Cognition, Technology & Work
https://doi.org/10.1007/s10111-025-00797-7

RESEARCH

Using mixed-fidelity to create effective environments

Jeremy Opie[1,2] · Shireen Jaufurally[1] · Carmen Salvadores Fernandez[1] · Lakchana Mahendran[1] · Adrien Desjardins[1] · Dimitrios Siassakos[1] · Anna L. David[1] · Manish K. Tiwari[1] · Ann Blandford[1,2]

Received: 6 November 2024 / Accepted: 27 March 2025

Abstract
User testing is crucial for ensuring that systems meet user expectations, involving both low- and high-fidelity prototypes. Recently, mixed-fidelity prototypes, which blend various levels of detail, have gained popularity but are mainly used for evaluating single digital prototypes. Typically, medical devices are tested only after engineering is complete, using high-fidelity prototypes in realistic environments. This paper explores when different fidelity levels can be applied and which levels are suitable for creating realistic use environments in medical technology evaluations. We used the PRETAR framework to determine the appropriate fidelity for each artefact based on study constraints, resources, ethical considerations, and data analysis methods. Using a case study of a sensorised glove for vaginal examinations, we developed artefacts with varying levels of fidelity to replicate the fetal head and birth canal to test whether mixed-fidelity prototypes can replicate clinical environments effectively while being cost- and time-efficient and critically examine each design suggestion, discussing the benefits and limitations of our decisions.

Keywords Mixed-fidelity · User studies · Healthcare · Vaginal Examination

1 Introduction

Within design and development there is an emphasis on testing early and testing regularly (Dumas and Fox 2012; Rueda et al. 2013) to identify potential issues early in the design phase Jensen et al. (2017). To achieve this, low-fidelity prototypes are typically developed that allow users to test a design idea, communicate how a system may work, validate designs, and identify issues without incurring high costs (Virzi 1989; Kelley 2001; Blackler 2008). For example, within web design, paper templates or mock pages created with online tools such as Figma can be used to demonstrate to a user the flow of a website to identify user concerns (Yew et al. 2020).

As a project progresses, there is a desire to develop higher-fidelity prototypes to assess realism (Rueda et al. 2013) or to justify stakeholders' investments (Deininger et al. 2019). Nevertheless, developing high-fidelity products can sometimes be cumbersome, costly, or overly restrictive for the requisite studies needed to evaluate the system (Rudd et al. 1996; Dieter and Schmidt 2021). Moreover, in the medical device community, there is an urge for higher ecological validity studies to be conducted as evaluations move from lab testing to the bedside (Borycki and Kushniruk 2010), and

✉ Jeremy Opie
j.opie@ucl.ac.uk

Shireen Jaufurally
s.jaufurally@ucl.ac.uk

Carmen Salvadores Fernandez
carmen.fernandez.15@ucl.ac.uk

Lakchana Mahendran
lakchana.mahendran.20@ucl.ac.uk

Adrien Desjardins
a.desjardins@ucl.ac.uk

Dimitrios Siassakos
d.siasakos@ucl.ac.uk

Anna L. David
a.david@ucl.ac.uk

Manish K. Tiwari
m.tiwari@ucl.ac.uk

Ann Blandford
a.blandford@ucl.ac.uk

1 UCL Hawkes Institute, University College London, London, UK

2 UCL Interaction Centre (UCLIC), University College London, London, UK

for researchers to report compromises made to ecological validity in their studies (van Berkel et al. 2020). To ensure researchers report on all aspects of a usability study, van Berkel et al. (2020) developed a framework comprised of seven dimensions; (1) user roles, (2) environment, (3) training, (4) scenario, (5) patient involvement, (6) software, and (7) hardware. However, this framework was not designed to assist with determining the appropriate level of fidelity within each dimension. Instead, this framework was designed to prompt researchers to consider all aspects when designing their usability studies.

The fidelity of a prototype is considered by most a direct relationship with the level of resolution or refinement of the object. McCurdy et al. (2006) describe mixed-fidelity as '*a prototype which is high fidelity in some aspects and low fidelity in others*'. They translate fidelity from a low- to high- spectrum of five orthogonal axes evaluating the levels of visual refinement and interactivity, the depth and breadth of functionality, and the depth of the data model for a single prototype. However, this definition of mixed-fidelity was designed to evaluate the fidelity of a digital prototype. In our study, we are exploring the use of mixed-fidelity in the use of both physical and digital artefacts not only in the prototype being developed but also in artefacts used to provide context of use during user evaluations.

Within HCI research, the use of low- and high-fidelity prototypes is commonplace and has been explored and compared extensively (Derboven et al. 2010; Lim et al. 2006; Brooks and Lopes 2023; Chen and Yoon 2024; Meck et al. 2022; Sim et al. 2016). Within the domain of healthcare, low-fidelity devices are often used in simulation environments to provide training on technical skills (Massoth et al. 2019). Studies have scrutinized the effectiveness of low- versus high-fidelity training simulators. For instance, Melvin et al. (2023) conducted a study in which they explored whether higher-fidelity prototypes outperformed lower-fidelity prototypes for achieving competency in fiber-optic intubation with resident clinicians. This involved residents performing 20 fiber-optic intubations with either an ORSIM airway simulator (high-fidelity) or a wooden-block task trainer (low-fidelity). Then, four months later their participants attempted a fiber-optic intubation on an anaesthetised patient whose airway was considered preoperatively normal. They found that using the high-fidelity device led to participants reaching fiber-optic intubation competency quicker and with fewer attempts during training. However, there was no statistically significant difference in their performance in regards to their checklist score, global rating scale, and the time to reach the carina, which is a ridge of cartilage located where the left and right bronchi split. A similar study was conducted by Chandra et al. (2008), who also found no significant difference when comparing a low- and high-fidelity simulators. Massoth et al. (2019) conducted a similar study with 135 medical students using low- and high-fidelity Advanced Life Support trainers and found that training medical students with high-fidelity simulators can lead to overconfidence in their abilities, which may have adverse learning affects. These studies highlight that similar levels of effectiveness can be obtained with different levels of realism. Similarly, Marcilly et al. (2024) evaluated a pain monitor using both low- and high-fidelity ecological validity to determine whether the level of fidelity affected participants' ability to identify areas of concern. They concluded that regardless of the fidelity level, they still identified the same number of use errors.

