effort is required to recreate a prototype in a recreated environment to yield meaningful results to inform product development early in the HCD process? Physicality is deemed to be more than the physical manifestation of prototypes, but a way of thinking—by utilizing the tools related to physical artifacts they become thinking tools and contribute and aid problem solving (Hallgrimsson, 2020); Crawford (2010) refers to metacognition, the means of stepping back and *"thinking of your own thinking"*; Castaneda (2012) refers to practical thinking as being a state that is more than theoretical or contemplative, but by the act of intentional doing we are applying thinking; this is also supported in the creativity literature, where Loudon & Deininger (2014) discuss the LCD model for creativity, premised on listening, connecting and doing. The LCD model notes that one component of the 'Connect' phase concerns divergent thinking and the role of exploration of ideas so connections are made, and the phase of *"doing relates to taking action, exploring, experimenting and prototyping"* (Loudon & Deininger, 2014).

Physicality is defined by Gill *et al.* (2008) as *"a broader term that encompasses our entire interaction with the physical world".* Therefore, physicality is recognized as much more than a means of visually representing a two-dimensional drawing and its intended human interactions of a product in development. While sketches are a critical component of the design process, a design of a 3D object on paper can be misleading—for example, Escher's drawings of impossible objects (figure 14).

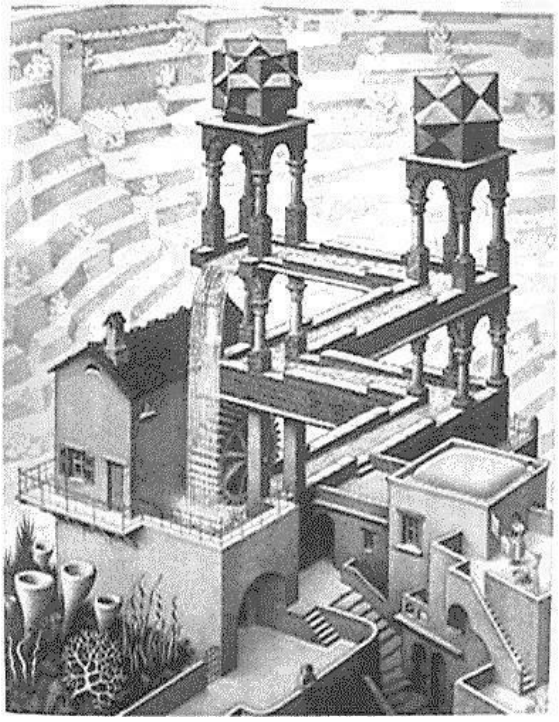


Figure 14 Waterfall (1961) by Escher

Only when a design is represented in 3D at a fidelity appropriate to the stage of the design process can a 3D design be properly evaluated: *"it is the prototype that brings to life the experience behind user experience"* (Cao *et al*., 2015). As Muratovski (2016) notes, prototypes develop as the research progresses. There are key processes that facilitate physicality as part of the HCD process: sketching and storyboarding, Computer Aided Design which facilitates 3D modelling, rapid prototypes and those that are built by hand. Warfel (2009) established some key principles that exemplify the purpose and role of prototyping in design. These were:

- communication and collaboration
- gauging feasibility while reducing waste
- selling your idea
- testing usability early

- selling your design priorities

Hallingrimsson (2020) noted specific design processes associated with roles of a prototype. These include exploring ideas for one's self-assurances when exploring ideas on paper, then testing ideas; verifying designs; and the technical exploration and standard testing when translated into a physical form. Houde and Hill's (1997) paper, '*What do Prototypes Prototype?*' identifies that the role of prototyping is to *"represent different states of an evolving design".* This sentiment is echoed by Muratovski (2016), in that the development of a prototype should echo the progress status of the design process. However, it is the fidelity level of a prototype that determines when and how a prototype is used during the HCD process.

To understand the relative intention and meaning of the term 'fidelity', Römer *et al.* (2001) classified low-fidelity prototypes as simple models/representations. Hallingrimsson (2020) recognizes that these simple models are made out of materials that are fast to work with and allow a designer to explore the 'bigger picture' before engaging in detail, as seen in figure 15. Warfel (2009) notes that prototypes can be rapid, low-investment works while Milton & Rogers (2019) refer to low-fidelity models as sketch models that are accessible 3D representations of an intended design early in the design process and are utilized during the Inspiration and Ideation phase of the HCD process.

> *"Low-fidelity prototypes are put together quickly and are usually crude and unpolished. [….] made out of cardboard, or perhaps digital but limited in functionality and a basic interface."*
>
> Saffer (2010, p177).

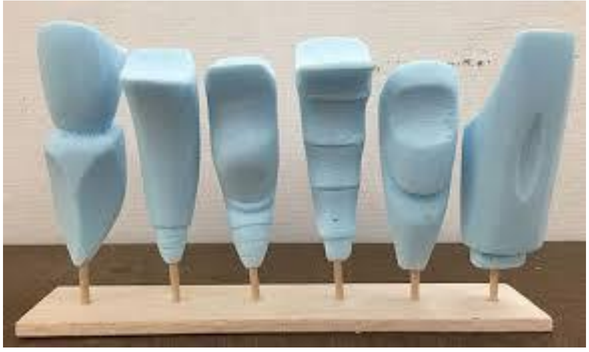


Figure 15 Low fidelity prototype

Buxton (2007) later describes the investment of time spent on each design activity (figure 16) as the design process progresses, noting that as the activity nears usability testing the time spent on prototyping increases.

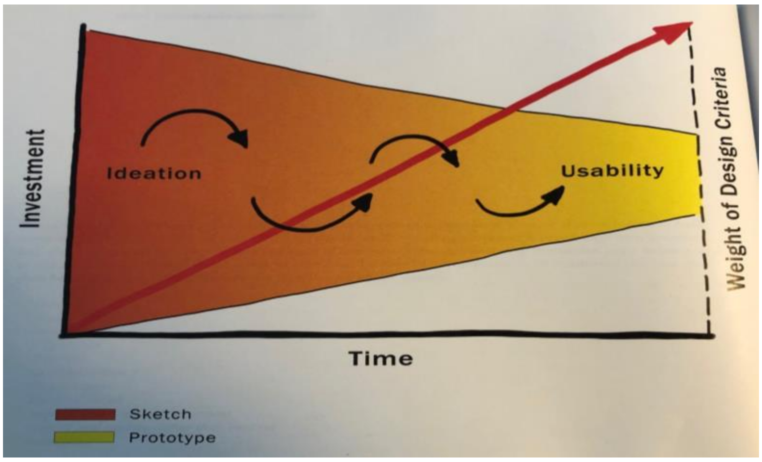


Figure 16 Buxton (2009) diagram of investment vs time

Medium-fidelity prototypes include more detail that supports the Ideation phase of the HCD process. Milton & Rogers (2019) refer to this type of model fidelity as 'mock-ups' that enable an individual to test and interact with the design. Figure 17 is an example of a prototype with mechanical capability in preparation for usability testing as part of the Ideation phase of the HCD process. They are more sophisticated representations of the intended design, although the choice of material can be simple. However, with sophisticated use of card and affordable and accessible material, and in parallel with each design iteration meaning, more functionality can be resolved, and can therefore be included in the prototypes (Muratovski, 2016).

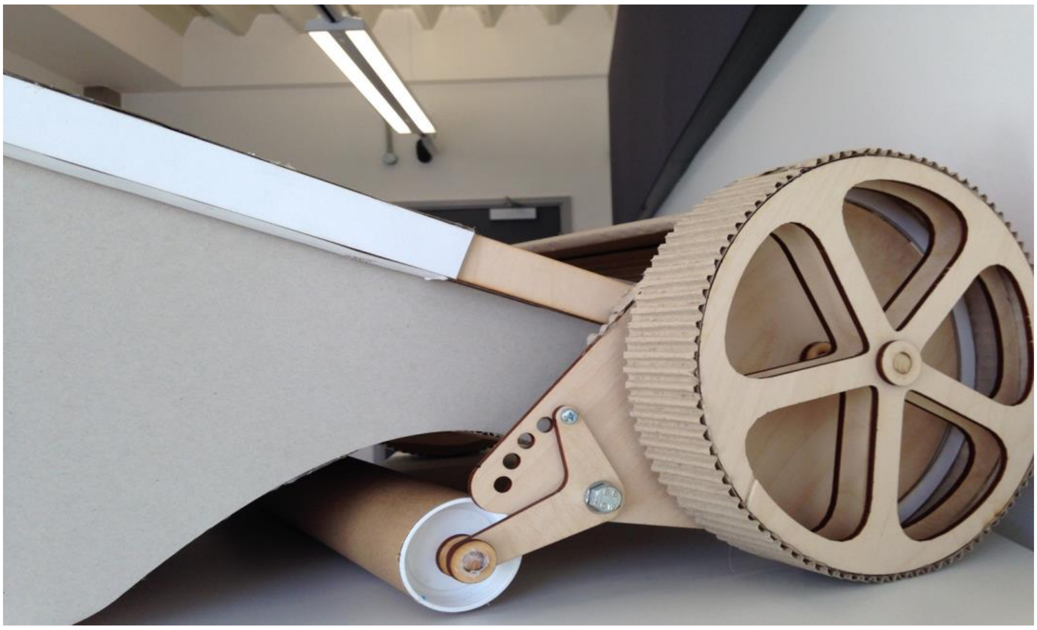


Figure 17 Example of Medium fidelity prototype

High fidelity is a prototype identified as being a refined representation of an intended design that includes functionality and communicates complex detail that has been executed often via rapid prototyping. This prototype aims to illustrate aesthetic and functionality (Hanington, 2019) and would normally be found at the Implementation phase of the HCD process. Recognizing that high-fidelity models can take *"double the time and cost"*, in comparison to a low-fidelity model (Gill *et al*., 2008), the level of investment recognizes the refinement of the design at the Implementation phase.

All three fidelity levels are important and have a purpose at each stage of the HCD process: *"Prototype early and often, making each iterative step a little more realistic."* Moggridge (2007, p643).

Houde and Hill's (1997) early work in prototyping established a very clear triangle that mapped the key components that need to be considered when creating a prototype (figure 18). It identifies 'role' as the role of the prototype, which also corresponds with purpose, then 'look and feel', which concerns the sensory components of a prototype, and then 'implementation', which refers to the functionality of a prototype.

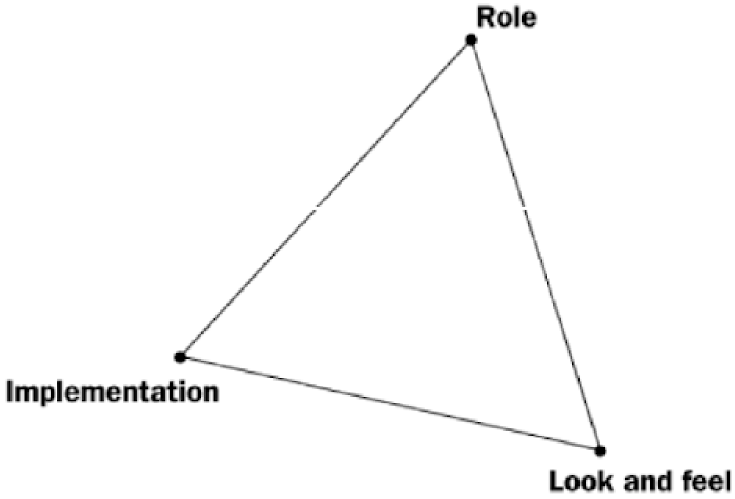


Figure 18 What prototypes Prototype (Houde & Hill 1997)

This work has been later addressed by Hare (2015) on active and passive physicality, noting that passive physicality is *"perceived affordance based on the visual appearance and tangibility of the prototype",* and active physicality is *"perceptible experience of interacting with the prototype".* Figure 19 maps the role of active and passive physicality and how meaningful the feedback is when using prototypes to glean user feedback. This work is important as it recognizes time invested and yield of results.

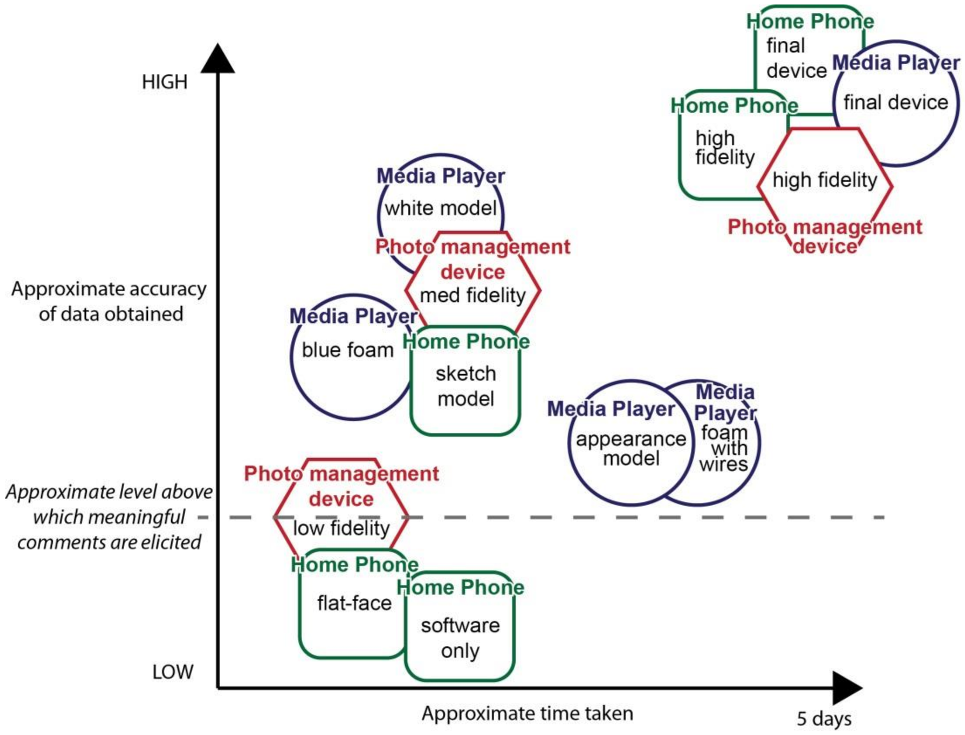


Figure 19 Relative success of the prototype versus the time taken to create the prototype. Hare (2015)

These principles identified by Hare *et al*., and as early as Houde and Hill, can also be applied to simulating the context, in that the fidelity of the context prototype should match the fidelity of the product prototype being tested. This is echoed in Dahl *et al*. (2010) who recommend that the fidelity of simulation should increase as the fidelity of the prototype increases, and needs to be planned, targeted and correlated with the prototype under test. Otherwise, the result yielded will not replicate the complexities of the product's intended context of use. Therefore, we can see the principles applied to prototype fidelity translate into simulating/prototyping environments in which to test the product.

Verification of ergonomics is usually conducted via usability testing, when the assumptions of the designer are challenged and areas for development are highlighted. This can only be conducted efficiently if there is enough understanding of physicality in the design process and the most appropriate usability testing method is deployed (Ramundy-Ellis, 2011). Ramundy-

Ellis *et al*. note that humans are naturally evolved to interact with physical objects. Interactivity, in the context of the HCD process, has been defined as a response to an action conducted by a user (Svanaes 2010) and a means of testing design principles (Saffer 2010).

2.2.4 Ergonomics

The study of ergonomics is recognized by the Health and Safety Executive (HSE) as Human Factors Engineering or Human Factors Integration, i.e. the *"application of scientific information concerning humans to the design of objects, systems and environment for human use"* (CIEHF 2019). Ergonomics (and/or human factors) is a scientific term used in the design industry. As defined by the International Ergonomics Association (IEA) (2000 & 2018), the term ergonomics derives from the Greek words *ergo,* meaning 'work', and *nomos,* meaning 'natural law' (Pheasant & Haslegrave, 2006).

> *"Ergonomics (or human factors) is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance."* Dul *et al*. (2012).

Pheasant & Haslegrave (2006) explain in their book, *'Bodyspace'*, that the term ergonomics was coined in 1949 by Professor Hywel Murrell when he started the human factors research group as a response to the growth of what was then called the 'man machine age'. By that point, it had become apparent that more work was required as machines became increasingly complex and sophisticated, and the interaction between human and machines required further work and analysis to better define their relationship.

When referring to the literature, it is evident that the definition 'ergo' and its origins were developed for safe working practice and this is supported in the HSE; therefore when researching the literature, ergonomics is often cited as a manual work related area, driven by the legislation in the government HSE. For example, Saetren *et al.* (2016) refer to ergonomics in the development of a large drill and explored why users' 'technology acceptance' was high in an offshore rigging environment. This reference is not alone; the datum found in literature is often a work-related product and this is due to the inherent safety issues that need to be addressed in the workplace, thus giving it more prominence in the literature than the ergonomics of everyday objects. One of IEA's reports, cited by Dul *et al*. (2012), highlights the need for ergonomics as falling into two categories: the need for ergonomics for marketing purposes (this is a very shallow outlook), and the need for ergonomics in 'safety critical industries'. The industries dominated by safety, therefore, embed human factors as standard, and they include the automotive, aviation and medical industries. The automotive and aviation industries have an established human factors culture embedded into their design and development process, while the medical profession developed an understanding of human factors' importance later when it was realized that human error can be a result of poor consideration of human needs, and can have a high impact on patient safety if it is not considered or discovered via usability testing when developing systems and products (Carayon, 2012). Again, it is the necessity of safety that forces the inclusion of human factors rather than it being generally desirable. Saetren's (2016) research into human factors in an industrial setting considers the 'forgotten factors' and relates to *"when human mental or physical actions do not lead to the planned outcome"*. This is not dissimilar to the work of Blandford *et al*. (2010) in the medical environment, where they concluded that factors that distract the user contribute to human error and that these types of errors are the result of a failure to identify flaws in a product at the prototype stage. For example, what happens when a nurse is interrupted while programming an infusion pump, and cannot remember their last action and has no, or limited, ability to interrogate the product for the relevant information due to its 'hidden state'? Blandford (2010) particularly makes reference to the context the product is used in and the impact it can have on a user's interaction with a product when in context. Although the previously mentioned ISO standard in Section 2.3 makes reference to medical computer embedded products as a means of ensuring human factors are considered in the development of the devices in their 'environment', it does not stipulate the consideration as required throughout the HCD process as it does with human involvement. This means that the impact an environment can have during usability testing can be left to the Implementation phase of the HCD process. Environment testing early is still problematic due to ethics and accessibility, but also, as noted by Gill (2009), a company's size and structure can impede on an organization's ability to engage in early context usability testing due to the speed a product needs to get to market. The IEA acknowledge the greater environment but do not stipulate the conditions required for better usability testing environments.

A more generalized approach to human factors has now been recognized by the IEA, complementing the earlier work of Pheasant (2006) and later Norman (2013). The IEA and Human Factors Ergonomics (HFE) characteristics found in Dul *et al.*'s (2012) report are developed from a common understanding that ergonomics is characterized by three descriptors:

> *1. HFE takes a systems approach.*
>
> *2. HFE is design driven.*
>
> *3. HFE focuses on two related outcomes: performance and well-being.*

The systems approach is all encompassing in that it considers the environment as well as the human task. This holistic approach concerns itself with each individual component and is underrepresented in current usability testing scenarios (Dul *et al.*, 2012). HFE is design-driven from the very heart of the process and encourages human factors specialists to play a key role in the design process, especially in the later phase of the development process. This approach focuses on systems to improve performance and wellbeing at every stage of a product's life, from maintenance to disposal. The third descriptor encompasses systems, performance and wellbeing and, by considering the environment, it consequently acknowledges productivity (Dul *et al.,* 2012).

When relating ergonomics to the HCD process of Inspiration, Ideation and Implementation, it is evident that human factors are considered during Ideation but usually only tested in context during Implementation.

2.3 Usability Testing

Usability testing is a method used during the HCD process to evaluate the effectiveness of a design with a representation of the intended end user (Barnum, 2011). It is traditionally associated with the refinement of a design in the latter end of the Ideation phase in preparation for implementation. Usability testing can also be conducted to gain insights during the Inspiration phase by exploring usability issues in existing products (Milton & Rogers, 2019). How usability tests are conducted varies significantly and during this section the author will explore the usability opportunities afforded in the HCD process, but also lessons learnt from the HCI community that focuses predominantly, but not exclusively, on mobile devices.

Usability has been defined by the ISO for Ergonomics of human-systems interaction as *"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use"* (BS EN ISO 9241-210, 2010, p3). The purpose of usability testing in the HCD process is to ascertain the visibility, mapping, feedback and the physical, semiotic, cultural and logical constraints of a design. This includes challenging the assumptions of the designer on the actual and perceived affordances and the design grammar, as described by Norman (2013).

Nielson is one of the founding pioneers in user experience and penned his book of Usability Engineering in 1993 where he noted the significance of usability testing in the software industry (Nielson, 1993). Around the same time, Norman was working for Apple and he too recognized the significance of usability testing, and later went on to apply the HCI thinking to broader design activity and authored the work *'The Design of Everyday Things'* (Norman, 2013). Norman made the connection between HCI software principles and physicality in design, noting that both need to coexist for a product to be intuitive to use (Nielsen Norman Group, 1998-2020).

The role of usability testing is *"the activity that focuses on observing users working with a product, performing tasks that are real and meaningful to them"* (Barnum, 2011, p13). As highlighted by Boothe *et al*. (2013, p1033), usability testing should begin early in the design process during the Ideation phase. Barnum (2011) identifies these studies as being formative usability studies that are not necessarily statistically valid, but offer insight to further develop a product during the Ideation phase. Barnum later describes summative usability testing as an activity that is conducted when a product is in development and involves larger scale usability testing. This stage of testing is more aligned to the Implementation phase of the HCD process. Note that it is about engaging participants in the process and is a tool to support objective decision-making in design. Usability testing is intended to engage participants in the process so objective insights can be gained, but when this is not possible, then a heuristic evaluation process should be conducted (Barnum, 2011). Heuristic evaluation was originally developed as a means of validating systems software (Barnum, 2011). It appears at slight odds with the intention of usability testing; however, it is used as an option when participants cannot directly engage with prototype testing. The process involves the designer and those who work for the organization 'walking through' the functionality of a product, challenging usability at each step of the way. The main disadvantage of this process is the perception blindness and investment in the design of the product (Saffer, 2010). In the 1990s, Rolf Molich and Jakob Nielsen developed a set of principles that enabled a more guided approach to evaluating a user interface, later refined by Nielsen in 1994 (Nielsen, 1994a). The principles are there to support the external expert's 'inspection' of a product and, prior to the evaluation starting, each 'evaluator' must agree a persona that is used when applying the principles (Barnum 2011). The design must comply with these principles:

1. Visibility of systems status
2. Match between system and real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognise, diagnose, and recover from errors
10. Help and documentation

Nielsen (1994a).

Nielson attempted to set up the above principles to formalize the process for software development. The heuristic (which means rule of thumb) approach recognizes that you evaluate your design against the principles and, when the principles are not met, then usability testing is introduced. This process is considered as a plausible approach to objectively evaluating a design proposal; however, it does not account for contextual usability evaluation. Like usability testing, Nielsen and Norman's approach of using at least five participants as a quick and effective way of finding usability faults, also applies to the heuristic approach. Five evaluators are believed to find all the main faults in a design in development, and although this approach was originally for computer interfaces, it is now a recognized approach in general design practice (Barnum, 2011).

When considering usability testing as part of a human-centred design process, it is expected that user insights have already been collated and applied to the initial concept sketch work and that usability testing is used as a means of affirming and fault-finding in order to develop the most appropriate product. Carulli *et al*. (2013) highlight a need for a clear approach. They identify three requirements that should be established when testing a product with its intended user:

- basic requirements
- technical performance requirements
- attractiveness requirements

Another framework that offers structure and a set of criteria for evaluating a design is the AEIOU framework developed by Rick Robinson, John Cain and Julie Pokorny, to offer structure and document a product's performance against coded criteria (Martin & Harrington, 2019). AEIOU takes account of Activity, Environment, Interactions, Objects and Users. This framework encourages due consideration of the holistic product and acknowledges environment and is captured via data sheets consolidating ethnographic insights. AEIOU is a powerful tool that considers task analysis and is used as a visualization method to consolidate and visualize the insights gained. This testing method is also described by Milton & Rogers (2019) as scenario testing, that captures the intended context. Used usually during the latter stage of the 'insight' phase of the HCD process, scenario testing is intended to develop a picture of a product's intended context of use and does so by creating a narrative and visually portraying it to participants via storyboards and photography—a form of future gazing predicated on research findings. Scenario testing is intended to capture and consolidate research findings in storyboards: *"By devising a scenario carefully with characters, narrative and context, designers can evaluate whether their design ideas will work with their intended users."* Milton & Rogers (2019, p121). It is not as the title alludes, i.e. conducting usability testing using scenarios, but instead tries to capture enough insights to be satisfied that context of use has been considered early in the design process. The process is about anticipating context-specific issues that may impact a product in its intended use, rather than feeding them into evaluation usability testing.

2.3.1 Usability Testing Spaces and Activity

A number of usability testing environments have developed over the last 30 years. Some consider context of use, while some focus on a controlled space. Barnum (2011) gives a brief insight into usability testing and its origins in the early 1990s. Conducted by psychologists or specialist human behaviour or human factors experts, it was conducted as a formal research study in a laboratory, but was known for being too expensive and resource hungry; therefore, it was later refined by Nielson into an accessible method that required fewer participants. Laboratory Usability Testing is the most commonly understood usability testing space (figure 20). In its simplest form, a usability testing laboratory is a controlled environment that enables a moderator to set up and conduct a predesigned usability study. These laboratories were initially established to explore psychology (Barnum, 2011). These laboratories may have two-way mirrors, but with the development of technology it is increasingly likely that

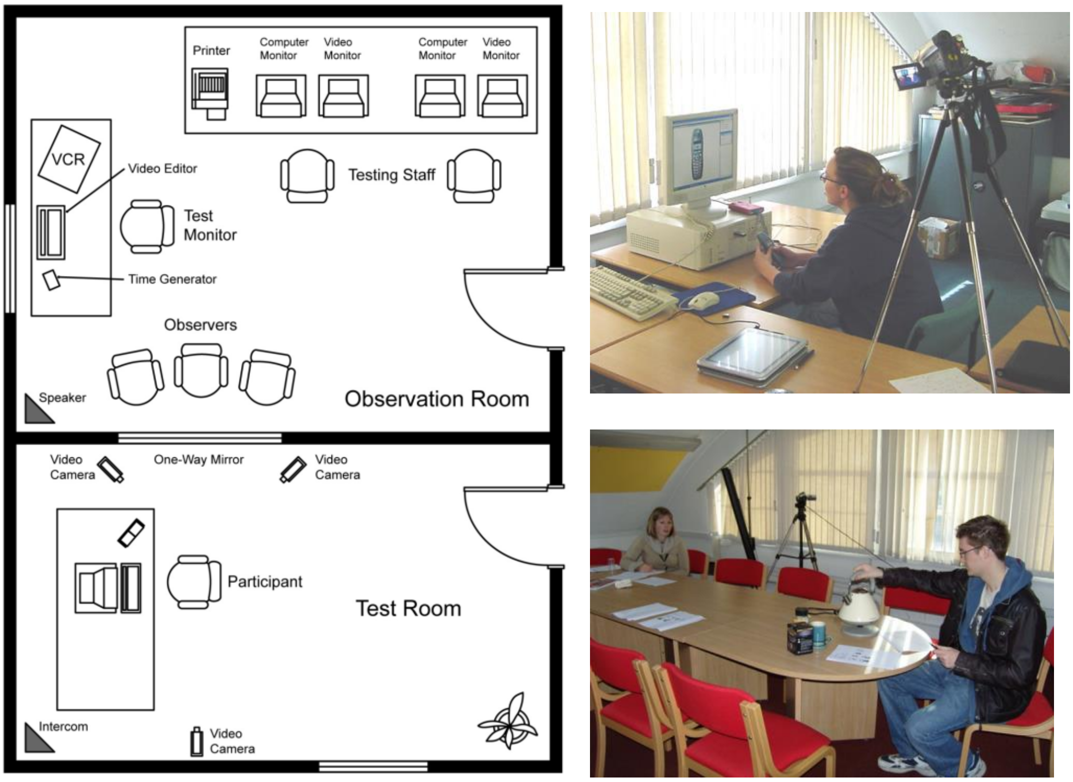


Figure 20 Usability Lab setup. Rubin (2008 p56)

observation will take place via cameras linked to observer software to capture the studies. In its simplest form, this space would have appropriate furniture and a means of capturing the study (Barnum, 2011; Rubin, 2008). It is also recognized that a laboratory can be a portable setup that enables the testing moderator to control the environment.

High-fidelity user trials, field studies and Gorilla usability testing[1], i.e. in-context usability testing, are associated with the Implementation phase of the HCD process, unlike earlier usability tests which are conducted in controlled laboratory settings (Brown *et al*. 2013). There is a presumption that usability testing is conducted throughout the design process and

[1] Gorilla studies involve taking the study to the participant in context. There are obvious advantages to the approach, but it can be difficult to engage participants in context due to the feeling of self-consciousness. It is also harder to control the environment, especially when involving multiple participants (Marsh, 2018; Brown *et al*., 2013).

this is a validation of a design, whereas additional resources are committed to developing a high-fidelity prototype for testing in context. Context testing, or field studies, has been found to reveal more valuable usability insights that are not found in a laboratory setting (Woolley *et al*., 2013). One of the more valuable findings is the design team's exposure to the real context during the test (Barnum, 2011). However, there are disadvantages to conducting studies in the field, namely logistics. However, the insights and education gained by the team conducting the studies outweigh the cost, but as noted by Gill (2009), company size can have an impact on a company's ability to engage in field studies. Field studies tend to be confined to the latter end of the design process as early, low-fidelity prototypes tend to be dependent on other devices to support them working. Woolley *et al*. (2013) concluded that studies with early mid-fidelity models can yield positive results in context, but would benefit from both laboratory and context usability testing. Living Labs (Georges *et al*., 2015) and simulated environments (Deniaud *et al*., 2015) have both been cited in the literature as a response to controlling the resource and research environment as an alternative to field studies for the purpose of research. For example, the Centre for eHealth at Agder University, developed an approach to involve participants in its design process of eHealth products. They integrate HCD principles into their development process and use a combination of longitudinal studies via a Living Lab setup, integrating role play to better understand user needs and to gain insights into participants' behaviour (Smaradottir *et al*., 2015). The insights gained in the Living Lab were deemed as evaluation and the findings were complemented by conducting functionality testing during field studies (Smaradottir *et al*. 2015). Their Living Lab is deemed high fidelity, with the studies providing ecological validity that were validated when the product was used in the field (Smaradottir *et al*. 2015).

An approach developed to explore context when evaluating ideas is Experience Prototyping. Developed by IDEO's Buchenau & Suri (2000), it has similarities to role play and bodystorming in that it encourages the team to experience their ideas (figure 21). It seeks to actively engage the participant, but through the eyes of the designer, in that *"Experience Prototyping involves exercises completed by design teams to foster a vivid sense of the user's potential experience"* (Hanington & Martin, 2019, p3750). Unlike role play, Experience Prototyping uses props and walk-through scenarios—but no media. The approach is intentionally low fidelity and accessible. For example, Buchenau and Suri mocked up the interior of an aircraft to evaluate their food-dispenser designs.

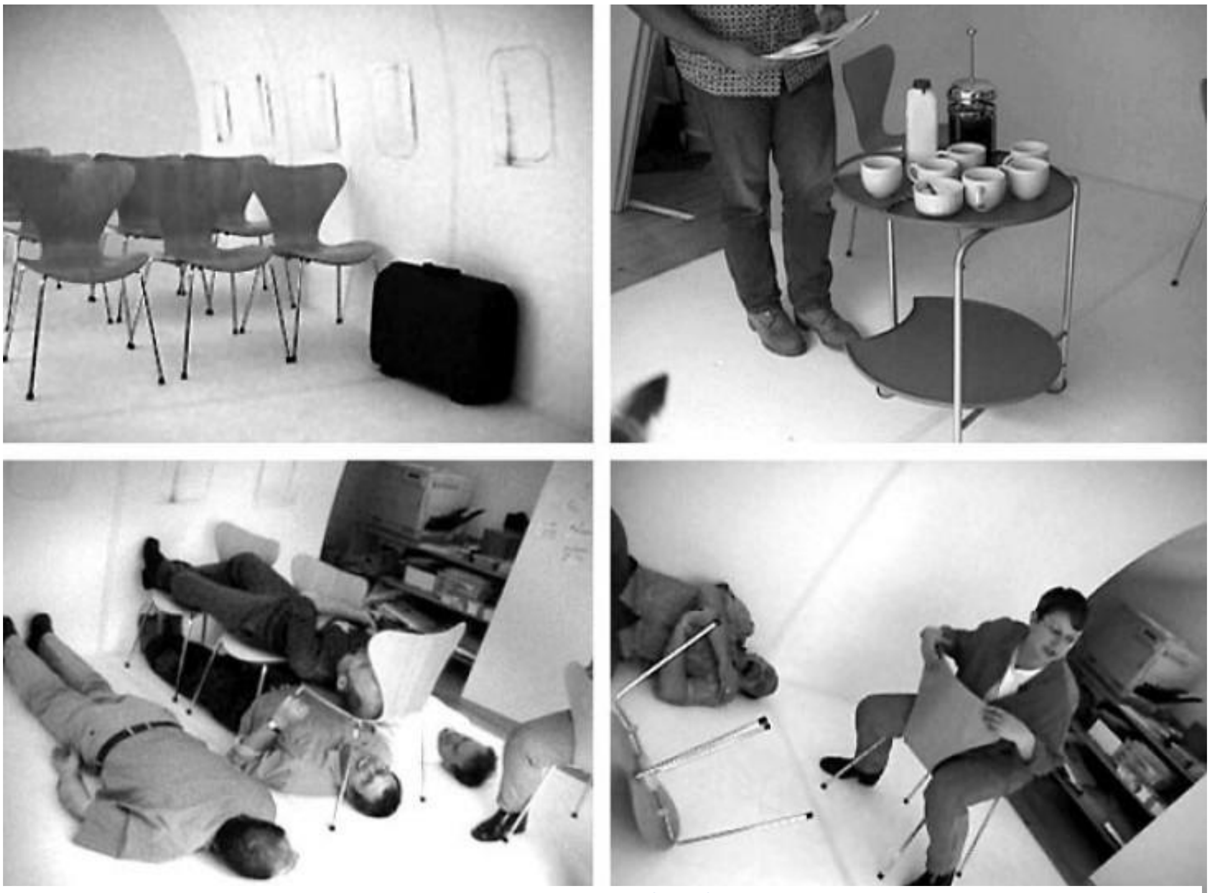


Figure 21 IDEO Experience Prototyping an aircraft interior. IDEO (2000).

Experience Prototyping proved capable of elucidating the user experience but has not gained much traction in the literature with reference to actively involving participants in the process

(Hanington & Martin, 2019). IDEO has superseded it with its role-play approach (Simsarian 2003).

Role play is a common method used in education scenarios, defined as "*pretending to be someone else, especially as part of learning a new skill: Role play is used in training courses, language-learning and psychotherapy*." Cambridge Dictionary (2020). This approach was introduced into design by IDEO, to user-test product ideas. It was subsequently developed into sub-sections or phases known as *Understanding, Observations, Visualization, Evaluation & Refinement* and developed into a form of bodystorming of the intended end users. IDEO use this approach with a team to ensure everyone understands the context and consciously places the design team in that moment, whereby the product or service in development is the sole focus (Simsarian, 2003). Bodystorming and role play are used as tools to collate insights and support usability testing during the Inspiration and Ideation phases of the HCD process. Buxton (2007) noted that, as early as 1985, More and Conn had recognized that designers could not assume they understood the needs of their intended audience. Role play and bodystorming are accessible methods to unpick the daily experience of different audiences, for example, the elderly.

Rapid ethnography and diary studies were developed as a response to the need to deploy research methods that allow designers to gain an appreciation of a product in its intended context of use (Woolley *et al*., 2013; Norman, 1998). Rapid ethnography differs from more conventional ethnography because it is about obtaining depth of understanding in a matter of hours or days rather than months (Rogers & Anusas, 2008). Diary studies is a technique used to obtain context-specific insights early in the development process. It is used to understand behaviours and needs over a long period of time rather than focus on repeated tasks, for example, purchasing a new washing machine (Ohly *et al*., 2010). Due to the volume of analysis as a result of diary studies, they are often coupled with contextual interviews so the data can be captured and controlled, and findings can be determined (Marsh 2018). Both methods are about early context-of-use appreciation; however, the work surrounding usability testing in context, early in the design process, is still limited.

Inspired by the *'The Wonderful Wizard of Oz'* (Baum, 1900), 'Wizard of Oz' is a method that enables the designers to recreate and 'fake' some of the interaction. Buxton (2007) helps us understand the relationship between the book and the concept used in usability testing. Buxton later notes: *"It is fidelity of the experience, not fidelity of the prototype, sketch, or technology that is important from the perspective of ideation and early design".* This method is used early in the design process and enables designers to gain user insights that help develop the product. To understand how best to utilize Wizard of Oz as a testing method, first we must appreciate the prototyping methods that lend themselves to Wizard of Oz. Paper prototyping is a method that allows designers to express designs, and the detail of the designs on paper, early in the design process, and allows for design developments to be rapidly implemented. Both Wizard of Oz and paper prototyping are low-fidelity prototyping approaches that enable rapid feedback on ideas in development and allow users to interact with designs to evaluate the concept's usability.

> "*Low-fidelity prototypes frequently don't work—that is they're usually static with no real interactivity at all. They require people to make them function, by faking any system behaviour."* Saffer (2010, p177).

The advantages of these approaches are their accessibility and ability to implement improvements early in the design process with minimal cost. However, some disadvantages include the lack of context and the requirement of the participant to use their imagination to anticipate what the overall product might look like (Cao, 2015).

An emerging question developed in the HCI community around twenty years ago, concerned appropriate evaluation methods and their role in capturing human interaction feedback (Kjeldskov & Skov, 2014). However, it was in Kjeldskov & Skov's paper *'Was it Worth the Hassle?*' (Kjeldskov & Skov, 2014) whereby we see the consideration for the space and place needed to conduct human evaluations of mobile devices. Driving this area for enquiry was the need to consider the wider complexities of a product's intended environment, for example the weather or sound. Findings from the 'Hassle' early papers, *'Is it Worth the Hassle?'* Kjeldskov, Skov & Hoegh (2004) set the scene for appropriate testing environment discourse and eventually concluded there was no clear answer (Kjeldskov & Skov, 2014); however, the messaging developed was that it was the when and the how, with a focus on 'truly in-the-wild and longitudinal' studies that held the key to refined evaluation methods. The paper, *'Was it Worth the Hassle'* offers up the debate of the values of the laboratory and the ability to control the setting and engage a number of participants in the study. However, Kjeldskov & Skov recognized that the level of ecological validity was low as a laboratory is a replicated/fake representation of the real intended context of use. Likewise, Kjeldskov highlights the same complexities as Deniaud *et al.* (2015) and Woolley *et al.* (2013) concerning in-context testing of multiple variables with less control over the environment and its influences on the testing conditions (Barnum, 2011). It was an apparent gap in the literature in the early 2000s that drove the author to conduct initial research into laboratory testing of products and the void of contextual attributes in the classic usability testing environment. It was during studies conducted as part of this work that there were hints that there was potential to explore the representation of context and its influence on the development of a product. For example, when a cup was dropped in the recreated environment setting, participants actively looked to see where it had gone, and when a dog ran into view, participants commented or gazed at the dog, breaking their concentration. The findings from the studies were preliminary and offered confidence in the direction of travel and the need for further research that established the optimum fidelity level of an environment and a detailed evaluation of the literature that spans product usability testing, the HCI community and the virtual environment community.

