IASDR 2019
International Association of Societies of Design Research Conference 2019
DESIGN REVOLUTIONS

Manchester School of Art
Manchester Metropolitan University
02-05 September 2019

Product user testing: the void between Laboratory testing and Field testing.

[*a] Gordon, Bethan, Cardiff Metropolitan University, Cardiff School of Art & Design, Wales.
[b] Loudon, Gareth, Cardiff Metropolitan University, Cardiff School of Art & Design, Wales
[c] Gill, Steve, Cardiff Metropolitan University, Cardiff, Wales
[d] Jo, Baldwin, Cardiff Metropolitan University, Cardiff School of Art & Design, Wales
* bgordon@cardiffmet.ac.uk

User 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 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 user 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 paper introduces research aimed at defining the optimum fidelity of mixed-reality user testing environments. It aims to develop knowledge enabling the optimisation of user testing environments by balancing effort vs. reward and thus developing critical and accurate data early in the design process.

Testing in a laboratory setting brings advantages such as the ability to limit experimental variability, control confidentiality and measure performance in great detail. Its disadvantages over 'in the wild' 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 focussed on high fidelity, technology-rich approaches that make them simultaneously more expensive, less flexible and less accessible. The final result is that they are less viable and hence somewhat counter-productive.

This paper presents the results of testing at a variety of fidelity levels within a mixed reality testing environment created by a team of artists and designers. It concludes with a series of recommendations regarding where and when fidelity is important.

Keywords: Products; product design; industrial design; mixed reality environments; user centred design; user testing environments; fidelity; usability

1 Introduction

It is generally recognised that it is important to iteratively create and test prototypes during the product design and development process (Boothe, Strawderman, and Hosea 2013). In some fields, such approaches are mandatory. For example, user 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 Design Process (PDP) requires user testing throughout the design process (Hare, Gill, Loudon, and Lewis, 2014; Rubin and Chisnell, 2008) including at the earliest low fidelity stages. Usability testing of low fidelity models usually takes place in a controlled laboratory setting (Kaikkonen, Kekalainen, Canker, Kallo, and Kankainen, 2005) both for ease of access and so 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 tends to limit the validity of any resultant usability data.

Literature on testing environments goes back a long way. Dahl, Andreas and Svanaes (2009), for example, found that the social and physical attributes of an environment are often ignored or given low priority when testing a prototype, and highlights the need to consider their implications in design testing. Even further back, Brehmer and Dorner (1993) found that while field research is often too complex to be practical, laboratory testing doesn't offer enough complexity for the often critical, fine grain conclusions. Later literature shows that testing prototypes 'in the wild' can be achieved e.g. Woolley, Loudon, Gill and 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, Kjeldskev and Skov's (2014) review 'Was it was 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) recognised 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 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? The answer is not yet known, 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 and Suri, 2000). Their 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 a surgeon and his team established that a low fidelity mixed-reality environment could be effective in training medical students in surgery techniques (Kassab et

al., 2011). 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 needs 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 focussed on how real visual cues are used to heighten participants' Presence and confidence in the simulated environment, with 'Presence' being defined as *"a user's subjective sensation of 'being there' "* in a simulated environment.

1.1 The Perceptual Experience Laboratory (PEL)

PEL is a mixed reality research laboratory created by a multi-disciplinary team of artists and designers. Its ability to project experiential perspective displays is literally unique – and will be the subject of another paper. Among other things, the laboratory was built to fill the void between traditional laboratory and 'in the wild' user testing. It is purpose-built to afford the restricted and consistent testing conditions of a standard usability laboratory within simulated context-of-use environments. It enables a mixed-reality method that couples physical objects with a 5K, 200° panoramic visual surround screen, artificially added smell and 3D ambisonics to engender an appropriate sense of presence. On the experimental monitoring side PEL has three video cameras to capture activity from various angles, omni-directional mics, heartbeat variability and 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. The current PEL is a development of an original prototype developed by the lead author. At that stage it consisted of a wooden frame and bedsheets that formed a back-projected surface. This was used as a proof-of-concept demonstrator via experiments investigating the extent to which low-fidelity virtual props might suffice in early product testing.