In usability evaluation, recent advancements in computer-aided design, 3D printing, and virtual environments have led to increased adoption of low-fidelity models (Joyner et al. 2022; Schembri et al. 2017). Despite this trend, there has been minimal exploration of mixed-fidelity in usability testing, where certain study aspects exhibit high fidelity while others remain low-fidelity. Existing studies mostly focus on combining virtual and real components (Barbieri et al. 2013; Speicher et al. 2021; Zhou and Rau 2019) or varying fidelity within a single component (Jensen et al. 2018). When evaluating the context of use for a novel medical device, phantoms or mannequins are commonly used as they provide a high-level of realism (DeWerd and Kissick 2014; Keenan et al. 2024; Gosselin et al. 2014). However, depending on the current state of the device and what aspect of the device is being evaluated, this level of realism may become a hindrance as high-fidelity prototypes may lead to the perception from users that they are interacting with a finished product and that their input is less valuable (Viswanathan and Linsey 2011; Jensen et al. 2018). Conversely, low-fidelity prototypes allow participants to create mental models of how they perceive the system could be, essentially a "deficiency compensation" effect (Sonderegger and Sauer 2010).

Within healthcare, several studies have reported on user evaluations of novel medical devices (Candidori et al. 2024; Dimitrakakis et al. 2022; Goldman et al. 2013; Li et al. 2012; Ramalhinho et al. 2023; Kotecha et al. 2021; Schnittker et al. 2016). However, for the most part these studies are carried out only after the medical device has reached a certain level of maturity, typically high-fidelity, and is evaluated in high-fidelity settings. This is a common theme, where the primary focus is on achieving the engineering requirements to develop a fully functional system, while neglecting to integrate human factors methods in parallel (Jerome and Kazman 2005).

In previous studies, we have shown that low-fidelity artefacts can be used in conjunction with high-fidelity medical devices to create a suitable use environment. In a study assessing the usability of a liver support device (Opie et al. 2024), the medical device itself was high fidelity;

however, to properly assess the device it needed to be connected to a patient. As the study was conducted prior to in-human studies, a jar of fluid was used to represent the patient, as this was the minimum requirement to assess the device.

In another study, we were assessing user perceptions of continuous AI in colonoscopy procedures (Van Berkel et al. 2022). For this study, we created a custom application to record participants' navigation behaviour and clinical assessment during a mock AI polyp examination. Initially, a custom 3D printed interface was developed to mimic the examination of the colon allowing participants to scroll through videos with two buttons they could use to classify the area highlighted by the AI as either a polyp or not a polyp. When this design was shown to the research group prior to commencing the study it was decided that the interface was too low-fidelity and there was concern that the device may look unprofessional to the study participants, all of whom were clinicians. The custom interface was scrapped and an Xbox controller was used in its place. However, during the study, it was observed that participants who were familiar with Xbox controllers used the device as expected, whereas those less familiar with Xboxes used the device in a variety of ways, with one participant positioning it on the table in front of them on its side and only interacting with the joystick with their fingers. In this instance, a lower-fidelity interface could have been utilised, on the proviso that the custom interface was designed to be more aesthetically pleasing.

To establish the appropriate level of fidelity of an artefact the design of a study should be scrutinised to ensure that it fits within the framework of that study. PRETAR (Blandford et al. 2008) is a framework that has been established to assist researchers in planning a user-centred study. Specifically, PRETAR focuses on the **P**urpose or goals of the user study, **R**esources and constraints the study has to work within, **E**thical considerations, **T**echniques used for collecting data, how the data will be **A**nalysed, and finally **R**eporting the findings.

The aim of the work reported here is to determine whether the use of mixed-fidelity artefacts in user evaluation can reproduce ecologically valid environments for the early assessment of medical devices. Our paper uses a case study of the evaluation of a sensorised glove used to assist clinicians in correctly identifying the position of the fetal skull when performing vaginal examinations during second stage labour. However, the effectiveness of the glove is not the focus of this paper. Therefore, to determine whether mixed-fidelity artifacts can reproduce ecologically valid environments, we focus on the level of abstraction suitable for user studies by using the PRETAR framework to help guide suitability. We found that incorporating mixed-fidelity into our artefacts allowed us to create an ecologically valid use environment whilst working within the constraints of our project. This environment was evaluated with participants (n = 18), who provided feedback on the study design through a think-aloud and a semi-structured interview after the user evaluation. To illustrate this, Fig. 1 depicts the relationship between the case study, the PRETAR framework, and the overarching principles of fidelity. Specifically, the case study objectives are assessed using elements of the PRETAR framework, which inform design recommendations for each artefact. Based on established design principles, each environmental artefact is then classified with some level of fidelity.