When conducting user studies, it is important to understand the number of participants that should be involved in a study to determine viable results. The Nielson Norman Group conducted research to understand the optimum number of participants for efficient fault-finding and have deduced that, although 15 participants is the optimum number to find all faults, beyond 5 participants the faults that are discovered are recurring, so that there is more value in running multiple studies with 5 participants (Nielson & Landauer, 1993; Nielson, 2018).

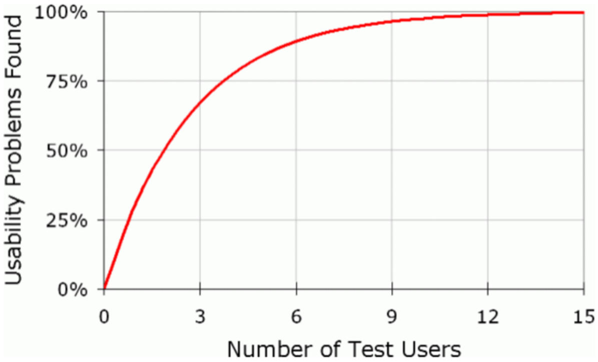


Figure 22 Nielson & Landauer (1993) user test faults found graph.

Nielson & Landauer note that even one user test is better than none but, to ensure finding all the key faults, they identified five participants as optimum (figure 22).

When referring to the literature on how to measure and capture usability faults, there are two highly cited approaches that can be used to analyse data gathered from studies. These are the Rolf Molich fault categorization (Molich, 2004) and Bangor's Systems Usability Score

(SUS) (Bangor 2008). Both offer something different in that the SUS applies a percentile to ascertain the overall usability performance of a product or system, whereas Rolf Molich's approach gives detailed usability issues. Both approaches are quantitative in nature (Barnum, 2011). There is some scepticism around post-completion questionnaires as they are predicated on the participant's memory of the activity rather than a real-time answer (Jokinen *et al*., 2015); however, when cross-referenced with the two above approaches, it can provide powerful data (Eccles & Arsal, 2017). One alternative, or supplement to the post-completion questionnaire is the think aloud protocol, developed by Ramey and noted by Barnum (2011) as a means of obtaining in-the-moment responses to usability issues.

2.3.2 Laboratory vs Context

There is much debate in the literature regarding laboratory vs field (in-context) usability testing. Kjeldskov & Skov (2014) reference Johnson (1998) supporting the need to conduct usability testing on products in context: *"the conventional usability laboratory would not be able to adequately simulate such important aspects [….] and could not easily provide for the wide range of competing activities and demands on users that might arise in a natural setting".*

Laboratory user studies are recognized as usability tests being conducted in a dedicated and controlled environment whereby product usability issues in general can be explored via a variety of methods (Kaikkonen *et al.,* 2005). Field studies are noted as set tests conducted in the real world—also known as Gorilla Testing. Both aim to achieve the same outcome, but the only point of difference is one contends with all the distraction of the environment and one has minimal noise as it is in a laboratory.

Kjeldskov & Skov (2014) reviewed ten years of literature concerning the laboratory vs field testing debate in a bid to definitively determine which approach should be used. They concluded that a definitive answer to the question regarding laboratory and field could not be arrived at. Instead, they note that it is not about whether we should test in one space or another but knowing when and how we test that are the key components of successful usability testing. These results are supported by Woolley *et al*. (2013) who concluded that there is a role for both laboratory and in-context usability testing of product prototypes early in the design process. The table below lists the findings arrived at by Woolley *et al*. (2013) in the debate around field vs laboratory studies.

Table 8 Summary of benefits and drawbacks for in-context and laboratory testing

Context	*Benefits*	*Drawbacks*
In-context	Uncover physical problems with a design, as well as physical and digital coordination problems early in the design process, that were not seen in the laboratory environment. Some minor problems experienced as more severe issues during in-context testing due to different social, physical and environmental situations affecting user performance.	Limited to scripted approaches for early stage prototype testing. Setup requirements for early stage prototyping can impose restrictions on mobility. Privacy and confidentiality harder to ensure when testing in some environments.
Laboratory	Promotes user reflection more readily to help uncover issues not seen in-context. Convenience of hardware setup for prototype testing and recording. Private environment to ensure confidentiality.	User behaviour and mode of interaction are limited. The impact of different contexts, including social, physical and environmental on physical problems with design is difficult to assess.

Woolley *et al*. (2013).

Barnum (2011) confirms Woolley's observations concerning field testing and notes that connectivity between participants and their natural setting is a powerful insight when evaluating a product. However, Barnum also notes that distraction, although beneficial when studying how a participant would use a product in context, can hinder the ability to collect and hear data. The lack of control of the environment adds the highest values in terms of 'realness', but causes complexities in data gathering and controlling the research.

Thimbleby (2013) conducted work into the user interaction of medical devices and concluded they were provoking user error because distraction, a constant feature of the real-world environment of medical products, was not tested for in the laboratory. It turned out that distraction was hindering a nurse's usability of a device to the point where patients were dying. Testing of the product prototype in context might have offered this critical insight.

The reason why the laboratory is still a prominent feature in any design usability testing process is because it is not always feasible to test *in situ.* Dahl (2009) highlights the complexity and ethics of usability testing in a hospital environment due to the health and safety of a busy environment, not to mention the red tape that is required to test in a medical ward. It can also be problematic due to the confidentiality of the product in development, which is why certain organizations tend to use heuristic evaluation methods.

Ultimately, when recreating context or emulating real-life context, ecological validity (Johnson, 1998; Kaikkonen, 2005) plays a key role in achieving realistic usability results. The need to control the conditions of a testing environment is critical, as is the need to ensure the fidelity of the prototype matches the fidelity of the context to obtain valid usability results—as identified in Section 2.2.2 on physicality and fidelity of prototypes in the design process.

2.3.3 Discussion

Human-centred design is not a new approach and, having been adapted and developed from the classic design process, it brings about focus concerning the entire human experience throughout the design process by involving and observing human behaviour. Having reviewed literature concerning variations of HCD approaches, for the sake of this research it is the IDEO Inspiration, Ideation and Implementation approach that will be referred to throughout this body of research. IDEO's approach, utilizing three fundamental components to capture the key phases in the HCD process, is a trusted process to adopt. In addition, its key philosophy of failing early and often in its approach to testing and validating reflects a core component of this body of research. Having researched usability testing in the context of early intervention in the HCD process and its associated benefits, it is evident that the mandatory nature of involving humans throughout the design process in the medical or equivalent high-risk sectors should be a recognized approach in all design practice. However, there is still work to be conducted in the general design practice that includes early intervention of usability in the HCD process that takes due account of the product's intended context of use and wider environment, as noted by Gill as early as 2009 and Dul *et al*. in 2012. The research literature suggests that prototypes should not be explored in isolation as they play a role in problem solving (Hallgrimsson, 2020). This means that usability testing needs to consider and include context of use with medium-fidelity 'active prototypes', as outlined by Hare (2015), being used as an accessible prototype, as they have sufficient complexity to yield usability feedback early in the design process, but can also be constructed from accessible material. The idea of failing early and often, as cited by IDEO, could then be accommodated, therefore reducing the risk of over-committing to a particular design proposal too early in the HCD process. This approach to prototyping levels will also be considered and echoed in any recreated environment that replicates context of use, in that they need to be accessible (Hare, 2015) with sufficient complexity to convey an environment (Kassab *et al.*, 2011).

Usability testing in the real context of use is recognized as a means of highlighting detail that otherwise would be missed when usability testing is carried out in a traditional laboratory environment. However, the 'lab vs context' debate allows for an opportunity to explore a hybrid of the two but recognizing the benefit of early intervention in the HCD process.

2.3.4 Conclusion

The first objective of the research was to understand the landscape and the literature concerning usability testing approaches. The purpose of this section of the literature review was to ascertain the landscape and establish prior knowledge in the field to justify the direction of travel. Having completed this work it is evident that early usability testing is a positive intervention as part of the HCD process, and although it can be conducted to gather feedback on competitor products as early as the Inspiration phase, it is early in the Ideation HCD phases where it has most benefit in discovering usability issues. Context setting when evaluating designs is noted as a key in the ergonomics definition and the significance of field testing is noted as gleaning valuable insight, but at a cost of time and resources and risking the unreliability of the environment. Both laboratory and field usability environments have negative and positive connotations on the development of the product, and it appears that both should be used. However, as field testing is often complex and costly to run it is often only utilized during the Implementation phase rather than earlier in the HCD process. In addition to the findings from the literature explored in this review, previous work conducted by the author also highlighted the need to explore context of use. Findings from that work pointed to the benefits of including context, but did not explore sufficiently how simulations could be created so that they were accessible and viable early in the HCD process. Early usability testing has benefits to the development of a product; a laboratory setting can facilitate this early design validation, however context is needed to yield results that benefit eventual *in situ* product use. Therefore, simulating an environment could marry the benefits of both laboratory and field testing early into the HCD process.

Having addressed objective 1, the next section of this review will progress to objective 2 and explore 'how' environments might be simulated and made accessible early in the HCD process to evaluate design proposals.

Objective 2: *To explore methods of recreating environments and how they can influence the requirements of the usability testing space.*

2.4 Simulated Environments

There is, at the time of writing, no design practice that uses virtual environments to test products in their context of use early in the Ideation phase of the HCD process (that the author is aware of). There are laboratories that are engaged in obtaining insight into human behaviour with regards to empathy and designing for the third age and specifically rehabilitation (Hanington, 2019). However, this thesis requires a review of practice in different contexts to establish whether they might apply to product design and development and how context can be recreated that is suitable and accessible in the HCD process.

The use of simulations is prevalent in sectors that deal with potential safety concerns, such as the automotive, aviation, nuclear, medical training and defence sectors. There are two main types of simulation: dynamic and static. In the context of a simulated environment, aviation simulations would be an example of a dynamic simulator and therefore used for training, whereas automotive simulations tend to be static and used predominantly for research and development purposes (Deniaud *et al*., 2015; Grey, 2002). The recreation of a virtual environment using fully immersive virtual reality and/or purpose-built simulation environments has been proven to carry too high a cost to be practical in most circumstances. For example, the work conducted by Dahl (2015) noted that there is a need to simulate an environment to ascertain participants' behaviour when evaluating HCI hospital devices, but he did so by creating a purpose-built replica hospital ward (Keller & Stappers, 2001; Deniaud *et al*., 2015; Dahl, 2010). Therefore, due to the resource implications, decisions are being made concerning 'what is good enough' in all virtual or recreated environments.

The automotive industry uses simulations to recreate various conditions and then test drivers' behaviour in a safe environment (Deniaud, 2015). These types of simulations are used to ascertain the impact alcohol, drugs and medical conditions have on a driver's ability. For safety reasons, these types of tests can only be conducted in a simulated environment so the participant and the general public are not put at risk, and the research conditions can be controlled in order to arrive at verifiable conclusions (Reimer *et al*., 2006).

2.4.1 Virtual and Simulated Task-orientated Environments: The Language

Kaur *et al*. (1998) describe virtual environments as providing *"a computer-based interface representing a real-life or abstract 3-dimensional space".* Meanwhile, simulated environments are recreated spaces that try to replicate a real context using physical and/or digital visual cues. The underlying principles required when developing a virtual or simulated task-orientated environment are the relationship between 'immersion' and 'presence'; and how to measure and increase these subjective properties. First, we will explore the terms and their definitions before exploring their collective relationship. Having distilled the differing opinions on the definitions of immersion, it is evident that both Slater (1999) and Witmer & Singer (1998) refer to content, be it physical, technological or, less tangible, memory. This is later confirmed by Mestre (2006), when he describes immersion as removing the real-world sensations and replacing them with virtual content. Witmer & Singer (1998) describe the conditions required to achieve presence and conclude that it can be captured in a basic formula:

$$\text{Involvement} + \text{Immersion} = \text{Presence}$$

Involvement or engagement relates to how meaningful a task is to the participant and how it captures their attention and consequently arousal[2]. Deniaud *et al*. (2015) note that when ascertaining ecological validity[3] of a driving simulator (i.e. the degree to which the simulated driving environment emulates the real driving environment), immersion and, subsequently, presence, can be measured utilizing data capture methods to determine the level of ecological validity. Involvement is measured via the 'attention' of the participant, the affect in an environment. Deniaud (2015) defines presence as *"a psychological state in which virtual objects are experienced as active objects".* Immersion concerns the psychological state of the participant as a consequence of the content they are exposed to (effect of content) and only with both these factors, the affect and effect, can presence be achieved. In Deniaud's thinking, to maximize the level of presence, a combination of factors should be considered:

- The activity needs to be meaningful.
- The environment needs to have consistent cues with the environment it is simulating.
- Attention is required.
- Social consideration is needed to enable the participants to 'believe' the simulated environment.
- Separation anxiety must be avoided, so the participant must not be thrust unknowingly into an alien environment or an environment with sudden changes.

Dillon *et al.* (2000), on the other hand, describe a simpler four dimensions required to achieve presence:

1. physical space

[2] Arousal is the term used in the literature to refer to involvement/engagement (Loudon & Deininger, 2016).

[3] Ecological validity: the degree to which a recreated environment emulates the real environment (Deniaud, 2015).

2. engagement
3. naturalness
4. lack of negative effects

These echo the work of Kassab *et al*. (2011) and Deniaud (2015), highlighting the importance of content, attention and ecological validity in creating simulated environments. One factor that cannot be used as a positive measure in a virtual environment is positive performance of the participant. This is because performance prior to entering the virtual environment must be measured relative to performance in a virtual environment. For example, if a participant could not play tennis prior to entering a simulated environment, then they should still not be able to play once in the simulated environment (Mastre, 2006; Slater, 1992).

Witmer & Singer (1998) define presence as *"a psychological, perceptual and cognitive consequence of immersion and involvement".* There are benefits to conflicting scholarly definitions of immersion and presence; each time a different application is explored in the context of virtual environments, a new variation of the term appears, thus offering a broader scope. The OmniPres report (2004) notes that this can *"provide valuable complementary perspectives and converging evidence, thus collectively overcoming weaknesses that any single measure will invariably have".* Presence can also be defined in social and physical terms, with social presence referring to the sense of 'being with others' in a virtual or simulated environment—this can relate to avatars or actors present within that environment. For example, Murray *et al*. (2015) conducted a study on female rowers aimed at understanding how social interaction changed performance. It was concluded that rowing training in a virtual reality environment improved the distance achieved vs a non-virtual environment. However, the added dimension of rowing against an avatar increased both performance and heartrate, highlighting the significance of social presence. Physical presence refers to the simulated space. The aim in this case is that participants should feel as if they have visited the simulated place as opposed to having seen some images or objects that hint at that place (Witmer & Singer, 1998). Deniaud *et al*. (2015) go one step further and refer to presence as a 'perceptual illusion'. They make reference to spatial presence as 'being there', a phrase originally coined by Steuer in 1992, when working on how to establish the construct and experience of an environment (Slater *et al*., 2013). Deniaud *et al*. (2015) also refer to presence as a manifestation of actions and location within an artificial environment, with the most widely experienced sense of presence being dreaming. Kneebone (2010) used physical pull-up posters with images of medical equipment on them to suggest actual physical equipment located in his simulated medical theatre (figure 23).

This is an example of heightening physical presence and, according to Slater's (2009) theory, if participants in Kneebone's (2012) environment refer to 'being in a theatre' then presence would have been achieved. Likewise, Kneebone's (2010) theory of how to select object representation links to Deniaud *et al*.'s (2015) work on attention, and how a hierarchy of

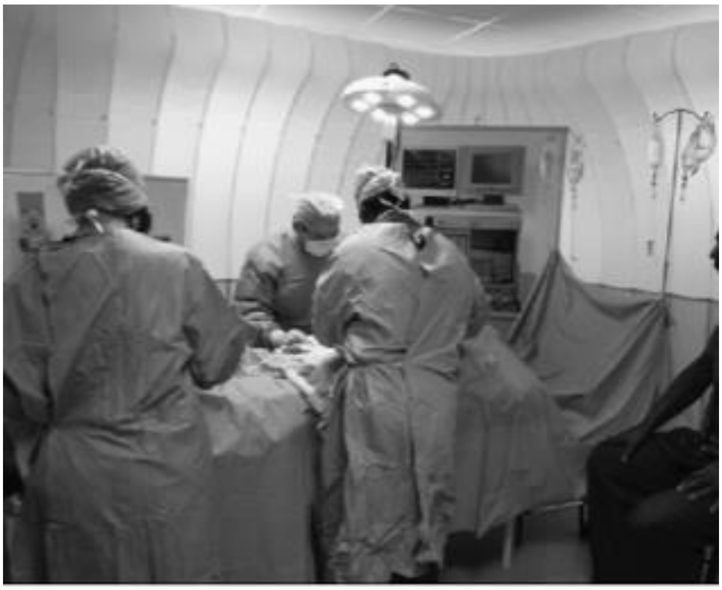


Figure 23 Simulated Operating Theatre. (Kneebone 2012)

attention can be formed to inform the design of a virtual or simulated environment that induces presence.

Slater (2003) argues that immersion should be defined as the technological content in a virtual environment, while presence should be thought of as human behaviour because of immersion. Mestre (2006) concluded that, to achieve immersion, one must replace real-world attributes with an equivalent compatible with the virtual environment. Later work conducted by Edinburgh Napier's PEACH research group supports this approach and notes that presence is the target, while the factor to be measured, albeit subjectively, is immersion. Slater (2013) is clear: how well a participant performs in an environment does not demonstrate positive presence. He illustrates the point with music and ball examples: If a music concert is simulated and a participant does not like the music being played, or a participant is asked to catch a ball and does not, this cannot be a measure of how well presence is achieved, as they are based respectively on an individual's taste and ability. However, if these personal traits or abilities were ascertained before a participant was placed in a virtual environment and the participant's performance was relative to their usual performance or likes and dislikes, then an argument could be made that this could be a measure of presence as well as a measure of the relative or absolute ecological validity of a virtual environment (Deniaud *et al.,* 2015). Validity is discussed in more detail in Section 2.5.4. Deniaud *et al.* (2015) use the Measurement-Effects-Condition (MEC) model to explain their approach to presence in the context of driving simulators.

> *"MEC model considers spatial presence, attention allocation and the construction of a spatial situation."*

The principle is derived from the knowledge of how humans experience spatial situation modelling and uses this capability to interpret the construction of spatial norms. For example, strategically placed lines could represent a room, potentially achieving presence. The concept of special cues and Deniaud's (2015) 'personal spatial memories and cognitions theory' can also be seen in other works. For example, to enhance creativity during the product design process (Keller & Stappers, 2001).

The recreation of an environment should, as far as is feasible, be simple. Boorstin (1995) notes that the person who is immersed is focusing on the task at hand and not on the virtual environment. This does not mean, however, that the user should be allowed to concentrate solely on interacting with the product. On the contrary, it is important to recreate the anxieties and natural distractions that would occur in real life (Thimbleby, 2013).

As early as 1997, Ballard *et al*. (1997) recognized that virtual environments could be developed to study behaviour. They alluded to the development of artefacts in virtual environments whereby tasks (cognition) could be controlled to study behaviour and later developed studies by Hayhoe & Ballard (2005) that studies behaviour via eye tracking. Although this was an early indication of the value of placing artefacts in virtual environments, its use at that time was limited to cognitive therapies, gaming, the leisure industry and training purposes.

Gray (2002) recognized the need to focus research on a simulated task environment that would enable researchers to develop inclusive environments that enabled them to focus on the study of behaviour in controlled environments to allow valid data to be collected. Gray cites the work of Ballard *et al*. (1997) that *"behaviour emerges from the constraints and opportunities provided by the task, the particular artefact designed to accompany the task, and embodied cognition*". This can only happen when results of research are the focus and not the environment. Gray: *"Simulated task environments enable laboratory research."*

Gray's work seeks to define the difference in simulated task environments using six categories:

1. high-fidelity simulation of complex systems
2. high-fidelity simulation of simple systems
3. scaled worlds
4. synthetic environments
5. microworlds
6. laboratory and simulated environments

Gray (2002).

Parallels can be drawn between Gray's definition of a synthetic environment and Kneebone's work on recreating a surgical environment, although all scenarios highlighted by Gray depict full virtual immersion, with less focus on the sense of presence that can be achieved as a consequence of the environment.

For the purposes of this research, immersion refers to psychological consequence of content, where real-world cues are replaced with a combination of virtual and physical cues in a simulated environment and should be subject to ecological validity. The definition of presence used is from Witmer & Singer (1998): *"a psychological, perceptual and cognitive consequence of immersion and involvement",* noting that the aim is for participants to express the sense of being there, being in, as noted by both Deniaud *et al.* (2015) and Kneebone (2012). Ecological validity will be measured to ascertain the level of induced presence.

2.4.2 Recreating Environments / Where can Presence be Achieved?

Donaldson (2008) concluded that product design usability testing needs to take place in its intended context of use and that the 'Parachute Design' effect is not enough to truly develop proper insight of how a product will be used in its real-world context. The term 'Parachute Design' refers to designing from a remote location with occasional visits to the context of intended use. Recreating environments is acknowledged by Tornros (1998) and referenced by Deniaud (2015) as a need for creating research environment spaces where conditions can be controlled and where a research question should accompany the simulated environment. When researching behaviour, be it in gaming behaviour or training spaces, with the accessibility of virtual design and CAVES, there is a need for clarity of understanding the human needs in all these newly designed simulated environments. Deniaud (2015) identified simulated environment development as being in response to the rising cost of field studies and the uncontrollable nature of them as a research space and goes as far as to note that *"from a methodological point of view there is no 'bad' or 'good' simulator"* (Deniaud, 2015, p2), because it is the behaviour output that is being scrutinized and not the environment design. Dahl (2010), on the other hand, argues the need to understand what training simulations and other virtual disciplines have to offer the usability testing space. His table below illustrates the differences:

	Usability assessment	**Training simulator**
Objective	Evaluate product performance	Enhance human skill performance
Knowledge recipient	Product evaluators	Simulation participants (Trainees)
Role of technology	Product to be evaluated	Training device and/or part of the simulator
Role of participant	Representative users	Trainee
Output	Product usability	Skill

Dahl (2010)

To further understand the simulation environment landscape, this next section explores research conducted on recreating environments for research purposes.

The lack of research in the domestic and social aspects of technology-based product development has led to the developments of the Living Lab research philosophy, developed by MIT in 2000 (Graczyk, 2015). The Living Lab is predominantly recognized as a vehicle to ascertain insights into human behaviour and how technology integrates into a domestic setting by integrating the context of use into design research. Living Labs have been described as *"an environment in which new solutions are evaluated or validated with all relevant stakeholders to create innovation"* (Graczyk, 2015).

Living Labs are also intended to play a part in co-design or participatory design early in the design process (Sanders & Stappers, 2008; Dell'Era & Landoni, 2014). Ley *et al*. (2014) describe the Living Lab as providing a specific research infrastructure that can be a mock-up of a living space or an actual living space as well as a means of obtaining real-context findings that can inform product development. Although this philosophy appears user centric, Ley *et al*. (2014) confess that the main driver in developing the Living Lab was actually technology, and that research methods were deployed according to the needs of the technology drivers rather than those of the participants (Keyson *et al*., 2017). However, by the very nature of their design, a Living Lab has become a strong user-centric tool. For example, the work of Agder University in eHealth and their use of a commercial and research-based Living Lab to develop computer embedded medical devices for home use. Agder's UCD approach (UCD is their choice of title) starts with scenario setting and includes tests in a Living Lab and testing at home. Agder define the Living Lab using a set of key principles that drive their UCD process: plan, do, study, act. Consequently, Agder are able to take concepts through to clinical verification for implementation (Smaradottir *et al*., 2015). The renewed focus on a user-centred approach rather than exclusively a technological approach of Living Labs is confirmed by the use of low-fidelity models in the environment as part of the iterative design process—a practice more closely aligned to a user-centred design or participatory design approach (Dell'Era & Landoni 2014). One example of a UCD approach is Lab4Living (Lab4Living, 2007), which was developed at Sheffield Hallam University, and is an environment used to trial design principles as live experiments with a user-centred approach. At the heart of Lab4Living's research is a drive to explore the health and social care agenda via developing future thinking environments for the elderly generation (Chamberlain & Yoxall, 2012). The concept was originated for product innovation with the intention of conducting studies that are repeatable, observable and able to extend over a long period of time to allow for the full validation of findings (Keyson *et al*., 2017; Ley *et al*., 2014). One area for concern with regards to the length of time a participant invests in a Living Lab study has been how motivated the participants can remain. Ley *et al*. (2014) used studies based on diary keeping but found that participants became less inclined to engage the longer the research went on.

Recreating space is not a new phenomenon, and neither is the concept of 'mixed reality' spaces, whereby actual physical objects and virtual content are used together to immerse participants and create a sense of presence. As early as 2002, Grey *et al*. (2002) conducted some scoping research on a 'simulated task environment' space highlighting the use of FlatWorld, a mixed-reality training environment developed by the University of Southern California (USC) between 2001 and 2005. FlatWorld was developed in response to the twin issues of expensive CAVE virtual environments and the limitations of head-mounted displays, with their lack of ecological validity and proprioception and their limited field of vision.

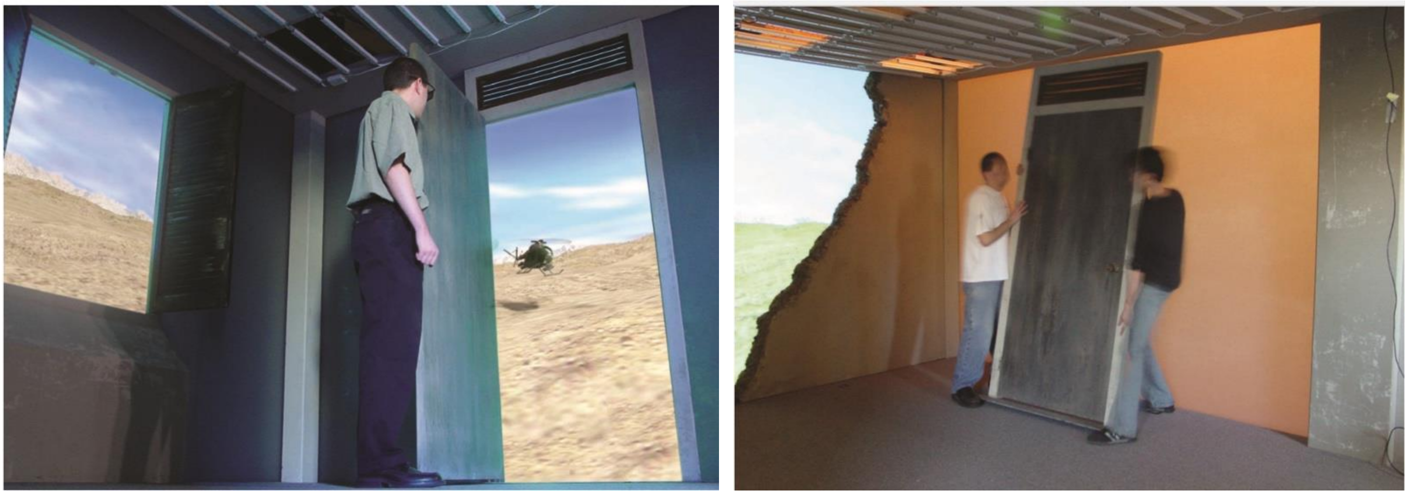


Figure 24 FlatWorld Environment and set up. (Pair et al., 2002)

FlatWorld (figure 24) was pioneering in that it capitalized on the fundamental flat digital walls used in the film industry to create environments that enabled participants to enter the space (Pair *et al*., 2003). Using an interdisciplinary team of filmmakers, the Army, and academics and using a human-centred approach, they developed multiple environments that supported health, wellbeing, education, reflex and decision-making, (Pair *et al*., 2003). The technology was a combination of augmented and virtual environments, with a focus on virtual rather than mixed reality. It was one of the first centres to explore this field. FlatWorld is described as a ‘task environment’ as it was designed for training purposes, with virtual reality used to heighten presence. This practice can also be seen in the medical profession (figure 25).

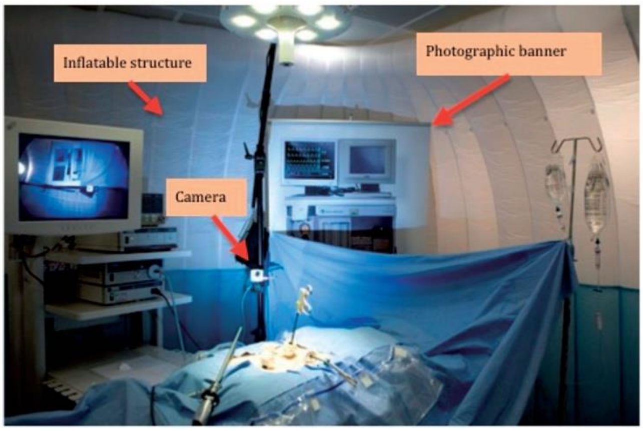


Figure 25 Simulated Surgical Training environment.   (Kneebone 2011)

Professor Roger Kneebone recognized the need to recreate the surgical theatre to improve the training of newly qualified surgeons. Rather than using virtual reality, he focused on simple visual cues to heighten presence, so his students felt like 'being there' (Kneebone, 2010). The research focused on where an individual's attention is required when in a medical theatre and, following these conclusions, a series of 'pop-up' printouts were created to offer visual cues. Kneebone went on to utilize presence-measuring methods to capture the results of the recreated environments. The research explored 'face validity'—an established measure in psychology which evaluates the extent to which an environment captures the intended construct (Reimer, 2006). If we align face validity with presence, then the question to be asked is, it is to what extent does the recreated environment concur with the intended real environment? Therefore evidencing ecological validity and achieving the sense of being in the space (Slater 2004). Kneebone and his team later developed a portable inflatable simulation that included the benefits of recreating presence but did not rely on costly and specialist virtual reality or technology-based simulators, focusing instead on research to establish what Burki *et al*. (2015) called ‘good enough’ when recreating an environment (Kneebone, 2011; Sadideen, 2012).

Kneebone and team used a 6-point Likert scale to ascertain the subjective nature of face validity comparing their simulated environment to the training surgical box used as industry standard in a normal laboratory. They found that the simulated environment had higher face validity. Kneebone’s work is also echoed in the work of Dahl *et al*. (2009) who conducted research assessing the correct fidelity of a simulated environment for full-scale laboratory simulations, closely aligned to IDEOs Experience Prototyping. He too concluded that a principle of ‘just enough’ should be applied to simulations.

Mastre (2006) and Slater (1992) only refer to an individual’s improved performance indicators in recreated environments based on prior benchmarking (Deniaud 2015). In this way, relative validity can be used to deduce whether the environment was a positive factor on the trainees’ output. Interestingly, they noted that trainees perform better in the simple laboratory setting, something Kneebone concluded was because it did not replicate the realism and stress of an actual surgical theatre and was therefore not helpful when preparing trainees for surgery.

Virtual reality has also been utilized in the design process, for example, by Carulli *et al*. (2013), Parks (2008) and Verlinden *et al*. (2007). Carulli *et al*. developed a system that would harness the ‘Voice Of Customer’ (VOC) early in the design process. They disposed of any physical prototypes, instead relying exclusively on virtual prototypes and a haptic device. The purpose was to integrate the customer voice early into the design process with an emphasis on facilitating personalization and customization via *“multimodal Virtual Prototyping environments for the capturing of the VOC regarding design solutions of products in development”.* The ‘customers’ (their choice of description for the intended end user) would be invited to wear a VR headset and the setup would enable haptic feedback by allowing the participants to touch a physical object that was manipulated into shape using hydraulics, allowing its shape and aesthetics to change (figure 26).

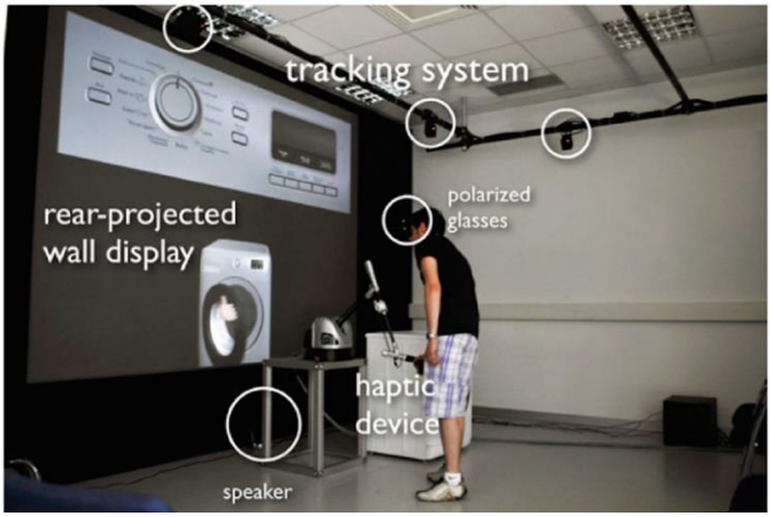


Figure 26 Multimodal virtual Environment Carulli et al., (2013)

This allowed 'customers' to be involved in the design process at an early stage while only using one physical prototype, thus reducing cost. Carulli *et al*. acknowledge the limitations of this approach, one being that it only allows for customization on a small scale. Another limitation could be the way participants are focused primarily on aesthetics, and the 'overarching' feel of a product rather than being able to test specifics in the design. Generally, there is a lack of research into the key principles required for successful testing of virtual prototypes in the product design development process.

The term 'augmented reality' is now widely recognized, primarily due to its accessibility via smart devices. First developed by NASA in the 1990s, it is the most accessible virtual technology available as it is not predicated on the need for headsets or large format screens. While virtual reality is predicated on immersing an individual in a digital space, augmented reality *"is taking digital or computer generated information, whether it be images, audio,* *video, and touch or haptic sensation and overlaying them over in a real-time environment"* (Kipper and Rampolla, 2013, p1). Moggridge (2006) describes augmented reality as requiring the use of a handheld computer and a camera, and interaction with the real world. For example, in gaming, participants are able to navigate actual terrain whilst accessing virtual content overlays, with the best-known example being Pokémon Go.

The growing trend in mixed interaction or reality systems has seen Couture *et al*. (2010) develop the term 'augmented virtuality' which they describe as *"embedding some physical information in a virtual world"*. Also known as tangible user interfaces, this involves augmentation of the world as we know it, by aligning information from the digital world to physical objects and environments (Ishii *et al*., 1997). This research was founded by computer scientists but has been co-opted for product development purposes. Its environments are usually small scale and focus on the manipulation of 3D data and have proved capable of integrating physical artefacts in digital space (figure 27). The digital content is displayed on screens or using small-scale table projectors, projecting CAD models and integrating physical buttons in the space.

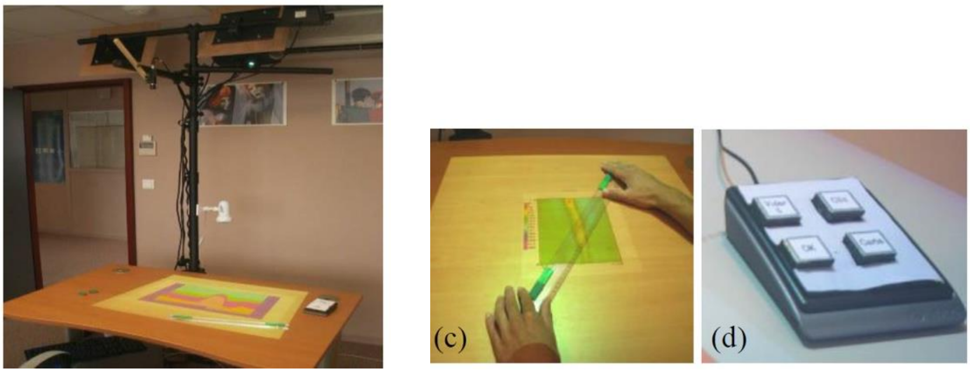


Figure 27 Virtual and physical objects found in Couture et al (2010) Augmented Virtuality Environment

The *'Tangible interaction in Mixed Reality Systems'* paper by Couture *et al*. (2010) describes work that is very similar to Carulli *et al*.'s (2013) VOC work, but with a better understanding of computer science. While the scale of Couture's work is smaller, it encourages designers to evaluate the design in terms of their needs—the heuristic approach (Barnum, 2011)—rather than the needs of the intended users. It is clear from the literature that when authors use the term 'virtual', they are almost exclusively referring to Computer Aided Design (CAD) models tested in isolation. Sometimes they are used for design review purposes (Aromaa *et al*., 2012; Aromaa & Väänänen, 2016), other times they are used to evaluate user need, but context of use is not considered. Instead, the focus tends to be on accessibility of a design idea via CAD, on a computer screen and rapid prototyping. Usability is often considered, but again they tend to be evaluated exclusively within the design team. Aromaa & Väänänen's later work (2016) does include the environment in the testing phase of their design process. They evaluated an augmented reality setup using tablets to augment detail over a physical object, and used a virtual reality setup by utilizing head-mounted displays. Their main aim was to evaluate ergonomics and they called it Virtual Prototyping (VP). In much the same way, Zhou *et al*. (2016) explored ergonomics in terms of 'maintainability' of a design. Environments are created in CAD, for example, heavy machinery or an oil rig platform. A participant wearing a head-mounted display evaluates the ergonomics of the designs in VR. The results showed that the augmented setup included less of the environment and did not yield the same value as those in the virtual environment: it was the context that encouraged feedback and observations, leading to product enhancements. Unfortunately, the results were deemed subjective as no ecological validity of the testing space was conducted, meaning it was not known how participants' behaviour in VR related to their real-world behaviour.

Early evidence of recreating 'context of use' during the design process can be found in the work by Keller & Stappers (2001). Their work utilized video collages to gain insights and support their Inspiration and Ideation phases of development, but also as a tool to immerse the designer in the inspiration gathered during the research phase of the design process. Also, Keller & Stappers (2001) tapped into the recreation of the ‘context of use’ via video collages as a tool to gain an understanding of the product’s requirements to inspire and inform the design process. They describe their process thusly

> *“Product designers can use these video collages to re-experience their observations in the environment in which a product is to be used, and to communicate this atmosphere to their colleagues and clients. For user-centred design, video collages can also provide an environmental context for concept testing with prospective user groups.”*
>
> Keller & Stappers (2001).