Figure 1: (left to right) St Athans hangar, PEL mock-up (exterior), PEL mock-up (interior)

Method

For this study PEL was employed to test the usability of a low-fidelity prototype of a product then under development for cleaning the flying surfaces and fuselages of a passenger aircraft. The product itself was being developed with the authors' input through a Knowledge Transfer Partnership (KTP). During the KTP the authors visited St Athens airfield, 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 simulated test environments – in this case simulations of St Athans aircraft hangar - each at a different level of fidelity to evaluate whether the fidelity of simulation affected the identification of key usability issues. The water tank prototype was a low-fidelity cardboard mock-up. The poles and cleaning head were also low fidelity but had moving parts and thus demanded a higher degree of interaction.

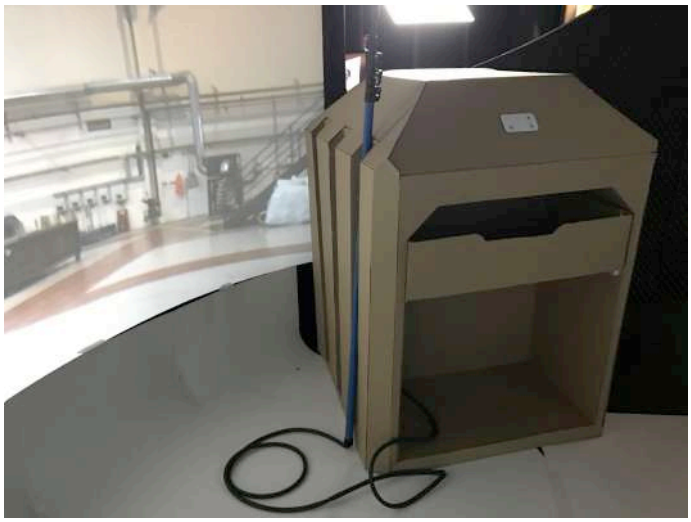

Figure 2: Low-fidelity prototype of the cleaning system

A key aim of the tests was to establish the *Face Validity*[1] (sometimes called *Physical Validity*) of *Presence*[2] (Slater, 2003) of the environment. 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, the authors used Face Validity to determine which fidelity levels of simulated environments provoke the strongest feeling of Presence (Deniaud et al., 2015). The four simulations are described briefly below:

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

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

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

***Environment 4 (E4)*:** Product prototype only.

The test design followed standard usability process with one facilitator controlling test conditions to ensure they remained constant. A mixed method approach was used; a

[1] Defined here as *"the extent to which experts agree that the measures capture the intended construct"* Reimer, D'Ambrosio, Coughlin, Kafrissen, and Biederman (2006).

[2] Presence is defined here as the degree to which a participant feels they are present within a simulation – without necessarily believing it to be real.

combination of Kassab et al's (2011) method with questions used to evaluate the Face Validity of the simulated environment combined with the Systems Usability Scale (SUS) approach developed by John Brook (Bangor, Kortum, and Miller, 2009). The SUS has been used extensively in the product, systems, web and media development industries to establish a perspective on overall usability. Used primarily as an Investigative tool, it uses 10 Likert questions, alternating between negative and positive. Results are converted to a positive ordinal number using the Brook's method and converted to a percentile. A value below 68 indicates usability issues, while:

> *"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)*

Note that normalized 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 used to normalize the data relative to the sector - in the case of the SUS at 68. The SUS is a means of summatively recognizing usability faults in a product or system as a whole. It is used to identify usability faults in a design at a global level and has to be used in conjunction with ethnographical techniques to identify detailed faults.