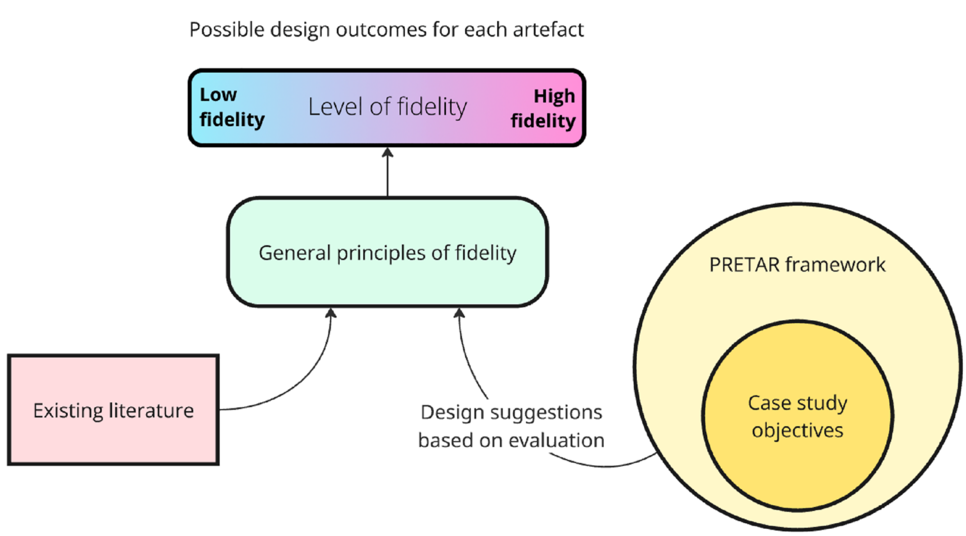

Fig. 1 The relationship between the case study, PRETAR framework, and level of fidelity for each artefact

2 Case study: Vaginal examination during second stage labour

Digital vaginal examination (VE) is a technique where a clinician inserts digits, i.e., their fingers, into the vagina of someone who is in labour to identify the position of the foetus based on anatomical landmarks on the foetal skull (Lipsett et al. 2023). The sutures are the joints between the five main skull bones of the baby's head: frontal bones, parietal bones, and the occipital bone. At the intersection of the sutures are fontanelles, with the anterior fontanelle at the front of the head having four sutures adjoining it, and the posterior fontanelle at the back of the head having three sutures intersecting (Lipsett et al. 2023). Using the tactile feedback from their fingers, clinicians can find suture lines, allowing them to locate the fontanelles and determine the position of the foetus based on the position of the anterior and posterior fontanelles. There are four positions the foetus may be, which are: occipito-anterior (OA) (head facing back), occipito-posterior (OP) (back to back), and occipito-transverse (either LOT (left occipito-transverse) or ROT (right occipito-transverse) depending on the direction the foetus is facing) (on their side), see Fig. 2. Correctly identifying the position of the foetus is essential so that appropriate means are used to assist with labour (i.e., manually or instrumentally rotating the foetus to the correct position, OA, for delivery). However, during labour, pressure from the birth canal can affect the shape of the foetal skull. As the bones of the foetal skull and suture lines are still soft and malleable the shape of the skull can deform, causing the bones to overlap slightly which is known as moulding. This pressure can also cause swelling on the scalp known as caput succedaneum, which can sometimes occur over the suture lines making them more difficult to identify. From this point forward we will refer to this as caput as this is the term commonly used in conversation.

Accurately determining the position of the foetus during second stage labour is critical to ensure that the correct method is used to safely deliver the baby. Currently, instrumental births (ventouse/forceps) are common, with up to 30% of first-time mothers undergoing such procedures (Anim-Somuah et al. 2018), and around 20% of women having caesarean sections (NHS 2022). In approximately one-third of instrumental births, incorrect instrument placement occurs due to inaccurate assessment of the fetal position, potentially causing trauma to both mother and baby (Ramphul et al. 2015). Failed instrumental attempts often result in a caesarean section when the woman is fully dilated. Performing a caesarean at full dilation can increase the risk of preterm birth in future pregnancies by about 15% (Hotton et al. 2019), a significant contributor to neonatal morbidity and mortality (Richter et al. 2019). In 2012–2013, preterm deliveries cost the UK public sector £3.4 billion (NENC Preterm Birth Group 2021).

Although ultrasound scans — considered the gold standard — can accurately diagnose fetal position (Ghi et al. 2018), they are often unavailable in emergency situations due to the machine's size, lengthy setup time, and operator dependency. As a result, the current standard practice is to rely on digital vaginal examinations (VE), where clinicians use haptic feedback to identify anatomical features of the foetal head. This method is challenging due to factors like swelling on the baby's head, overlapping skull bones, and the inability to directly visualize the examined area.

To aid in determining fetal position during this critical time, we are developing a novel wearable, sensorised surgical glove. This glove incorporates a triboelectric sensor onto the index finger, capable of detecting real-time changes in rigidity as it moves across surfaces (Jaufurally et al. 2023). For instance, the sensor can distinguish different surface textures when moving over materials of varying rigidity. In this study, we evaluated the glove — referred to hereafter as the '*suture glove*' — to assess its ability to detect when a suture has been crossed during vaginal examinations in the second stage of labour (full dilation).