Creating multimodal environments combining virtual and physical objects has been used in cognitive therapies for some time, with the early work of Walshe (2005) exemplifying this. His exploration of therapies concerning victims of road collision resulted in exposure therapy. Walshe created an immersive environment that consisted of a projected image of a moving road viewed through a windscreen. Meanwhile, a chair was placed on a platform housing a subwoofer to create the sense of road vibration. Walshe used this setup to create immersion and presence. The results were measured during the experience via verbal feedback and the use of a heartrate monitor. Results demonstrated that presence was achieved, and improvements were found in patients, therefore the environment was a success—the impact of immersion was not noted in Walshe’s work. In essence, Walshe used Kneebone’s approach in his work; he ascertained ‘what was enough’ to support patients into getting back behind the wheel of a car by establishing where attention would be captured. For example, viewing through a windscreen would make the patient feel as if they are in a car. This echoes the narrative in Dahl’s (2010) work, concerning what influences decision-making.

In the main, simulated environments are used as research tools to evaluate performance. Nilsson (1993) notes that simulated environments must be valid if they are to be used as research tools. To do so effectively requires in-depth understanding of the impact context has so that the environmental conditions can be appropriately set up and controlled. Gray (2002) developed three dimensions of a 'simulated task environment' that need consideration from the perspectives of the researcher, the task, and the participant.

2.4.3 Ecological Validity of Simulated Environments

A simulated environment should impact a participant so as to change their emotions and/or behaviour in alignment with expected real-world behaviour and/or emotion (Deniaud *et al.* 2015). This behaviour alignment is known as 'ecological validity' and measuring ecological validity in a virtual environment research laboratory is paramount (Deniaud *et al*., 2015).

Ecological validity is an important area, too often overlooked. What literature there is focuses predominantly on simulated automotive environments and has flaws (Godley *et al.*, *2002)*. For example, the conditions under test have not been validated in a real environment. An example in the literature is highlighted by Godley *et al.* (2012) whereby a study was conducted on road rumble strips. The study involved establishing whether a speed reduction achieved in a simulator would reflect the real-world scenario. Having completed the study and contrasted with the real environment they found that speed had been validated for simulator use and that each area of research required this level of validation prior to engagement in simulated research, i.e. seeking ecological validity. Deniaud *et al.* (2015) note that *"the literature highlights some disadvantages, including simulator sickness, the accurate replication of physical sensations, and most importantly, validity".* Typical validation has included using pre-existing driving events, for example, an event could be overtaking a car, to validate against performance in a virtual environment (Reimer *et al.* 2006). This is problematic, because the complexity in the real world means there is less control on the condition. Reimer *et al*. (2006) conducted research on self-reporting when assessing the impact medication has on the driving ability of people with long-term chronic disabilities. To collect this data would have meant putting participants at risk when driving a vehicle in a real-world scenario. Therefore, a simulator was a safe environment to test the impact, but in order to explore how to ascertain validity in the results, the team explored self-reporting via questionnaires as a viable option. This again has flaws, since it is predicated on memory, and any false recollection could potentially skew the data, although the research reported a significant relationship between data collected by questionnaires and the behaviour of the participants in the simulator.

Blaauw (1982) classified ecological validity in two ways: absolute and relative validity. His work has gone on to be cited by Goodley *et al.* (2002); Reimer *et al*. (2006); and, of late, Deniaud *et al*. (2015). Absolute validity would yield the exact results as found in a real environment, something very difficult to achieve in reality, and more of a goal. Relative validity concerns the direction of travel; are both environments displaying results that point to the same conclusions? If so, it would be classified as a strong positive indication whilst not being conclusive.

Face or physical validity are the most significant components when ascertaining relative participant behaviour in a simulated environment (Reimer *et al.*, 2006). Face validity refers to the comparison of a recreated setting in line with its intended construct, and in the context of this thesis, it relates to ecological validity. The crucial question is whether the simulated environment elicits the same emotion and behaviours as its real-world counterpart. Kassab *et al*. (2011) conducted participant interviews to ascertain the content and face validity of a portable simulated environment, with the eventual goal of ascertaining the degree of presence. The results indicated a high degree of presence and validated a positive approach to training new medics in a safe simulated environment.

2.5 Human Behaviour / Hacking the Human Brain

Human behaviour, or hacking the human brain, is an expression used to describe the relationship of human behaviour in computer science and was coined by Napier's PEACH research group (Benyon, Smyth & Helgason, 2009) and captures the nature of the research involved in creating presence in a range of simulated environments. First, to understand how the human brain interprets simulations, we must understand 'reality'.

> *"What is becoming clearer to researchers is that what we call reality is simply a mental construct, an educated guess arising from the exchange of information between our minds and the environment and mediated by our bodies."*
>
> Benyon *et al.* (2009).

> *"The maxim states that seeing is believing, but that it is touch that determines reality."*
>
> Benyon *et al.* (2009).

When usability testing a product there are certain human factors that need to be considered. The most prominent of these is mental load, as it replicates what happens in everyday life. Mental load is recognized as having three components: cognitive, visual and motor (Weinschenk, 2011 p65). They are called upon in different balances depending on the task and scenario an individual is in. Visual is the most intense sensory load, cognitive load involves thinking and problem solving, while motor load refers to a physical action like pushing a button (Weinschenk, 2011). The more loads an individual is under, the greater the need to consider trade-offs in how they conduct themselves in a given task. These human factors need to be considered in a testing environment, as every external factor that contributes to mental load impacts on how a user interacts with a product.

The ability to study how the brain functions has highlighted two networks that deal with tasks. The first is the task-positive network, whereby our brain is consciously involved in a focused task. The second is a task-negative network, whereby our brain is wandering from one thing to the next with no real task at hand (Evans, 2017). This is important because in life, humans have the intention of being focused on a task, but will often find themselves distracted, as recognized by Thimbleby (2013). To better understand how humans receive information, we need to look at the human visual system, because eye movements during a task are shaped by experiences and attention, so perceived irrelevant objects are ignored so a task can be executed properly. With this in mind, humans look in advance at the objects before handling the object, or even before the object is relevant; therefore, the eye is working in anticipation of the next task (Hayhoe & Ballard, 2005). *"Thus, eye movement patterns appear to be shaped by learnt internal models of the dynamic properties of the world."* Ware (2008, p13).

The visual system is premised on attention. Ware (2008) notes that this top-down visual messaging system is dictated by required cognitive tasks.

Object → Patterns → features → Retinal image Ware (2004).

In this context attention is how presence is measured. As noted in Section 2.5.1, involvement and engagement help shape the visual cues and participation in a simulated environment. This is relevant due to the nature of a usability test and where the attention is directed by the design of the product, prototype medium and user tasks. The eye has two types of vision: vision at its centre and vision at its periphery. Contrary to popular perception, it is peripheral vision that we actually primarily use to interpret the environment around us (Weinschenk, 2011) and so has a key contribution to make within both real and simulated environments. Much of the traditional laboratory-based practices of the usability testing of products focuses the eye with a task but does not account for peripheral vision's role in shaping the context of the product interaction.

The eye can interpret part images via 'visual memory making' (Ware, 2008, p11; Ware, 2004). In such a scenario, the visual cortex must work harder to imagine and reinterpret the rest of the image (Solso, 2005). Knowing this, we can start to correlate the way the visual system works with Kneebone, Sevdalis & Nestel (2010) and Burki *et al.*'s (2005) theories found in the literature on 'what is good enough' when recreating simulated environments. This involves acuity and visual field: both eyes have a 180-degree visual field, but our peripheral vision works better with motion. Hence why our static field of view is 180 degrees but motion sensitivity in the peripheral view can allow peripheral views of up to 270 degrees (Ware, 2004; Evans, 2017). Acuity declines rapidly beyond the fovea and Ware notes it is *"one-tenth the detail at 10 degrees from the fovea"* (Ware, 2004, p50).

Understanding the eye and how we interpret the world around us when under stress is also of value. Ware (2004) notes that when under extreme stress, the field of view becomes restricted, a phenomenon commonly known as tunnel vision. In 1985, Williams (1985) established that the processing of information from the peripheral field of view was not occurring under stress, resulting in this 'tunnel' of view. Further research found that under short-term stress the heartrate increases as the body enters the fight-or-flight mode (Jensen *et al*., 1995) and this impacts on a person, and in some cases their ability to tap into their cognitive mental mode effectively.

One way to optimize presence is by tapping into human factors and utilizing an immersive space to ensure presence does not have to be predicated on a complex environment. Early work of Keller & Stappers (2001) and Reeves & Nass (2002) have already demonstrated that presence can be achieved in low-tech environments because the brain's creativity can induce it with a minima of external stimuli required (Benyon *et al*., 2009). Reeves & Nass conclude that very basic illustrations or representations of an intended environment will suffice and cause a real emotional response from users, while Keller & Stappers took inspiration from David Hockney's use of partial and interrupted views to generate his photographic collages and created environments. Ware (2004) explains that *"when data is presented in certain ways, the patterns can be readily perceived".*

According to Boorstin (1995), when setting a scene, the use of olfactory, lighting, audio and narrative are factors that should be considered, although smell can have a short-lived impact as we adjust and adapt to it over time, reducing its impact (HIT lab, 1997; Leffingwell, 2002; and Powers, 2004). Previous work conducted by Ramic *et al*. (2007), Brikic *et al*. (2009), Chalmers *et al*. (2009a) and Brikic *et al*. (2013) confirm this, concluding that smell is not a significant contributing factor in attracting humans' attention in VR.

2.6 Conclusion

Referring back to objective 2 ('*to explore methods of recreating environments and how they can influence the requirements of the usability testing space'),* the answer is somewhat clearer from the literature review. The fidelity of the environment should be 'good enough' and should match the fidelity of the product at the stage of the design process where testing takes place. What this means is that, as more functionality is developed, the prototype fidelity should reflect that—and so should the fidelity of the environment. The best time and methods to test a product is an interesting question. Nielsen, Normal, Hare, Barnum and Woolley all cite that early testing is critical to the success of a new product development and that a combination of field and laboratory usability testing is required, since each have their limitations. Kjeldskov *et al*. (2014) on the other hand, note that the question that should be asked is not when in the design process should a product be user tested—because it should be ‘as early as possible’—but *how* should those tests be conducted. The literature on virtual reality and recreating and simulating environments demonstrates that there is an opportunity to develop the context of use in the usability testing of products early in the HCD process and that, correctly conceived, could become a ‘third space’ in the usability testing debate. Disciplines outside of design have been integrating VR or the recreation of context for some time and the lessons learnt from fields including, medical training could be transferred to the product design usability testing process. For example, the following key lessons are from a combination of the work of Deniaud *et al*. (2015) and Dillon *et al.* (2000):

- Ecological validity and relative validity are critically important. Known as ‘naturalness’, it impacts on the believability and concerns the psychological response as a consequence of content, be it virtual or physical, in a recreated environment.
- Understating where attention needs to be will ensure a simulation is ‘good enough’. This relates to purpose and engagement in a recreated environment and contributes towards presence. Smell is not as important a contributing factor as originally assumed.
- Using participants’ peripheral vision appropriately is an important component of helping create important context and/or appropriate distractions during a usability test.
- Physical space relates to the content and how to distinguish what content should be in an environment and knowing what should be physical and what should be virtual is key.
- Consideration of the negative effects or anxiety associated with recreating environments and acknowledging how to mitigate the negative effects in the design of any environment.

As noted in Section 2.5.1, for the purposes of this research, immersion will refer to the psychological consequence of content, where real-world cues are replaced with a combination of virtual and/or physical cues in a simulated environment and should be subjected to ecological validity. Presence is defined as the psychological effect/response/consequence of the content and task completed in the simulated environment; it should elicit the psychological feeling of being there / being in. This definition has been taken from a combination of authors including Deniaud *et al.* (2015).

Literature that combines design and the virtual space is still underdeveloped and lacks a coherent approach that integrates HCD with psychology and computer science to understand how to optimize environments relative to the needs of the task at hand. It follows that current design practice incorporating VR testing is limited. Virtual reality can be found in the design literature, but in reference to virtual prototypes, with an emphasis on haptic feedback, e.g. Carulli *et al*. (2011), or the replacing of physical objects with virtual ones. Missing from this virtual reality literature is an understanding of the importance of ecological validity and the means of measuring subjective behaviour in a virtual environment to measure presence. The social and physical aspects of usability testing are key human factors that need to be considered along with ergonomics and a full consideration of natural behaviour in everyday experiences. In addition, consideration must be given to the impact peripheral vision has on a participant's judgment call, premised on the surrounding environment and consequently any given task. Understanding how this impacts on the ways we interact with the products, systems and services around us is critical; and so to test solely in isolation is to miss an opportunity to optimize the user experience early in the HCD process.

It is evident the fundamentals of a usability test are valid and still apply: the use of a moderator, collecting data and designing an appropriate study with an appropriate number of participants to ascertain design faults are all important. However, there is work to be done to ascertain how to incorporate context-specific scenarios into the testing environment early in the design process. This research therefore focuses on running a series of studies that compare environments of different fidelities to identify the optimum fidelity usability context testing environment, aiming to recognize what is 'enough' in terms of usability scenarios. In order to validate these findings, ecological validity, relative to a usability design study, will be ascertained so optimum presence can be achieved, with the right amount of input so as not to overburden the HCD process and negatively impact on the time it takes to take a product to market. These results will then underpin a final study that aims to validate that recreating context early in the HCD process can identify meaningful usability issues that are identified in the real-world context.

Having reviewed the literature and responded to both objectives 1 and 2, to ascertain where in the HCD process different types of usability test should be conducted, it is evident there are synergies with the fidelity of the testing method and the fidelity and refinement of the product as it develops. The literature review has highlighted a need for early usability testing environments that can include the benefits of field studies, with simulated environments being a potential candidate to meet these needs. However, these simulated environments must have 'enough' visual cues to create a sense of presence and impact on a participant's behaviour.

Chapter 3 Requirements Gathering

Chapter 3 Requirements Gathering

This chapter deals with primary research undertaken to provide further insights to complement the literature review and offer further insight to objectives 1 and 2. This primary research also plays a role in informing the design requirements of the Perceptual Experience Lab (PEL). The following research exercises are covered:

- A two-day simulation laboratory 'Train the Trainers' session at the University Hospital of Wales Cochrane Simulation Laboratory to develop insights into the nuances of designing and using a simulated task environment that might be used to inform the design and setup of the Perceptual Experience Lab (PEL) for usability testing.
- A visit to Welsh National Opera (The Wales Millennium Centre), hosted by its Director, to understand how immersion and presence are created in a theatre setting, in order to further inform the design development of the PEL.
- 20 days in the design department of Frontier Medical, a design and manufacturing company, to develop direct understanding of contemporary human-centred design and usability testing practice in industry.
- Semi-structured interviews with designers from three different design consultancies, Kinnier Dufort, DCA and PDR, to understand their needs in terms of usability testing.

3.1 Cochrane Simulation Laboratory

As part of the contextual review, the author attended a 'Train the Trainers' event at the Cochrane Simulation Laboratory (CSL), University Hospital of Wales (UHW) (figure 28).

The CSL is a purpose-built simulated suite designed for medical training. Key features include:

- visual references to the real intended environment
- state-of-the-art simulation manikins that can respond to standard medical equipment and the actions of trainees

- setting the scene of believability by including actors to induce participant focus and emotional response to rehearse stress management

The CSL is a permanent fixture UHW rather than being portable. It is linked to computers via a two-way screen and uses actual equipment rather than banners to hint or suggest at a hospital ward (as found in Kneebone's work). In other ways however, the design of the CSL utilizes principles similar to Kneebone's work. For example, both are simulations not simulators. Simulators are defined as *"a machine designed to provide realistic imitation of the controls and operation of a vehicle, aircraft or other complex system, used for training purposes".* Simulations are defined as *"imitation of a situation or process, the action of pretending; deception".* Oxford Dictionary (2020).

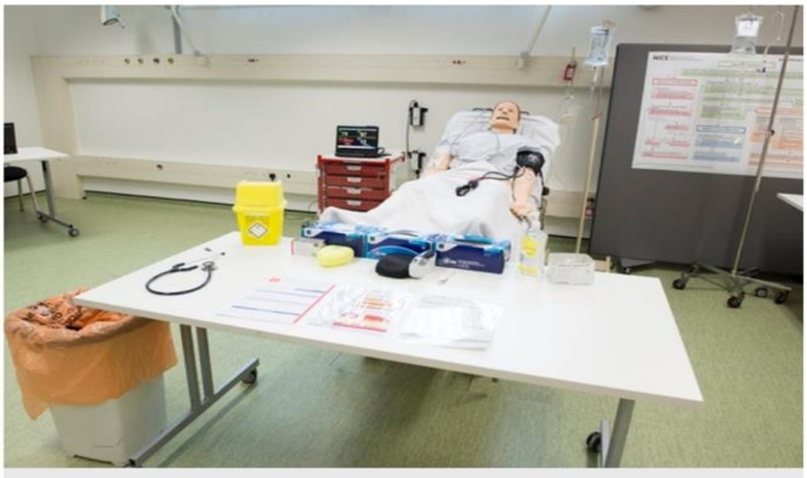


Figure 28 Cochrane Simulated Room Cardiff UW Hospital

Both Kneebone and the CSL use visual cues that require attention during the task. For example, the CSL purposely uses the same flooring found in hospitals and hospital pillows to directly reference hospital wards and Kneebone used banners with printouts of actual hospital equipment.

The aim of the course was to introduce the simulated environment to experienced doctors and train them to train others (Train the Trainers) and to explore how best to interact with the laboratory to facilitate learning. The aim of the author attending the session was to inform the design and build of the Perceptual Experience Lab (PEL). Findings included how the design of the simulation laboratory successfully utilized cues with physical objects to help replicate the real environment, so the scene is set for the participant (trainee) to be in the right frame of mind. A hierarchy of props was established, predicated on where the trainee's action would be. These included the bed, the curtains and the bedside table; everything else was neutral in colour so as not to sensory overload the trainee. The CSL trainer placed significant focus on the task and emotional responsibility of the participants in the simulated environments, so stress levels could be developed, and social interaction could be practiced. The support required to utilize the technology involved was very specialist and required technicians to service and manage the laboratory. In terms of lessons learnt, it reinforced the work in the literature in terms of recognizing where attention is required and subsequently what content is required to make a simulation believable and solicit the appropriate emotions and reactions.

3.2 Welsh National Opera

Theatre is well versed in recreating scenes, mood and emotions; therefore, the author visited the Welsh Millennium Centre (WMC) where the Welsh National Opera is resident, to understand how immersion and presence are created in a theatre setting. During the tour with the Director of the Welsh National Opera, key observations were gleaned. For example, the use of light is important, but not white light as it flattens the scene; the role of lighting is to create mood and focus the audience. Projections are from the front so as to maximize the stage space. Stages are angled to heighten audience immersion. Storage is found in the ceiling space but also in the construction of the flooring, by utilizing crates to hide cables and to create different floor levels. Height is achieved by using a scaffolding structure. Further areas for consideration are allowing time for set construction. Another issue raised during the visit was how to manage acoustics and portability if it is to be reconfigured. Although not all the findings translate directly to the design requirements of a usability laboratory, there were some insights gleaned that were particularly helpful. At the time of this visit, PEL was being constructed and observations from the visit to WMC and CSL directly informed the design of PEL. These included the use of a specialist crate modular floor to ensure all the cabling for the observational cameras and projectors could be hidden out of sight. The floor was chosen to be black, as in the WMC's stage configuration, so focus could be on the product and peripheral scene. Specialist acoustic foam was purchased to minimize sound pollution from outside the laboratory and a specialist cloth was purchased for the screen to enhance the image quality.

3.3 Case Studies from Within the Design Industry

In order to establish a deeper understanding of the current design philosophies being implemented in the commercial design industry, the author used an inductive research approach: an ethnographic research process was conducted from within Frontier Medical, a design and manufacturing company based in Wales, UK. The research was funded by a competitively won Strategic Insights Placement (SIP) grant funding for a 20-day placement. Frontier Medical have in-house design and manufacture capabilities and are situated in the southeast valleys of Wales, UK. Frontier Medical comprises three divisions, each with its own design team: sharps boxes, needle exchange and pressure pads for bedsore prevention.

Findings: The product development process used at Frontier Medical had been developed and implemented by the Product Design Manager. It was customized to the organization's needs but two of the design teams precluded end-user involvement and there was limited opportunity for implementing the 'Inspiration' phase of the design process, although they did give exceptional consideration for medical research and empathetic research. There was considerable reliance on tacit knowledge and in-house experience and expertise regarding buyers' thinking and perceived user needs. The Product Manager was instrumental in feeding the design teams new ideas but had no design process training and so tended to take a business-led approach. The three design teams had very mixed practices. The needle exchange project team was very proactive, engaging with end users very early on in the design process. As a consequence, they developed very subtle, yet profound, design features. For example, allowing each drug user to identify their needle by the scratching of a letter to minimize the spread of infection. The pressure pad design team sent questionnaires to end users, but these focused on feedback on existing products, with the questioning conducted by a salesperson rather than a design researcher conducting ethnographic research. They did conduct in-context user trials but only with an over-committed design. No early stage usability trials were conducted. The sharps box design team focused on reducing cost, differentiating markets by adding or subtracting to existing moulds. Design was conducted in isolation of the context of end-user needs, although growing competition in the market meant the team were reviewing the feedback obtained from their existing product range. Limited consideration was given to the context that the product would be operating in, the ambient sound types and levels of a hospital environment and their impact on users' ability to safely determine whether the bin had been successfully made safe after use (although safety was paramount and not having access to used needles was at the forefront of the design team's mind).

With the new insights gained from the placement, a knowledge transfer partnership (KTP) bid with Odoni Elwell Ltd (a manufacturer of multi-purpose steel buildings and cycle storage solutions) was written and granted, to explore how to strengthen HCD practices inside the company. This was followed by a second successful KTP bid with Window Cleaning Warehouse along a similar theme. The KTPs were two-year projects that allowed for implementing ideas and testing them out in an organizational setting. Subsequent interviews were conducted in three design consultancies and one design and manufacture organization. These interviews consisted of semi-structured interview questions to ensure continuity. The purpose of this work was to gain insight into current practice with regards to usability testing during the design process. Here is a snapshot of all the organizations consulted during this phase of the research (table 1).

Table 1 Snapshot profile of the organizations consulted during this work

	Odoni Elwell Ltd	Window Cleaning Warehouse Ltd	Frontier Medical Ltd	Flexicare Group Ltd	DCA	Kinnier Dufort
Involvement	KTP	KTP	Placement/ ethnography. Interviewed	Interviewed *in situ*	Interviewed	Interviewed
Company designation	Manufacturer with in-house design capability	Retailer with in-house design capability	Medical device manufacturer with in-house design capability	Medical device manufacturer with in-house design capability	Specialized design activity	Specialized design activity
Manufacturer/Service	Manufacturer	Retail, design capability and outsource manufacture	Medical device manufacturer	Medical device manufacturer	Design service	Design service
Staff employed	71	Unreported	235	111	Unreported	Unreported
Net worth	6.4m	362.98K	22.61m	6.8m	1.12m	2.74m

Endole Ltd (2016).

Table 2 Commonalities and differences of the organizations captured in table

Commonalities	Differentiations
The organizations have a design provision	Maturity of the company specific design processes, in particular HCD.
Similar size organizations: similar worth in assets approx. £14 million and a net worth of approx. £6 million.	Polar practices that enabled the author to see more differences and gain breadth of understanding.
Successful companies with a healthy turnover and a £ multimillion turnover.	Sectors: medical products, cycle storage facilities and window cleaning equipment.
Accessible locations allowed for repeat longitude visits.	Odoni Elwell is a family-run organization, Frontier Medical once was, but is no longer.
A mix of onsite and outsourced design and manufacture.	
Private Limited Companies	

Semi-structured interviews with staff from DCA, Kinnier Dufort and Flexicare made it apparent that design consultancies have embraced HCD and usability testing both as optimal ways of working but also as commercial opportunities. In contrast, the family-established design and manufacturer company mentioned above have been slower at integrating humans in the design and product testing processes. DCA and Kinnier Dufort placed people at the core of their practice, with usability testing used as stage gates to secure the next course of funding or next project from the client. DCA were also integrating virtual technologies to communicate 3D prototypes, although only for visualization purposes. Staff from DCA and Kinnier Dufort were introduced to the Perceptual Experience Lab and were able to offer feedback on its potential use and value. Dr Alex Woolley, formerly DCA, noted that when he conducted usability tests in the field, individuals behaved and felt self-conscious, particularly if the model was a low-fidelity mock-up. Consequently, their behaviour changed, skewing the user experience data. One key finding from the interview was that PEL could be used as part of the Ideation stage to immerse the design team in visual representations of the research context. In addition, PEL could be used to communicate research findings to the design team, enabling the design team to contextualize their work with the user insights in mind; as often those involved in the research are not those involved in the design work, therefore more needs to be done to join this work together. It was noted that PEL could also be used to showcase a proposed design to a client to contextualize the work and secure the next stage of the project. Lastly, they noted that PEL had a role to play in usability testing studies early in the design process with the added benefit of maintaining confidentiality while controlling a study. It was noted that often when field usability studies are conducted, participants feel self-conscious and it changes their behaviour and subsequently the usability results; therefore a space that can recreate the conditions and benefits of a field study, whilst offering privacy, would remove this element of self-consciousness. Ian Culverhouse from Kinnier Dufort noted that the moderator plays a key role in running a successful usability testing scenario. Therefore, it is important to make sure that they are able to function without distracting the participants. In addition, he highlighted the importance of always running pilot usability studies so modifications can be made to the environment, if required.

3.5 Conclusions

Design praxeology will always be an area for development as technology and design sophistication develops with time. Involving humans in the design process and in early usability testing has been found to yield positive results and testing with as few as five participants. Using Nielson's principles of 'little and often' will deliver enhancement to designs in the product development process. It is evident that design organizations need to embrace a philosophy of involving people in the design process before they can even consider the role of usability testing, but there is an opportunity to adapt the usability testing environment so organizations new to this approach will see more positive results from its inclusion. While many design practices have long recognized the significance of usability testing, there is still opportunity for enhancement. The use of simulations of an environment in a laboratory setting, which recreates the context of use specifically for the usability testing of products early in the design process, is an underdeveloped area in the literature.

Based on the insights gathered from the literature review and from first-hand insights from a range of key sources, a number of key salient points need to be considered when creating a simulated environment for usability testing. These include:

- The environment, be it technologically complex or simple image-based, should achieve a feeling of physical presence—the sense of 'being there, being in'.
- Visual cues should be used to contribute to the sense of presence.
- A participant's field of view should provide relevant distraction that enables the participant to feel as if they are 'there'.
- It is important to identify where the participant's attention is (and should be).
- Images can hint at the background. They can be simple representations that make the background images accessible but not resource-hungry.
- The prototype fidelity should match the fidelity of the environment of the product. This balance should be maintained throughout.
- It is not where in the design process, but *how* usability testing should be conducted.
- Establish what is 'good enough'. Do not over-commit to technology that adds no additional benefit.
- Be mindful of ecological validity but focus on achieving relative validity in the immersive content.
- Consider human factors, and how to create a realistic atmosphere, so participants feel as if they are in the product's intended context of use.

The secondary and primary research have allowed the following conclusions regarding the requirements for a simulated environment. They include:

- Use visual cues that relate to the intended context-of-use environment . For example, the shoes and scrubs a surgeon would wear in the real context should be worn in the simulated environment, as seen in Kneebone's work. The hospital-issue pillowcases in the Cochrane Simulation Lab are another good example.
- It is important to determine what elements of the simulated environment require recreating in a form that can be interacted with, and those elements that can be static and used as context in order to create the best sense of presence. Kneebone observed the environment they intended to recreate to ascertain what needed to be interacted with, and used during the training, and what could be static.
- Studies must be replicable and reliable. It follows that the space needs to be controlled and confidentiality maintained.

Chapter 4 Methodology

Chapter 4 Methodology

4.1 Introduction

It was noted in the contextual review that, in current practice, the social and physical attributes of a product are often ignored during the usability testing process, especially early in the HCD process, during the Ideation phase. It also found that laboratory-based usability testing alone does not offer enough insight, and that field testing has a role. Two issues with the field testing of prototypes are the difficulties in controlling the environment and a jarring of prototype fidelity—i.e. a low-fidelity prototype in a real environment can hinder user behaviour (Behember, 2011; Dorner, 1993; Woolley *et al.*, 2011). It was also found that knowing what is 'good enough' to recreate an environment is critical. The combination of usability testing in a recreated context, early in the HCD process, is a novel approach. However, to address objectives 3 and 4 of this research (*to ascertain the fidelity of a simulated environment that is most effective in discovering product design flaws early in the design process; and to validate the findings and detail the outcome of this research exploration*), a methodological approach needs to be designed and implemented. As a response to objectives 3 and 4, four studies were designed (figure 29).

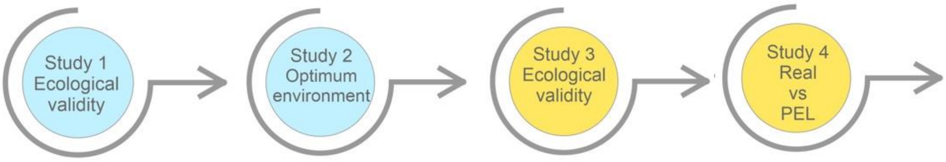


Figure 29 Four research studies

Studies 1 and 3 were born from the computer science literature that noted that any simulated environment that is used for research purposes, i.e. the collection of data, needs to be ecologically validated prior to conducting the studies. In these studies, relative validity was ascertained to establish the level of immersion and consequently naturalness, as noted by

Dillon *et al.* (2000) and Deniaud (2015) in their principles of recreating environments. PEL was developed and relocated in between studies two and four, therefore there was a need to conduct two validation studies (studies 1 and 3). Only having ecologically validated the environments could studies 2 and 4 be conducted. The data collected in studies 1 and 3 were not only used for validation, but also as a means of asserting if infrastructure improvements in PEL were adding value.

Study 2 involved establishing an optimum usability study environment by creating four usability conditions for a product prototype under test. These conditions strived to establish what was 'good enough' and the effort required to establish critical usability flaws in a design. Study 4 aimed to ascertain the true value of usability results from a medium-fidelity prototype tested in a recreated environment (at a 'good enough' level) compared with usability data collected on a final product in its real environment, i.e. trying to establish whether key usability issues could be discovered early in the HCD process (using a recreated environment).

4.1.2 Epistemology and Ontology

Social Sciences have underlying philosophical principles that guide researchers to understand the social research in question: ontology and epistemology (Moon, 2014). These terms are also recognized in the natural sciences as a means of interpreting social conventions that influences empirical data. Kunzle's (2019) review of Tuli's (2010) qualitative and quantitative research in social science takes an ontological and epistemological perspective of their influences and chosen methodology in relation to the research question. As identified in the literature review, Cross (1999) makes reference to design in relation to these areas, particularly epistemology, which Kunzle (2019) called *"the nature of knowing"*. Epistemology outlines a question: *"what is the relationship between the knower and what is known? How do we know what we know?"* (Kunzle, 2019; Tuli, 2010). This PhD develops an epistemology that will later inform other research by using the very nature of knowledge and understanding to recreate the findings. Ontology, however, was not visited by Cross, but has significance in the qualitative realm. Kunzle (2019) referred to it as the *"nature of reality"* and, like Moon (2014), captures the same essence when asking: *"what exists in the human world that we can acquire knowledge about?"* An ontological perspective is significant in the context of this body of research as it relates to the perception of reality and subjective nature of social constructs, a key focus of this research. This also links directly to the literature concerning recreating environments and the role of presence and ecological validity when inducing a sense of reality (Benyon, 2009). According to Brynman (2001), to understand epistemology, one must first understand Positivism and Constructivism. Positivism is aligned with a quantitative method; it seeks to deploy research methods that conform to those used in the natural sciences and will confirm and predict patterns premised on empirical data collection (Bassey, 1995); however, it can promote one method as being more superior than the other (Tuli, 2010). Constructivism on the other hand, is concerned with real-world scenarios as they unfold. It does not produce generalized data relating to a population, but rather it is a qualitative method that explores experiences and interactions of people in order to understand phenomena. Since both empirical and observational data are analysed during usability studies, this body of research uses a mixed methods approach utilizing a Constructionist or, as it is also known, an Interpretivist approach.

Ontology concerns the mental constructs that are dependent on an individual's human context (Biggs *et al*., 2013). When considering the nature of reality, assumptions are cast and then tested in the context of an individual's wider socio-experience. As identified in the literature review, Blanford *et al*. (2010) focused their research on medical staff's user errors with infusion pumps. In doing so, they recognized that product interaction does not happen in a vacuum, and by applying ontological philosophical assumptions, they derived findings concerning distraction and its role in reality. There are also ontology parallels in the research conducted by Slater (2013). He acknowledges that a participant's ability to achieve presence is related to their past experiences and worldviews. These shape their experience in virtual reality. This is not dissimilar to Moon (2014) and Biggs *et al.*'s (2013) description of the Relativist view of ontology.

There are two important branches of ontology: Constructionism and Objectivism. Objectivism is recognized as an independent reality or actual reality, whereas Constructionism deals with theories of reality premised on social process (Tuli, 2010).

4.2 Methodology Options

Friedman (2003) noted that design does not work in isolation and can be associated with several areas. In keeping with that approach, the research methodology used here transcends behavioural science, psychology, human-centred design, computer science and medicine; and is a transdisciplinary research methodology.

> *"Transdisciplinary ways of working call for a fusion of disciplines'—a way of working in which designers have 'transgressed' or 'transcended' their own disciplinary norms and have adapted ways of working from other disciplines."*
>
> Muratovski (2016: 20).

There are a number of research methodologies that can be called upon in design research and, in particular, in HCD:

- quantitative research
- qualitative research
- ethnography
- action research
- grounded theory

- case studies
- mixed methods

In order to best understand the methodology selected, it is key to explore the above methodologies and justify the approach taken.

Quantitative and qualitative approaches can be used independently, and indeed were; quantitative approaches were seen as a means of removing the researcher from the temptation of subjectivity to ensure they remained independent and bias-free via an empirical approach founded in the natural sciences (Kunzle, 2019). Meanwhile, advocates of the qualitative approach recognize that data alone is not enough to interpret understanding of cause and effect and that reality tends not to occur in a single, logical, empirical paradigm (Johnson & Onwuegbuze, 2004).

> *"Qualitative methodology … attempts to increase understanding of why things are the way they are in social world and why people act the ways they do."*
>
> Kunzle (2019).

Ethnography is a study of collective behaviours via observations, participation and interviews. It is used to develop deep insights into people's behaviour, including unconscious acts, and is focused on action and object (Dawson, 2013). Ethnography involves close observation, an activity that is also at the heart of usability testing. Key attributes of ethnography are utilized in the mixed methods approach used in this research.

Action Research is the methodology most closely aligned to the mixed methods approach. Like ethnography, it places people at the heart of the research in a form of research collaboration where there are research partners rather than passive participants. The very word 'action' indicates that the ultimate goal is to act as well as observe and reflect (Muratovski, 2016).

Grounded Theory is an approach that grounds theory in the data, meaning that rather than beginning with a hypothesis, the theory is developed as the data is collected. Grounded Theory is recognized as an approach that has flexibility at its heart as it allows the use of the literature and data to emerge from findings (Dawson, 2013; Leedy & Ormrod, 2010). Leedy & Ormrod's (2010) approaches differ from the mixed methods approach utilized in this body of research as it questions the need, benefits and enhancements from integrating context of use into laboratory usability studies.

A case study approach enables for an in-depth study that can happen over a period of time, enabling the researcher to observe change premised on time, and offers a framework to gather qualitative research data (Muratovski, 2016). It too requires careful documentation and is predicated on a clearly defined context prior to commencement.

4.3 Methodology Mixed Methods

The mixed methods approach is a means of mitigating identified weaknesses in quantitative or qualitative approaches by combining them. It is a pragmatic and flexible approach that can be applied sequentially and concurrently (Terrell, 2011; Tashakkori & Teddlie, 2008). As this work is grounded in social science, there is a debate as to whether a social research approach (based on a qualitative research methodology) vs a social scientific or natural science approach (based on a quantitative research methodology) can be mitigated by working the quantitative and qualitative approaches together. Therefore, acknowledging the role of a quantitative research methodology and what it identifies (the 'what'), together with the role of a qualitative research methodology (helping to also identify the 'why'). This offers a balance of opinion and behavioural observations and data capture to strengthen the research and quell the anxiety of the scientific world that one approach is stronger than the other, and that the strength lies in both. Using both a quantitative and qualitative approach is called triangulation (Dawson, 2019). Both quantitative and qualitative can be used as independent approaches, and indeed were, when quantitative was seen as a means of removing the researcher from the temptation of subjectivity and ensuring they remain independent and bias-free via an empirical approach founded in the natural sciences (Kunzle, 2019), whilst qualitative recognized that data alone was not enough to interpret understanding of cause and subsequent effect; that reality does not happen in a logical empirical single paradigm; and that interpretation can be made that, too, is free of subjectivity (Johnson & Onwuegbuze, 2004).

4.4 Methodological Approach

All four studies conducted in this body of research have used a triangulation approach, also known as mixed methods approach (Muratovski, 2016; Dawson, 2019). Further detail and the specific characteristics will be outlined in the individual studies detailed in Chapter 5.

4.4.1 Research Context

The research described below was conducted in PEL2 and PEL3 at the Cardiff School of Art & Design, Cardiff Metropolitan University, between 2016 and 2019. Where a comparator study was required, for example in the final study and the ecological validity studies, then a real-world location was selected for proximity and access to participants. Validity studies were conducted in both PEL iterations as per the literature (Deniaud, 2015) allowing the research results to be analysed for relative validity as an indication of the environmental ecological validity.

4.4.2 Participant Recruitment

The role of participants is to represent the intended population impacted by the research. In product design terms, they might also be described as the intended target audience. The participant selection process must result in as random a sample as possible, within the resource envelope, to produce a valid research outcome (Rowntree, 2018). Convenience sampling, a nonprobability sampling method (Etikan, 2016), was used to select participants, due to accessibility to the participants. A statistically viable 30 participants were recruited in each ecological validity study, 24 participants in the optimum environment study and 30 participants in the Real vs PEL study. The usability studies utilize Norman & Nielsen's optimum principles of usability fault findings, ensuring no less than five participants were used in studies 2 and 4.

In each study, a bio questionnaire was used to screen out any participants that did not fit the required user profile. For example, if a participant had already used the product tested in study 4 (a 'nextbike'), they were omitted from the study as their learnt behaviour of the bike interface would have skewed the data. In both ecological validity studies, studies 1 and 3, each participant experienced three different environments using a randomized order, a process also known as 'counterbalance design' (Kneebone, 2011; Field, 2016). Counterbalance mitigates against a collective biased result as a consequence of practice effect, enabling the researcher to control the impact of order in the studies. For example, if environments were visited in the same order, participants might tend to become lethargic by the time they visited the third environment and if this were the same environment each time the results risk being skewed (Field 2016).

> *"Counterbalance is a technique used to eliminate sources of systematic variation. One source of systematic variance is practice effects"*
>
> Field (2016, p27).