A pilot study was run to ensure the tests were appropriately set up (Rubin and Chisnell, 2008; Leedy and Ormrod, 2010). Each participant was asked for consent to complete the relevant ethical participation 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 utilised low fidelity prototypes or "horizontal representations" of a concept that allow 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 and Chisnell, 2008). Participants were 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. While post-completion questionnaires rely on memory and latent experience (Jokinen, Silvennoinen, Perala, and Saariloma, 2015), questionnaire completion in real time is impractical. It was felt that their use was valid in combination with the think aloud protocol to capture real time conscious thoughts (Eccles and Arsal, 2017) and video recordings to capture unconscious action. Taken together, these methods allowed for some measure of triangulation.

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 hangar's concrete floor – with cardboard used for the purpose. It was also evident the prototype needed more functionality to get full value from the tests and so a hose was added to make more visual sense of the water tank and the cleaning poles.

1.2 Participants

A total of 24 participants were recruited using *Convenience Sampling* - a nonprobability sampling method (Etikan, 2016). 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.

Participants (17M, 7F) 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.

2 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 and Kassab et al's (2011) question set were used to ascertain Face Validity of the environment. The SUS was used to determine the usability of the low fidelity prototype product, while the Kassab et al questionnaire assessed the validity of the simulated environment. Face Validity questions were presented in Likert Scale format with a 6-point scale used because studies have found they offer more discrimination on attitude, perception or opinion than 5-point scales. Chomeya's (2010) study also highlighted reliability as a factor in opting for 6-point scale. Participants were asked to circle one number, 1 = not at all, and 6 = very much. Mode and Median were used to identify preferences and opinions. This use of ordinal data is used rather than analysing the mean, which should only be used for ratio data. Non-parametric statistics were used because they are more appropriate for data founded on opinion. Note that while capturing ordinal data allows for directionality in preference it does not afford proportional analysis – e.g. that something is 'twice as good'. In other words, the results cannot measure in-between the intervals (Dawson, 2013) and the data captured in this study affords descriptive statistics due to the sample size.

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

Table 1: This table is representative of the Face Validity questions asked during the test

While the sample was the appropriate size for a user test, as noted by Nielson Norman (2012), it was evident that it was not big enough for credible statistical analysis. It was therefore imperative that the qualitative data was captured to help make sense of the data collected.

3 Results

1. The aim of the study comprised of three objectives:
2. Ascertaining the product's overall usability using the SUS questions based on product testing.
3. Testing the Face Validity of each environment using the Likert Scale questionnaire to ascertain the characteristics of an optimum product-testing environment.
4. 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.

3.1 Questionnaire Results

Face Validity analysis was first conducted on the level of Presence achieved in the simulated environment. The 1 to 6 Likert Scale used made a score of 4 and above positive and 3 and below negative indicators respectively. The Mode and Median results showed Presence as being most marked in Environments 1 (E1) and 2 (E2), while the lower fidelity Environments 3 (E3) and 4 (E4) returned negative indicators (see *Table 2* below).

Measure	MODE	MEDIAN
E1	4	4
E2	4	4
E3	3	3.5
E4	1	2.5

Table 2: Collective Narrative: Presence Central tendencies.

SUS
67.1
57.5
76.7
72.5

Table 3: Systems Usability Score

While the prototype under test remained constant in all four environments, the SUS results (see Table 3) 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 Figure 3 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 (see Figure 3 below).