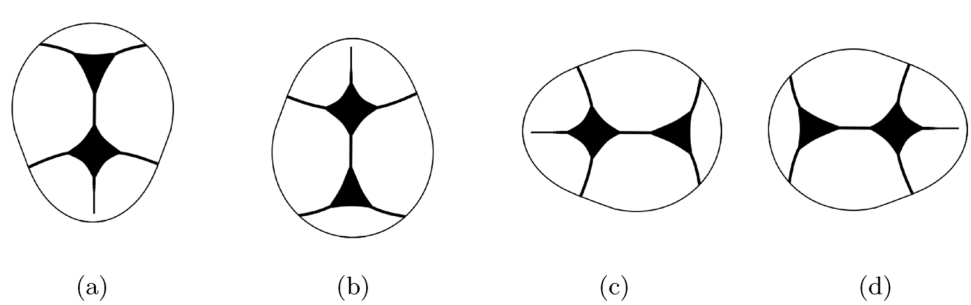

Fig. 2 Positions of the foetal skull during second stage labour using their abbreviated terms **a** '*OA*' **b** '*OP*' **c** '*LOT*' **d** '*ROT*'

3 Using PRETAR as a guide

PRETAR has been established as a framework used for structuring user-centred evaluation studies (Blandford et al. 2008) and has been used in various domains (Edwards et al. 2013; Hess 2015; Makri et al. 2011). As noted above, PRETAR focuses on the **P**urpose or goals of the user study, **R**esources and constraints the study has to work within, **E**thical considerations, **T**echniques used for collecting data, how the data will be **A**nalysed, and finally **R**eporting the findings.

In this study, we are assessing whether mixed-fidelity artefacts can provide realistic environments when assessing early-stage medical devices. To establish the appropriate level of abstraction for each artefact, we employed the PRETAR framework to systematically guide our decisions on fidelity design for each environmental component, incorporating the suture glove in alignment with the study's objectives. We have included Table 1 to summarise how each element of the PRETAR framework relates to the aims of the case study, and the following subsections provide a detailed discussion of each of its key elements. However, the reporting column has been omitted as the results of this study are designed to be reported to a regulatory body in preparation for potential clinical trials.

3.1 Purpose

The vision is that the suture glove is a device that can be used during vaginal examinations by a trained healthcare worker to recognise obstructed labour earlier, improve safety, and reduce the number of vaginal examinations required. The suture glove is designed to be used globally, including in rural communities, and is estimated to have a cost of approximately US$1 per set of gloves.

Currently, the suture glove is in early development, and this is an early formative study to help in the future design and implementation of the device. To determine the effectiveness of the glove, how to improve the glove in future iterations, and how a broad range of users may use the suture glove, we designed this study with three core aims:

1. To ascertain whether the suture glove could accurately and reproducibly determine the foetal position on a phantom neonatal head.
2. Whether intended users are comfortable using such a device, to determine how they perceive it amalgamating into their current workflow, and identify design requirements to improve the system.
3. To understand the various techniques that intended users currently employ when conducting vaginal examinations.

Essentially, we are investigating whether the sensor can detect the change in surface texture between the foetal bone and the sutures when participants with varying techniques and healthcare disciplines use the glove.

3.2 Resources and constraints

It was possible to conduct this study due to the multidisciplinary team that developed the study and the available resources within our lab. Our team included a clinical fellow with 7+ years of training in obstetrics, an engineer specialising in nanotechnology, a human factors expert specialising in healthcare, and a medical student with a passion for women's health. This team made it possible to understand the clinical factors associated with second stage labour and

Table 1 Relationship between the case study objectives using the PRETAR framework

Research objective	Effectiveness of the suture glove	Participant feedback	Vaginal examination method
Purpose	Determine if the suture glove is accurate and findings are reproducible	Thoughts on suture glove, Integration into current workflow, and design improvements	Understand the various techniques
Resources and Constraints	High-resolution data required from participant examination of mock foetal skulls	Level of participant experience, recruitment method, duration of study, and timeframe to complete study	Ability to view the movements of the hand at all times
Ethics	Could only be evaluated in a mock scenario	The data would be anonymised	Could only be evaluated in a mock scenario
Techniques for data collection	Record signal readings from the suture glove via a data acquisition device	Think-aloud, semi-structured interview, and a questionnaire	Video recordings of hand movements on the foetal skull
Analysis	Review of electrometer readings in conjunction with video data	Thematic analysis and using the questionnaire protocol	Analysis of the recorded video capturing hand position and finger techniques

the difficulties with correctly identifying the position of the foetal skull. We had the knowledge and resources to develop the sensorised gloves within our lab and to validate the accuracy and reproducibility of the suture glove, create artefacts, conduct user studies, evaluate qualitative data, and produce design requirements that can improve future iterations.

Additionally, our lab provided the resources to design and develop most of our artefacts, having 3D design software and printers, access to modelling materials, electronic instrumentation, and software. We also had access to classrooms where studies could be conducted that were easily accessible.

Despite the high level of resources available, two of the team members were conducting this research as part of their PhDs, allowing only a few months to conduct the studies. Additionally, even with our access to 3D design software and printing capabilities, we were unable to generate some of the artefacts due to the complexity of their design. However, we did have access to external suppliers who have the capability of developing and 3D printing high-fidelity parts.

3.3 Ethics

Based on the level of maturity of the suture glove, that is having only been lab tested, using an early prototype of the glove that is still physically connected to the signal acquisition device by a wire, and software that is also still in early development, this study was designed for this stage of evaluation. With a more mature iteration, an in-human study, that is conducting a VE on a patient during second stage labour with the suture glove, may be feasible; at this time we are unsure of the potential risk this system may have on the patient and unborn child. Therefore, a mock VE was deemed most appropriate.