4.4.3 Study Protocols

In studies one and three, the ecological validity studies, participants were asked to sit on a designated log seat and to stay seated for a total of 5 minutes. Participants were asked to focus on their surroundings. Study two is the first of the usability studies. Participants were asked to 'walk through' the whole user experience of the cleaning device rather than being given specific tasks (Rubin & Chisnell, 2008). Each participant was asked to assemble the product, clean the fuselage and put the product away. Where a fuselage was not present in the study, participants were asked to role-play the cleaning on any available surface, namely the floor. The fourth and final study, also a usability study, involved three distinct tasks that enabled the use of Rolf Molich's analysis tool (Molich, 2004). Both usability studies used a prototype with active physicality (Hare, 2015), ensuring usability was explored rather than passive aesthetic attributes, although this did not preclude participants on commenting on appearance, particularly if it hindered usability.

4.4.4 Data Collection

Repeated measures, also known as 'within-subject testing', was used for the ecological validity studies, since the aim of the research was to compare within the dataset as each participant experienced each condition. However, the usability studies did not use this measure; instead, it used an independent study method, also known as 'between subjects' and participants only experienced one condition. This was because the purpose of these studies was to focus on usability as a measure. Had a participant experienced both real and PEL conditions it would have skewed the usability results and overall study due to learnability. When developing a controlled environment that is purposely designed to test certain conditions, it is important that validity is achieved. Ecological validity has been explored predominantly in the context of automotive human factors development and studies. Deniaud *et al*. (2015) explored these issues in depth, in particular the methodologies that should be implemented when using virtual environments for experimental purposes and when aiming to achieve simulation validity (Blaauw, 1982). In these cases, the tasks should always seek to verify a simulated environment vs a real-world environment, and two variables should always be explored: physical and behavioural validity.

4.4.3.1 Face Validity / Presence

Face validity was used in all four studies. The decision on what data to collect in the studies was influenced by the literature. ITC–SOPI immersion questions (Lessiter *et al*., 2001) (an independent television commissioned questionnaire developed for cinema, screen-based activity, and later applied to gaming) were used to cover four key themes. This is because the questionnaire embodies the work identified in the literature review, particularly Deniaud (2015) and Dillon (2000), on key areas of consideration in the design of a simulated environment:

1. sense of physical space
2. engagement
3. ecological validity
4. negative effects

Lessiter *et al*. (2001); van Baren & IJsselsteijn (2004).

For the purpose of the studies found in this body of research, the ITC-Sense of Presence Inventory (ITC-SOPI) (Lessiter *et al*., 2001) was cross-referenced with Kneebone's (2011) questions used to identify face validity in his simulated training environments to develop a set of questions that were used to measure presence in each environment. The reason for a combined use of the two questionnaires was because the environment under test in these studies was not only screen based and neither was the screen the focal point. Therefore, Kneebone's integration of objects into his test environments and the ITC-SCOPI approach were deemed appropriate. A 6-point Likert scale was used in the face validity questions, because studies conducted into the use of the Likert scale have found that 6 points offered more discrimination than 5-point scales in psychological tests aimed at obtaining an attitude, perception, or opinion. Abdul's (2010) study also highlighted reliability as a factor in opting for a 6-point scale. The Systems Usability Scale (used in studies 2 and 4) is a set questionnaire designed by Bangor (2009) and was not altered from the 5 point Likert scale, as it has been calibrated in terms of results according to its original design.

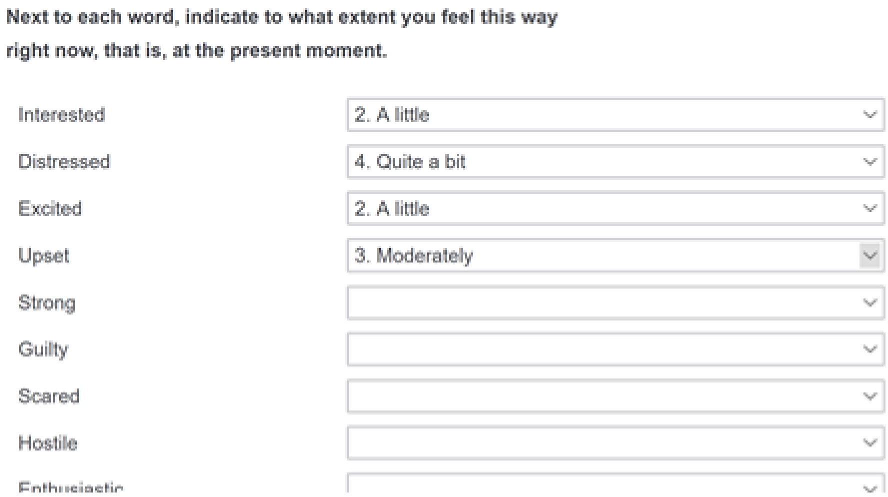
When using a Likert scale to capture data founded on opinion, it is appropriate to represent the data using mode and median to identify the central tendencies that describe preferences and opinions rather than using arithmetic data such as mean and parametric statistics. Capturing ordinal data allows for directionality in preference when usability testing a product but does not afford analysis that highlights something as being twice as good, i.e. the results cannot be measured in between the intervals (Dawson, 2013). There is also common practice of using Likert scales to imply intervals and the use of parametric statistics are common practice; therefore, the studies with smaller sample size will use ordinal data and the larger studies will utilize parametric data (Wu & Leung, 2017).

For the purpose of this study, relative validity is used as a method to obtain data on the setup of the virtual environment. Unlike absolute validity, tests can be conducted on both the real world and the virtual environment and the differences achieved in each test can be assessed for magnitude and points of difference. If the differences appear correlated, then relative validity has been achieved between the two environments.

4.4.3.2 Positive and Negative Affect Schedule (PANAS)

The PANAS method of self-reporting is used by psychologists to ascertain an individual's emotional state (Magyar-Moe, 2009; Merz *et al*., 2013) and was developed by Watson *et al.*

(1998). PANAS is a key tool when measuring emotion because it translates emotion and feeling into data, so it can be numerically analysed. The PANAS questionnaire is a method deployed by Witmer & Singer (1998) and later used by Falconer *et al.* (2014), in works that included Mel Slater, one of the leading authors in virtual environments, having developed simulated environments for training the US army. The PANAS scale was used in all four studies. Its role in these studies is to understand the impact of different environments on participants' behaviour when usability testing a product and, in particular, to ascertain their emotional state during a study. PANAS has a predetermined analysis tool that balances the positive and negative scores to determine an overall state of emotion from determining the mean. It is used to measure input, in this case the individuals' state of emotion, unlike the usability results and the presence questionnaires that measure outputs. Reliability of data has been validated by Magyar-Moe (2009), whereby the conducted studies ascertained the validity of the PANAS results and deemed them reliable and a valid measure of emotional state. A positive momentary mean of 29.7 and standard deviation (SD) of 7.9 for positive is used to calculate a positive disposition, and for negative, a mean of 14.8 and SD of 5.4 is deemed a negative disposition (Watson, 1998). The descriptor 'momentary' refers to a PANAS study that is conducted at a moment in time and uses a different set of reliability data to data collected over a week or longer period of time that in turns generates a weekly mean (Watson *et al.*, 1988). One limitation with PANAS is that participants need to speak English fluently to understand the meaning of the words used, otherwise they might misinterpret certain words and skew the PANAS score. An example of the PANAS can be found in figure 30.



Next to each word, indicate to what extent you feel this way right now, that is, at the present moment.

Interested	2. A little
Distressed	4. Quite a bit
Excited	2. A little
Upset	3. Moderately
Strong	
Guilty	
Scared	
Hostile	
Enthusiastic	

Figure 30 PANAS (section of)

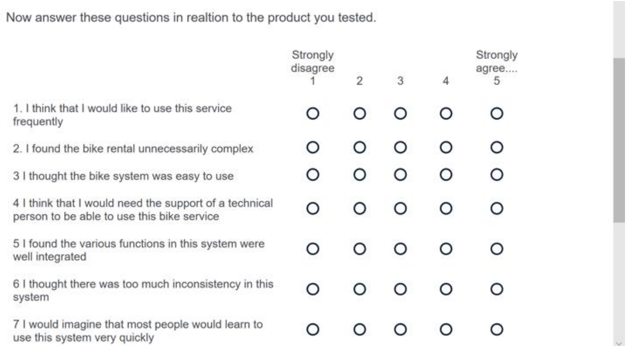
4.4.3.3 Heartrate

Heartrate was used in the ecological validity studies 1 and 3. As determined in the literature by Guger *et al*. (2005), heartrate reflects the physiological state of a participant and is a means of utilizing repeated measures when engaging participants. The approach enables the comparison of data related to emotional response within subjects and was therefore used in the two ecological validity studies. The heartrate measure was not used in the usability studies because the data would have been measured across subjects and, since everyone's heartrate differs, and different participants experienced different environment conditions, the data would have no value.

4.4.3.4 Systems Usability Scale (SUS)

SUS was used in the usability testing studies 2 and 4. The SUS was developed in 1986 by John Brook (Bangor *et al*., 2009). Brook argues that it is not enough to determine the cause of the fault on its own, hence the SUS is used in support of observational findings. SUS's purpose is to ascertain a product's overall usability rather than individual faults with a design. It works by alternating questions between a negative and positive inflection using a 5-point Likert scale as seen in figure 31 (Barnum, 2011). Brook developed a method of converting these results to ascertain the usability of an overall system or product. The results are converted into a positive ordinal data set using Brook's calculation and converted into a percentile. SUS percentiles are explained by Brook, who states: *"Products that scored in the 90s were exceptional, products that scored in the 80s were good, and products that scored in the 70s were acceptable. Anything below a 70 had usability issues that were cause for concern."* Bangor *et al*. (2009). A score of 68% or below denotes severe usability issues.

Note that normalization in terms of a SUS scale is not a statistical calculation arrived at by the mean and standard deviation of a particular result, but a sector-agreed score that is used to normalize the data relative to the sector: in this case, a SUS of 68%. The SUS is a summative means of recognizing usability faults in a product or system as a whole. It is used to signify whether an entire design has usability faults rather than highlighting detailed usability issues, hence the value of pairing with ethnography to additionally pinpoint detailed faults (Gordon *et al*., 2019).



Now answer these questions in realtion to the product you tested.

	Strongly disagree 1	2	3	4	Strongly agree.... 5
1. I think that I would like to use this service frequently	○	○	○	○	○
2. I found the bike rental unnecessarily complex	○	○	○	○	○
3 I thought the bike system was easy to use	○	○	○	○	○
4 I think that I would need the support of a technical person to be able to use this bike service	○	○	○	○	○
5 I found the various functions in this system were well integrated	○	○	○	○	○
6 I thought there was too much inconsistency in this system	○	○	○	○	○
7 I would imagine that most people would learn to use this system very quickly	○	○	○	○	○

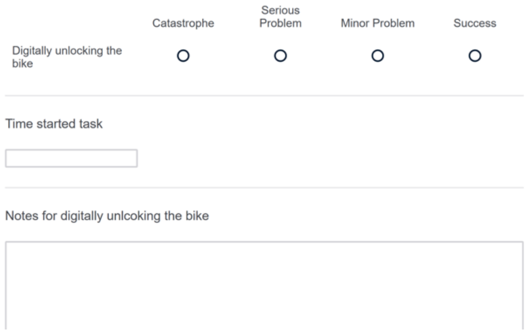
Figure 31 Section of the SUS

4.4.3.5 Rolf Molich

Rolf Molich's performance scale was used in the fourth and final verification study: PEL vs Real. Molich's performance rating scale highlights specific usability faults in a design, unlike the SUS that gives a result that indicates the overall usability of a product or system. It utilizes a scale and set criteria captured by Molich and Dumas (2006):

Catastrophe: *user is unable, refuses or solves the task incorrectly.*
Serious problem: *user is significantly delayed (1-5 minutes) but manages to complete the task.*
Minor Problem: *user is briefly delayed (the user experiences a problem but corrects themselves reasonably quickly—less than 1 minute).*
Success: *user completed the task without problems or delay.*
Woolley (2008).

Barnham (2011) captures the same principles, but refers to the performance rating as 'route task results'; however, Redish *et al*. (2002) found that the categories used were too complex and needed categorization, hence the development of the Rolf Molich performance categorization (figure 32). These refer to results captured by the usability test moderator either live or via video footage, and categorized post-test into quantitative data. In the case of studies 2 and 4, time taken to complete the task was also captured and analysed as it is also an indication of usability performance.



	Catastrophe	Serious Problem	Minor Problem	Success
Digitally unlocking the bike	○	○	○	○

Time started task

Notes for digitally unlcoking the bike

Figure 32 Section of Rolf Molich Performance Rating

4.4.3.6 Eye Tracking

Eye tracking was used in study 4, PEL vs Real. A Tobii eye tracker was used to measure the gaze of participants by capturing data from 'Areas of Interest' (AOI). AOI is a means of analysing a participant's attention. By assigning areas of the environment as AOI, researchers are able study and analyse gaze fixation data (Orquin *et al*., 2015). The eye tracker was used in the 4th study to ascertain how many times an individual looked beyond the product under test, and whether they were distracted the same way in both conditions: Real and PEL. One limitation of the eye tracking is that it is not particularly inclusive. If participants wore glasses, for example, they could not be used for the study, as the tracker could not calibrate to their eyes.

4.4.3.7 Questionnaire

General open-ended questions were included in all the studies to better interpret the data. There is some degree of contention around post-task questionnaires due to their reliance on participants' memory of the task (Jokinen *et al*., 2015). The alternative, however, is to complete a questionnaire during the studies, which would cause more disturbance and would impact on the ecological validity. The definition of ecological validity in terms of statistics is: *"evidence that the results of a study, experiment or test can be applied, and allow inferences, to real-world conditions"* Field (2016, p706). The questionnaires used in these studies have been designed for use in the field of Product Design and Virtual & Simulated Environments following close reading of the literature; however, at the end of each study, additional open questions were asked relating to the environment to explore the 'what' and 'how' in a bid to further contextualize the findings.

> *"Ending a usability test with an interview provides another way for the participant to share their experience in their own words."* Barnum (2011, p187).

5.4.3.8 Audio & Visual Recording

Research moderators have a lot of information to deal with: the think aloud protocol, visual observation, participant questioning and controlling the experiment. For this reason, Rubin (2008) recommends capturing user tests on video so that post-study analysis can be conducted. The think aloud protocol is useful in this context since it allows researchers to really understand what a user is thinking about when interacting with a product. Paired with video, it is possible to analyse the user's thoughts and interactions side by side. In the literature, we can see that Rubin (2008) and Barnham (2011) refer to this method as a means of ascertaining real-time thoughts during a study and this can mitigate against the limitations of post-test questionnaires.

4.4.3.9 Statistics

This table highlights the key attributes of each study.

Table 3 Details of all studies

Study	1. Ecological Validity 1	2. Optimum Environment	3. Ecological Validity 2	4. Real vs PEL Validation
Number of participants total study	30	24	30	30
Number of participants in each condition	30	6	30	15
Conditions	3	4	3	2
Research Method	Face validity (questionnaire) PANAS Heartrate Open comments	Face validity SUS Open comments Observation Think aloud	Face validity PANAS Heartrate Open comments	PANAS SUS Rolf Molich* Observation Think aloud Time
Statistical type	Repeated measures within subjects	Independent / between subjects	Repeated measures within subjects	Independent / between subjects
Analysis method	ANOVA	Mode & Median	ANOVA	*Independent t-test.

The two ecological validity studies compared multiple conditions within subjects in a population of 30 participants. Therefore, to ascertain variance in the data sets, an Analysis of

Variance (ANOVA) was used on the heartrate data and the face validity questions. ANOVA results were captured using a post-hoc Bonferroni test that allows researchers to determine whether a significant difference exists.

The analysis method used for the fourth study, Real vs PEL, was an independent t-test. T-tests are common in virtual environment literature. For example, Kneebone (2011) used the method to validate his distributed simulation environment in his paper *'Blowing up the Barriers'*. The t-test is designed for use on study populations of less than 30 participants—populations of 30+ are tested via a z-test. In this case, the t-tests were applied to the data from two groups of 15 participants who had interacted with two conditions. These were analysed against each other, as in this case, the study's focus was on usability in two environments rather than whether presence was achieved, as had been the case in study 2: Optimizing PEL. The fourth study was also an independent study or between subjects, meaning different participants experienced different environments to avoid learnability (Field, 2016). The fourth study also used Cohen's D test to look at effect size (Field, 2016). When analysing the data, Cohen suggested that a 'D' of 0.2 should be considered a 'small' effect size, with 0.5 representing a 'medium' effect and 0.8 a 'large' effect. D is calculated by taking the difference between the two group means and dividing by the Standard Deviation (Field, 2016, p382).

4.4.5 Ethics

> *"Many people are willing to disclose a lot of personal information during our research, so we need to make sure that we treat both the participants and the information they provide with honesty and respect. This is called Research Ethics."*
>
> Dawson (2009, p149).

Cardiff Metropolitan University has an ethics process that ensures the researcher treats the participant with respect and ensures the participant is informed as to the contents and intended communication output of the studies they intend to be involved in, and is willing to be involved. Participants have the option to withdraw at any stage if they feel uncomfortable. Ethics approval was gained for each of the studies, with paperwork consisting of an information sheet explaining the purpose of the study and what was about to happen, and a consent form explaining how the data would be captured and used. Samples are available in Appendix 1. Each study was piloted to snag any issues with the design of the study via conducting walkthroughs (Barnum, 2011). Each participant was informed of the nature of the study prior to starting and was not asked to do anything that would make them feel uncomfortable. The moderator took a scripted approach to each study, which ensured that no information was left out and that each participant felt at ease. All of the above contributed to ensuring participants were treated with respect, with their wellbeing at the forefront of the researcher's thoughts as they elected to participate (Dawson, 2009).

4.4.6 Challenges of the Chosen Methodologies

Limitations of this work include population sample size. It is recognized in the usability literature that sampling 5–8 participants is sufficient when establishing usability issues in a design proposal:

> *"According to Nielson, you should stop after the fifth user because you are seeing the same thing repeated, and you will have researched the optimal return of 85% of the findings to be uncovered."*
>
> Barnum (2011, p16).

The limitations are due to the research nature of the work, with statistical viability said to be with a population beyond 30 participants. To mitigate this limitation, when conducting the analysis of the means in study 4, the T-test is utilized, because it is a method that has been designed to determine statistical credibility in population sizes of less than 30. It too was an approach adopted and published by Kneebone (2010) when he conducted his studies on his medical training environment.

Another limitation of this research is determining the optimum 'relative' validity, as there is no reference in the literature concerning the threshold of the relativity in order to determine if an environment is ecologically valid. Therefore it is the responses in the studies themselves that determine if the environment elicits an emotional and behavioural response of being in and being there, also known as presence (Steuer, 1993).

The tools utilized for these research studies are detailed in Chapter 5. However, it is important to recognize the ontological position of this research, since assumptions about the nature of reality and their impact on usability testing and consequently product development are key to the research. By deploying a mixed methods approach, the research is afforded the benefits of both Positivism and Interpretative Constructivism.

Chapter 5 Studies

Chapter 5 User Studies

5.1 Aim & Objectives of Research

During this chapter, objectives 3 and 4 are addressed:

3. *To ascertain the fidelity of a simulated environment that is most effective in discovering product design flaws early in the design process.*
4. *To validate the findings and detail the outcome of this research exploration.*

5.2 Introduction

The studies described in this chapter seek to establish key findings relating to the gap in the literature concerning the role of simulated environments and their potential contribution to testing products early in the design development process. Their focus is to identify the role simulated environments play when filling the void between the laboratory and field studies, as identified by Woolley *et al*. (2011). Although the literature has identified an opportunity to introduce more complexity into the laboratory environment, Kjeldskov *et al*.'s (2014) finding that it is not 'when in the design process' but 'how' user tests should be conducted has been a key message. Kneebone *et al*.'s (2011) research, premised on attention and involvement when recreating environments for training needs, has also been influential. The research thus far described above has made clear that the correctly focused combination of physical and image-based environments can support the creation of fully immersive virtual environments that induce a sense of presence, or 'being there'. In theory, it ought to follow that if participants at some level accept that the environment they are in is real, that product testing within such an environment ought to find similar results as those found in field testing. Another key message of the research thus far concerns the fidelity of the prototype: it is clearly important that the fidelity of the simulated environment and the prototype under test should be of similar levels, as to do otherwise appears to have a jarring effect on participants.

Ecological validity is another key target in any simulated environment, especially when mimicking real-world elements, as is relative validity.

5.3 Study 1: Ecological Validity

The purpose of this study was to ascertain if participants respond and behave in PEL as they would in the real world to ascertain the level of ecological validity (Blaauw, 1982) and determine the relative validity of PEL. This study is needed as a prerequisite to fulfil objectives 3 and 4.

5.3.1 Conditions

This study involved evaluating PEL in comparison to the real world and a laptop. The scene used in all three settings was identical. This study consisted of three different conditions:

1. Outside environment (figure 33).
2. PEL recreating the outside scene (figure 35).
3. Laptop with an image of the outside scene (figure 34).

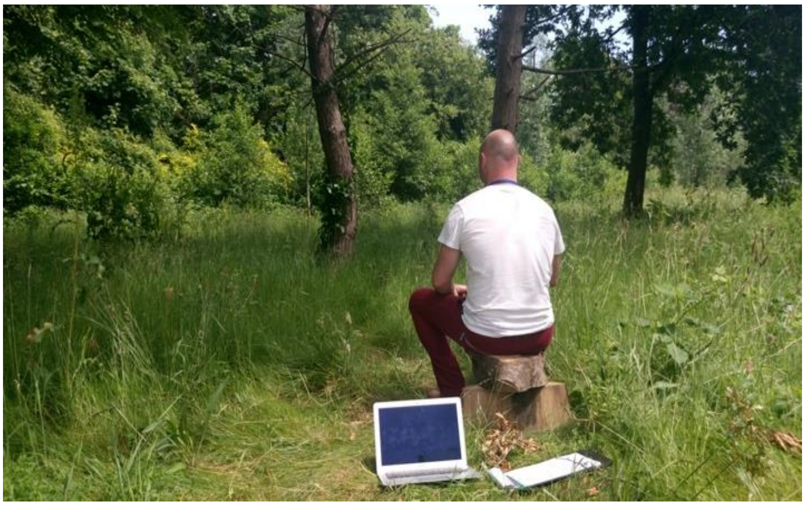


Figure 33 Outside study

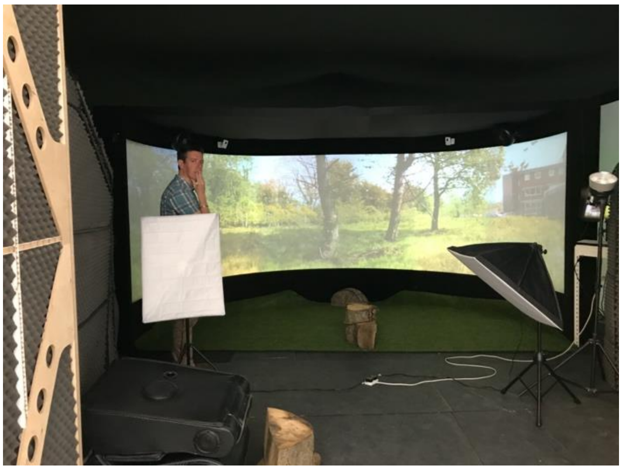


Figure 35 PEL study

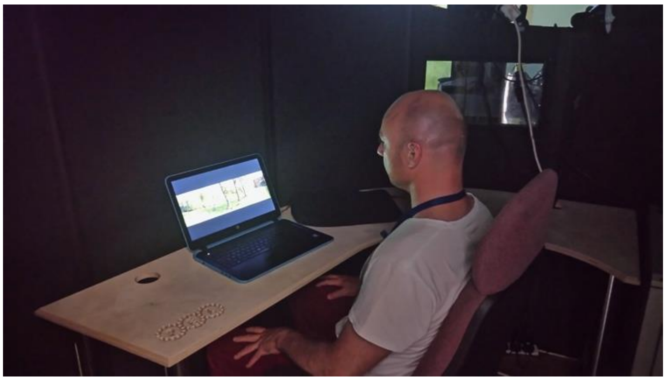


Figure 34 Laptop study

PEL used fans to simulate breeze, and artificial grass, logs and wood for scene setting, to immerse the participant and heighten the sense of presence. The same log seat was used in both PEL and the outside environment; and the image projected in PEL, and used on the laptop, was the same scene experienced in the outside condition. The outside location was selected because of its proximity to PEL, allowing participants to be led from one environment to the next with ease. Beyond simple practicality and convenience, one reason to do this was that too much cardiovascular exercise would have had an impact on participants' heartrate data, and as the data was used to reflect the physiological state of the participants (Guger *et al.*, 2005), strenuous movements would skew the data or require a longer rest in between the studies. Therefore, to minimize the impact of an increased heartrate due to climbing stairs to reach the next study, a location with minimal distance was used.

The study was piloted prior to engaging with the participants. Piloting a study is a fundamental requirement of any research study as it enables the researchers to establish any oversights or errors in the study. For example, to check if questionnaires identify any ambiguities (Dawson, 2019). In addition, a pilot study allows the researchers to check all equipment is working properly. For this study, a participant experienced the study and completed the relevant paperwork. On conclusion of the pilot it was agreed that more wood was needed in PEL; the position of the log seat needed to change; and the process of collecting data in the outside space was modified. For example, a laptop needed to be repositioned to pick up the Bluetooth heartrate device on the participant, but put out of view so it did not become a part of the scene.

5.3.2 Methodology

The study used a mixed methods approach as outlined in Chapter 5. The quantitative methods used were face validity, PANAS and absolute heartrate, with open comments providing qualitative data. There was no observational research in this study, as participants did not physically interact with any product or prototype. This was because the focus of the study was establishing the extent of the ecological validity rather than behavioural validity.

5.3.3 Methods

In total, 30 people participated in the study. Participants were selected using convenience sampling. Of the 30 participants, 22 were male and 8 were female. Their average age was 25.

A counterbalance method was used to ensure the data was not skewed by the order in which the participants experienced the different conditions (as shown in table 4 below). table 5 below details the different research methods and analysis methods used in the study.

Table 4 Counterbalance applied in study.

Participant	Testing order
1-5	Laptop, Real, PEL
6-10	PEL, Real, Laptop
11-15	Real, PEL, Laptop
16-20	Real, Laptop, PEL
21-25	Laptop, PEL, Real
26-30	PEL, Laptop, Real

Table 5 Study Detail

	No of Participants	No of participants in each condition	No of conditions	Research methods	Statistical analysis	Analysis method
Study 1 Ecological Validity 1	30	30	3	-Face Validity Questionnaire -PANAS -Heartrate -Open Comments	Within subjects	ANOVA

Arousal is one measure of a participant's involvement in a virtual environment, and one objective physiological indicator of presence (Meeham, 2001). *"Arousal has generally been conceived of as a drive state or a non-specific energiser of behaviour, something that describes the intensity of an experience but not its quality."* Dillon *et al.* (2000). Arousal is also an important factor relating to a person's emotional state and can be empirically detected by monitoring the autonomic nervous system. For example, Guger *et al*. (2005) showed that heartrate reflects the physiological state of the participant and relates to emotional response. Such monitoring can detect the level of participant activity and its intensity; however, it cannot measure the quality of the involvement (Duffy, 1962). To address this limitation, questionnaires can be used to supplement the physiological data in order to discover the quality of a participant's involvement, and hence measure the level of presence (Guger *et al.,* 2005; Dillon *et al.,* 2000).

Heartrate is defined as the rhythm and time between heartbeats. It is part of the autonomic nervous system and is regulated by the sympathetic and parasympathetic aspects of that system (Loudon & Deininger, 2016; Schafer *et al*., 2013; Dillon *et al.*, 2013; and Bernston *et al.*, 1991). Slater, a leading researcher in the field of virtual reality, describes the necessary conditions for optimizing presence, concluding that they principally concern focus and attention (Slater, 2004). Slater was an early adopter of heartrate as an objective measure to ascertain a participant's level of arousal and presence when placed in a new virtual environment. Capturing heartrate is a non-invasive process, with participants simply using a device worn on the wrist or around the chest.

Mestre (2006) notes that many different factors can impact on an individual's heartrate. It is recognized, for example, that a sudden increase in heartrate is associated with negative/defensive emotions such as anxiety, stress or fear. An increase in heartrate is associated with reorientation due to an unexpected event or sudden change that diverts a participant's attention (Guger *et al.,* 2005).

For the purpose of this study, heartrate data obtained in each environment was used as an indication of 'direction of travel' to help substantiate, or not, relative validity (Godley *et al*., 2002; Deniaud *et al*., 2015). There are a number of factors that affect heartrate, such as caffeine, the time of day, whether a person is standing or sitting, and current or recent activity. A protocol was therefore adopted whereby testing only took place in the morning, participants were asked to avoid consuming caffeine prior to the study and remained in a seated position throughout, in accordance with the literature (Allen *et al*., 2006). Five minutes of data was obtained from each participant, with the first two minutes discounted to enable the participant to orientate themselves in the environment, giving three minutes of data per participant.

Deniaud *et al*. (2015) recognizes the complexity of differentiating subjective behavioural presence and emotional presence, where the first can be captured via a questionnaire, while the latter requires more complex, potentially dynamic tasks.

As described above, face validity questionnaires are a construct that can be used to evidence the level of participant presence in an environment. The questions below were taken from Kassab *et al*. (2011) and adapted for the purposes of this study (table 6). The same set of questions were used in each condition.

Table 6 Face Validity questionnaire

Please circle, 1 = not at all. 6 = very much

1. I felt I was visiting the place in the displayed environment

   1 2 3 4 5 6

2. I felt like I was just watching something

   1 2 3 4 5 6

3. I felt like I was outdoors

   1 2 3 4 5 6

4. I had a sense of being in the scenes displayed

   1 2 3 4 5 6

Note that question 2 contributes to negative affect, using language such as ‘just watching’, and questions 1, 3 and 4 contribute to a positive experience utilizing language such as ‘visiting’, ‘I was’ and ‘being in’.

The face validity questions are recognized in the literature as a means of interpreting validity in a simulated environment, regardless of the medium used (Kneebone, 2010; Witmer & Singer, 1998; Lessiter, 2001; Kassab, 2011; and Slater, 1999). They contribute to the factors of participants ‘being there, being in’ a space and attempt to quantify a subjective experience (Deniaud, 2015). When combined with a PANAS questionnaire, specifically designed by the Psychology Association to ascertain momentary emotional state (Watson *et al*., 1988), a picture can be formed of the contributing factors that can measure presence.

The PANAS questionnaire was used to capture a participant's emotional state triggered by the different environment condition; the participant is asked to use the values that are aligned with the descriptions and add one value per word so as to capture the emotional state of the participant and then translate it into data, so it can be numerically analysed. The analysis, as identified by Watson *et al.* (1988), will determine a negative emotional state and a positive emotional state. See the PANAS questions below in table 7:

Table 7 PANAS Questionnaire

1	2	3	4	5
Very Slightly or Not at All	A Little	Moderately	Quite a Bit	Extremely

__________ 1. Interested

__________ 2. Distressed

__________ 3. Excited

__________ 4. Upset

__________ 5. Strong

__________ 6. Guilty

__________ 7. Scared

__________ 8. Hostile

__________ 9. Enthusiastic

__________ 10. Proud

__________ 11. Irritable

__________ 12. Alert

__________ 13. Ashamed

__________ 14. Inspired

__________ 15. Nervous

__________ 16. Determined

__________ 17. Attentive

__________ 18. Jittery

__________ 19. Active

__________ 20. Afraid

5.3.4 Data Analysis

The study utilized a repeated measures approach, also known as 'within subjects', to establish relative validity between the three different conditions (settings)—real-world environment, PEL and laptop—the main interest being to ascertain the ecological validity of PEL.

Heartrate, face validity and PANAS data were analysed using a parametric analysis method to obtain the analysis of variance (ANOVA) across multiple conditions for different environment conditions, using a post-hoc Bonferroni correction method in SPSS (Field 2016). The f-test was used to ascertain the F ratio[4], an estimate of population variance, and P value, a measure of the probability that an observed difference between two or more conditions could have happened just by chance (Rowntree, 2018). The P level was set at 0.05 for the classification of significant difference between conditions. Note that while capturing ordinal data allows for directionality in preference, it does not afford proportional analysis—i.e. it is not possible to say that something is 'twice as good'. In other words, the results cannot measure in between the intervals (Dawson, 2013). Open comments were analysed thematically to identify the frequency and type of comments, with quotes evidencing supporting reasoning found in the data captured. Themes were identified from the literature. Open comments were coded against themes, with the resulting data used to offer further insight that could be triangulated with the quantitative data, as per the triangulation approach noted in Methodology Section 4.2 (Terrell, 2011).

5.3.5 Results

Table 8 Mean (SD) results for the Face Validity questions

Question	**Real**	**PEL**	**Laptop**	**F -Ratio**	**Sig.**
Q 1	5.70 (.837)	3.33 (1.155)	1.6 (1.003)	231.454	< 0.001
Q 2	2.03 (1.56)	3.5 (1.196)	5.13 (1.196)	36.162	< 0.001
Q 3	5.97 (.183)	3.07 (1.285)	1.30 (.702)	276.691	< 0.001
Q 4	5.63 (.850)	3.57 (1.073)	1.90 (.995)	142.199	< 0.001

[4] The F-test offers the same post-hoc analysis as the Z- or t-test; however, the Z-test is for more than 30 participants, the t-test for less than 30 participants, but both apply to two conditions, whereas the F-test applies to multiple conditions (Rowntree, 2018, p141).

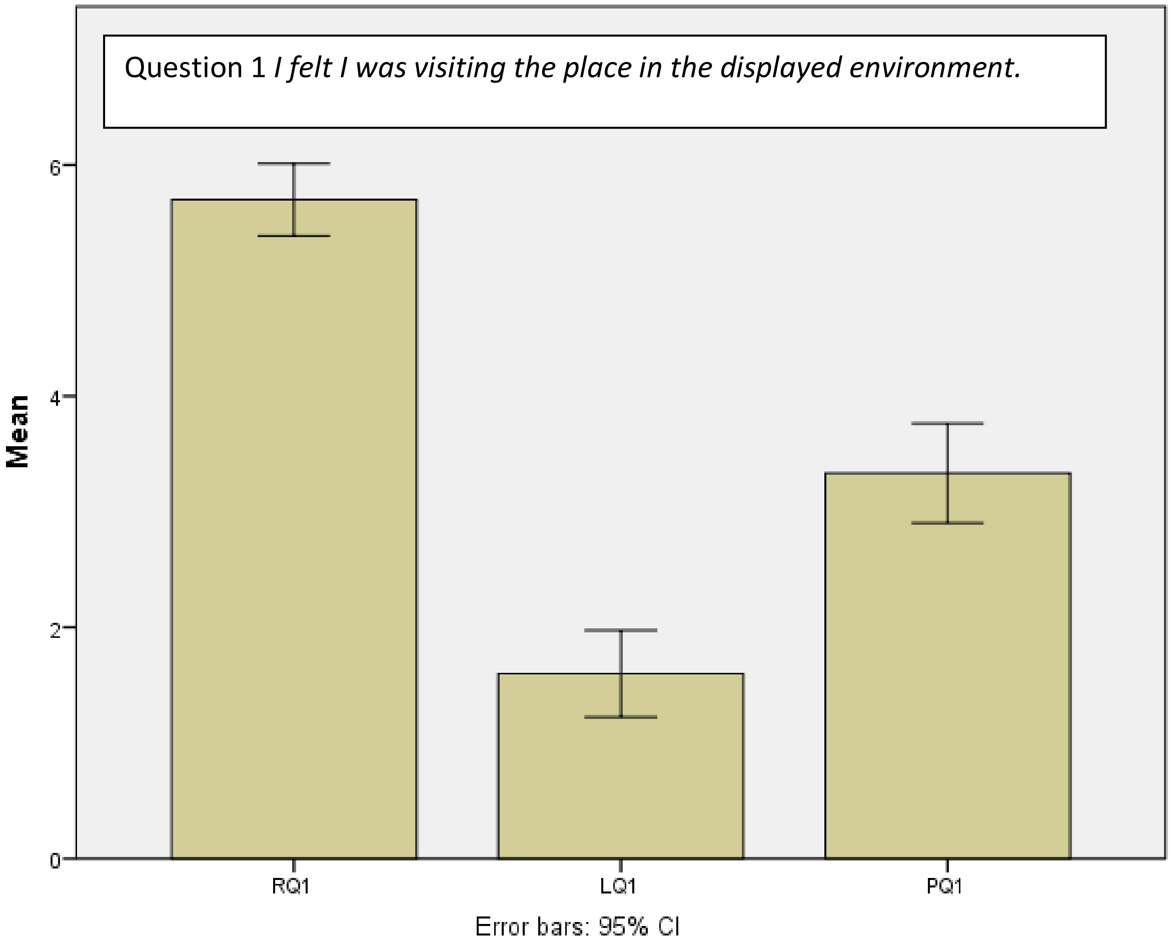


Figure 36 Question 1

Figure 36 shows the graph of the mean ratings of the answers to question 1 (‘*I felt I was visiting the place in the displayed environment*’). RQ1 is the Real environment, LQ1 is the Laptop and PQ1 is PEL.

The Mauchly’s Test indicated that the assumption of sphericity had not been violated, χ2(2) = 5.993, p = 0.050. The results show that the response to question 1 was significantly affected by the condition, F(2,58) = 231.454, p < 0.001. Post-hoc pairwise comparison highlighted that there were significant differences between all pairwise comparisons.

The mean rating of 5.7 in the Real environment indicated that participants, on average, felt they were 'visiting' an outside space (which of course they were). However, relative validity can be evidenced in the results, with a sliding scale of 3.33 mean rating in PEL and 1.6 mean rating in the Laptop conditions, resulting in participants 'feeling' less as if they were visiting the outside space. Relative validity is defined as an *"alternative method … to compare performance differences between experimental conditions in the simulator and a real car. In this approach, if the differences that are found between the two systems are in the same direction and have a similar magnitude, then it is possible to claim 'relative' validity".* Kaptein *et al.* (1996), as referenced by Deniaud *et al.* (2015).