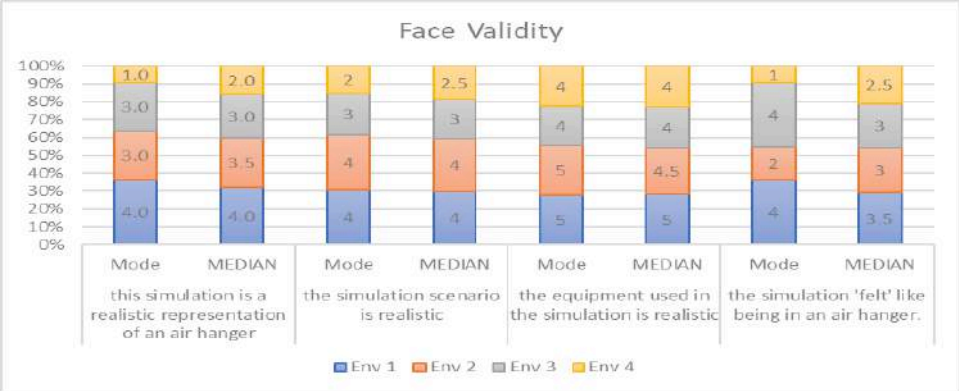

Figure 3: Central Tendencies of the Face Validity Likert Scale

The average time taken for a participant to complete the study is shown in Table 4 below. They show similar results for E1 - E3 (Range = 108 to 132 seconds), while the user testing in the low fidelity E4 was markedly different, with tests taking an average of 367 seconds to complete.

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

Table 4: Average task completion time

3.2 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 Environments 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 apparently innocuous 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 categorised 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 didn't 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. In fact, the video footage showed all participants cleaning the fuselage and actively mimicking real world actions with the prototype in every environment apart from E4.

Figure 4: Fuselage prop in PEL

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 made found that they offered no insight into functionality that had not been explicitly built into the low fidelity prototype. 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.

4 Discussion

The Face Validity results demonstrated that the higher the fidelity of the user 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 underperformance of the prototype in the SUS tests in the highest fidelity environment E1. While it created the greatest sense of Presence, its SUS performance was actually 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 just as highly as those in E1 and E2, with 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 supported by constructive comments, so the low score might be a result of participant frustrations. The SUS scores show a more reliable – if unexpected - 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). E3 yielded similar Face Validity results and time taken to complete results as the higher fidelity E2 environment, which further confuses the issue. 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 in E3, evidencing that E3 yields better prototype usability findings. However, E2 identified design flaws that should have been found at this stage of the design process, distinguishing E2 as a more informative environment.

Figure 5: User test E2 in PEL.

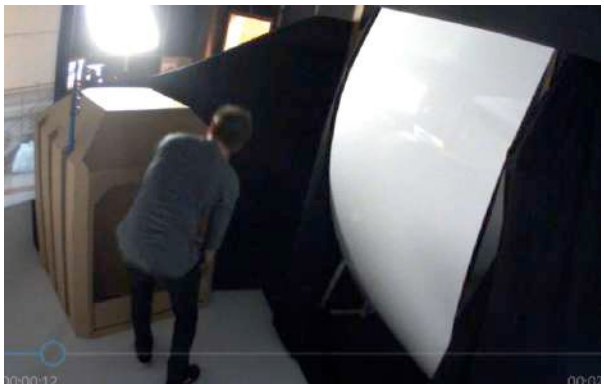

Figure 6: User test E3 in PEL

The trial proved inconclusive regarding the importance of high-fidelity environment details such as the use of the 'high viz' vests and smell in heightening Presence. However, E1 did achieve a higher Face Validity response to the statement '*the simulation feeling 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.

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. If reference is made to Nielson Norman Group's work on effort against financial benefit, then E2 offers the return required in identifying usability flaws early in the design process, against effort made to produce the 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 Conclusions

The purpose of this study was to establish the optimum context fidelity of a user-testing environment. The findings showed that the highest-fidelity environment yielded improved Presence and an assumption can be made that when presence is heightened, the context of use influences the amount of prototype usability issues found. The high-fidelity environments also clearly highlight more usability issues than the low fidelity environments, but time spent on extra detail such as smell and props such as vests will not necessarily deliver benefit.

The authors therefore concluded that simulating the context of use while testing a low fidelity design prototype early in the design 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 no major difference in Face Validity scores between the E1, E2 and E3 environments. The big drop in Face Validity was associated with E4, the lowest fidelity environment and effectively the experiment's datum, being a representation of a standard laboratory test. This 'context-less' 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 better data coming from the product trials. In the event, more usability flaws were identified in E2, 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 Kjeldskev and 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 next study will be conducted in three different environments, the real context, the laboratory and another laboratory equipped to the fidelity of Environment 2 as identified in this work. The focus will be on product usability with 15 participants in each environment and will aim to differentiate the role of each environment at different stages of the design process. This next study will not focus on which environment is most appropriate, but when each environment should be used.