Additionally, as this study was collecting a significant amount of qualitative data from audio recordings of the participants, from the outset we informed them that this information would be anonymised. To achieve this, after each session all recordings, both video and audio, were immediately transferred to a secure password protected university research drive. The audio was then transcribed and participants were given new identifiers, P1-P18, so that data between the various methods could be synchronised, and the original recordings deleted. During the interviews and think alouds, when specific details were mentioned, e.g., names, institutions, etc., these were generalised in the transcriptions, e.g., [patient], [researcher], [hospital], etc.

Based on these constraints, ethics for this study was obtained from UCL, 22491/001, which allowed for the acquisition and analysis of participants using the suture glove during mock VEs. After being recruited, participants read the participation information sheet and signed a consent form to state that they were happy to be involved in the study.

3.4 Techniques for gathering data

This study employed a mixed methods approach to gathering data, which included sensor data acquisition, audio recordings, video recordings, and a questionnaire.

The sensor data was captured via LabVIEW, the software used to program the glove, through an electrometer (Keithley 6517B), as it can measure low currents and is controlled via software written in LabVIEW.

The audio recordings were taken from think-alouds as the participants were using the devices and from semi-structured interviews. Think-aloud enabled participants to verbalise their thought processes including what they were trying to do, what they were feeling, and any other comments about the study. The focus of the semi-structured interview was to determine their thoughts on using the glove and how it affected their tactile feedback, how it could be adopted into clinical practice, barriers that may arise, and their opinion on the interface. Additionally, we explored the ecological validity of the study design, with a focus on the setup and the fidelity of the phantom foetal heads.

To identify how people performed vaginal examinations, and to validate if they were using the glove correctly, video recordings of the participant's fingers on the phantom foetal skulls were captured.

Lastly, a System Usability Study (SUS) questionnaire was also used, for the participant to report on the usability of the suture glove in its current state.

3.5 Analysis

The data from the sensor was analysed by reviewing the data and looking for changes in the current. A drop in current represented a transition from one material to another. These moments were then compared to data captured from the video recordings, to ensure that the participant was moving over a suture and not over a different part of the foetal skull. This provided a complete picture of the effectiveness of the suture glove.

The audio recordings were transcribed verbatim and anonymised (i.e., all identifiable information was removed). Then, using NVivo, a software program designed specifically for qualitative research, the data was coded and labelled. Once completed, the labelled sections of the transcript were grouped based on their coding into themes to identify participants' perceptions on the use of the suture glove, how it fits within clinical practice, and any barriers that exist to moving this device to clinical trials.

The SUS scores were all generated and analysed following the protocol for this questionnaire (Brooke 1996).

3.6 Reporting

This study was conducted as a formative evaluation of the device as part of regulatory requirements before beginning clinical trials. This was to ensure that the device was safe to use, as well as being reliable and repeatable in its use.

4 Pilot study

4.1 Participant

Before conducting our study, we ran a pilot study to test our setup and identify any areas of improvement. An obstetrician with 6 years' experience who had no prior experience with the suture glove was recruited by the clinical researcher on the team.

4.2 Artefacts

4.2.1 The suture glove

The suture glove was developed at UCL and was the focus of the study; it consists of a flexible triboelectric sensor attached to the index finger of a surgical glove (Jaufurally et al. 2023). The sensor is made using a metal-oxide nanocomposite spray, allowing for flexibility and, using the properties of triboelectrification, it generates a current when two originally uncharged materials are separated after making contact. The resultant current can be measured to identify when a change in surface texture occurs; this has been previously lab tested (Jaufurally et al. 2023). To ensure sterility when using the suture glove a second standard surgical glove will be used to cover it, known as double-gloving.

Currently, one of the constraints of the suture glove is that it needs to be physically connected to the signal acquisition device to read the signals. The final proposed version of the suture glove is intended to be wireless, via Bluetooth, with the ability for the output to be viewed from a different device, potentially a mobile phone. However, at this time the hardware is still undecided until further design considerations are understood, i.e., context of use, hospital regulations, and intended users.

4.2.2 The phantom foetal skulls

To evaluate the suture gloves under various conditions, two replica foetal skulls were hand-made using 3D CAD software based on images of foetal skulls; (1) a 'normal' phantom with no suture deformation, and (2) a phantom with 'moulding'. The normal foetal skull had average width suture lines whereas the other had wider suture lines representing a foetal skull with moulding. From the designs, negative moulds were created and using a silicone elastomer to fill the moulds the foetal skulls were created. During the pilot study, the phantom skulls were attached to a stand so they could be rotated to different positions.

4.3 User study

4.3.1 The setup

Prior to conducting the study, the participant was provided with the participation information sheet, which they read, and then signed the consent form. Following this, they proceeded to put on a suture glove with a secondary regular glove on top and attached it to the signal acquisition device. The participant was then asked to sit down beside the setup to hold the foetal skull, positioning themselves next to the foetal skull as they would in a normal VE. Then one of the two phantoms was placed onto the setup for them to begin exploration. So that the participant could not see the position of the foetal skull during the study, they were asked to look away from the model and close their eyes. An image taken during the study can be seen in Fig. 3.