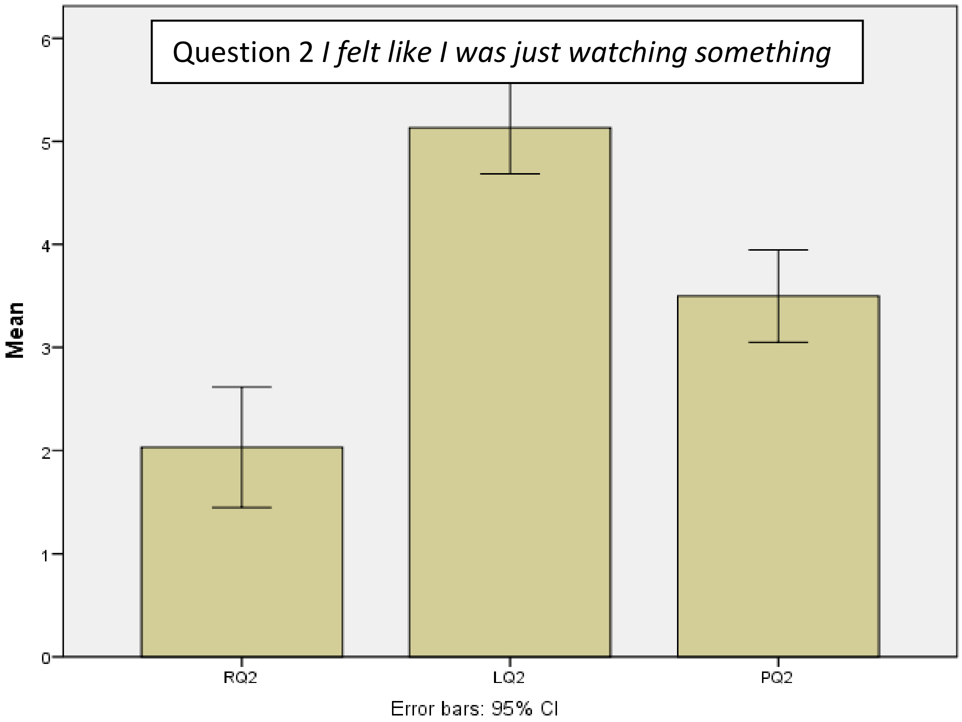


Figure 37 Question 2

Figure 37 shows the graph of the mean ratings of the answers to question 2 ('*I felt like I was just watching something*'). The Mauchly's Test indicated that the assumption of sphericity had not been violated, $\chi^2(2) = 3.379$, p = .185. The results show that the response to question 2 was significantly affected by the condition, F(2,58) = 36.162, p < 0.001. Post-hoc pairwise comparison highlighted that there were significant differences between all pairwise comparisons. As expected, the mean rating is low in the Real environment (2.03), i.e., on average, the participants did not feel as if they were just watching a scene. The mean rating increases in PEL to 3.5 and then to a high mean of 5.13 for the Laptop condition. Again, this suggests that relative validity has been achieved. The ratings for PEL are significantly different to the Laptop and the Real environment conditions; however, there is evidence in a direction of travel, whereby more ecological validity is evident in the PEL condition compared to the Laptop condition.

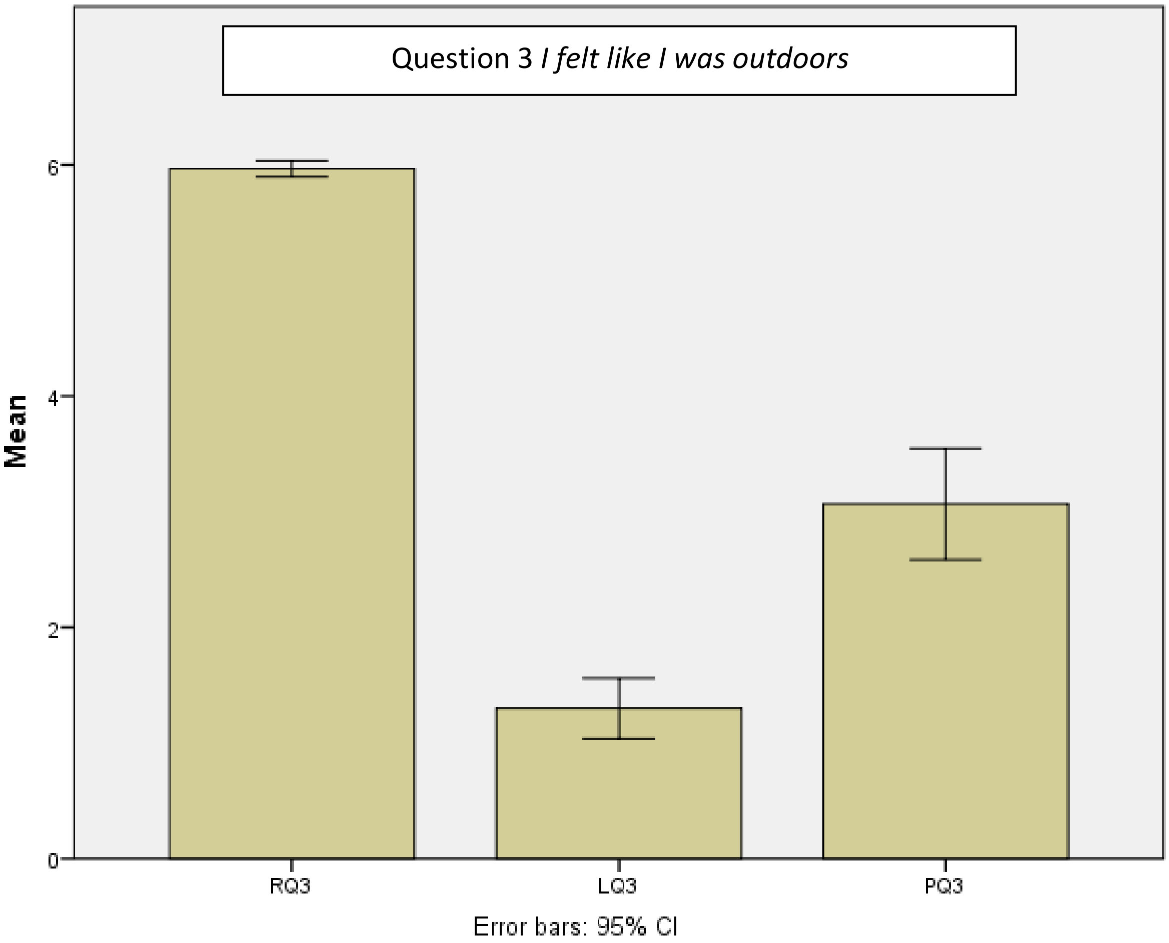


Figure 38 Question 3

Figure 38 shows the graph of the mean ratings of the answers to question 3 (*'I felt like I was outdoors'*), with a mean rating of 5.97 for the Real environment, 3.7 for PEL and 1.3 for the Laptop conditions. The Mauchly's Test indicated that the assumption of sphericity has been violated, χ2(2) = 12.104, p = 0.002; therefore, Greenhouse-Geisser corrected tests were reported (ε = 0.740). The results show that the response to question 3 was significantly affected by the condition, F(1.480,42.932) = 276.691, p < 0.001. Again, this suggests that relative validity has been achieved. Post-hoc pairwise comparison highlighted that there were significant differences between all pairwise comparisons.

Question 4 *I had a sense of being in the scene displayed*

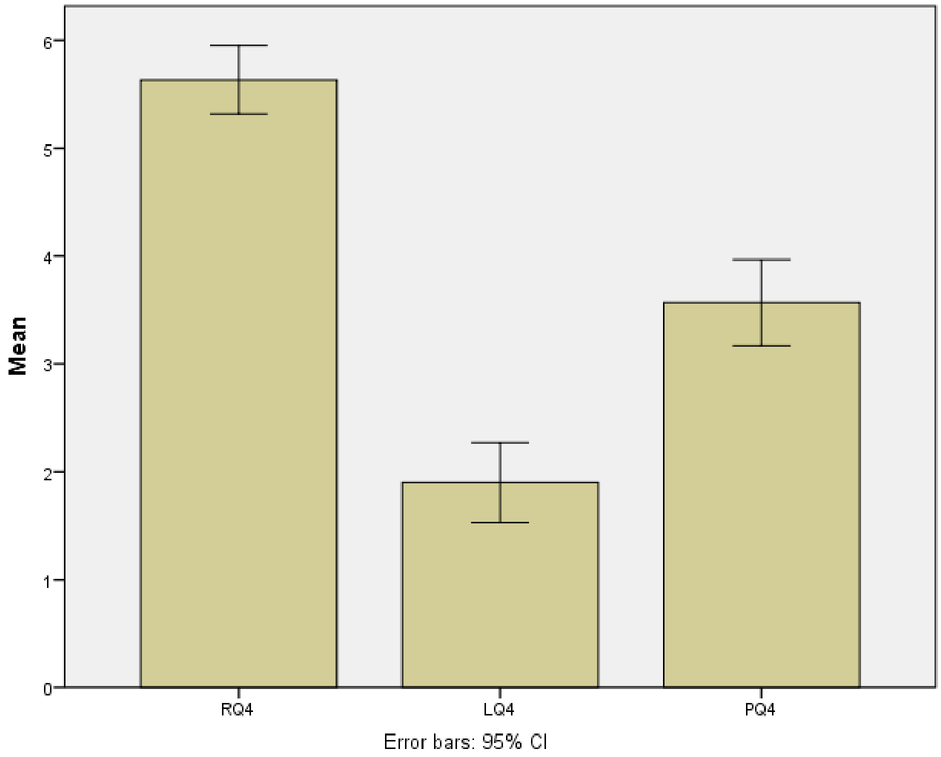


Figure 39 Question 4

Figure 39 shows the graph of the mean ratings of the answers to question 4 (*'I had a sense of being in the scenes displayed'*), with a mean rating of 5.63 for the Real environment, 3.57 for PEL and 1.9 for the Laptop conditions. The Mauchly's Test indicated that the assumption of sphericity had not been violated, χ2(2) = 3.601, p = 0.165. The results show that the response to question 4 was significantly affected by the condition, F(2,58) = 142.199, p < 0.001. Again, this suggests that relative validity has been achieved as PEL sits clearly between the real environment and the Laptop. Post-hoc pairwise comparison highlighted that there were significant differences between all pairwise comparisons.

Table 9 PANAS results

PANAS	**Real**	**PEL**	**Laptop**	**F-Ratio**	**Sig.**
Positive Mean	27 (8.133)	24 (7.541)	20 (8.044)	21.212	< 0.001
Negative Mean	11 (1.524)	12 (2,184)	13 (3.677)	10.825	< 0.001

Table 9 shows the PANAS results for the three conditions, both for the positive and negative means. For the positive mean, the Mauchly's Test indicated that the assumption of sphericity had not been violated, χ2(2) = 0.367, p = 0.833. The results show that the positive mean was significantly affected by the condition, F(2,58) = 21.212, p < 0.001. Post-hoc pairwise comparison highlighted that there was no significant difference in the positive means of the PEL and the Real environment conditions (p = 0.12). However, there was a significance difference in the positive mean between the Real environment and the Laptop condition (p < 0.001); and a significant difference in the positive mean between PEL and the Laptop condition (p = 0.004).

For the negative mean, the Mauchly's Test indicated that the assumption of sphericity has been violated, χ2(2) = 9.315 p = .009; therefore, Greenhouse-Geisser corrected tests were reported (ε = 0.779). The results show that the negative mean was significantly affected by the condition, F(1.559,45.207) = 10.825, p < 0.001. Post-hoc pairwise comparison highlighted that there was no significant difference in the negative means of the PEL and the Real environment conditions (p = 0.442); however, there was a significant difference in the negative means of the PEL and the Laptop conditions (p = 0.007) and between the Real environment and the Laptop condition (p = 0.001).

According to Watson (1988), who developed the PANAS, positive means above 29.7 suggest that participants are in a positive state, while negative means above 14.8 suggest that participants are in a negative state. With positive means below 29.7 across all conditions, this suggests that, on average, participants were not in a positive state in any condition. However, with negative means below 14.8 across all conditions, it also suggests that, on average, participants were not in a negative state either (in any condition).

The difference in the heartrate data for both the first two 'acclimatizing' minutes compared with the following three minutes showed an increase of approximately the same rate: 2.179 bpm in the Real Environment, 1.93 bpm in PEL and 2.24 bpm in the Laptop condition. This suggests that the change of a participant's heartrate is consistent across all three conditions with a mean difference of approximately 2.12 bpm between the two-minute acclimatizing period and following three minutes. During the three minutes when participants were engaged, there was no significant difference between conditions in relation to a participant's heartrate.

The consistent heartrate results across conditions suggests that participants were in the same 'state of being' regardless of the testing condition /environment. It therefore follows that no one condition induced an adverse reaction as per recommendations in the literature, which notes that a sudden increase in heartrate valence would indicate a negative and/or defensive emotion (Guger *et al*., 2005).

The open comments were analysed thematically to offer further explanation of the data. In total, 78 comments were made by the 30 participants. They fell into four distinct themes:

- technological content: 33 negative comments
- technological content: 19 positive comments

- physical presence / props: 1 negative comment
- physical presence / props: 15 positive comments

When writing these observations, participants were comparing each environment and describing a relative experience. Because the Laptop condition has evidenced the least ecological validity in comparison to PEL and the Real environment, it is not surprising that 33 negative comments were received. These comments focused on the overall differences between the Real environment and the PEL and Laptop conditions. For example, *"outside isn't a controlled environment, you can't predict what is going to happen*". However, comments also revealed that PEL did induce relaxation and calm, for example, *"I felt calm and relaxed in both outside and PEL".* In all, there were 17 positive comments in this vein and 10 positive comments concerning the addition of props and sound to heighten the experience: one participant noted, *"The Log seat in PEL, breeze and field of vision was more like outdoors than Laptop".*

When analysing the face validity questionnaire responses and comparing each condition and the direction of travel, significant differences were found to exist between the three conditions. However, only by reviewing the nature of the data can direction of travel be detected and, on this measure, PEL was found to be closer to the Real environment, compared to the Laptop condition, in all four questionnaire responses. The data thus indicates that relative ecological validity was achieved, and that the results were moving in the right direction as per Deniaud *et al*. (2015) and Godley *et al*. (2002). It is evident that PEL does not create the same level of presence to the real-world environment; however, PEL achieves significantly higher presence than the Laptop condition, so it can be deemed as a halfway house when determining presence.

5.3.6 Discussion

The aim of this research is to explore the potential of using simulated environments early in the design process to enhance the effectiveness of prototype-based usability testing, and to uncover key requirements in the design of such simulated environments. The emphasis in this study concerns validation of an environment in preparation for usability studies rather than proof that a recreated environment will enhance usability studies.

The objectives covered in this study are:

1. *To understand the landscape and the literature concerning usability testing approaches.*
2. *To explore methods of recreating environments and how they can influence the requirements of the usability testing space.*

5.3.6.1 Objective 1: To understand the landscape and the literature concerning usability testing approaches.

The purpose of this study was to ascertain ecological validity in preparation for conducting usability testing in a simulated environment. Early work of Kaptein *et al.* (1996), then Godley *et al.* (2002), and more recently Deniaud *et al.* (2015), noted that, to ascertain ecological validity, there are two measures: absolute and relative validity. Absolute validity is recognized as being almost impossible to achieve when using technology to replace real-world experience. However, if results can demonstrate a direction of travel that is in line with achieving ecological validity, and validated against a real environment, then it can be used to ascertain if relative validity has been achieved. The results from this study suggest that relative validity has been achieved for PEL, and that ecological validity is higher in PEL (a simulated environment) compared to the Laptop condition. Ecological validity is important as it helps to ascertain whether a simulated environment can add value and or not when conducting usability tests

5.3.6.2 Objective 2: To explore methods of recreating environments and how they can influence the requirements of the usability testing space.

In terms of relative validity, the results from the four face validity questions showed that the Real, PEL and Laptop conditions were significantly different from one another. This highlights that PEL did not exactly replicate, nor induce the exact, absolute sense of presence as the Real environment. However, as this study was not concerned with validating absolute validity, this is not a concern at this stage. To understand why PEL was positioned in its ecological validity relative to the real environment, we can identify that this is partly because it is an artificial environment and not as responsive as the real world. For example, when the sun goes behind a cloud there is no way of predicting how long that is going to last. Therefore, to replicate it in PEL during a study, when all the scenes are captured beforehand, would be too complex. We were not able to capture the weather conditions on that day as the scene is prepared in advance and tested over a series of days. Nevertheless, any participant involved in future usability studies would not have these comparisons to draw upon, therefore this does not hinder the next stage of this research. When participants are asked to compare environments, they do so by looking at every detail, but ecological validity is about replicating enough detail to induce certain emotions. Therefore, the question is not 'Is PEL an identical replication of the real world?' Instead, we need to consider relative validity, i.e. is there a direction of travel from Laptop to PEL to Real. Based on the participant responses to the four face validity questions, there is a positive direction of travel between each condition, with PEL moving towards eliciting an emotional response similar to the Real environment.

The PANAS emotional state questionnaire revealed that relative validity had been achieved in the participants' emotional state, and that the Real and PEL conditions were not significantly different to each other; but that the Real and Laptop conditions were significantly different from each other. This suggests that PEL can induce similar feelings as the Real environment, but that different feelings are created in the Laptop condition. The open comments noted the sense of feeling surrounded—another term for immersed—in the Real and PEL environments and attributed this sense to the screen wrapping round them just like in the real world: *"PEL felt more immersive vs the laptop, due to the larger screen, detailed sounds"; "The PEL was a better experience as you can see the scenery around you"; and* "*Due to a larger screen you feel more a part of the environment in PEL; also the sounds/temperature etc. make it a lot more realistic".* There were also comments about a sense of calmness and being relaxed in both PEL and Real conditions, for example, *"PEL—the log seat, breeze and field of vision was more like outside".*

When translated to a usability study, this suggests that participants would be reacting in PEL in a similar way as they would in the real context of use, but not in the Laptop condition. This provides further evidence of the importance of conducting usability tests in a simulated environment and not just in a traditional laboratory. The findings of this study offer positive validation and suggest merit in recreating the context of use for usability testing conditions.

The open comments offer another interesting insight. There were exactly 34 negative comments and 34 positive comments attributed to both PEL and the Laptop environment conditions. These focused primarily on immersion, with comments highlighting presence being achieved because of the technological content. One key finding from the comments is that when asked to focus on the display/screen/environment, as was the case in this study, then the image and contents must, unsurprisingly, be high resolution. However, in Kneebone's distributed simulated environment, he created popup images of medical equipment, in areas where the participants' attention was not expected to be. He found that these peripheral images did not have to be of high resolution, but rather just good enough to hint at the surrounding environment; Keller & Stappers (2001) also noted similar findings related to what they called partial views.

5.3.7 Conclusion

The need to validate a virtual and or simulated environment was triggered by the literature that notes that simulations should be validated for the tasks they are intended to be used for. For example, in the automotive simulation research, Godley *et al.* (2012) notes how speed reduction was validated in a simulated environment, meaning that any future speed-related research would be ecologically valid in their designated simulator. Likewise, Deniaud *et al.* (2015) notes: "*each simulator must be validated for a specific use, as each experiment has its own requirements"* (Deniaud *et al*., 2015). The purpose of this first study was to ecologically validate PEL in terms of presence achieved, in preparation for conducting usability studies, with the first study demonstrating that relative ecological validity has been achieved. This suggests that findings from the next study will be of value,. Having conducted this study, it can be concluded that PEL is not able to achieve the same level of presence as the real-world environment; however, PEL achieves a significantly higher level of presence compared to the Laptop condition. What needs to be established in the next study is whether that level of presence is good enough for conducting usability studies. Therefore, the next study will respond to the third objective of this body of research, objective 3: *To ascertain the fidelity of a simulated environment that is most effective in discovering product design flaws early in the design process.*

5.4 Study 2: Optimum Environment

5.4.1 Aim

The broader aim of this study is to understand how much content is needed to achieve ‘enough presence’, as noted by Kneebone *et al*. (2011) and Burki *et al*. (2005). In particular, this study focuses on usability studies in a simulated environment to evaluate whether the fidelity of simulation affects the identification of key usability issues.

For this study, PEL was employed to test the usability of a low-fidelity prototype of a product, then under development, for cleaning the fuselage of a passenger aircraft. The product itself was being developed with the author’s input through a Knowledge Transfer Partnership (KTP). Window Cleaning Warehouse is an established distribution company with design and manufacture ambitions and a limited, but developing, product line of their own. Again, the focus of the KTP was on developing an innovation-led human-centred design process capability within the organization, founded on ethnographic research approaches. The lack of a deeply embedded design development process made it easier to develop something with the organization from the ground up. The KTP required the development of an aircraft cleaning system with low water waste—an ideal candidate for developing real-world test simulations. During the KTP, the author visited the St Athan airfield (in the UK) to gain a close insight into the problems with existing aircraft cleaning systems. Security was strict, with a sponsored escort, photographic ID and extensive paperwork required. The experience highlighted the complexity of conducting research ‘in the wild’ for environments of this nature. The prototype was tested in a series of four simulations of the St Athan aircraft hangar—each at a different level of fidelity.

A key aim of the tests was to establish the face validity of presence (Slater, 2003) for the four simulated environments, each with a different fidelity level, by conducting usability tests in each of them.

The study targeted three objectives:

1. ascertaining the product's overall usability using the SUS questions based on product testing
2. testing the face validity of each environment using the Likert scale questionnaire to ascertain the characteristics of an optimum product-testing environment
3. gathering qualitative data about the user experience of the product, based on observations and audio and video recordings of the task, to explore the '*why*' of the results

5.4.2 Conditions

The four simulations were:

Environment 1 (E1)**:** Projected images of the St Athan hangar; smell; sound recorded at St Athan; a 1:1 scale mocked-up fuselage section; and a prototype of the aircraft cleaning product. Participants wore a high visibility jacket (as required by all users of the St Athan hangar).

Environment 2 (E2)**:** Projected images of the hangar; a 1:1 scale mocked-up fuselage section; and a prototype of the aircraft cleaning product. No sound, smell or high visibility jacket.

Environment 3 (E3)**:** As *Environment 2,* but without projected imagery.

Environment 4 (E4)**:** Prototype of the aircraft cleaning product only

Table 10 Environment contents

High fidelity ↔ Low fidelity

Environment	1	2	3	4
Low-fidelity prototype	●	●	●	●
Fuselage prop	●	●	●	
Air hangar scene	●	●		
Ambient smell	●			
Ambient sound	●			
High-vis jacket	●			

A total of 24 participants were recruited using convenience sampling. Each participant experienced the same prototype and completed the same task, but after every sixth participant, the environment was changed. This allowed observations to be made on the impact different fidelity testing environments had on the behaviour of a participant and the types of usability issues identified and focused on, during and after the test.

The participants (17 males, 7 females) were undergraduate students aged between 18 and 25. English was the first language of all participants. A biography questionnaire ascertained prior cleaning experience ('Yes': 19) and aircraft hangar experience ('Yes': 3). Lack of aircraft hangar experience was not deemed to be problematic as it was probable the index product in question would be used by students in a summer job, i.e. the task was assessed as a low-skilled job, with specialist expertise or tacit knowledge not required.

5.4.3 Methodology

A mixed methodology was utilized in this study as outlined in Chapter 4.

5.4.4 Methods

The table below notes the particular characteristics of this study.

Table 11 Characteristics of the Optimum usability testing environment study

	No of Participants	No of participants in each condition	No of conditions	Research methods	Statistical analysis	Analysis method
Study 2 Optimum Environmen t	24	6	4	-Face validity -SUS -Open comments -Observation -Think aloud	Independent	Mode & Median

The test design followed a standard usability process with one facilitator controlling the test conditions to ensure they remained consistent. Each participant was asked for consent to participate via the relevant ethics paperwork before completing a biography questionnaire. The product test was designed to reflect the types of exploratory tests typically used early in the design process and so utilized low-fidelity prototypes or 'horizontal representations' of a concept that allowed for surface interaction usability to be ascertained.

Participants were asked to 'walk through' the whole user experience of the cleaning device rather than being given specific tasks (Rubin & Chisnell, 2008). Each participant was asked to assemble the product, clean the fuselage and put the product away. Participants completed two post-test questionnaires: the SUS, to ascertain prototype usability, and a questionnaire on the environment to assess the level of presence.

A pilot study was run to ensure the tests were appropriately set up (Rubin & Chisnell, 2008; Leedy & Ormrod, 2010). Changes made, as a consequence of the pilot study, included covering the support frame of the fuselage mock-up (as it was too distracting); re-positioning PEL's cameras to accommodate the unusually large props; and using panoramic rather than fish-eye images. It also became evident that PEL's floor needed to be covered to better simulate the concrete of the hangar floor—with cardboard used for the purpose. It was also apparent that the prototype needed more functionality to get full value from the tests and so a hose was added to make more visual the sense of the water tank and the cleaning poles.

5.4.5 Data Capture

Data was captured using quantitative and qualitative research methods. A think aloud protocol was used during the tasks and participants were audio and video recorded, with Observer XT software being used to capture the data on a common timeline for correlation and analysis.

Post task, the SUS was used to determine the usability of the low-fidelity prototype product, while the modified version of Kassab *et al.*'s (2011) questionnaire assessed the validity of the simulated environment. Face validity questions were presented in a 6-point Likert scale format, with participants asked to circle 1 for 'not at all', and 6 for 'very much'. Mode and median were used to identify preferences and opinions. The sample size was appropriate for a user test, as noted by Nielson Norman Group (2012), with qualitative data captured to help make sense of the quantitative data collected.

Face validity is a well-tested technique developed by Kelly (1927) to ensure answers relate to questions in the way they are intended to. In this case, face validity was used to determine which fidelity level (in the different simulated environments) provoked the strongest feeling of presence (Deniaud *et al.*, 2015).

Table 12 Face validity questions asked during the test

This simulation is a realistic representation of an aircraft hangar?	1	2	3	4	5	6
The simulation scenario is realistic?	1	2	3	4	5	6
The equipment used in the simulation is realistic?	1	2	3	4	5	6
The simulation 'felt' like being in an aircraft hangar?	1	2	3	4	5	6

5.4.6 Results

The study targeted three objectives:

1. ascertaining the product's overall usability using the SUS questions based on product testing
2. testing the face validity of each environment using the Likert Scale questionnaire to ascertain the characteristics of an optimum product-testing environment
3. gathering qualitative data about the user experience of the product, based on observations and audio and video recordings of the task, to explore the '*why*' of the results

Face validity analysis was first conducted on the level of presence achieved in the simulated environment from the questionnaire results; this was conducted utilizing a series of questions used by Kneebone (2011) to determine immersion and involvement in the environment, resulting in data that indicates the level of presence. The 1 to 6 Likert scale used made a score of 4 and above positive indicators and 3 and below negative indicators respectively. The mode and median results showed a positive inflection in terms of presence achieved in environments 1 (E1) and 2 (E2) when analysing the combination of face validity questions for each environment, while the lower fidelity environments 3 (E3) and 4 (E4) returned negative indicators (see *t*able 13 below).

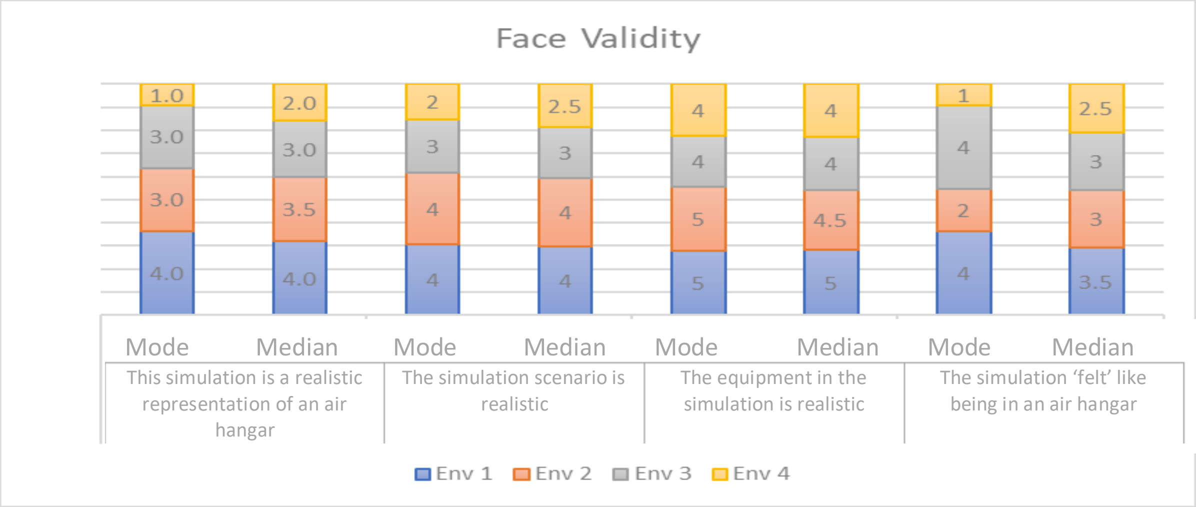
Table 13 Central Tendencies of the Face Validity questions and SUS by environment

Measure	**MODE**	**MEDIAN**	**SUS**
Environment 1	4	4	67.1
Environment 2	4	4	57.5
Environment 3	3	3.5	76.7
Environment 4	1	2.5	72.5

While the prototype under test remained constant in all four environments, the SUS results (see table 13) for E1 and E2 fell into the category of cause for concern' by the SUS scale definition (67.1 and 57.5 respectively). In contrast, while the level of presence was reduced in E3 and E4, overall usability scores of ‘acceptable’ were achieved at 76.7 and 72.5 respectively.

Analysing the overall face validity results, by summing the mode values across the four face validity questions for each environment, showed E1 to have the highest value, with a total mode value of 17 (see table 14 below). E2 and E3 both achieved slightly lower modes of 14 while E4 was clearly differentiated with a mode of 8. A review of the sum of median values across the four face validity questions, for each environment, showed the midpoint reference in E1 as 16.5, E2 as 15, E3 as 13 and E4 as 11 (table 14).

Table 14 Face Validity results by question



The average time taken for a participant to complete the study is shown in table 15 below. The average times were similar for E1–E3 (Range = 108 to 132 seconds), while the usability testing in the low-fidelity E4 was markedly different, with tests taking an average of 367 seconds to complete.

Table 15 Average task completion time

	Seconds
Env. 1	132
Env. 2	134
Env. 3	108
Env. 4	367

Observation and Audio Analysis

Observations of participants' interaction with the prototype and analysis of their qualitative comments highlighted a number of 'errors'—or, more accurately, ways in which the prototype's intended use was not understood by the participant. Careful analysis of the video and audio data revealed a series of key themes concerning how the environment affected participants' interaction with the prototype. For example, 'Confusion' was a term used by 50% of participants in E4 and 33% in E3 but only 17% in E1 and E2. E1 was also found to be better for uncovering serious design flaws. For example, participants in E1 found issues with the design of the product cleaning head. Audio data using the think aloud protocol included phrases such as *"the bottom squidgy kept flipping up and down".* This referred to the fact that the head would sometimes flip, exposing its edge, a problem categorized as a catastrophic design fault that might damage the aircraft's pressure hull. Analysis of the written comments and audio playback showed that E1 clearly caused participants to uncover usability issues around the cleaning head and the poles, while the main body / water tank component (the component requiring the least interaction) dominated the observations made in the lowest fidelity environment, E4. Perhaps even more interestingly, participants in E2 and E3 interacted with and commented on the entire prototype. Only one participant in E4 even 'turned the device on', with one other saying that they should, but not doing so. Three participants in E4 did not interact with the product prototype at all, but instead just talked about it from what they could see. In contrast, all participants in E1 and E2 'turned the device on’ and all participants also ‘cleaned the fuselage’, actively mimicking real-world actions, like actively cleaning a surface, as shown in figure 40.

The critical design fault with the cleaning head of the prototype was identified by 50% of participants in E1; 67% of participants in E2; 17% of participants in E3; and 0% in E4. A review of the audio data found only positive comments about E4, but a more careful examination of the type of comment found that they offered no insight into usability faults. Instead, they commented on potential design issues that had not been built into the low-fidelity prototype and observations tended to be speculative and lack depth. For example, the “*product was simple and easy to use so no complaints, maybe improve the aesthetics of the main unit*”. In E1–E3 the fuselage was the clear focus of attention, with all participants actively using the prototype to simulate cleaning the fuselage.

5.4.7 Discussion

The purpose of this study was to address objective 3 of the thesis.

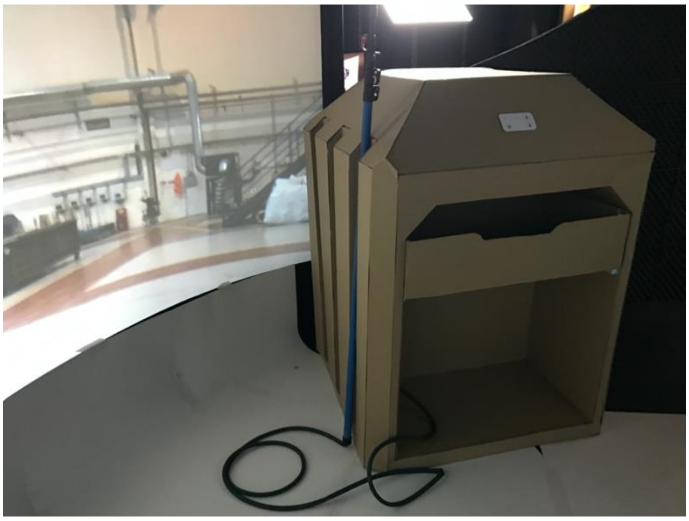


Figure 41 Low-fidelity prototype of the cleaning system

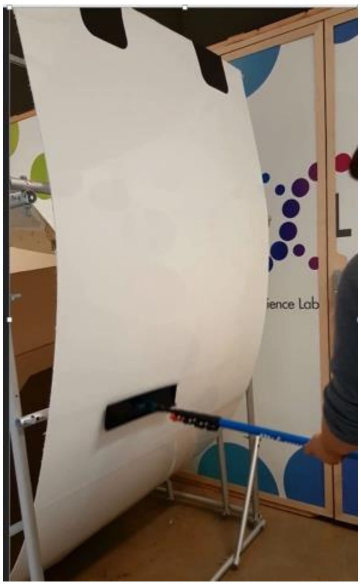


Figure 40 Fuselage prop in PEL

5.4.7.1 Objective 3: To ascertain the fidelity of a simulated environment that is most effective in discovering product design flaws early in the design process.

The face validity results demonstrated that the higher the fidelity of the usability testing environment, the stronger the measure of presence. This was somewhat as expected since it confirms Kassab *et al*.'s (2011) findings of how to heighten the sense of presence in a simulated environment via the selection of key objects from the real world. The difference between the results achieved in E4 and the higher fidelity environments E1, E2 and E3 is also less than surprising, but a nonetheless useful confirmation of expectation.

More surprising was the comparative high score of the prototype in the SUS tests in the highest fidelity environment, E1 compared to E2, as it was expected that a higher fidelity environment would identify more usability issues, and therefore have a lower SUS score. Nevertheless, while E1 created the greatest sense of presence, its SUS performance was worse than the much lower fidelity environment, E3. While E4 also scored relatively well in the SUS, analysis showed that participants tended to focus solely on the prototype rather than the interaction between the prototype and—in their case imagined—context to make their judgements on the product's performance. This resulted in participants looking rather than interacting. It would appear that E4's reliance on participants' imagination about the real-life scenario led to over-speculation and presumption. Therefore, while E4 did help gain user insights about the physical design of the product, it was less useful for gaining insights into key usability issues in context. As the lowest fidelity environment, E4 was easier to set up, but it delivered fewer insights and took much longer to deliver results. The average amount of time taken to complete a task in E4 was 367 seconds, compared to 108 seconds in E3—and with less useful results. It would appear that participants' lack of attention and focus in E4 was closely linked to the lack of simulation, which seemed to impede their ability to engage meaningfully with the product and thus give useful feedback. For example, E4 participants tended to get confused about where they were in the test, repeating actions they had already completed or failing to acknowledge that the product would need to be turned on. This confusion was not at all apparent to the participants themselves, however. In response to the face validity statement *’the equipment used in the simulation is realistic’,* participants in the lower fidelity environments E3 and E4 scored them positively at a mode and median of 4. One explanation might be ambiguity in the statement: participants may have been referring to the model as ‘equipment’ rather than the testing environment as a whole. Another statement, *‘the simulation felt like being in an aircraft hangar’*, garnered a different and unexpected result. Participant replies for E1 and E3 scored a mode of 4, while those in E2 scored a mode of 2. One explanation could be that the product had the worst usability feedback in this environment, so the low score might be a result of participant frustrations. The SUS scores show a more reliable pattern: as the fidelity of the simulated environment decreases, the overall perception of usability of the prototype increases with both E3 and E4 achieving SUS scores in the 70s (i.e. within the ‘acceptable’ range). The SUS scores showed usability as ‘acceptable’ in E3 when there was still a critical fault in the design (the potentially dangerous flaw with the cleaning head). This fault was not identified in E3 even though the fuselage was used, evidencing that E2 yields better prototype usability findings. E2 identified design flaws that needed to be found at this stage of the design process, distinguishing E2 as a more informative environment.

The trial proved inconclusive regarding the importance of high-fidelity environment details such as the use of the 'high-vis' vests and smell in heightening presence. However, E1 did achieve a higher face validity response to the statement '*the simulation feels like being in an aircraft hangar*' than E2. Since the vest and smell were the only points of difference, it is tempting to conclude that they contributed positively to the overall experience, but there is no specific evidence of this.

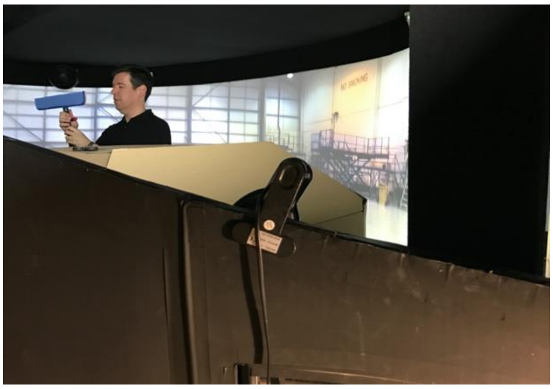
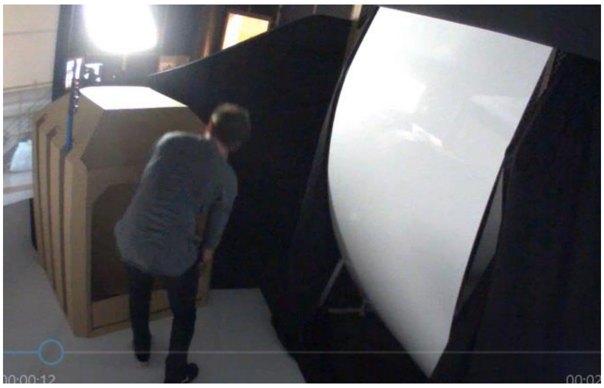


Figure 42 Usability test E2 in PEL



Usability test E3 in PEL

In summary, E4 is clearly an outlier and the usability results are not as powerful because of the lack of reference to the context of use. E1 and E2 achieved similar presence results and raised similar usability issues; however, the time for setup is arguably less for E2, but the results are equally powerful, noting that effort invested is better rewarded in E2. As noted by HIT lab (1997), Leffingwell (2002) and Powers (2004), smell has a short-term impact as participants adjust to it, meaning the time it takes to include it will not have a marked impact on the outcome of the usability issues found at this stage. With reference to Nielson Norman Group's work (2007) on effort against financial benefit, E2 offers the return required in identifying usability flaws early in the design process but saves on time in setting up a higher fidelity environment, such as E1. This is also in line with Kneebone's 'good enough' principle for creating a simulated environment.

The trials had one further interesting outcome: the designer of the low-fidelity prototype watched the product trials via a live video feed and made changes to the concept in real time and in response to user interactions with the prototype. This was unanticipated, but clearly potentially powerful, since he was literally designing in direct response to user feedback, unfiltered, and in real time. In other words, the line between testing and design became distinctly blurred.