6 Reference

Bangor, A., Kortum, P. and Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale, *Journal of Usability Studies*, 4(3), pp. 114-123

Boothe, C., Strawderman, L. and Hosea, E. (2013). The effects of prototype medium on usability testing. *Applied Ergonomics*, 44(6), pp. 1033-1038.

Brehmer, B. and 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), pp. 171-184.

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

Buchenau, M. and Suri, J.F. (2000). Experience prototyping. *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, ACM, pp. 424-433.

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.

Dahl, Y., Andreas. A. and Svanaes. D. (2009). Evaluating Mobile Usability: The Role of Fidelity in Full-Scale Laboratory Simulations with Mobile ICT for Hospitals, *International Conference on Human-Computer Interaction*, Springer, Berlin, Heidelberg, pp. 232-241.

Dawson, C. (2013). *Introduction to Research Methods: A practical guide for anyone undertaking a research project*. 4th Edition. How to Books, Oxford.

Deniaud, C. Honnet, V. Jeanne, B. and Mestre, D. (2015). The concept of "presence" as a measure of ecological validity in driving simulators, *Journal of International Science,* 3(1), pp. 1.

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. 514-531.

Etikan, I. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics,* 5(1), pp. 1.

Hare, J., Gill, S., Loudon, G. Lewis, A. (2014). Active and passive physicality: making the most of low fidelity physical interactive prototypes. *Journal of Design Research,* 12(4), pp. 330-348.

Johnson, P. (1998). Usability and Mobility; Interactions on the move, *Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices*.

Jokinen, J., Silvennoinen, J., Perälä, P. and Saariluoma, P. (2015). Quick Affective Judgments: Validation of a method for primed product comparisons, *Proceedings of the 23rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, pp. 2221-2230.

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,* 1(1), pp. 4-16.

Kassab, E., Tun. J.K., Arora. S., King. D., Ahmed. K., Miskovic, D., Cope, A., Vadwana, B., Bello, F., Sevdalis, N. and Kneebone, R. (2011). Blowing up the barriers in surgical training: Exploring and validating the concept of distributed simulation. *Annals of Surgery*, 254(6), pp. 1059-1065.

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

Kjeldsov, J. and Skov, M.B. (2014). Was it worth the Hassle? Ten years of mobile HCI research discussion on lab and field evaluations, *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services*, ACM, pp. 43-52.

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

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.

Muratovski, G. (2016). *Research for Designers. A guide to methods and Practice.* SAGE. London.

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

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.

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

Woolley, A., Loudon, G., Gill, S. and Hare, J. (2013). Getting into context early: A comparative study of laboratory and in-context user testing of low fidelity information appliance prototypes, *The Design Journal,* 16(4), pp. 460-485.

About the Authors:

Bethan Gordon: is the Deputy Dean of the Cardiff School of Art & Design (CSAD) and a Human Centred Designer. Research areas cover, HCD design thinking and education. She has lead on a number of design projects, the latest involving the New Curriculum for Wales.

Prof Gareth Loudon: research focusses on creativity and the innovation process, combining ideas from anthropology, psychology, engineering and design. He has over 30 years academic and industrial research experience, with several patents and over 80 publications to his name

Prof Steve Gill: is Director of Research and Professor of Product Design at Cardiff Metropolitan University. He has delivered international keynotes, published 80+ papers and 3 patents and co-authored an Oxford University Press book, 'Touch-IT' on physicality & design to be published in 2019.

Dr Jo Baldwin: is a designer and research scientist based in the Perceptual Experience Lab (PEL). Joe's doctorate involved research informing 'Fovography™', a revolutionary theory by FovoLab™, CSAD which discards linear perspective in favour experiential perspective – as used in PEL

Acknowledgment: Peter Mundy the lead designer and associate on the Knowledge Transfer Partnership. The Partner Company Window Cleaning Warehouse and their Directorate Stephen Fox and Julian Davies. These results informed the development of a product intended production.