4.3.2 Assessment

With a foetal skull in place, the participant was asked to conduct a VE on the skull as they normally would to identify its position while using the think-aloud method

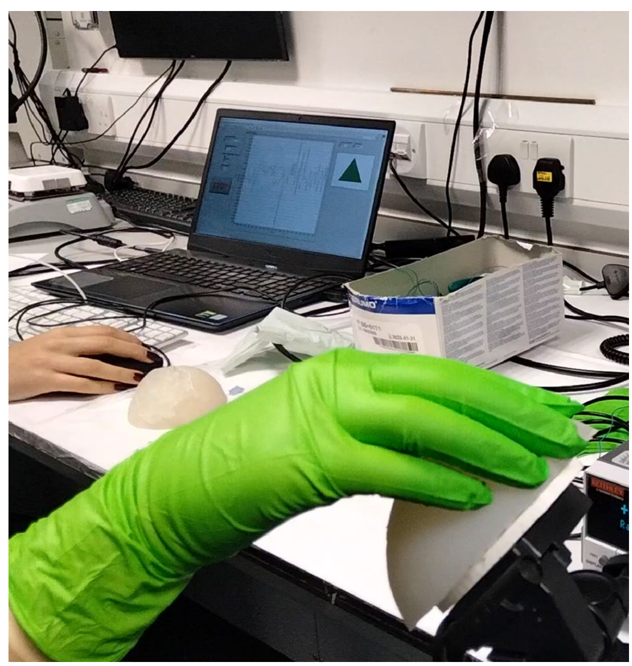

Fig. 3 A still image taken from the pilot study, showing the participant conducting a mock VE on the low-fidelity foetal skull cap

to verbalise their thought processes (Ericsson 2003). The participant was asked to evaluate the position of the foetal skull as they would in their clinical practice. When the participant was confident with their decision regarding the position of the foetal skull, they were asked to move to one of the two fontanelles and then use the suture glove to move around the fontanelle while the engineer captured the readings from the glove. Using this approach, we could compare (i) the true position of the foetal skull by the way it was placed, (ii) the participant's assessment of the foetal skull's position, and (iii) the suture glove's estimation of the foetal skull position. The study was repeated eight times, with the position and type of foetal skull changing between each examination. For two of the recordings, two pieces of cling film with gel between them was placed over the 'normal' foetal skull to represent caput.

During the assessment, the participant did not view the computer monitor with the outputs of the gloves readings so as not to influence their perception of whether they could feel a suture. However, after these assessments, they conducted two unassessed examinations where they could observe the output from the monitor. In this evaluation, the participant continued to use the think-aloud method, so we could identify how they interpreted the data from the suture glove.

Each researcher had a specific role in coordinating the study and helping it run smoothly. One of the researchers was video recording each of the assessments, another was logging the data on the computer, and a third was managing the position and type of the foetal skulls. During the evaluation, all of the researchers were actively watching the participant and asking questions relating to their evaluation to better understand their thought processes, which were captured as part of the think-aloud.

Additionally, during the studies body language and the position/posture of the participant were observed and recorded to identify moments when the participant appeared frustrated or unsure as to what they were feeling/seeing. When these moments were identified the researchers would remind the participant to verbalise their thoughts to gain insight into the frustration. Furthermore, the researcher who conducted the semi-structured interview included some of these observations in the interview to explore potential solutions to overcome those frustrations.

4.3.3 Post-assessment

After the examination with the suture glove, a semi-structured interview was conducted with the participant and a System Usability Study (SUS) questionnaire, as described in Sect. 3.4.

4.4 Analyses

The data from the study was analysed using the approach described in Sect. 3.5. However, for this pilot study we were specifically exploring areas of the study that needed improving so future studies could yield robust results.

4.5 Results

From the pilot study we were able to identify several key elements that needed further refinement; (1) a screen to hide the foetal head, (2) a means to restrict hand movements, (3) a higher fidelity phantom foetal skull, (4) an easier method to switch and rotate phantoms, and (5) changing the colours of the fontanelle indicators on the UI.

Our study was specifically designed to observe the participants' hand movements during the mock VEs, ensuring the foetal head remained completely unobscured. However, from the pilot study, this was found to create a lack of realism, thus impacting the ecological validity of the environment of use.

One of the main observations was that, as they were required to look away and have their eyes closed, the participant was facing away from the skull, which was not representative of how they would be sitting in practice. Therefore, a screen to hide the foetal head was required.

Another concern from the participant was the freedom to touch the foetal head without any barriers, stating "*You haven't always got the full range of movement or the full feel of the full foetal head. But to make it more realistic, it should be for me a sort of tube that you put your hand into... and then assess it in a tight spot*". When discussing the phantom foetal skulls, the participant shared that they thought the feel of the foetal skulls was generally good, but "*less realistic*" than they would expect, particularly the caput.

During the study, a lot of time was spent between VEs switching and repositioning the phantom foetal skull due to how they were mounted and the limited range of motion of the setup. To rotate the phantom foetal skull 90 degrees, from either OA or OP to LOT or ROT, required loosening two parts of the setup to swivel it into the right position.

Lastly, during the semi-structured interview, the pilot participant remarked that "*[the fontanelle colours] may be alarming to have one that's green and one that's red. I think if they were... more neutral like blue and yellow, then maybe that would be less confusing than green and red*".

From our pilot study, we identified several artefacts that required improvement before proceeding with the user study. The following sections will focus on the user study, emphasising the modifications made to both the methods and artefacts as a result of the pilot study.

5 Method

5.1 Participants

As our team included an obstetrician and a medical student, recruiting participants from within obstetrics, midwifery, and student groups was relatively easy. Using mostly word of mouth, we were able to recruit participants for the study (n=18). These included: obstetricians (n=4), midwives (n=6), student midwives (n=3), and medical students (n=5), see Table 2. These participants were included because they were identified as intended users or potential future users, that is student midwives and medical students with an interest in women's health. Ideally, we were hoping to have five participants per user group. However, despite our personal connections, we found it difficult to recruit obstetricians due to their busy schedules and the location for the studies being off-site; also, during the time we conducted the studies, student midwives were in the middle of their exam period.