5.4.7 Conclusion

The purpose of this study was to establish the optimum fidelity of a   simulated environment for the usability testing of a low-fidelity, active prototype. The findings showed that the two higher fidelity environments yielded improved presence compared with the two lower fidelity environments, as well as highlighting more usability issues with the prototype under study. However, time spent on extra detail such as smell and props, e.g. vests, will not necessarily deliver benefit.

It can therefore be concluded that simulating the context of use while testing a low-fidelity active prototype early in the HCD process significantly affects the type of usability issues identified, but that the effect of fidelity is far less linear than previously thought. The first surprise was that, although face validity was strongest in the higher fidelity environments, there was a marginal difference between E1, E2 & E3. A clear difference could be seen in the face validity associated with E4, the lowest fidelity environment, with E4 effectively being the experiment’s datum and a representation of a standard laboratory environment. This ‘contextless’ approach had the least impact when participants were trying to identify flaws in the physical aspects of the proposed design, as opposed to the interactions between the user, product and environment. The benefit of including visual cues as to the product’s intended environment of use was clear, but, as noted above, it was the fidelity of those cues that provided the study's most unexpected result. E1 achieved the greatest sense of presence and the expectation had been that this would translate into more and meaningful usability faults being identified from the product trials. From analysing the open comments and observations, more critical usability flaws, associated with the cleaning head, were identified in E2 and this is echoed in the low SUS score of 57.5. Although 11 comments were made concerning usability in E1, it fared slightly better than E2 in the SUS overall usability score of 67.1. However, this SUS score would still raise usability concerns, according to Bangor (2008), with the inevitable conclusion that higher levels of presence are not necessarily required in product testing scenarios. Future studies should seek to test the reproducibility of this intriguing result at greater scale, enabling a revisiting of Kjeldskov & Skov's (2014) conclusions on the 'When and How' of laboratory vs field testing, as well as Deniaud *et al*.'s (2015) 'Why'. Such work will give test moderators the appropriate tools to appropriately assess effort versus reward and select the most appropriate testing environment to identify critical usability issues early in the design process. The usability results from different environments in study 2 highlight that presence is important, even the level of presence reported in PEL relative to the transitional laboratory usability study setting.

5.5 Study 3: PEL Ecological Validity V2

5.5.1 Aim

Having ascertained the optimum fidelity level of a recreated environment for usability testing of low-fidelity, active prototypes in the previous study, the aim of this study was to once again ascertain the ecological validity of PEL so that future data gathered can be benchmarked as ecologically valid by ascertaining the relative validity of the environment vs the real-world context. This study is a replica of the first ecological validity study because during this body of work, the PEL underwent a move and complete rebuild that included a large new screen and enhanced, 4K back projection. In line with Deniaud *et al*. (2015) and Grey (2002)'s recommendations that research environments should be validated before relative ecological validity can be established, it was deemed appropriate that a second ecological study should be conducted to ensure that, like PEL2, PEL3 would produce data that could be used as a benchmark of ecological validity. Therefore, a revised ecological validation study was required, not only to compare results with the previous iteration of PEL, but to ensure the final study data was robust. The study was conducted in three different environments: the real context, viewing a laptop, and a simulated environment with the fidelity of Environment 2 in study 2 used as a guide for the setup characteristics of the PEL.

5.5.2 Conditions

This design of the study is identical to the first ecological validity study, with 30 new participants recruited using the convenience sampling method. The study again used repeated measures and involved participants experiencing three different environments: PEL, Real environment and Laptop, as per the first ecological validity study. Each simulation (PEL and Laptop) displayed the same dynamic image of the view experienced in the outside environment (figure 43,44,45). The study was piloted prior to engaging with the participants.

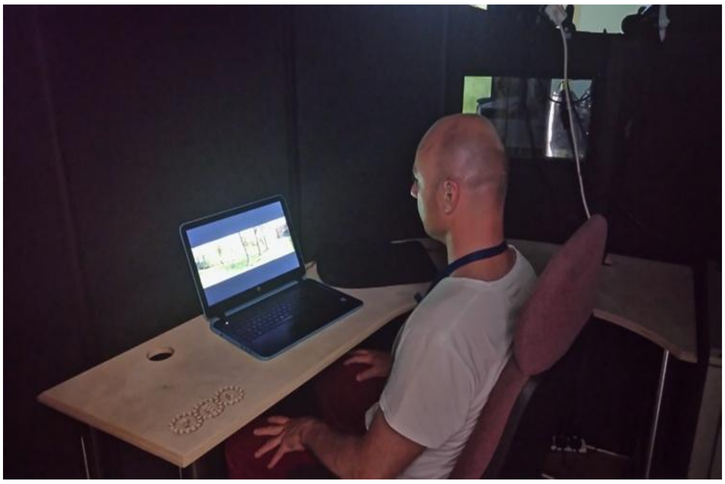


Figure 43 Laptop Study

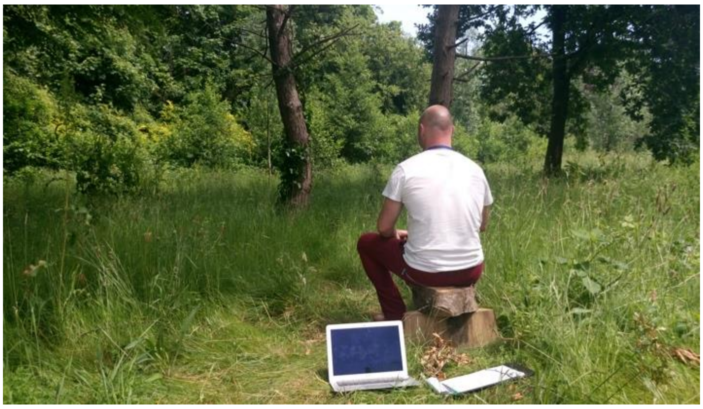


Figure 44 Outside study (Real environment)

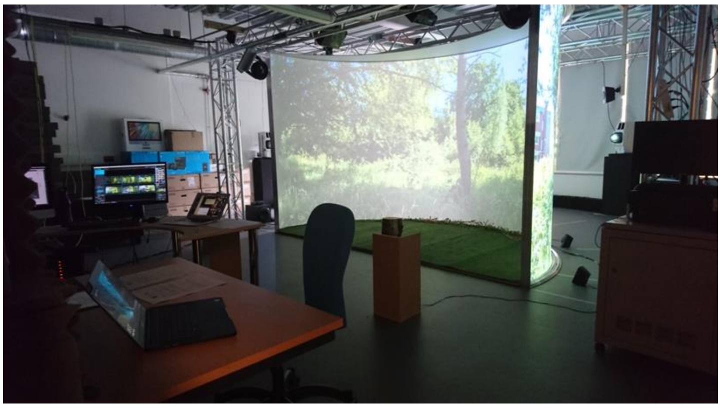


Figure 45 New configuration of PEL

Participant engagement was static. The same log seat was used in both PEL and the outside environment and artificial grass was again installed in PEL. The outside location was selected premised on its proximity to PEL and its new location was closer again, with no stairs to climb in order to avoid increasing participants' heartrates prior to studies. However, the Laptop condition mimicked a more traditional laboratory testing method of desk, chair and laptop.

5.5.3 Methodology

This study used a mixed methods approach as outlined in Chapter 5. Each participant sat in the environment for five minutes while connected to a heartrate monitor. At the end of each study, the participants were asked to complete a brief questionnaire on perception in Likert scale format and answer questions from a PANAS questionnaire (Watson *et al*., 1988). Open comments enabled the researcher to better interpret the quantitative data. As with study 1, there was no observational research, because participants did not physically interact with any product or prototype. This was because the study was focused on establishing ecological validity rather than usability testing in an environment—the focus of study 4.

5.5.4 Methods

The study was conducted with the assistance of 30 participants, 16 of whom were male and 14 female, with an average age of 40 years old. A counterbalance method, as captured in table 16, was used to ensure the data was not skewed by the order in which the participants experienced the different conditions. The order was as follows:

Table 16 Counterbalance approach

Participant	Testing order
1-5	Laptop, Real, PEL
6-10	PEL, Real, Laptop
11-15	Real, PEL, Laptop
16-20	Real, Laptop, PEL
21-25	Laptop, PEL, Real
26-30	PEL, Laptop, Real

The table below outlines the detail of this study.

Table 17 Study Detail

	No of Participants	No of participants in each condition	No of conditions	Research methods	Statistical analysis	Analysis method
Study 3 Ecological Validity 2	30	30	3	-Face validity questionnaire -PANAS -Heartrate -Open comments	Within subjects	ANOVA

Heartrate was again used as a physiological indicator of presence (Meeham 2001) because capturing heartrate is a non-invasive process. Five minutes of data was again obtained from each participant, with the first two minutes discounted while participants orientated themselves in the environment and the final three minutes used for dataset comparison. The same protocols around caffeine consumption were applied as in study 1 and, again, testing only took place in the morning. Participants also remained in a seated position in the tests. Repeated measures enabled the comparison of data within subjects.

The same set of questions (as in study 1) were used in each condition.

Face validity was studied using the following four questions:

1. I felt I was visiting the place in the displayed environment
2. I felt like I was just watching something
3. I felt like I was outdoors
4. I had a sense of being in the scenes displayed

PANAS questions were also used to help ascertain presence.

Table 18 PANAS

1	2	3	4	5
Very Slightly or Not at All	A Little	Moderately	Quite a Bit	Extremely

__________ 1. Interested

__________ 2. Distressed

__________ 3. Excited

__________ 4. Upset

__________ 5. Strong

__________ 6. Guilty

__________ 7. Scared

__________ 8. Hostile

__________ 9. Enthusiastic

__________ 10. Proud

__________ 11. Irritable

__________ 12. Alert

__________ 13. Ashamed

__________ 14. Inspired

__________ 15. Nervous

__________ 16. Determined

__________ 17. Attentive

__________ 18. Jittery

__________ 19. Active

__________ 20. Afraid

In this study, the aim was to ascertain ecological validity. As in study 1, the aim was not to try and achieve absolute validity, i.e. the results yielded in PEL being the same as the results yielded in the Real environment. Instead, this study should demonstrate if there is relative validity with clear direction of travel in terms of PEL performing better than the Laptop condition.

5.5.5 Results

Table 19 shows the mean ratings (and standard deviations) for all four face validity questions across the three conditions, along with their differences captured via the F-Ratio and P-Value.

Table 19 Mean (SD) results for the face validity questions

Question	**Real**	**PEL**	**Laptop**	**F -Ratio**	**Sig.**
Q 1	5.47 (1.332)	3.37 (1.326)	1.57 (.898)	74.154	< 0.001
Q 2	2.23 (1.794)	3.37 (1.608)	4.80 (1.690)	20.263	< 0.001
Q 3	5.83 (0.913)	2.93 (1.363)	1.53 (0.860)	115.070	< 0.001
Q 4	5.37 (1.520)	3.67 (1.470)	1.90 (1.185)	45.185	< 0.001

Figure 47 shows the graph of the mean ratings of the answers to question 1 (*'I felt I was visiting the place in the displayed environment'*). Note that RQ1 is the Real environment, LQ1 is the Laptop and PQ1 is PEL.

Question 1: *I felt I was visiting the place in the displayed environment.*

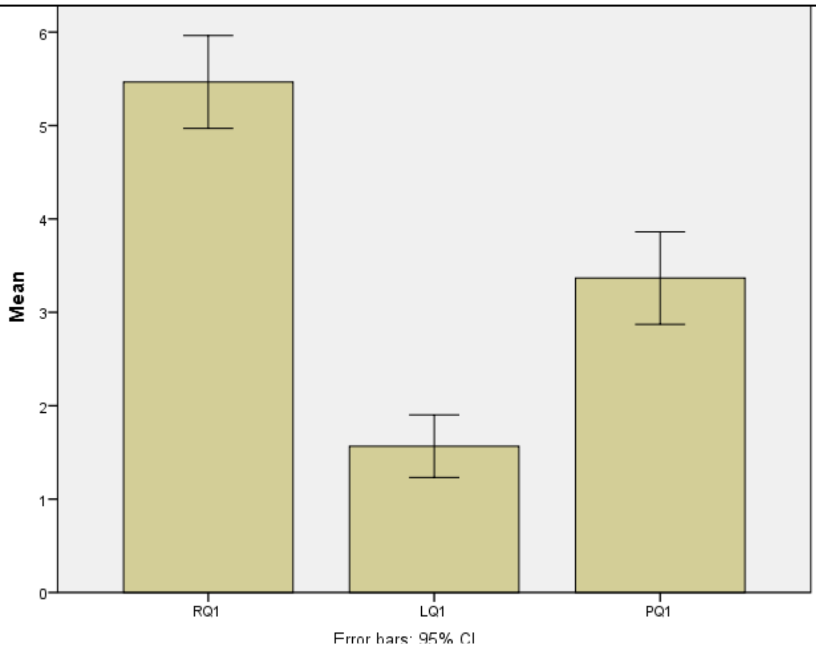


Figure 46 Question 1

The Mauchly's Test indicated that the assumption of sphericity has been violated, χ2(2) = 12.482, p = 0.002. Therefore, Greenhouse-Geisser corrected tests were reported (ε = 0.735). The results show that the response to question 1 was significantly affected by the condition, F(1.471,42.657) = 74.154, p < 0.001. Post-hoc pairwise comparisons highlighted that there were significant differences between all pairwise comparisons. However, the relative validity can be evidenced in the results, with a sliding scale of 5.47 mean rating for the Real environment, 3.37 mean rating in PEL and 1.57 mean rating in the Laptop conditions, resulting in participants 'feeling' less as if they were visiting the outside space.

Figure 48 shows the graph of the mean ratings of the answers to question 2 ('*I felt like I was just watching something*'). Note that RQ2 is the Real environment, LQ2 is the Laptop and PQ2 is PEL. The Mauchly's Test indicated that the assumption of sphericity has been violated, $\chi^2(2)$ = 10.275, p = 0.006. Therefore, Greenhouse-Geisser corrected tests were reported (ε = 0.765).

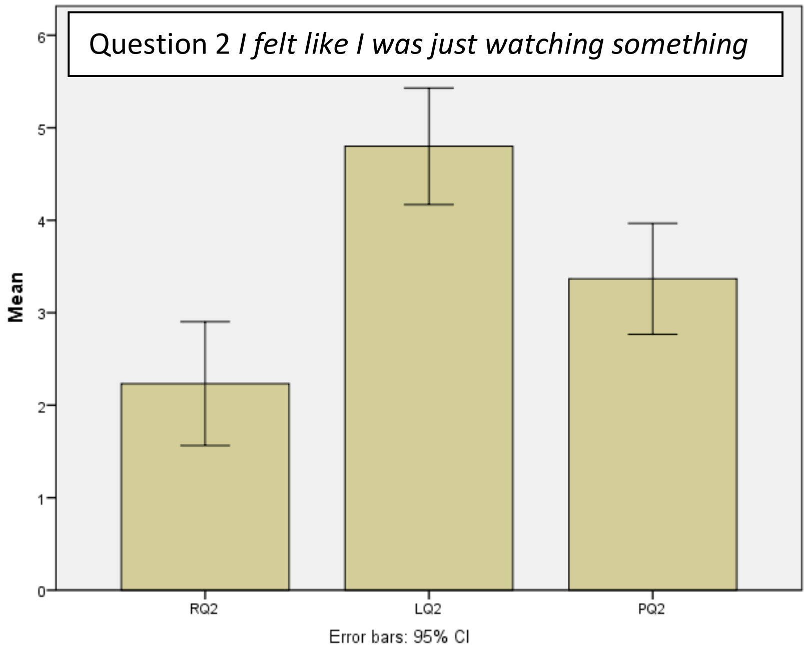


Figure 47 Question 2

The results show that the response to question 2 was significantly affected by the condition, F(1.530,44.371) = 20.263, p < 0.001. Pairwise comparisons highlighted that there were significant differences between the Real and Laptop conditions (p < 0.001), and between the PEL and Laptop conditions (p < 0.001). However, there was marginally no significant difference (p = 0.059) between the Real and PEL conditions.

As expected, the mean rating is low in the Real environment (2.23), i.e., on average, the participants did not feel as if they were just watching a scene. The mean rating increases in PEL to 3.37 and then to a high mean of 4.80 for the Laptop condition. Again, this suggests that relative validity has been achieved with evidence in a direction of travel, whereby more ecological validity is evident in the PEL condition compared to the Laptop condition.

Figure 48 shows the graph of the mean ratings of the answers to question 3 ('*I felt like I was outdoors*'), with a mean rating of 5.83 for the Real environment, 2.93 for PEL and 1.53 for the Laptop conditions. The Mauchly's Test indicated that the assumption of sphericity has been violated, χ2(2) = 10.207, p = 0.006. Therefore, Greenhouse-Geisser corrected tests were

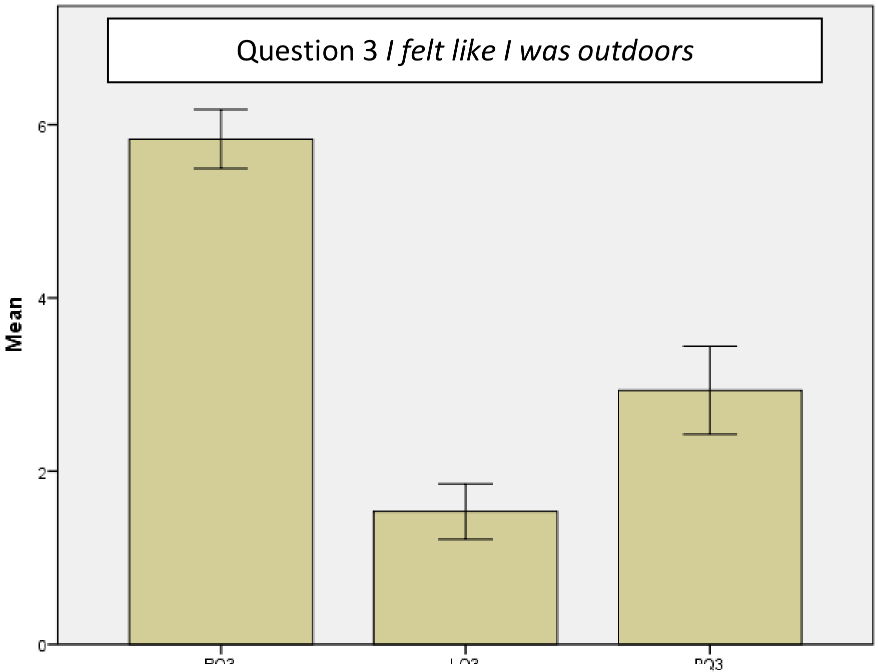


Figure 48 Question 3

reported (ε = 0.766). The results show that the response to question 3 was significantly affected by the condition, F(1.532,44.428) = 115.070, p < 0.001. Pairwise comparisons highlighted that there were significant differences between all pairwise comparisons. Again, this suggests that relative validity has been achieved.

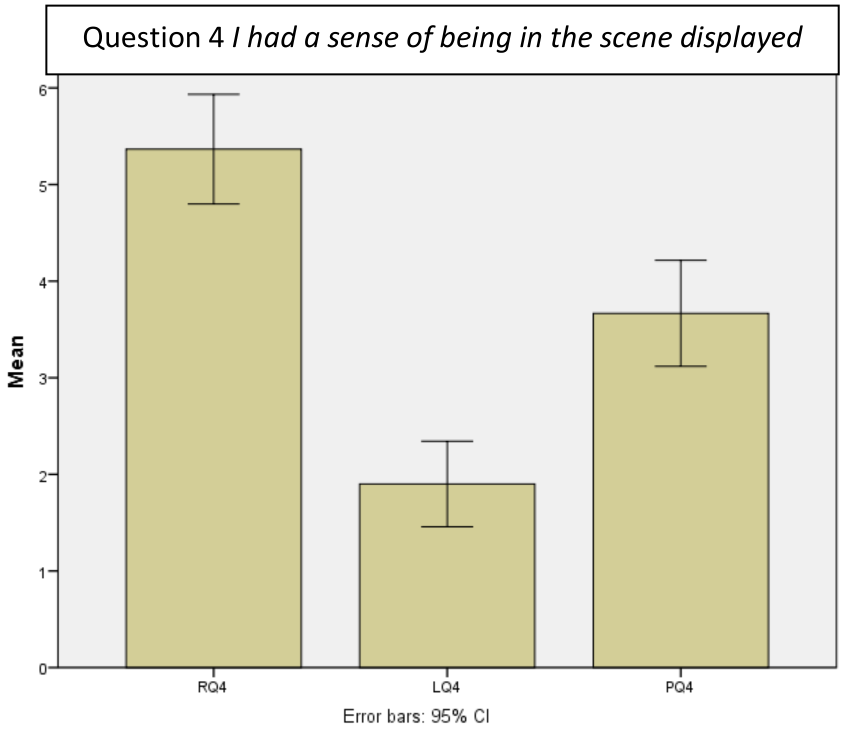

Question 4 *I had a sense of being in the scene displayed*

Figure 49 Question 4

Figure 49 shows the graph of the mean ratings of the answers to question 4 (*'I had a sense of being in the scenes displayed'*), with a mean rating of 5.37 for the Real environment, 3.67 for PEL and 1.90 for the Laptop conditions. The Mauchly's Test indicated that the assumption of sphericity has been violated, χ2(2) = 6.979, p = 0.031. Therefore, Greenhouse-Geisser corrected tests were reported (ε = 0.800). The results show that the response to question 3 was significantly affected by the condition, F(1.639,47.517) = 45.185, p < 0.001. Post-hoc pairwise comparison highlighted that there were significant differences between all conditions. Again, this suggests that relative validity has been achieved.

Further data used to measure presence is the emotional state of the participants in each condition, as captured by the PANAS.

Table 20 PANAS results

PANAS	**Real**	**PEL**	**Laptop**	**F-Ratio**	**Sig.**
Positive Mean	24 (8.02)	21 (7.133)	17 (7.23)	12.504	< 0.001
Negative Mean	11 (1.63)	12 (2.93)	14 (4.30)	8.058	0.001

Table 20 shows the PANAS results for the three conditions, both for the positive and negative means as does the graph in figure 51. For the positive mean, the Mauchly's Test indicated that the assumption of sphericity has been violated, χ2(2) = 6.570, p = 0.037. Therefore, Greenhouse-Geisser corrected tests were reported (ε = 0.827). The results show that the positive mean was significantly affected by the condition, F(1.654,47,967) = 12.540, p < 0.001. Pairwise comparisons highlighted that there was a significant difference in the positive means of the Real and PEL conditions (p = 0.007); and a significant difference in the positive means of the Real and Laptop conditions (p = 0.01). However, there was no significant difference in the positive mean between PEL and the Laptop condition (p = 0.071).

For the negative mean, the Mauchly's Test indicated that the assumption of sphericity has not been violated, $\chi^2(2) = 2.010$, p = 0.366. The results show that the negative mean was significantly affected by the condition, F(2,58) = 8.058, p = 0.001. Pairwise comparison highlighted that there was a significant difference in the negative means of the Real and Laptop conditions (p = 0.02); but no significant difference in the negative means of the Real and PEL conditions (p = 0.428) or in the negative means of the PEL and Laptop conditions (p = 0.061).

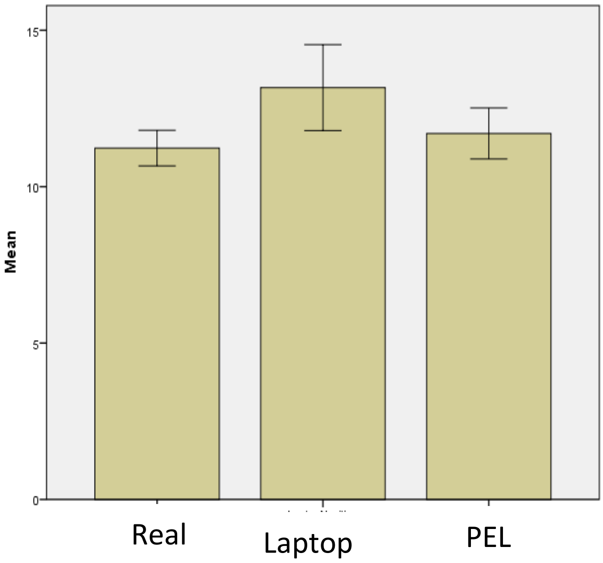
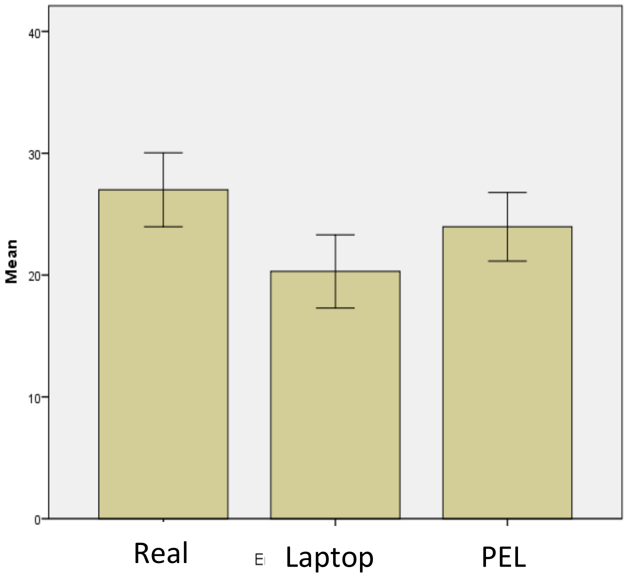


Figure 50 PANAS Negative



PANAS Positive

As in the first study, the positive means in the Real, PEL and Laptop conditions fall below Watson's (1998) benchmark mean for displaying positive momentary emotion (29.7). However, the negative means in all conditions fall below Watson's (1998) benchmark mean for displaying negative momentary emotion (14.8). This suggests that participants were in a relatively neutral emotional state in all conditions.

As with study 1, this study with 30 new participants has also highlighted no significant difference (p = .480) in terms of heartrate between any of the conditions, with 70.57 bpm in PEL, 69.00 bpm in the Laptop condition and 68.80 bpm in the Real condition. The mean difference in heartrate between the first two minutes (for acclimatization) and the subsequent three minutes showed a difference of 1.23 bpm in the Real environment, 1.57 bpm in PEL and 1.03 bpm in the Laptop condition. In other words, participants were in a similar state of arousal in each condition and acclimatized to each environment in roughly the same way.

The 83 qualitative participant comments were thematically analysed and placed in the four categories below:

1. technological content: 48 negative comments
2. technological content: 23 positive comments
3. physical presence / props: 5 negative comments
4. physical presence / props: 7 positive comments

Whilst there was a high occurrence of negative comments concerning technology or the lack of responsiveness to outside conditions, the positive comments also highlight a more positive picture of how the senses are fulfilled in the outside condition and how PEL compared. For example, *"outside I was able to listen and breathe in fresh air, though the breeze in PEL was nice"*. 23 positive comments were made concerning PEL and its relative performance to the Real condition. Participants commented that it made them feel relaxed, and the longer a participant was in the space the more they reported feeling like 'being there'. This is in line with the results in question 4, when the notion of 'being there' was asked in relation to the three conditions, noting a mean rating of 3.67 in PEL compared to 1.90 in the Laptop condition. The open comments repeatedly referenced characteristics that define presence: *"PEL made me feel like I was there and the sound of PEL made it more real"*; *"PEL felt like I was in the outside environment, rather than inside with the laptop"*; and *"The PEL screen wrapping around created a sense of being in a space".*

5.5.6 Discussion

The purpose of this study was to re-run study 1 for PEL3 to ascertain relative ecological validity of the newly configured laboratory. The notion of revalidation is taken from Deniaud *et al*.'s (2015) reference to validating an environment for a specific use. As PEL's new configuration was significantly modified from the previous iteration, it was deemed appropriate to make sure it was able to provide relative validity.

The face validity questions resulted in an interesting set of results. Unlike study 1, where there was a significant difference between the three environments across all questions, study 3 highlighted no significant difference between PEL and the Real environment in question 2: *"I felt like I was just watching something".* (Note that for questions 1, 3 and 4 there were significant differences between the three environments.) This result suggests that there is a notion of similar ecological experiences in the PEL and Real environment. The result, therefore, is important to this study, allowing the conclusion that PEL3 has an even closer ecological validity than PEL2. One participant stated that *"PEL felt like I was in the outside environment, rather than inside with the laptop"*.

The emotional state of participants, as measured by the PANAS questionnaire, showed no significant difference in the negative emotional response between PEL and the Real environment. The same was not true of the positive emotions. This possibly suggests that if stress or anxiety needed to be created, then PEL could engage this emotion, much like in the Real environment. Of the 83 comments, the 53 negative comments were primarily aimed at the Laptop and how it appeared to be a static image (although it wasn't) which made it feel unrealistic. The image quality of the 4K loop of film footage was noted again, but this time participants commented that it was poor in both PEL and Laptop, so participants were drawing closer comparisons to these two conditions during this study. Other negative comments concerned equipment and props, like the fan used in PEL to simulate wind, with participants complaining that it was too noisy and that it distracted them. Normally these negative comments would be indicative of a significant difference between environments; however, they were balanced by comments about how PEL was closely aligned to the Real.

In contrast to the differences between studies 1 and 3 found in the PANAS, the two studies reached very similar results when examining heartrate data with no significant results found. As in study 1, it is evident that PEL does not create the same level of presence as the real-world environment; however, PEL achieves significantly higher presence than the Laptop condition.

5.5.7 Conclusion

Relative validity in terms of presence achieved in each environment was achieved in both study 1 and in the newly configured PEL3 in study 3. There were evident enhancements found in the data in study 3. For example, in the face validity questions, one returned a reliable result that aligned PEL and the Real environment with no significant differences in the results. The positive inclination in one of the face validity questions does offer some insight that the experience of 'watching in PEL' was not significantly different to the real-world experience, and this is supported in the open comments: *"Outside I felt relaxed and alert. PEL—I felt relaxed and a bit more assertive, even though it was a screen"; "The PEL felt like I was in the outside environment, rather than inside with the laptop".* What is interesting is the reaction and volume of comments in this third study vs the first ecological study. They are very detailed and there are more of them. The key finding from this study is that relative validity has been achieved in PEL3 and the following study is of value as certain levels of presence have been achieved. Study 4 does not include a laboratory setting as data yielded in study 2 has already demonstrated that the optimum fidelity environment for PEL can highlight usability issues of low-fidelity, active prototypes not found in environments similar to traditional laboratory settings. The next study will ascertain if presence is 'good enough', and whether the same critical and key usability issues can be identified in PEL and the real-world context.

5.6 Study 4: Context, Optimum PEL vs Real

5.6.1 Introduction

Studies 1 and 3 established the level of relative ecological validity needed while study 2 investigated the optimum environment fidelity for identifying usability faults of low-fidelity, active prototypes early in the HCD process. The results from the first three studies indicate that the reconfigured PEL3 setup, based on insights gained from studies 1 and 2, is sufficient to create a sense of presence for participants. Therefore, when conducting usability tests of a product (prototype) in the simulated environment, PEL3, similar usability issues should be discovered to those found when testing a product in the real context of use. This final study seeks to validate these findings by comparing a working product in its real context of use with tests of a medium-fidelity, active prototype in PEL. A medium-fidelity, active prototype was chosen for the tests in PEL because they are typically produced early in the design process, where the focus of this research lies, while having more functionality than a simple mock-up typically used to explore aesthetics and basic interactions. Because medium-fidelity, active prototypes can facilitate reasonably complex levels of interaction, they allow research of sufficient complexity to make the usability findings informative, as recommended by Brehmer & Dorner (1993). Comparisons between real-world tests of a real product and the reverse-engineered medium-fidelity, active prototype in a matched fidelity environmental simulation (PEL3) will offer insights into the effectiveness of such testing environments early in the design process. The specific areas tested are those areas noted as the forgotten factors in laboratory usability testing (Dahl, Andreas & Svanaes, 2010):

- social
- physical
- psychological

5.6.2 Conditions

Selecting a product to use as a vehicle in these studies was a challenging task. Initially, a car-parking machine was going to be the focus of the usability testing process; however, these once complex machines have been subject to a number of usability refinements, meaning there were no longer sufficient usability states to allow adequate opportunity for fault finding. This study instead focuses on the usability testing of the nextbike, a national public cycle hire scheme that allows commuters access to a network of bikes around the city. It was selected for this study as it involves both physical and digital interaction and offered enough complexity to compare it in a prototype state with a fully manufactured product state.

The study is a comparison between two different conditions:

1. outside environment / real context of use with real product (nextbike)
2. medium-fidelity prototype of the nextbike in PEL

15 participants undertook testing in each environment, therefore a total of 30 participants were involved in this study. A static image was projected into PEL3 with appropriate traffic sounds (obtained from the real scene), and broadcast using the 3D ambisonics array to create a dynamic soundscape. The view projected on PEL3's screen ensured the prototype was in perspective relative to the other (virtual) bikes, with the appropriate line of sight to enhance the sense of being in the real environment.

5.6.3 Methods

A mixed methods approach was used, as outlined in Chapter 4. A convenience sampling approach was taken with an independent population for each condition. Table 21 captures the specific characteristics of this study.

Table 21 The Characteristics of the Real vs PEL study

	No of Participants	No of participants in each condition	No of conditions	Research methods	Statistical analysis	Analysis method
Study 4 Real vs PEL Validation	30	15	2	-PANAS -SUS -Rolf Molich -Observation -Think aloud -Time*	Between subjects	Independent t-test.

Methods deployed in this study included eye tracking to allow a clearer observation of the user experience with a focus on areas of interest. It also enables the Rolf Molich performance data to be captured during the study while detailed analysis could be conducted post-study. The think aloud protocol was also used, with participant comments captured on the eye tracking audio recorder. This allowed for more nuanced user experience to be captured during the study and offered in-the-moment insights into the user's thoughts and feelings. Unlike in previous studies, face validity was not utilized in study 4 since it was not the purpose of the study to ascertain presence. Likewise, heartrate was not measured. Because usability studies are skewed by participant learnability (Field, 2106), study 4 measures between rather than within subjects. Because the two participant populations are independent, absolute heartrate offers no usable insight and so all emotional recording was captured via PANAS. The SUS proforma was also implemented to identify overall usability issues in a given design, offering a clear percentile that identifies whether or not a significant difference exists between relevant data. The Rolf Molich performance rating was used to complement the SUS, enabling the researcher to delve down into the detail to ascertain exactly where the faults lie. Molich's method enables the researcher to translate the qualitative observations into quantitative data that can be analysed via a T-test analysis method to identify if significant difference can be found.

During this study, participants were asked to complete three tasks:

1. locate the code to unlock the bike
2. digitally unlock the bike
3. physically unlock the bike

Prior to commencing the study, ethics consent was gained and a bio questionnaire completed so data could be gathered on the demographic of the sample size, but also to ascertain if any of the participants had interacted with the product prior to the study. If they had, they risked skewing the results with learnt behaviour, and therefore were excluded from the study. Once the participant had completed the bio data, the eye tracking was calibrated to each participant. They were then handed a bag as a prop and asked to complete the study, whilst talking through their actions. As the moderator, actions were viewed via the eye tracking data, but also by observing the participant at a distance. On completion of the study, the participants were handed an iPad and were asked to complete the PANAS and SUS. The Rolf Molich performance rating was completed post study via the footage obtained from the studies. The Qualtrics survey software was used to collate the data collected to allow easier storage and management. Rolf Molich's performance rating was utilized as described in section 4.4.3.5; and as noted by Molich and Dumas (2006):

Catastrophe: *user is unable, refuses or solves the task incorrectly.*
Serious problem: *user is significantly delayed (1-5 Minutes) but manages to complete the task.*

Minor Problem: *user is briefly delayed (the user experiences a problem but corrects themselves reasonably quickly—less than 1 minute).*

Success: *user completed the task without problems or delay.*

Woolley (2008).

The task performance of each participant was put into categories (catastrophe, serious problem, minor problem, success) using observation along with how long each task took to complete. The categories were then converted to interval data based on the following assignment of numerical values: 0 = success, 1 = minor, 2 = serious, 3 = catastrophe.

The results were analysed using the AVOVA method. With 30 participants in total, and two conditions, the t-test was used to highlight the calculated differences between the two conditions and to determine any significant differences (Rowntree, 2018).

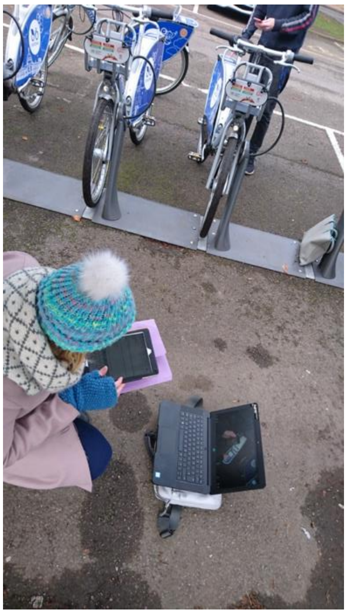


Figure 51 nextbike Usability testing

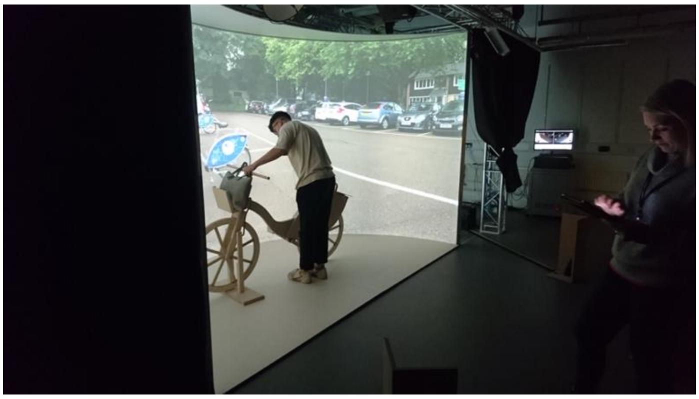


Figure 52 Simulated environment in PEL with prototype of nextbike.

Figures 51 and 52 capture the set up.

5.6.4 Results

Table 22 captures the raw performance rating data by participant for the three tasks:

1. finding the code on the application
2. inputting the code to digitally unlock the bike
3. physically unlocking the bike

Table 22 Rolf Molich's performance raw data

RM performance rating		Catastrophe	Serious Problem	Minor Problem	Success
Finding code	PEL	4	3	1	7
	Real	0	1	1	13
Digitally unlocking	PEL	6	1	0	8
	Real	1	4	1	9
Physically unlock	PEL	11	0	0	4
	Real	12	0	2	1

Table 23 shows the mapping of the raw performance rating data as numerical values (0 = success, 1 = minor, 2 = serious, 3 = catastrophe).

Table 23 Performance rating captured as numerical values

	RM	Real	PEL	Sig (2 tailed)	T -Test
Mean (SD)	Code finding	0.20 (0.561)	1.27 (1.335)	0.010	2.854
	Digital unlock	0.93 (1.100)	1.4 (1.454)	0.331	-0.991
	Physical unlock	2.53 (0.990)	2.20 (1.37)	0.452	0.762

The 'time to complete' task was also measured in seconds and captured in table 24.

Table 24 Time in seconds to complete the task

	Task	Real	PEL	Sig. (2 tailed)	T -Test
Mean (SD)	Code finding	31.33 (42.132)	32.67 (30.684)	0.922	0.099
	Digital unlock	53.67 (78.762)	52.00 (48.384)	0.945	-0.070
	Physical unlock	22.80 (22.365)	21.20 (13.029)	0.813	-0.239

The independent t-test results highlight that in task one (finding the code in the application), there was no significant difference between the time taken to locate the code in PEL (M = 32.67s, SD = 30.684) and the Real environment (M = 31.33s, SD = 42.132), t(28) = 0.099, p = 0.922. Figure 53 is an example of this task in PEL.

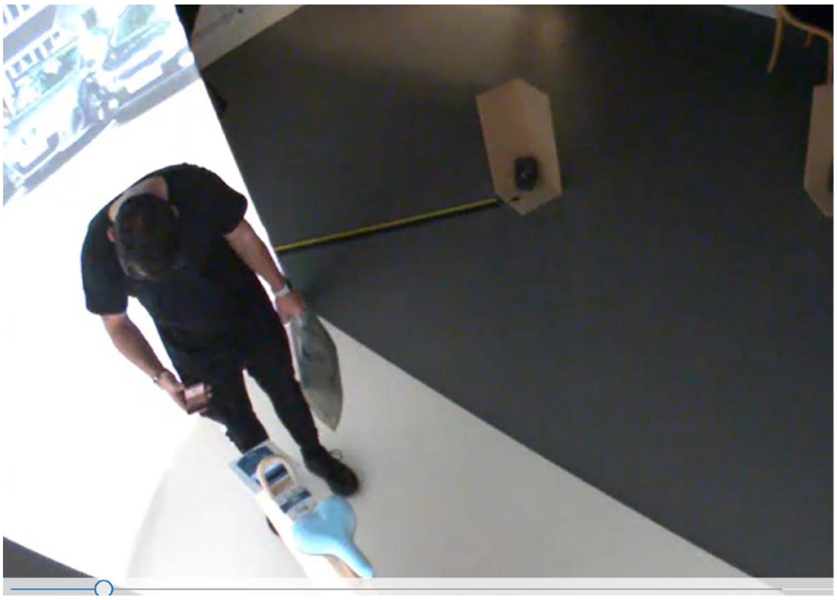


Figure 53 participant inputting code in PEL

When analysing the RM performance rating for the same task in PEL (M = 1.27, SD = 1.335) and in the Real environment (M = 0.20, SD = 0.561), there is significant difference between the ratings, t(28) = 2.854, p = 0.010. When analysing the observations as to where the usability issues arose in PEL, more usability issues were identified with locating the code in the application; and participants were quick to assume the bike number was the code they should input. Only two participants made this error in the Real environment. Participants who made errors in both environments assumed the task was complete and were ready to move to inputting the code digitally; however, more errors were unnoticed in PEL.

The independent t-test results highlighted that in task two (inputting the code to digitally unlock the bike), there was no significant difference between the time taken to input the code in PEL (m = 52.00s, SD = 48.384) and the Real environment (M = 53.67s, SD = 78.762), t(28) =

-0.070, p = 0.945. When analysing the RM performance rating for the same task for PEL (M = 1.4, SD = 1.454) and the Real environment (M = 0.93, SD = 1.100), the independent t-test notes no significant difference, t(28) = 0.991, p = 0.331. From the observations made, there were more usability errors identified in PEL concerning the usability of the interface when inputting the wrong code. However, six participants made the same error in the Real environment but were able to amend their actions and input the correct code, whereas in PEL, seven of the participants were unsuccessful in the task.

The independent t-test results highlighted that in task three (physically unlocking the bike), there was no significant difference in the time taken to physically unlock the bike in PEL (M = 21.20s, SD = 13.029) compared to the Real environment (M = 22.80, SD = 22.365), t(28) = 0.239, p = 0.813. When analysing the RM performance rating for PEL (M = 2.20, SD = 1.37) against the Real environment (M = 2.53, SD = 0.990) there was no significant difference, t(28) = 0.762, p = 0.452. From the observations from both environments there were many usability faults found in both environments, with the main usability issue centring around participants not locating the bike lock in the designated holder, instead opting to let it hang down and become a safety hazard.

When comparing the performance of the three tasks between PEL and the Real environment, the stark contrast was in the code-finding task. Here there were seven incidences noted as serious/catastrophic in PEL compared to only one in the Real environment. Analysis of observations made via the eye tracker found that faults fell into four key themes:

- clicked the delete button instead of the OK button
- hit the ‘wake screen’ but thought they were entering a number, and subsequently registered the wrong code
- did not locate the physical lock in the lock holder

- entered the bike number not the digital code

27% of participants in PEL entered the number located on the bike itself as the unlock code rather than the supplied secure code that is unique to each scenario. This contrasts with 13% of participants who did this in the Real environment. While the percentage of participants making the error was very different, the same design flaw was discovered in the PEL environment.

A significant difference was also noted in the task to digitally unlock the bike, with seven 'severe faults identified in the PEL environment against five in the Real environment, while eight participants in PEL and nine in the Real environment were successful in completing the task. A common usability issue was clicking the OK button rather than the delete button, with participants in both PEL and the Real environment making this mistake (27% participants in PEL and 13% in the Real environment). Another clear usability issue identified in both environments was associated with 'waking' the digital display: participants would enter the number without realizing that the first tap of the number pad 'wakes it up'. Consequently, the first number received was in fact the 2nd digit of the 4-digit code, leading to the participant receiving an error message without knowing why. More incidences of this were captured in the Real environment (20%) versus PEL (13%), but again, PEL uncovered the same usability errors that occurred in the Real environment using a medium-fidelity prototype, typically used early in the design process.

From a usability perspective, the task of physically unlocking the bike disclosed a 'catastrophic' failure in both environments with no significant difference between PEL and the Real environment. By analysing the observations, it is evident the fault was the same in each task. It consisted of unlocking the bike and removing the lock cord, but not knowing how to store it away appropriately ready for transit. This error occurred 11 times in PEL and

12 times in the Real environment. Each time the participant left the lock cable hanging, which might prove dangerous in real use.

On average, participants spent a similar length of time completing the task in PEL and the Real environment, with no significant difference found in each of the tasks when comparing both environments. It can therefore be stated with confidence that the PEL/medium-fidelity prototype accurately predicted the time within which a user might be expected to complete the task in the Real environment.

The details of the PANAS results are shown in table 25. The PANAS data was analysed using an independent t-test, as well as by Watson *et al.*'s (1998) method.

Table 25 PANAS results

Mean		+	-
	Real	26	16
	PEL	27	15

The independent-sample t-test was conducted to compare the positive and negative emotions as a consequence of conducting a usability test in the Real environment and PEL. There was no significant difference in scores for negative means (M = 15.07, SD = 6.431) for the Real environment and the negative means (m = 16.47, SD = 5.939) for PEL, t(28) = 0.619, p = 0.541. There was also no significant difference in the positive means (M = 25.67, SD = 6.863) for the Real environment and PEL (M = 26.80, SD = 7.739), t = 0.424, p = 0.675.

The PANAS results indicate a similar emotional response to both environments. PEL achieved a mean of 27 in a participant's positive emotional state compared to a mean of 26 in the Real environment; and a mean of 15 in a participant's negative emotional state compared to a mean of 16 in the participants positive emotion state in the Real environment. Using Watson *et al.*'s (1998) analysis method, the results suggest that participants did not exceed their benchmarked mean for displaying positive momentary emotion (set at 29.7) in either the Real environment or PEL. However, participants did exceed the benchmarked mean for displaying negative momentary emotion (set at 14.8) in both the Real environment and PEL. The fundamental aspect here is that results are evidencing absolute validity in terms of emotional state in both environments.

Participants were also asked to engage in the think aloud protocol. However, all but two participants in the Real environment spoke out loud during the tasks; and no tangible quotes were gathered to help understand behaviour in that environment. Nevertheless, having captured the eye tracking data, observations were recorded and these recordings provided an insight into participant behaviour in lieu of participants thinking out loud. Inside PEL, the protocol was utilized by all participants, with some interesting quotes gathered that helped the researcher understand participant behaviour. For example, in the code-finding task, one participant noted *"I don't know what I'm doing"* and *"I think I have done it"* on physically unlocking. In fact, they had not completed the task and had left the lock hanging down, as highlighted in figure 54.

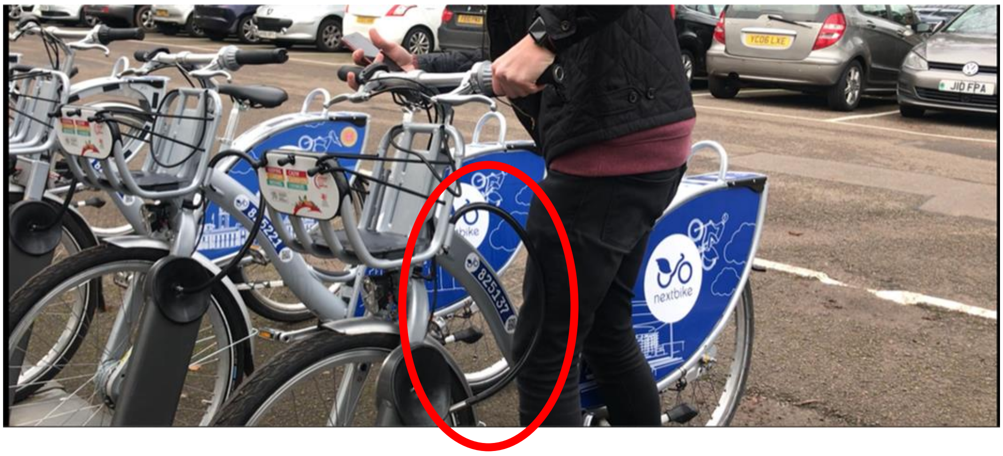


Figure 54 Lock left hanging down

The SUS scores for overall usability were 50% in PEL and 61% in the Real environment. Therefore, usability results in both environments produced a SUS score below the threshold score, defined by Bangor *et al.* (2009) as 70%, with Bangor *et al*. highlighting that *"anything below a 70 had usability issues that were cause for concern".*

5.5.5 Discussion

Of the three usability tasks, two of them noted a significant difference between the Real environment and PEL in terms of performance rating and, at first glance, those would appear disappointing. However, once a comparison of the types of errors identified had been undertaken, a different picture emerged. The starkest significant difference was in the first task of finding the security unlock code via the app to input into the bike. By analysing the think aloud protocol data it became evident that during the studies in PEL, participants did not engage with the app first, instead going directly to the bike interface. Therefore, all 4 catastrophic failures were accompanied by think aloud comments that illustrated their confusion. Further analysis in both environments also found that catastrophic faults included entering the bike identification number, found on the bike frame, rather than the app security code, indicating a usability problem. Four participants made this error in PEL and two in the Real environment. This is, therefore, a positive outcome; the reason there was no catastrophic fault in the Real environment was because the bike system would not physically move on to the next step, so the participants knew they had made a mistake and would try via trial and error to rectify the mistake. Therefore, it was the constraints of the actual product that enabled the corrective action; however it did not mitigate against the level of anxiety a participant demonstrated when trying to complete the tasks in both the Real and PEL environment. It may be that the prototype fidelity provoked participants to make errors because the primitive nature of the model compared to the working product meant they were unable to take appropriate corrective action. Again, however, PEL correctly identified a real-world usability fault at the medium-fidelity prototype stage.

Participants had the same positive and negative emotional responses to both environments noting no significant difference. As relative ecological validity had been established in the third study, and now by ascertaining that the emotional state of the participant in both PEL and the Real environment (when conducting a usability study) is the same, we can gain confidence that the level of ecological validity is a contributing factor in this positive outcome. From a participant behaviour perspective, the think aloud protocol worked well in PEL, but much less so in the Real environment. These findings complement the work of Woolley (2008) who found that laboratories are more reflective spaces. PEL appears to maintain this quality. Only two participants engaged in the think aloud protocol in the Real environment. A potential reason for this might be the level of self-consciousness in the Real environment. Although the participants took, on average, the same amount of time to complete the task as they did in PEL, with no significant difference in the time, their body language was more 'head down' and rushed, possibly because the eye tracking equipment made them self-conscious. This was the only difference in participants' behaviour between PEL and the Real environment.

5.5.6 Conclusion

The areas under consideration included the forgotten factors in laboratory usability testing (Dahl, Andreas & Svanaes, 2010):

1. social, including interruptions and self-consciousness
2. physical, the interaction with the prototype
3. psychological factors, presence and emotional state of being

The question is: were these factors teased out in the PEL environment?

The answer is yes. This study was conducted to identify if PEL added value to the usability testing and in short it did. The nextbike is a fully manufactured, functioning product that has inherent usability issues, which the PEL setup was able to identify using a medium-fidelity, active prototype and matching environmental simulation.

Both emotional response and engagement with task were like for like, as noted by the time to complete each task and the PANAS results. On balance, more could have been done to include interruptions in the PEL environment to add another layer of complexity, such as having someone ask for assistance with a nextbike. This should probably be conducted as a separate usability test, however, so that one set of participants would work through a task with interruptions while a different set attempted the usability task in more straightforward circumstances. This could be effectively integrated into a usability testing scenario guided by Nielson (1993) of testing five participants and employing the principle of 'little and often'.

The study took longer to complete in the outside environment as the weather needed to be clear and the study was postponed on a number of occasions (it rains frequently in Wales!). In addition, the equipment kept malfunctioning in the cold weather, which appeared to have an impact on the connection between the eye tracking and the laptop. Further limitations of the Real environment included the think aloud protocol, with participants being reluctant to engage. Unlike the data capture process designed into Qualtrics, whereby participants progressed through the questions until all were completed, think aloud was mainly ignored or forgotten about. Regardless of the number of errors identified in each condition, it is evident that the same errors were picked up in PEL and the Real environment.

In responding to the final objective of this thesis: *To validate the findings and detail the outcome of this research exploration*, it can be concluded that a recreated environment was validated against the real context of use. It emerged from the findings that PEL did deliver more usability flaws, and although they were disproportionate to the flaws found in the real context of use, they were nonetheless still identified as usability issues in the final product. Therefore, although PEL exaggerated the faults, it did find them and as they were found using a medium-fidelity, active prototype, it can be noted that this would have been found in the ideas phase of the HCD process. The one telling result was the identification of the flaw with the physical lock in the real context of use. This is a serious design flaw that was evident in the real bike in the real context, and was also identified as a fault in PEL. Therefore, although significant differences appear in the amount of usability issues found in PEL (compared to the Real environment), it still validates that there are benefits to recreating the environment and usability testing early in that environment. This study did not include a comparison against a laboratory, i.e. no context scenario testing environment, as that had been achieved in the second study and it was recognized as not identifying key usability issues early enough due to the fidelity of the context. Another key outcome of this study is the emotional state of the participant when engaging in a usability study, there is no significant difference found in a participant's emotional state between PEL, an artificial environment and the real environment. This suggests that you can create a simulated environment to induce the same level of anxiety or emotion as a real environment. Creating the same level of emotional stress for a usability test scenario early in the HCD process, to reflect how a person would interact with a product in the real context of use, is extremely important as this can reveal usability errors often missed in traditional laboratory settings.

To conclude, PEL identified the same critical usability flaws as the real context using a medium-fidelity, active prototype; therefore, PEL plays an important role in facilitating context of use in the early stages of the HCD process.

Chapter 6 Discussion / Findings

Chapter 6 Discussion / Findings

6.1 Discussion

The aim of this research was to explore the potential of using simulated environments early in the human-centred design process to enhance the effectiveness of prototype-based usability testing. In addition, it was to uncover the key requirements needed in the design of such accessible mixed-reality simulated environments to facilitate the identification of critical usability issues with an 'active prototype' early in the HCD process.

Behaviour and attitudes (Abra *et al.*, 2005) that impact upon culture, and the social and physical environment (Dahl, 2010), are important components of usability testing, so creating an environment that teased these elements out was the purpose of the collective studies. To do so, it was necessary to find out what was 'good enough' to engage participants' emotions through a set of studies that assessed emotion and behaviour. The fundamental difference between a laboratory usability setup and PEL was the level of presence achieved, noting that presence is the psychological, perceptual consequence of being immersed in an environment and the involvement in a task when situated in a space (Witmer & Singer, 1998).

Maintaining physicality in the design and testing processes was another key factor in this work, as it relates to involvement and the task conducted in the environment. When married with immersion, this increases believability and engagement, two of the dimensions required to achieve presence. Physical interaction with objects and environments forms part of our thinking process and so through usability testing it helps induce the emotions and behaviours expected in the real context of use. In design practice, physical 3D outputs of various forms are integral decision-making tools that enable designers to continually and iteratively enhance a design in development (Booth *et al.*, 2013). What was evident from the literature was the acknowledgment of prototypes that had active physicality attributes (Hare, 2015), and how this definition enables designers to benchmark an appropriate prototype fidelity according to the needs of the usability study, but also how prototypes are an essential part of the HCD process (Hallgrimsson, 2020). However, what was missing was the role simulated environments (such as PEL) could play in finding usability issues in active prototypes, only previously highlighted during tests in the real context of use. There was also a need to understand how these simulated environments should be designed to increase the sense of presence.

Studies 1 and 2 found that PEL did not create the same level of presence as the real-world environment; however, PEL achieved significantly higher presence than the Laptop condition. Therefore, simulated environments such as PEL could be deemed as a halfway house between a traditional laboratory environment and a real-world environment. PEL provided more cues (compared to a traditional laboratory) to enhance ecological validity of a real-world scenario, therefore it appeals to three of the dimensions of achieving presence: the physical space; naturalness; and reducing the negative effects. These studies achieved relative ecological validity for PEL.

Study 2 found the optimum fidelity environment context that is sufficient, i.e. has ‘enough’ presence, so participants are able to identify critical design faults in a low-fidelity, active prototype early in the HCD process. Study 2 provided the guide for recreating context, while study 4 enabled the refinement of that study and a validation of the effectiveness of simulated environments for uncovering critical usability issues.

Study 3, like study 1, reinforced the findings that PEL3 does not create the same level of presence as the real-world environment. However, PEL3 achieved significantly higher presence than the Laptop condition and was still ecologically valid in terms of its enhancements. The data collected from this study paved the way for the final study, a validation of the findings.

Study 4 applied the attributes of presence learnt in the previous studies and then added the factor of usability testing. The purpose of this study was to determine if the same level of critical usability issues were found in PEL using a medium-fidelity, active prototype, compared to the actual product in the real world. Study 4 did not need the inclusion nor comparison of the laboratory, as this has been evaluated in study 2, where it was determined that there was insufficient presence in a laboratory that did not allude to context. Consequently, valuable usability issues were missed by participants as they focused on anticipating imaginary faults rather than identifying experienced faults.

A key insight learnt from these four studies is that to effectively recreate context of use, there has to be an appreciation of how the fidelity of an environment can enhance or negatively affect a person's sense of presence. Too much, and it can hinder believability and subsequently naturalness and a participant could experience negative effects (and consequently skew the usability results); too little and critical usability issues are missed early in the HCD process.

Below, each one of the original objectives are addressed in turn.

6.1.1 Objective 1

To understand the landscape and the literature concerning usability testing approaches.

Context of use is a key concept in the literature and is even enshrined as a key component for consideration in the legislation for medical device product development. However, in order to fully embrace Norman's (1998) findings that it is the product at fault and not the user, when a product is difficult to interact with or leads to error, it was appropriate to revisit the usability testing process with context of use early in the HCD process as the focus. Doing so builds on the recommendations of Hare *et al*. (2014) and Rubin (2008) that product designs should be tested with intended users early and repeatedly throughout the design process. Both these theories align very well with Nielsen's (2018) recommendations: little (as in participants numbers) and often (as in throughout the design process). This research took those recommendations on board and sought to develop the knowledge needed to build context of use into simulated spaces that would allow these approaches to be applied with the added value of real-world believability in a laboratory. The research also used the insights gained by Dillon *et al.* (2000) and Deniaud *et al.* (2015) where they define the four dimensions of achieving presence, one of them being 'naturalness'.

The literature concerning laboratory and field usability testing attests to the fact that field testing has its place, but also explains why it was not the answer to testing during the early stages of the design process where this thesis's focus lies. Brehmer & Dorner (1993), for example, report on the complexities of field testing while also reflecting on the lack of complexity in laboratory testing. Woolley *et al*. (2013) noted that both field and laboratory usability testing have a role and there are strengths and limitations of both approaches, allowing the author of this research to recognize an opportunity to create a hybrid approach to laboratory and field testing. The laboratory setting is a safe space to explore product development enhancements, but does not offer sufficient complexities to truly test a participant's response to uncontrollable contextual experience; whereas field studies are too complex for prototypes in deployment, meaning there is a 'jar' of fidelity, which impacts on the participant's ability to engage in a meaningful way. A 'jar' could be described as fidelity of context environment being at odds with the fidelity of the active prototype, consequently impacting on a participant's expectations of the prototype and, therefore, skewing the usability findings.

In summary, key attributes of the literature are:

- Context is significant and although noted in literature, it can be difficult to recreate in an accessible manner.
- There is an absence of an appreciation of presence when recreating environments for usability testing in the Product Design literature. This is significant as more of the industry seeks to include virtual reality in their development process, without acknowledging how to maximize the use of this world.
- There are four dimensions of presence, and presence is induced when immersion and involvement are engaged (Dillon *et al.*, 2000).
- Woolley (2008) noted that field and laboratory usability testing scenarios have a role to play and that one is not better than the other, but instead should be accessed at different stages of the design process.
- Field studies offer the advantages of ecological validity; however, Kjeldskov & Skov (2014) highlighted that it should not be about which environment is better than the other when usability testing, but instead when the studies are conducted and how they are conducted that is significant. This meant there was an opportune moment to identify early context usability testing but to redefine the 'how' and introduce key components of context into the usability study at an earlier stage.

6.1.2 Objective 2

To explore methods of recreating environments and how they can influence the requirements of the usability testing space.

When exploring the notion of recreating environments, it was evident that the practice was limited in the design discipline. For example, Living Labs—normally high-fidelity spaces designed for longitudinal studies later in the HCD process or the experience prototyping methods of IDEO. However, those in fields such as computer science, medical training, and cognitive science are well versed in the psychology and human factor implications of manipulating technology and/or props to recreate environments. Thus, borrowing from other disciplines helped form the foundations from which the researcher could develop new approaches for the design sector, using practical knowledge of how to integrate simulated environments into the design process. This knowledge developed by other disciplines is largely absent from current design thinking. The theories of immersion and presence, for example, are not generally used in design circles.

Findings from the studies in this thesis highlight key attributes required to conduct usability studies in a recreated context, and when coupled with Deniaud *et al.'s* (2015) work on the factors that contribute to the believability of the recreated environment (known as the four dimensions to achieve presence: physical space; engagement; naturalness; lack of negative effects), then an optimum recreated context of use can be created. In this case 'optimum' refers to 'what is enough' to induce sufficient presence to ensure participants believe they are 'there' in the intended context and are using the product as they would in the real context of use. Study 2 highlighted that although high-fidelity environments can appeal to the participants' senses, like smell, sound, props etc., when these attributes are pared back to some extent, the impact on the usability findings were negligible. The other advantage of paring back the level of fidelity of the simulated environment is that it is cheaper and faster to create. The literature (Ramic *et al*., 2007; Brikic *et al*., 2009; Chalmers *et al*., 2009a; Brikic *et al*., 2013) also notes that smell does not add value when inducing presence especially when offset with the complexities of setting up and sustaining smell, as participants adjust and no longer pick up the scent. Having conducted interviews with experts from industry, it was evident that whatever was developed had to be accessible so that rapid usability testing can be conducted early and often during the HCD process. Findings from these interviews highlighted that relatively straightforward mixed-reality setups, like that facilitated in PEL, had the ability to capitalize on the advantages of context with the confidentiality of a laboratory and the conditions of a controlled research environment.

The research by Dillon *et al.* (2000) and Deniaud *et al.* (2015) were interesting and informative but less relevant because the research did not compare the simulated environment with the real environment and therefore did not include a study on ecological validity (so relative results could be yielded). However, it allowed an opportunity to contextualize the computer science research in the discipline of HCD and apply it to ascertain if a recreated contextual environment can offer improved insights to improve the development of a product early in the HCD process. The findings from the studies suggest that it can. Attributes taken from other disciplines included the data capturing methods to assessing face validity and the level of presence achieved.

The four dimensions of presence were explored through further literature research, through the requirements gathering research work and then evaluated in the four studies. Physical space was researched by attending a workshop at a simulated medical laboratory. Attributes that included key visual and audio cues that grounded the participant in the environment were applied to PEL. For example, the use of a fuselage in study 2 and the use of a bag and the sound of traffic in study 4. A tour of the Welsh Millennium Centre coupled with the literature on creating medical training environments (Kneebone, 2011; Kassab *et al.*, 2011) ensured the hierarchy of props was established, predicated on where the trainee's action and/or attention would be. In the studies, this translated to fresh grass cutting and logs in studies 1 and 3, with a log seat that was used in all conditions so there was a consistency of message in each environment. In study 2, where different fidelities of an environment were tested to explore the optimum fidelity (or 'enough' as cited by Walshe (2005)), attention was on the fuselage and the prototype. The final study put the emphasis on the image of the nextbikes and the sound of moving traffic as well as a bag that was incorporated into the task. The emphasis was on using physical props when they were a part of the task, or virtual images if they were significant to creating immersion, but not required to be interacted with. As seen in Kneebone (2011) and the Cardiff Simulation Laboratory (CSL), both used visual cues that require attention during the task. For example, the CSL used hospital pillows to directly reference hospital wards and Kneebone used banners with printouts of actual hospital equipment. These attributes of attention relate to the physical space and naturalness, i.e. immersive properties of an environment contribute to the naturalness of the physical space, two of the dimensions of achieving presence.

6.1.3 Objective 3

To ascertain the fidelity of a simulated environment that is most effective in discovering product design flaws early in the design process.

The findings from study 2 demonstrate that results from laboratory usability testing, which are devoid of context, do not find all critical usability issues (that exist in real-world usage) early in the design process, as noted by Woolley (2008). This is because the stimuli that would normally impact on a user's mental load in the real world, changes their behaviour and how they interact with products. Study 3 was positioned to identify the optimum ('enough') stimuli to induce the participant's response to a real-world scenario and this was achieved by establishing the amount of presence that could be created in a simulated environment. The literature noted that experience prototyping was a tool that could enhance the creative process premised on hinting at real-world experiences (Buchenau & Suri, 2000). In addition, the literature on virtual environments and psychology therapy (Slater (1999, 2002, 2003, 2004, 2013); Witmer & Singer (1998); Deniaud (2015); and Walshe (2005)) noted that simulated environments are able to induce presence. However, there was no literature on the optimum environment to induce presence when usability testing a product in development. Study 2 established the optimum setup by analysing data that contributed to achieving presence. This data included the face validity questions to ascertain that the content was as expected in the recreated environment, thus measuring immersion. This was combined with the use of observation, a think aloud protocol, heartrate analysis and a questionnaire on participants' emotions to estimate the level of involvement in the environment. Collectively the immersion and involvement data highlighted the level of presence achieved in each environment. As a similar number of critical usability issues were identified in both the two highest fidelity environments, it was concluded that the environment that required the least resource (of the two) was deemed 'accessible' and 'enough'; however, it was the fourth study that would validate these findings.

To ascertain the validity of any study in a recreated environment, it is important to ascertain the ecological validity of the space, as per Deniaud (2015). Each iteration of PEL was validated utilizing the face validity question and PANAS, as it was in these studies we could ascertain the level of presence that could be achieved relative to the Real environment and a simple laboratory setup. The literature is clear that relative ecological validity (relative to the real-world experience) is a fundamental component for assessing the effectiveness of a recreated environment and how presence can be achieved, so that participants behave as they would in the real world. Studies 1 and 3 of PEL2 and PEL3 both found relative ecological validity and, as expected, both offered more believability compared to the Laptop condition. They both also solicited positive feedback from participants regarding the similarity between PEL and the real-world environment. From these findings, the studies that followed (studies 2 and 4) were conducted on the premise that relative ecological validity was able to induce sufficient presence in PEL to conduct usability testing in a replicated context-of-use environment. However, the literature does not offer a golden value to determine ultimate ecological validity, only that it is in the direction of the real-world experience (Deniaud *et al.*, 2015). These studies did not wholly replicate the ecological validity results of the real environment, as expected, and absolute validity was not therefore achieved, but nor did it need to be. This is because relative benchmarking is the critical criterion. The ecological validity studies demonstrated that the emotions provoked in the Real environment were not significantly different to those invoked in PEL and this was later verified in the final study, when the positive and negative emotions of the participants in the real and PEL environment were not significantly different. The emotional state of a participant, as determined by the PANAS data gathering, was an important component of this study as it offered the point of difference in the design usability literature and was used by Witmer (1998) to ascertain presence in army training virtual environments.

Another interesting finding was participants' response to the face validity questions in the different environments. Face validity questions were used to ascertain the impact on presence achieved in each environment. The ecological studies, 1 and 3, were conducted approximately one year apart, because PEL was relocated and rebuilt. In that time there were some changes; the outside environment had changed slightly with the grass having been cut and, although it was around the same time of year, the temperature fluctuates. The same image had been used to validate both PEL2 and the newly built PEL3 with the rationale that it offered consistency, but by then it was slightly out of date. On reflection, this was probably an error that contributed to variance in the data.

Study 2's exploration of what was 'enough' also resulted in key findings. When just in the laboratory setting, with no indication of context, participants took longer to complete the tasks and the results were not as meaningful, suggesting that the laboratory alone was not enough. When PEL included smell and props such as a 'high-vis' jacket, no value was added to the usability result. This confirmed findings by Ramic *et al*. (2007), Brikic *et al*. (2009), Chalmers *et al*. (2009a) and Brikic *et al*. (2013), that smell does not add value, but does add complexity to the setup time. It can therefore be concluded that an environment providing 'enough' context does not require smell. Study 2 also made it evident that a certain level of complexity is required to allow critical usability faults to be found.

These findings from the first three studies were used to set an appropriate level of fidelity and complexity for study 4, which found that PEL was able to find real-world design issues using a medium-fidelity prototype in a matched environment.

6.1.4 Objective 4

To validate the findings and detail the outcome of this research exploration.

Study 4 was positioned as a validating study, whereby it sought to identify if critical usability findings could be identified in a simulated environment using an active prototype. The results highlighted that you can identify the same critical product usability flaws in a simulated environment with a medium fidelity, active prototype as those found in a final product in the real context of use. Therefore, study 4 did validate the findings identified in study 2.

Kassab *et al.'s* (2011) and Daniaud *et al.'s* (2015) work found that content, attention, and ecological validity are key considerations when creating a simulated environment. This research has borne those findings out. By conducting a usability study and utilizing these requirements, more profound and critical usability results can be uncovered early in the design process because the added complexity makes participants feel like 'they are there, they are in' the real environment; therefore, they experience presence, as noted by Steuer, Slater and Kneebone—a contributing factor towards ecological validity.

It is noted in the literature review that Carulli *et al*. (2013) highlighted a need for a clear approach in usability testing and identified three requirements that should be considered when testing a product with its intended user. They are:

- basic requirements
- technical performance requirements
- attractiveness requirements

With these in mind, the recreated environment can be configured to facilitate these needs, with the caveat that for low- to medium-fidelity prototypes at least, technical aspects may not be fully or exhaustively replicated. An advantage identified in PEL versus the field study was that the think aloud protocol was not used by the participants in the field study, even when prompted. This result echoes the findings from the early interviews conducted with industry experts, who said that participants feel self-conscious in the real environment. This is important, as the purpose of a usability study is to collate research and enhance the product in development. Therefore, any lack of engagement in the protocol would hinder the collection of data.

The second study involved cleaning of a commercial passenger jet, with the attendant security restrictions meant that obtaining participants for a field study would be complex and resource hungry, if not impossible as each would need to pass security checks. This meant the 'little and often' testing that Nielson and Norman refer to would be almost impossible if not unsustainable and definitely not accessible. The medical industry too is dogged by the complexities of ethics and *in situ* resourcing, so simulated settings are an ideal alternative for recreating context for usability testing in both spaces.

The studies that involved the real environment, studies 1, 3 and 4, all encountered issues due to the unpredictable nature of the weather. When conducting user studies outside, in the real context of use, it was evident that the researcher was at the mercy of the Great British Weather, which frequently delayed the studies—a challenge when participants have been recruited in advance of a study. PEL offered a stable environment that can be controlled with the added values of context, which meant that every study scheduled in PEL was able to be conducted and with no interruption of schedule.

It is predominantly study 4 that meets objective 4. Participants were asked to conduct a usability task which, as noted in the computer science literature, would require attention in a recreated environment and have the potential to enhance a participant's emotional state if Dillon *et al.* (2000) and Deniaud *et al.*'s (2015) principles (as mentioned in section 6.1.2) on achieving presence were deployed. It aligned not only between PEL and the Real environment, but also with the literature for validity of results. These results highlight the importance of attention in a recreated environment and potentially the need for using a more focused task in usability studies rather than the commonly used 'walk through' approach. This is because the structure encourages attention, albeit at the expense of allowing participants' minds to wander, which is also useful for gathering insights.

Findings established that participants' experience in PEL aligned with their experience of the real world, effectively validating PEL for usability testing purposes. This finding is important as Blanford *et al.* (2010), who note that it is behaviour in context that impacts on how products are used *in situ*, therefore to ascertain findings that can replicate emotion and consequently behaviour early in the HCD process is in itself insightful.

6.2 Guiding Principles

Having completed the studies, a set of guiding principles were established on the requirements of recreating an environment for usability testing purposes.

- ➢ Establish the task(s) to be completed.
  - o The tasks should be determined by the usability to be scrutinized under the test conditions. This could be looking at the product usage as a whole (as in study two) or breaking down the overall task (e.g. renting and unlocking a nextbike, as in study four) into smaller sub-tasks to evaluate usability issues in detail for particular sub-tasks.
- ➢ Consider the content used for immersion.
  - o For example, where is the attention? What is being interacted with? The device where the interaction happens should be a physical representation,

while other areas of attention can be either digital or physical as long as they are not overstated. For example, study two used a physical prototype of a cleaning device as the main object to be interacted with. However, as the cleaning device was to be used with an aeroplane, a supporting physical prop that had the same visual cues as a fuselage was also used. The background environment was captured digitally through sound and a still image of the aircraft hangar.

  - Consider the use of sound to create an atmosphere and to increase the intensity when married with a static image. For example, study four utilised a static image of a road scene, accompanied by background traffic sounds to show a typical context of use, as the nextbikes are normally available for rent by a roadside.
  - Consider video if a still image is not sufficient. For example, naturally quiet settings may require movement to enhance immersion. You can mitigate the negative dimension of presence by ensuring there is a hierarchy to the mixed media. Hierarchy is related to where the participant's attention is in the recreated environment. If direct interaction is required then a physical representation should be made available. For example, in a recreated hospital, if a product under test is to be used on a patient lying on a bed, then a physical bed may be used as well as the physical prototype. However, equipment that is not interacted with directly could be image based through either card, printed or digital representations.

- Consider the use of back projection.

  - Back projectors prevent shadows, but if you have to use front projection consider the positioning of the projected media to reduce the effect. With back and front projectors, you also need to try and mitigate the possible negative effects of image distortion through the careful positioning of the projectors.

- The moderator should be out of sight.
  - The sight of a moderator could inadvertently distract or emotionally affect a participant and affect the usability testing. Therefore, it is important that the moderator is out of sight and cameras and audio equipment are used to capture the activity in the space.
- Always run a pilot study first.
  - Always pilot a recreated environment first before running your main study to help in working out how best to refine the content (immersion) of the space. For example, consider at what point the context is disproportionally distracting from the task. A pilot study can also highlight possible weaknesses in your usability testing process.
- When to use simulated/recreated environments.
  - There is a time and place for utilizing a simulated/recreated environment for usability testing, and that is typically when mid-fidelity prototypes have been developed. However, as a product prototype is refined throughout the development process, a fully working, high-fidelity prototype should ideally be tested in a classic field-testing scenario.

Chapter 7 Conclusion

Chapter 7 Conclusion

This body of research contributes to design praxeology: the study of process and methodology with the intention to enhance the designer's experience of the human-centred design process by integrating context and complexity in a meaningful and accessible way that complements field usability studies and replaces the traditional laboratory usability study. At the initial stages of this research a question was posed: "*What is the optimum fidelity for a user-testing environment in order to meaningfully inform design decisions early in the design process, before a design team has committed to a particular design path?"* That question can now be answered. The optimum fidelity of a usability testing environment is one that acknowledges the product's intended context of use and does so by being matched to the fidelity of the prototype. As the design process progresses so should the fidelity of the design until the prototype is resolved to a level of a high-fidelity prototype that can be tested in the real context of use. A recreated context should offer enough stimuli to impact on a participant's mental mode that increases as the HCD process progresses. To achieve this, the simulated environment should be constructed with consideration for attention and the creation of an environment that enables participants to feel like they are situated in the context of use, by capitalizing on how to enhance presence in a usability context of use.

This research has benefitted greatly from other fields, including computer science, psychology, medicine and Living Labs. Literature in these fields have really helped the researcher to understand the requirements of recreated environments that could not have come from the design field alone.

Early on, it became clear that highly sophisticated laboratory setups are not required. Rather, there are a number of basic key setup considerations that allow a laboratory to be accessible and effective and to comply with Kjeldskov and Skov's (2014) theories of 'how' to usability test. Key findings are that:

1. To achieve presence requires an image of the context, ambient sound that dominates the real context, an active prototype under test and props that would require attention from the participant.
2. Presence is best established by identifying where participant attention is required and generating appropriate background imagery good enough to suggest the real space.
3. Including consideration of the participant's entire field of view will aid immersion.
4. Controlling the user testing space and making it feel safe and confidential is important, so setting up instruments to unobtrusively capture data for post-test analysis is critical.
5. Using sound makes a space dynamic and provides appropriate real-world complexity without being too onerous or time-consuming to set up.
6. Working with a minimum of five participants and testing often is best practice.
7. Investing in key attributes of an environment and then adding complexity to the simulated space as the design process progresses is the most effective way to proceed, and key attributes include sound, projection, observation equipment.
8. Matching the fidelity of the environment and the fidelity of the prototype is critical.

The eight areas above are what Walshe (2005) referred to as 'enough' to recreate an environment in which to usability test products early in the human-centred design process.

Recreated environments are not a panacea, but they have a role in the design process. This research shows that, properly set up, they work effectively early in the design process and can uncover the types of critical usability issues that in the past have often not been found until a product is produced and used in context.

7.1 Limitations

There are limitations to the work, some of which will form the basis for future research, while others are simply observations.

There were a number of challenges with this body of research. For example, in order to ascertain an optimum environment for usability testing there needs to be a comparative study with the real environment. However, true to the literature, the limitations imposed by a need to obtain access and maintain confidentiality have meant that the products selected as usability testing vehicles had to be reconsidered on a number of occasions. One study was going to include a medical product; however, the limitations of testing in the actual context of use meant that the final validation study could not happen and so the partially complete data had to be left out of this body of work.

By way of an observation, over the duration of this research project, the way information was captured improved. For example, study 1 involved capturing all the participant data on paper, while the final study was conducted on Qualtrics so all the data was held in one place, making analysis and organization and management vastly easier.