Table 2 Participants used in the study, highlighting their role and years of experience within that role

ID	Role	Years XP
P01	Midwife	7
P02	Midwife	6
P03	Midwife	6
P04	Midwife	6
P05	Obstetrician	4
P06	Obstetrician	4
P07	Student midwife	3rd year
P08	Obstetrician	9
P09	Student midwife	3rd year
P10	Medical student	3rd year
P11	Medical student	3rd year
P12	Medical student	2nd year
P13	Medical student	3rd year
P14	Medical student	3rd year
P15	Midwife	9
P16	Student midwife	3rd year
P17	Obstetrician	2
P18	Midwife	7

5.2 Artefacts

The artefacts used for this study, specifically those in relation to the simulated environment, differ from those described in Sect. 4.2. These new artefacts are described below. However, the suture glove remained the same between the pilot and the main study.

5.2.1 The phantom foetal skulls and caput

As with the foetal skulls used during the pilot study, we developed two foetal skulls with one representing a 'normal' skull and the other 'moulding'. These new foetal skulls were 3D printed from 3D scans of a real foetal skull. As we were unable to produce these parts in our own labs, we had these developed by an external supplier. Different consistencies of silicone were used so that the skull bones were rigid, but the sections replicating the suture lines and fontanelles were a softer silicone. The caput used during the pilot study was also inadequate and was difficult to hold in place on the foetal skull during examination. To improve this a caput was made using soft silicone to create a membrane that acted like a cap that rested on top of the 'normal' skull.

These models satisfied the first goal by creating high-fidelity models to generate high-resolution data to evaluate

(a) (b) (c)

Fig. 4 Various high-fidelity skulls used during the study: **a** '*Normal skull*' **b** '*Normal skull with caput*' **c** '*Skull with moulding*' the suture glove, as well as satisfying the feedback from the pilot study. All of these variants can be seen in Fig. 4.

5.2.2 The phantom birth canal

During our pilot study two issues arose relating to the birth canal, (1) there were no constraints to the hand movements making it feel unrealistic, and (2) too much time was spent changing the position of the foetal skull.

Currently, in obstetrics, mannequins are commonly used in training scenarios when learning child birthing techniques (e.g., forceps, ventouse, manual rotation, etc.). These mannequins are lifelike and display the anatomy from the top of the belly down to the middle of the thigh. One of the downsides of using these mannequins is that it is difficult to observe exactly what the trainees are doing with their hands, even with the mannequin's belly skin removed, and training is given via verbal instructions based on the trainee verbalising what they are feeling rather than observing their hand movements and providing direct feedback. This is in contrast to one of our study goals, as we wanted to observe how participants used their hands to conduct VEs to gain an understanding of common practices. Therefore, it was important that the participant's hand be visible during the study.

When designing the phantom birth canal, many ideas were put forward, including using an existing silicone mould of a birth canal, creating our own mould, or creating an abstract opening to represent a birth canal. The issue with existing moulds is that the study is designed for patients who are in the second stage of labour, essentially fully dilated, which these moulds are not. Additionally, to make a mould of a fully dilated patient would necessitate taking the shape during a live birth, which would be unsafe for the unborn child and impractical for the mother, as full dilation only occurs just before birth. Also, to create and modify an existing model to be fully dilated is impractical. We also contemplated 3D printing a transparent birth canal so that the hands could always be seen, yet providing a high level of realism. However, the time and financial resources required to achieve a robust setup exceeded the project's constraints. Thus, we were left with creating an abstract opening to represent a birth canal; we cut a hole in a piece of plywood to represent the size of a fully dilated patient, approximately 90 mm in diameter. The hole was large enough for the participant to insert their hand, but small enough that it obscured the participant's view of the foetal skull and limited their hand movement.

As we were using a phantom birth canal rather than a patient or mannequin there was no pelvic bone to secure the foetal skull into a set position. We designed our system to allow for quick adjustments of the foetal skulls and positions using a camera tripod stand that provided a large range of rotation and a quick release as a base for the setup. Connectors designed in Autodesk Fusion 360 were 3D printed to attach to the tripod stand to hold the board representing the phantom birth canal, and quick release connectors were added to the bases of the foetal skull 3D prints so they could be interchanged quickly between examinations. This setup allowed us to quickly and easily switch the positions and foetal skulls, allowing us to explore more positions within the allotted time. A screenshot taken from one of the recordings can be seen in Fig. 5 that shows the setup of the device in use.

Table 3 Example order of assessments of the foetal skulls showing the position of the foetal skull and the type of foetal skull (Participant 6)

Position	Foetal Skull
occipito-posterior (OP)	Normal
occipito-transverse (ROT)	Normal
occipito-anterior (OA)	Normal (Caput)
occipito-transverse (ROT)	Moulding
occipito-anterior (OA)	Moulding
occipito-posterior (OP)	Normal (Caput)
occipito-transverse (LOT)	Normal (Caput)
occipito-posterior (OP)	Moulding

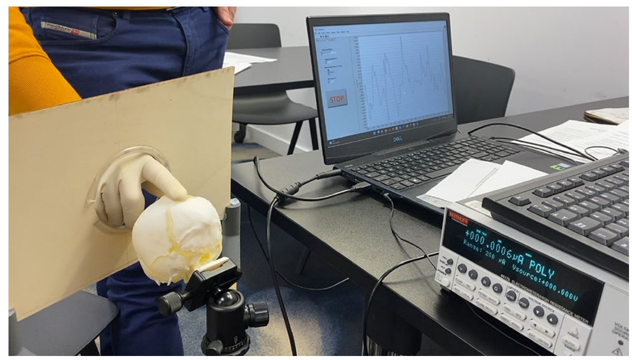

Fig. 5 A snapshot of the device being tested with a participant

5.3 User study

Following the same process as the pilot study, see Sect. 4.3, the user study was made up of three parts: the setup, assessment, and post-assessment. The only differences were a consequence of the inclusion of the phantom birth canal, and a method to remove bias due to the number of participants within the study. These are both described within the part of the study where they were conducted. Additionally, the data from the study was analysed using the approach described in Sect. 3.5.