When viewing these results, it should be noted that the participant populations were different and also that some studies were conducted between subjects and others within subjects. While this means that no cross analysis can be robust in statistical terms, in qualitative terms it does offer insights that form the foundations of future work.

Another limitation was the lack of literature that indicates the threshold of ecological validity, making the analysis of data predicated on a subjective measure of relativity. Although the product's selection for study 4 worked well in terms of its proximity to participants and for not being an obtrusive or difficult replicate, it was not the first-choice product. A product that would have induced clearer stress in context would have been preferable to further demonstrate emotional state and consequently support the notion of presence achieved. For example, a medical product would have been preferable, but due to the complexities of ethics, testing in context and subsequently gaining access to the appropriate target group for usability testing, it was not possible. Further studies should include multiple variations of environments to reinforce the findings.

Another limitation of the work was a lack of actors to put under test the 'social' factors of a recreated space. Such studies would have added too many variables, but might benefit from future study, now that an optimum environment has been established.

7.2 Future work

The role of actors to create social constructs in a recreated environment is an area of interest for future research. Having distilled clear results concerning the identification of critical faults, adding variables, such as an actor, would be a useful next step to further understand social interaction's impact on usability results, including in high-stress environments. Another interest is connected to how to make recreated environments portable. PEL was a static setup, and although it had the benefit of being a flexible space that can be set up for any context, a portable setup might be a useful development. Such work would build on Kneebone's distributed simulation lab for training surgeons but used to develop a series of guiding principles for a portable setup. Further work might also include fully trialling Nielsen's (2018) 'little and often' theories and benchmarking fidelity vs results on a sliding scale. This could be tested by conducting larger studies using a medical product as the vehicle and comparing PEL, laboratory and real setting. There is likely always to come a time in the design process whereby 'in the wild' testing is the best approach. Exactly what these conditions are will also be the subject of future work.

7.3 Contributions to Knowledge

Contributions to knowledge include:

- emerging evidence that simulated environments have a key role early in the development process, uncovering critical usability issues in a product within its context while still at the early prototype stage
- emerging evidence that simulated environments are capable of delivering a degree of presence close enough to real-world environments to deliver critical contextual design testing with many of the advantages of real-world scenarios and all of the advantages of a laboratory
- a guide to recreating context through the use of a simulated environment that will induce presence in a recreated environment

References

References

2010. BS EN ISO 9241-210:2010: Ergonomics of human-system interaction. Human-centred design for interactive systems. British Standards Institute.

A

Abdul. (2010). Quality of Psychology Test Between Likert Scale 5 and 6 Points. *Journal of Social Sciences,* **6**(3), pp. 399-403.

Abras, C., Maloney-Krichmar, D., Preece, J. (2004) User-Centered Design. In Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications.

Allen, J.J.B., Chambers, A.S. and Towers, D.N., (2007). The many metrics of cardiac chronotropy: A pragmatic primer and a brief comparison of metrics. *Biological Psychology,* **74**(2), pp. 243-262.

Aromaa, S., Leino, S., Viitaniemi, J., Jokinen, L. and Kiviranta, S., (2012). Benefits of the use of virtual environments in product design review meeting, 2012, University of Zagreb.

Aromaa, S. and Väänänen, K., (2016). Suitability of virtual prototypes to support human factors/ergonomics evaluation during the design. *Applied Ergonomics,* **56**, pp. 11-18.

Aspelund, K (2015) (3rd Ed) The Design Process. Bloomsbury Publishing ltd. London.

B

Ballard, D. H., Hayhoe, M. M,. Pook, P. K.,&Rao, R.P.N. (1997). Deictic codes for the embodiment of cognition. Behaviour Science, 20(4), 723-742.

Bangor, A. (2008). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale.

Barnum, C, M. (2011) Usability Testing Essentials, Ready Set Test. Morgan Kaufmann Elsevier UK.

Bassey, M., (1992). Creating Education through Research. *British Educational Research Journal,* **18**(1), pp. 3.

Benz, P. (2014) Experience Design: Concepts and Case Studies kindle edition. Bloomsbury Academic; UK ed. edition

Benyon, D., Smyth, M., Helgason,. (2009) Presence for everyone. A short guide to presence research. Published by the Centre for Interaction Design Edinburgh Napier University, UK. ISBN 978-0-9562169-0-8

Berntson, G. G., Bigger,J.T.J.,Eckberg,D.L.,Grossman,P.,Kaufmann,P.G.,Malik,M.,*et al*.,(1997). Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology 34 (6), 623–648.

Blaauw, GJ. (1982) Driving experiments and task demands in simulators and instrumented car”. *Human Factors, 24, 473-486)*

Blanford, A. Buchanan, G. Curzon, B. Furniss, D. Thimbleby, H. (2010) Who’s looking? Invisible problems with interactive medical devices. Proceedings of Workshop on Interactive Systems in Healthcare, Atlanta GA.

Brehmer, B, Dorner, D. (1993). Experiments with computer-simulated microwolds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field of study. *Computer in Human Behaviour,* 9(2-3), 171-184.

Boothe, C., Strawderman, L. and Hosea, E. (2013). The effects of prototype medium on usability testing. *Applied Ergonomics,* **44**(6), pp. 1033.

Brehmer, B, Dorner, D. (1993). Experiments with computer-simulated microwolds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field of study. *Computer in Human Behaviour,* 9(2-3), 171-184.

British Design Council (2007). *Eleven lessons: Managing design in eleven global brands*. 2007. Available at: http://www.designcouncil.org.uk/sites/default/files/asset/document/ElevenLessons_Design_Council%20(2).pdf [Accessed: 14 July 2016].

Brown, B N., Reeves, S. and Sherwood, S. (2011). Into the wild: Challenges and opportunities for Field Trial Methods. In Proc. Of CHI'11, pp 1657-1666. ACM, New York. Doi:10.1145/1978942.1979185.

B. R. Brkic., B.R., Chalmers, A., Sadzak, A., Debattista, K., and Sultanic, S. (2013). Exploring multiple modalities for selective rendering of virtual environments. In *Proceedings of the Spring Conference on Computer Graphics (SCCG'13)*. ACM, New York, NY, 91–98.

Buchenau, M. and SURI, J.F. (2000). Experience prototyping. Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS, , pp. 424-433.

Burki Cohen, J, & Go, TH. (2005). The effect of simulator motion cues on initial training od airlines pilots. San Francisco, CA: AIAA-2005-6109. Proceedings of American Institute Aeronautics and Astronautics Modelling and Simulation Conference.

Buxton, B (2007). Sketching the user Experience. London: Morgon Kaufmann.

Buxton, W. (2001). Less is More (More or Less), in P. Denning (Ed) *The Invisible Future: The seamless integration of technology in everyday life.* New York Hill 145-179.

C

Cantarella, L., HegeL, C. and Marcus, G.E. (2019). Ethnography by design. London. New York: Bloomsbury Academic.

Cao, J., Zieba, K., Ellis, M. (2015). The ultimate Guide to Prototyping. The best prototyping methods, tools & process. UxPin Inc.

Couture, N., Riviere, G. and Reuter, P., (2010). Tangible Interaction in Mixed Reality Systems. *MIXER - The Engineering of Mixed Reality Systems.* Springer-Verlag, pp. 101-120.

Cambridge Dictionary. (2020). Cambridge university Press Accessed online 17.02.20 https://dictionary.cambridge.org/dictionary/english/role-play

Castaneda, H.N (2012). Thinking and Doing: The Philosphucal Foundation of Institutions (Philosophical Studies Series Book 7) Springer 1975 USA.

Chalmers. A., Debattista, K., and Ramic-Brkic. B., (2009a). Towards high-fidelity multi-sensory virtual environments. *Visual Computer* 25, 12, 1101–1108. DOI: http://dx.doi.org/10.1007/s00371-009-0389-2

Chartered Institute of Ergonomics & Human factors. (2017-29). What is Ergonomics? https://www.ergonomics.org.uk/Public/Resources/What_is_Ergonomics_.aspx Accessed May 2020

Carayon, P. (2012). Handbook of Human Factors and Ergonomics in Health Care and Patient Safety. 2nd Ed. CRC Press. Taylor and Francis group, Pub.. Boca Raton US.

Carlgren, L. *et al*. (2016). Framing design thinking: The concept in idea and enactment. *Creativity and Innovation Management* 25(1), pp. 38–57.

Carulli, M., Bordegoni, M. and Cugini, U. (2013). An approach for capturing the Voice of the Customer based on Virtual Prototyping. *Journal of Intelligent Manufacturing,* **24**(5), pp. 887-903.

Chamberlain, P., Yoxall, A. (2012). 'Of Mice and Men': The Role of Interactive Exhibitions as Research Tools for Inclusive Design. The Design Journal 15, 57–78.

Charter Institute of Ergonomics & Human Factors. (2017). https://www.ergonomics.org.uk/Public/Awards_Accreditation/Awards_list/Hywel_Murrell_Award.aspx accessed 19.03.18

Chisnell, D. and Rubin, J. (2008). Handbook of Usability Testing How to Plan, Design, and conduct Effective Tests. 2nd Ed. John Wiley and Sons, Canada.

Chomeya, R. (2010). Quality of psychology test between Likert scale 5 and 6 Points. *Journal of Social Sciences,* 6(3), pp. 399-403.

Crawford, M. (2010). The Case for Working with your hands or Why Office Work is Bad for us and why fixing thigs feels Good. Penguin books. London.

Creswell, J. (2003). Research design: qualitative, quantitative, and mixed methods approaches (2nd ed.). Thousand Oaks, CA: Sage.

Cross, N., (2018). A brief history of the Design Thinking Research Symposium series. *Design Studies,* **57**, pp. 160-164.

Cross, N. (1999). Design Research: A Disciplined Conversation. *Design Issues,* **15**(2), pp. 5-10.

Cross, N. (1980), The Recent History of Post-industrial Design Methods. Hamilton (ed.) Design and Industry. London, The Design Council.

D

Dahl, Y., Oct 16, (2010). Seeking a theoretical foundation for design of in sitro usability assessments, Oct 16, 2010, ACM, pp. 623-626.

Dahl, Y., Alsos, O.A. and Svanæs, D., (2010). Fidelity Considerations for Simulation-Based Usability Assessments of Mobile ICT for Hospitals. International Journal of Human-Computer Interaction: Evaluating New Interactions in Health care: Challenges and Approaches, 26(5), pp. 445-476.

Dawson, C. (2019) Introduction to Research Methods A practical guide for anyone undertaking a research project 5th Ed. How to Books Ltd Oxford.

Dell'era, C. and Landoni, P., (2014). Living Lab: A Methodology between User-Centred Design and Participatory Design. *Creativity and Innovation Management,* **23**(2), pp. 137-154.

Deniaud, C. Honnet, V. Jeanne, B. Mestre, D (2015 3:1). The Concept of 'Presence" as a measure of ecological validity in driving simulators. Journal of International Science a Springer open Journal.

Design Council (2007). Eleven Lessons: managing design in eleven global brands. A study of the design process. https://www.designcouncil.org.uk/resources/report/11-lessons-managing-design-global-brands accessed Nov 2020.

Dillon, C., Keogh, E., Freeman, J., Davidoff, J. (2000). Aroused and Immersed: The Psychophysiology of Presence.

Donaldson, K. (2008). Why to be wary of Design for Developing Countries.   Ambidextrous p5-37.

Duffy, E. (1962). Activation and Behaviour. Wiley. New York NY.

Dul, J., Bruder, R., Buckle, P., Carayon, P., Falzon, P., Marras, W.S., Wilson, J.R. and Van Der Doelen, B., (2012). A strategy for human factors/ergonomics: developing the discipline and profession. *Ergonomics,* **55**(4), pp. 377-395.

E

Eccles, D.W. and ARSAL, G., (2017). The think aloud method: what is it and how do I use it? *Qualitative Research in Sport, Exercise and Health,* **9**(4), pp. 514.

EDWARDS, S.D., (2015). HeartMath: a positive psychology paradigm for promoting psychophysiological and global coherence. *Journal of Psychology in Africa,* **25**(4), pp. 367-374.

Endole LTD (2016). Endole company insights. https://www.endole.co.uk/ accessed 11.10.2016.
Etikan, I. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics,* 5(1), pp. 1.

Evans, C., D. (2017). Bottlenecks: Aligning UX Design with User Psychology. Apress. Springer Science New York.

F

Falconer, C.J., Slater, M., Rovira, A., King, J.A., Gilbert, P., Antley, A. and Brewin, C.R. (2014). Embodying Compassion: A Virtual Reality Paradigm for Overcoming Excessive Self-Criticism. *PloS one,* **9**(11), pp. e111933.

Filed, A. (2016 p27). An Adventure in Statistics the reality enigma. SAGE edge London.

Friedman, K., (2003). Theory construction in design research: criteria : approaches, and methods.

G

Gaver, W., Boucher, A., Pennington, S. and Walker, B., (2004). Cultural probes and the value of uncertainty. *interactions, ACM* **11**(5), pp. 53-56.

García, M., Constantino A. (2017). Heartrate Variability Analysis with the R Package RHRV, Springer. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/cardiffmet/detail.action?docID=5049922.

Georges, A., Schuurman, D., Baccarne, B. and Coorevits, L. (2015). User engagement in living lab field trials. *info,* **17**(4), pp. 26-39.

Giacomin, J. (2014). What is human Centred design? *The Design Journal* 17(4), pp. 606–623.

Gill, S. (2008). A Dissertation Exploring Issues Surrounding the rapid Development of Information Appliances by Design. PhD by Publication University of Wales.

Gill, S. (2009). 'Six challenges facing user-oriented industrial design', The Design Journal, 12(1), pp.41-67.; 1460-6925; 1756-3062.

Gill, S., Loudon, G., Woolley, A., Hare, J., Walker, D., Dix, A. and Ellis, D.R. (2008). Rapid development of tangible interactive appliances: achieving the fidelity/time balance. *International Journal of Arts and Technology,* **1**(3/4), pp. 309.

Godley, ST. Triggs, TJ, & Fildes, BN. (2002). Driving Simulator validation for speed research. Accident Analysis and Prevention, 34(5), 589-600.

Gordon, B. (2007). Emulation of Real-life Environment via Augmented Virtual Environment. MSc Thesis. Swansea Institute of Higher Education.

Gordon, B. S., Loudon, G,. Gill, S., Baldwin, J., (2019). Product user testing: the void between Laboratory testing and Field testing. IASDR 2019 conference. Manchester School of Art.

Gordon, P. Fuge, M. Agogino, Alice. (2017). Examining Design for Developing online: An HCD Analysis of Open IDEO using HCD/UCD Metrics Belgrade: Association of Economists and Managers of the Balkans.

Government UK (2020). Staying at home and away from others (Social distancing) Guidance May 2020 accessed https://www.gov.uk/government/publications/full-guidance-on-staying-at-home-and-away-from-others/full-guidance-on-staying-at-home-and-away-from-others

Graczyk, P. (2015). Embedding a Living Lab approach at the University of Edinburgh. Social Responsibility and Sustainability.

Gray, W D. (2002). Simulated Task Environments: The role of High-Fidelity Simulations, Scaled Worlds, Synthetic Environments, and Laboratory Tasks in Basic and applied Cognitive Research. Cognitive Science Quarterly 2, 205-227. Pub Lavoisier.

Greenburg, S. (2001). Context as a dynamic construct. *Hum-Computer. Interact.* 16, 2 (Dec. 2001), 257-268.

Guger, C., Edlinger, G., LEEB, R., Pfurtscheller, G., Antley, A., Garau, M., Brogni, A., Friedman, D., Slater, M. (2005). *Heart-Rate Variability and Event-Related ECG in Virtual Environments.*

Gupta, R., Whitney, D. and Zeltzer, D. (1997). Prototyping and design for assembly analysis using multimodal virtual environments. *Computer-Aided Design,* **29**(8), pp. 585-597.

Hare, J. Gill, S. Loudon, G. Lewis, A. (2014). Active and passive physicality: making the most of low fidelity physical interactive prototypes. Journal in Design Research, Vol. 12, No 4, 2014.

H

Hallgrimsson, B. (2020). Prototyping and Modelmaking for Product Design. 2nd Edition. Laurence King Pub. London.

Hanington, B., Martin, B. (2019 p3750). Universal Methods of Design and Expanded and Revised. Rockport Publishes.

Hare, J. (2015). Physicality in the design and development of computer embedded products, Cardiff Metropolitan University.

Hare, J. Gill, S. Loudon, G. Lewis, A. (2014). Active and passive physicality: making the most of low fidelity physical interactive prototypes. Journal in Design Research, Vol. 12, No 4, 2014.

HAYHOE, M. and BALLARD, D. (2005). Eye movements in natural behaviour. Trends in cognitive sciences, 9(4), pp. 188-194.

Health & Safety Executive. (ND). Human Factors Design. Accessed on the Jan 2017 & May 2020 www.hse.gov.uk/humanfactors/topics/design.htm

Hettinger, L.., Haas, M. (2003). Virtual and Adaptive Environments, Application, Implications and humans Performance Issues. Lawrence Erlbaum Associates Publishers London.

Holzinger, K. (1928). *The Elementary School Journal, 28*(5), 387-388. Retrieved from http://www.jstor.org/stable/995617

Houde, S., Hill. C. (1997). Apple Computer, inc. Cupertino, CA, USA S.houde@ix.netcom.com and Hillc@ix.netcom.com, *What do Prototypes Prototype?*

Human Interface Technology Lab. (1997). Olfactory Interfaces. Retrieved May 13, 2014, from http://www.hitl.washington.edu/.

I

IDEO. (2015). The Field Guide to Human Centred Design. ISBN 978-0-9914063-1-9. Pub in Canada.

IEA. (2000). Definition and Domains of Ergonomics [WWW Document]. http://www.iea.cc/whats (accessed 19.03.18)

Ishii, H., Ullmer, B. (1997). Tangible bits: towards seamless interfaces between people, bits and atoms. In: CHI’97 the 15th SIGCHI conference on Human factors in computing systems, pp. 234--241. ACM.

J

Jansen, A.S.P., Nguyen, X.V., Karpitskiy, V., Mettenleiter, T.C. & Loewy, A.D. (1995). "Central command neurons of the sympathetic nervous system: Basis of the fight-or-flight response", Science, vol. 270, no. 5236, pp. 644.

Johnson, R. B. & Onwuegbuzie, A. J. (2004). Mixed-methods research: a research paradigm whose time has come. Educational Researcher, 33(7), 14-26.

Johnson, P. (1998). Usability and Mobility; Interactions on the move. Proc. Mobile HCI’98, GIST Technical Report G98-1 Accessed from http://www.dcs.gla.ac.uk/~johnson/papers/mobile/HCIMD1.html

Jokinen, J., Silvennoinen, J., Perälä, P. and Saariluoma, P., Apr 18, (2015). Quick Affective Judgments, Apr 18, 2015, ACM, pp. 2221-2230.

JUMISKO-PYYKKÖ, S. and VAINIO, T., 2010. Framing the Context of Use for Mobile HCI. *International journal of mobile human computer interaction,* **2**(4), pp. 1-28.

K

Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T. (2005). Usability testing of mobile applications: a comparison between laboratory and field testing. Journal of Usability Studies, volume 1, number 1, pages 4-16.

Kaptein, NA., Theeuwes, J & Van der Horst, R. (1996). Driving simulator validity: some consideration. Transportation Research Record, 1550, 30-36.

KARWOWSKI, W. (2006). International encyclopaedia of ergonomics and human factors. 2nd ed. edn. Boca Raton: CRC Press.

Kassab, E., Tun, J.K., Arora, S., King, D, Ahmed., K, Miskovic., D, Cope., A, Vadwana., B, Bello., F, Sevdalis., N., Kneebone, R. (2011). Blowing up the Barriers in surgical Training: Exploring and Validating the concept of Distributed Simulation. Annals of Surgery. Vol 254(6) p1059-1065.

Keller, I., Stappers, P. (2001). Presence for Design: Conveying Atmosphere through Video Collages. CyberPsychology & Behavior. Vol. 4, No 2, 2001, pp215-223.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: Macmillan.

Kelly, T., Littman, L. (2004). The Art of Innovation. Bookmarque ltd. Surrey.

Kipper, G., Rampool, J. (p1: 2013). Augmented Reality: An Emerging Technology Guide to AR. Elsevier. USA.

Keyson, D.V., Guerra-santin, O. and Lockton, D., (2017). Living labs. Cham: Springer.

Kneebone, R., Arora, S., King, D., Bello, F., Sevdalis, N., Kassab, E., Aggarwal, R., Darzi, A., Nestel, D. (2010). Distributed simulation – Accessible immersive training. Medical Teacher v32 P65-70 Taylor & Francis Ltd.

Kjeldskov, J. and Skov, M. (2014). Was it worth the hassle? Sep 23, 2014, ACM, pp. 43-52.

Kjeldskov, J., Skov, M.B., ALS, B.S., Høegh, R.T. (2004). Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. Springer.

Kneebone, R., Arora, S., King, D., Bello, F., Sevdalis, N., Kassab, E., Aggarwal, R., Darzi, A. and Nestel, D. (2010). Distributed simulation--accessible immersive training. Medical teacher, 32(1), pp. 65-70.

Kneebone, R. (2012). A Novel Approach to Contextualized Surgical Simulation Training. DOI 10.1097/SIH.0b013e31824a86db.

Kunzle, D. (2019). Review Article. European Comic Ar*t,* **12**(2), pp. 106-113.

L

Leedy, P.D. and Ormeod, J.E. (2010). Practical Research: Planning and Design. Boston MA:Pearson.

Leffingwell,J., C. (2002). Olfaction—A Review. Retrieved May 13, 2014, from
http://www.leffingwell.com/olfaction.htm.

Lessiter, J., Freeman, J., Keogh, E. and Davidoff, J.B. (2001). A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory, Presence: Teleoperators & Virtual Environments, 10(3), pp. 282-297.

Ley, B. Ogonowski, C. Mu, M. Hess, J. Race, N. Randall, D. Rouncefield, M. Wulf, V. (2014). At Home with Users: A Comparative View of Living Labs. Oxford University Press. The British Computing Society.

Lim, Y., Pangam, A., Periyasami, S. and Aneja, S. (2006). Comparative analysis of high- and low-fidelity prototypes for more valid usability evaluations of mobile devices, Oct 14, 2006, ACM, pp. 291-300.

Lab4Living. (2007). https://lab4living.org.uk/projects/uti-testing-in-primary-care/ Sheffield Hallam accesses Nov 2020

Loudon, G. (2006). Nice technology, shame about the product, Communications Engineer, Vol. 4, Issue 4, pages 12-15.

Loudon, G.H. and Deininger, G.M. (2017). The Physiological Response to Drawing and Its Relation to Attention and Relaxation. *Journal of Behavioral and Brain Science,* **7**(3), pp. 111-124.

Luma Institute. (2012: p1). Innovating for People: Handbook of Human-Centered Design Methods. Pub Luma Institute USA.

M

Ma, J. Nickerson, J. (2006). Hands-On, Simulated, and Remote Laboratories: A Comparative Literature Review. ACM Computing Survey Journal I 3 V 38 P1-1-24.

Maesh, S. (2018). User Research. A practical guide to designing better products and services. Kogan Page Ltd. UK.

Mack, A., Rock. I. (1998). Inattention Blindness. MIT press. Cambridge.

Magyar-moe, J.L. (2011). Incorporating positive psychology content and applications into various psychology courses. *The Journal of Positive Psychology: Positive Psychology in Higher Education,* **6**(6), pp. 451-456.

Maths is Fun. (2018). Definition of Mean. https://www.mathsisfun.com/definitions/mean.html accessed 05.03.2020.

Marshalla (2010) HADRIAN: a virtual approach to design for all. Journal of Engineering Design Special Issue on Inclusive Design Vol. 21, No. 2-3, April-June 2010, 253–273.

Mcadams, D.P., Bauer, J.J., Sakaeda, A.R., Anyidoho, N.A., Machado, M.A., Magrino-failla, K., White, K.W., Pals, J.L. (2006). Continuity and Change in the Life Story: A Longitudinal Study of Autobiographical Memories in Emerging Adulthood. *Journal of Personality,* **74**(5), pp. 1371-1400.

Meehan, M.J. (2001). Physiological reaction as an objective measure of presence in virtual environments, The University of North Carolina at Chapel Hill.

Merz, E.L., Malcarne, V.L., Roesch, S.C., KO, C.M., Emerson, M., Roma, V.G. and Sadler, G.R. (2013). Psychometric properties of Positive and Negative Affect Schedule (PANAS) original and short forms in an African American community sample. *Journal of Affective Disorders,* **151**(3), pp. 942-949.

Mestre, D. & Fuchs, P. (2006). Immersion et Présence. In Traité de la Réalité Virtuelle, Troisième Edition (P. Fuchs, Ed.).Paris: Presses de l'Ecole des Mines. pp 309-338.

Milton, A., Rodgers, P. (2019 p 121). Research Methods for Product Design, 3rd Edition. Laurence King Publishers London.

Molich, R., Ede, M.R., Kaasgaard, K. and Karyukin, B. (2004). Comparative usability evaluation, *Behaviour & Information Technology,* **23**(1), pp. 65-74.

Molich, R., & Dumas, J. S. (2006). Comparative usability evaluation (CUE-4), *Behaviour & Information Technology*, pp. 1–19, December.

Moggridge, B. (2007). *Designing Interactions.* Cambridge, MA.: The MIT Press.

Moore, P. & Conn, C.P. (1985). Disguised: A True Story. Waco, Texas: Word Books.

Moon, K. and Blackman, D. (2014). A Guide to Understanding Social Science Research for Natural Scientists. *Conservation Biology,* **28**(5), pp. 1167-1177.

Muratovski, G. (2016). Research for Designers a guide to methods and practice. SAGE London.

Murray, GE. Neumann, DL, Moffitt, RL. Thomas, P. (2015). The effects of the presence of others during a rowing exercise in a virtual reality environment. Psychology of Sport and Exercise. Elsevier Vol 22 P 238-336.

N

Nilsson, L. (1993). Behavioural research in an advanced driving simulator-experience of the VTI system. Proceedings of the Human Factors Society 37th Annual Meeting 37(1), 612-616.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. Proc. ACM CHI'94 Conf. (Boston, MA, April 24-28), 152-158.

Nielsen, J. (2018). Why you only need to test with 5 users. NN/G Nielsen Norman Group https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Nielsen, J. (2007). High-Cost usability sometimes Makes sense. NN/G Nielsen Norman Group https://www.nngroup.com/articles/when-high-cost-usability-makes-sense/ accessed 10/20

Nielsen, J, and Landauer, T K. (1993). "A mathematical model of the finding of usability problems," Proceedings of ACM INTERCHI'93 Conference (Amsterdam, The Netherlands, 24-29 April 1993), pp. 206-213.

Nielsen Norman Group. (2012). Usability 101: Introduction to Usability. Retrieved March 2019 http://dockerby.com/web/Unit%206%20Validating/Usability%20101_%20Introduction%20to%20Usability.pdf

Norman, D.A. (2013). The design of everyday things, revised and expanded edition MIT Press

Norman, D. A. (1998). The Invisible Computer: Why good products can fail, the personal computer is so complex and Information Appliances are the Solution. London: MIT Press.

Norman, D. A, & Verganti, R. (2014). Incremental and radical Innovation: Design research versus technology and meaning change. *Design Issue, 30, 78-96.*

O

Ohly, S., Sonnentag, S., Niessen, C. Zapf, D. (2010). Diary Studies in Organizational Research. *Journal of Personnel Psychology,* **9**(2), pp. 79-93.

Orquin, J.L., Ashby, N.J.S. Clarke, A.D.F. (2016). Areas of Interest as a Signal Detection Problem in Behavioral Eye-Tracking Research. *Journal of Behavioral Decision Making,* **29**(2-3), pp. 103-115.

P

Pair, J., Neumann, U., Piepol, D., Swartout, B. (2003). FlatWorld: combining Hollywood set-design techniques with VR. IEEE Computer Graphics and Applications, **23**(1), pp. 12-15.

Park, C., Ko, H. and Kim, T. (2003). NAVER: networked and augmented virtual environment architecture; design and implementation of VR framework for Gyeongju VR theatre. Computers & Graphics, 27(2), pp. 223-230.

Powers. W. (2004). The Science of Smell Part 1: Odor Perception and Physiological Response. Retrieved May 13, 2014, from http://www.extension.iastate.edu/Publications/PM1963A.pdf.

Pheasant, S. Haslegrave, C.M. (2006). Bodyspace Anthropometry, Ergonomics and the Design of Work. Third Edition. Taylor & Francis London.

Psychology Dictionary. (2013). Professional Reference. https://psychologydictionary.org/ecological-validity/ Accessed 06.05.20.

R

Ramduny-Ellis, D., Dix, A., Evans, M., Hare, J. and Gill, S. (2010). Physicality in Design: An Exploration. The Design Journal, 13(1), pp. 48-76.

Ramic, B., Chalmers. A., Hasic, J., and Rizvic., S. (2007). Selective rendering in a multimodal environment: Scent and graphics. In Proceedings of the Spring Conference on Computer Graphics (SCCG'07).

Redish, J., Bias, R., Bailey, R., Molich, R., Dumas, J. and Spool, J. (2002). Usability in practice, Apr 20, 2002, ACM, pp. 885-890.

Redstorm. (2006). Towards user design? On the shift from object to user as the subject of design. Design studies, 27(2), pp. 123-139).

Reeves, B., Nass, C. (2002) The Media Equation. How People Treat Computers, Television, and New Media Like Real People and Places. California. CSLI Publications.

Reimer, B., D'Ambrosio, L., Coughlin, J., Kafrissen, M. and Biederman, J. (2006). Using self-reported data to assess the validity of driving simulation data. *Behavior Research Methods,* **38**(2), pp. 314-324.

Rodgers, P., A. Anusas, M. (2008). Ethnography and Design. International conference of Engineering and Product Design Education. Barcelona.

Rodgers, P. Milton, A. (2011). Product Design. Laurence king Publishing Ltd. London.

Roto, V. (2006). Web Browsing on Mobile Phones – Characteristics of User Experience. Doctoral dissertation. Helsinki University of Technology. Finland.

Rowntree, D. (2018). Statistics without Tears. An Introduction for non-mathematicians. Penguin Random House UK.

Rubin, J. Chisnell, D. (2008). 2nd Ed Handbook of Usability Testing How to Plan, Design, and conduct Effective Tests. Pub. John Wiley and Sons Inc. Canada.

Römer, A., Pache, M., Weißhahn, G., Lindemann, U., & Hacker, W. (2001). Effort-saving product representations in design -results of a questionnaire survey. *Design Studies, 22*(6), 473-491.

S

Sætren, G., Hogenboom, S. and Laumann, K. (2016). A study of a technological development process: Human factors—the forgotten factors? *Cognition, Technology & Work,* **18**(3), pp. 595-611.

Sadideen, H., Hamaoui, K., Saadeddin, M. and Kneebone, R., (2012). Simulators and the simulation environment: Getting the balance right in simulation-based surgical education. *International Journal of Surgery,* **10**(9), pp. 458-462.

Saffer, D. (2010). Designing for interaction 2nd edition. New Riders. Berkeley.

Sanders, E.B.-. and Stappers, P.J. (2008). Co-creation and the new landscapes of design. *CoDesign,* **4**(1), pp. 5-18.

Sanders, E. B.-N., & Stappers, P. J. (2014). Probes, toolkits and prototypes: three approaches to making in codesigning. CoDesign, 10(1), 5–14. http://doi.org/10.1080/15710882.2014.888183

Sauer, J., Seibel, K. and Rüttinger, B. (2010). The influence of user expertise and prototype fidelity in usability tests. *Applied Ergonomics,* **41**(1), pp. 130-140.

Schäfer, A., & Jan Vagedes, J. (2013). How accurate is pulse rate variability as an estimate of heartrate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram, International Journal of Cardiology, 166, 15–29.

Simons, D.J. and Chabris, C.F. (1999). Gorillas in our midst: Sustained In attentional blindness for dynamic events. *Perception* [Online] 28(9), pp. 1059–1074. Available at: http://pec.sagepub.com/content/28/9/1059.short [Accessed: 13 July 2016].

Simsarian, K., (2003). Take it to the next stage: The Roles of Role Playing in the Design Process. Apr 5, 2003, ACM, pp. 1012-1013.

Slater, M., (1999). Measuring presence: A response to the Witmer and Singer presence questionnaire. *Presence: Teleoperators & Virtual Environments,* **8**(5), pp. 560-565.

Slater, M. (2004.) Presence the sixth Sense. Presence: Teleoperators and Virtual Environments. Vol 11 , No. 4.

Slater, M., Steed., A. Usoh, M (2013). Being There Together: Experiment on Presence in Virtual Environments (1990s) Technical Report, Department of Computer Science, University of College London.

Slater, M. (2002). Presence and the sixth sense. Presence: Teleoperators and Virtual Environments, 11(4), 435-439.

Slater, M. (2003). A note on presence terminology. Presence connect. Retrieved March 2019, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.800.3452&rep=rep1&type=pdf

Slater, M., Lotto, B., Arnold, MM., & Sanchez Vives., MV. (2009). How we experience immersive virtual environments : the concept of presence and its measurement". Anuario de Psicologia, 40(2), 193–210. Meeting 37(1), 612–616.

Slater, M., Steed, A., Usoh, M. (2013) Being There Together. Experiments on Presence in Virtual Environment. Department of Computer Science. University of College London.

SmaradottiR, B., Gerdes, M., Fensli, R. and Martinez, S., (2015). Usability Evaluation of a COPD Remote Monitoring Application. Studies in health technology and informatics, 210, pp. 845.

Solso., R., Maclin, K. MacLin, O. (2005) Cognitive Psychology, 7th ed. Boosten Allyn and Bacon.

Spicer, R., Evangelista, E., Yahata, R., Campbell, J., Richmond, T. (2015). Innovation and Rapid Evolutionary Design by Virtual Doing: Understanding Early Synthetic Prototyping (ESP). University of Southern California Institute for Creative Technologies.

STEUER, J., (1992). Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication,* **42**(4), pp. 73-93.

Svanaes, D., Alsos, O.A. and Dahl, Y. (2010). Usability testing of mobile ICT for clinical settings: methodological and practical challenges. International journal of medical informatics, 79(4), pp. 24.

Svanæs, D., Alsos, O.A. and Dahl, Y. (2008). Usability testing of mobile ICT for clinical settings: Methodological and practical challenges. *International Journal of Medical Informatics,* **79**(4), pp. e24-e34.

T

Tashakkori, A., & Teddlie, C. (2008). Introduction to mixed method and mixed model studies in the social and behavioral science. In V.L. Plano-Clark & J. W. Creswell (Eds.), The mixed methods reader.

Terrell, S. (2011). Mixed-methods research methodologies. The Qualitative Report, 17(1), 254-280. Retrieved from http://www.nova.edu/ssss/QR/QR17-1/terrell.pdf

Thimbleby, H. (2013). *Improving safety in medical devices and systems.* IEEE International Conference on Healthcare Informatic.

Tornros, J. (1998). Driving behaviour in a real and simulated road tunnel- a validation study. Accident Analysis and Prevention, 30(4), 497-503.

U

Ulrich, T., Eppinger, S.D. (2003). Product Design and development International Edition (Fifth Ed) McGraw-Hill Education. New York.

Ulrich, T., Eppinger, S.D. (2016). Product Design and development International Edition (Sixth Ed) McGraw-Hill Education. New York.

Ulrich, T., Eppinger, S.D. (2009). Product Design and development International Edition (Seventh Ed) McGraw-Hill Education. New York.

Unger, D.W. and Eppinger, S.D., (2009). Comparing product development processes and managing risk. *International Journal of Product Development,* **8**(4), pp. 382-402.

Utterback, J. and Vedin, B., (2006). *Design-inspired Innovation.* River Edge, NJ, USA: World Scientific Publishing Company.

V

Van Baren, J & Jsselsteijn, W.I. (2004). Measuring Presence: A guide to Current Measurement approaches. OmniPres project IST-2001-39237.

Van Der Bijl-Brouwer, M. and Dorst, K. (2017). Advancing the strategic impact of human-centred design. *Design Studies,* **53**, pp. 1-23.

Van Manen, M. (2014). Phenomenology of Practice. New York: Routledge.
https://doi.org/10.4324/9781315422657

Vassallo, S. (2017). The case against empathy. In Co.Design. New York: Fast Company.

Verlinden, J., Suurmeijer, C. and Horvath, I. (2007). Which Prototype to Augment? A Retrospective Case Study on Industrial and User Interface Design. *Virtual Reality.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 574-583.

Vines, J., Clarke, R,. Wright, P., McCarthy, J., and Olivier, P. (2013). Configuring Participation: On How we Involve People in Design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI' 13,p 429. ACM Press, New York, New York, USA. Doi:10.1145/2470654.2470716/.

Virzi, R., Sokolov, J. and Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes, Apr 13, 1996, ACM, pp. 236-243.

W

Walshe, D. Lewis, E. O'Sullivan, K. (2005). Virtually Driving; Are the Driving Environments "Real Enough" for Exposure Therapy with Accident Victims? An Explorative Study. CyberPsychology & Behavior. Vol. 8, No 6, 2005, pp532-537.

Walters A.T., Evans J. (2011). *Developing a Framework for Accessible User-Centric Design*, 18th Int. Product Dev. Mgmt Conference, Netherlands, 6-7 June 2011.

Warfel, T, Z. (2009). Prototyping: A practitioners Guide. Rosenfeld Media. New York.

Ware, C. (2004) (2nd Ed.). Information Visualization. Perception for Design. San Francisco. Elsevier.

War, C. (2008). Visual Thinking for Design. Morgan Kaufmann Pub. Elsevier. Burlington.

Watson, D. Clark, L.A., Tellegan, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scale. Journal of Personality and social Psychology, 54 (6), 1063-1070.

Weinschenk, S, M. (2011). 100 Things Every Designers needs to know about people. Pub New Riders. Canada.

Williams, L. J. (1985). Tunnel Vision induced by a foveal load manipulation. Human factors 27 (2): 221-227

Witmer, B.G & Singer, M.J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire, Presence: Teleoperators and Virtual Environments, 7(3), 225-240.

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioural responses to attitude objects. Journal of Social Issues, 25(4), 41–78.

Wilson, C. (2020). Handbook of user-centred Design Methods. Morgan Kaufmann.

Woolley, A. (2008). Contextual testing of interactive prototypes at the early stages of the design process. University of Wales.

Woolley, A. Loudon, G. Gill, S. Hare, J. (2013). Contextual Testing of Interactive Prototypes at The Early Stage Of The Design Process. The Design Journal vol 16. I2. pp460-485. Bloomsbury Pub plc.

Woolley, A. Loudon, G. Gill, S. Hare, J. (2013). Getting into Context, early: a Comparative, study of laboratory, and in-context, user testing, of low-fidelity, information and appliance prototypes, the design journal volume 16, issue 4 reprints available photocopying © Bloomsbury pp 460–485 directly from the permitted by publishing plc 2013 publishers license only printed in the uk.

Wu, H. and Leung, S. (2017). Can Likert Scales be Treated as Interval Scales? A Simulation Study. Journal of social service research, 43(4), pp. 527-532.