5.3.1 Assessment

A foetal skull was attached to the phantom birth canal setup, and when the participant was ready, they were asked to insert their hand into the phantom birth canal and begin their evaluation using the sensorised glove.

To remove bias, each participant performed eight randomly selected case scenarios, which were randomised using the Latin Square method so that as many possible variations to the order could be tested, with each assessment changing the position and type of foetal skull, see Table 3 for an example. Additionally, participants were not given an exhaustive list of positions to assess to reduce the time of the study.

6 Results

As this study took a mixed-methods approach, we were able to capture data specific to each of our research objectives. As these outcomes are not the focus of this paper, which is instead exploring the use of mixed-fidelity in user evaluation to reproduce ecologically valid environments, they are not reported in detail here and instead are being reported in other outlets. However, we offer some of the overarching results:

- Participants were impressed by the simplicity of the device and that it did not impact their tactile feedback
- The method of using the suture glove was well received, with some participants stating they would adopt this technique in their current practice
- The use of a suture glove was perceived as having little impact on the existing workflow
- The Design of the glove needs to be improved for use in clinical practice, i.e., feedback modality, wireless connectivity
- The suture glove did not perform as well as expected from a technical perspective and requires further development and evaluation
- The suture glove's intended users should only be obstetricians, as midwives have different clinical objectives

These outcomes, plus others, will be taken into consideration for future iterations and studies with the device. For this paper, we assess each of the artefacts used to create an ecologically valid environment based on the results from the semi-structured interview, where participants were asked to comment specifically on the artefacts. This resulted in participants discussing their likes and dislikes for each item and whether in provided a level of realism that mimicked that of an actual vaginal examination.

6.1 The suture glove

Despite the suture glove being the medical device evaluated in this study, we included participants' thoughts on the use of the suture glove as the way it was situated within the environment impacted the ecological validity of a vaginal examination.

One of the major concerns, and a cause of frustration during the study, was that the suture glove was tethered to the signal acquisition device. During the first couple of studies, participants would complain that their arms were getting tired or they would pull their hand out from the phantom birth canal, as they would in clinical practice, resulting in them either disconnecting the glove from the device or pulling the wire off the sensor requiring them to put on a new suture glove. To alleviate this, participants were asked to rest their hands either inside or on top of the phantom vagina.

6.2 The phantom skulls

One of the most notable achievements of this study was that we were able to design a system that provided participants with enough realism to mimic VEs in the real world. When discussing the setup of the study, specifically the skulls and the caput, for the most part participants had similar comments with P04 sharing "*I thought they were like really quite realistic compared to like, the [ones used with the mannequins] that I'm used to.*" The moulding skull was unanimously considered to be realistic, "*This is like really realistic… This is what it always feels like.*" (P01).

Despite the majority of the feedback being positive (n = 14), three participants indicated that the skulls seemed unrealistic. When asked to explain they expressed their thoughts on specific skulls articulating that the normal skull felt to have unrealistic landmarks, specifically, "*the fontanelles felt bigger than they really would be in real life*" (P01), the "*suture lines were a bit far apart*" (P02), and "*you didn't have the same movement of the bones that you have in real life*" (P05).

However, there was one standout comment raised by each of the obstetricians that was unexpected in that the foetal skulls lacked ears with P8 expressing "*Your foetal heads don't have ears. So that threw me a bit initially because I tend to use ears*".

6.3 The caput

The participants' reactions to the caput were mixed between unrealistic, difficult and confusing, through to realistic. For the most part, participants with little to no experience with VE found identifying fontanelles through the caput difficult, as it created another barrier between their tactile feedback and the sutures. However, those with more experience with caput shared their frustration stating, "*this [caput] feels more firm than [normal] caput and feels more diffused over the foetal head rather than in real life where the caput feels more like it's on one focal point of the head that has been pushing down on the cervix rather than spread over the whole head*" (P17) and "*I think the cap was a lot less true to real life*" (P05).

When using the caput during this study it quickly became apparent that the design of the caput created moments of frustration because the caput was designed to sit on top of the foetal skull with nothing to hold it securely in place. This required one of the researchers to hold it in place with their fingers while the VE was being carried out to stop it moving, see Fig. 6. This in turn resulted in participants either moving beyond the caput and onto the foetal skull or feeling the fingers of the researcher who was holding the caput in place and thinking it was something else. However, when this occurred the researcher would inform the participant that it was their fingers they were feeling and not another part of the foetus' anatomy.

6.4 The phantom birth canal

When describing the phantom birth canal, participants reported that it provided them with a sense of realism despite the abstract nature of the design, "*it is clearly not real, but I think it is quite high fidelity in the sense that is the same struggle to try and identify what position the baby's in and I was using the same technique that I would use clinically*" (P08). Additionally, "*it's not quite like how it feels like in real life, but it does approximate to the fact that there is some kind of obstruction*" (P01) and "*you don't have the patient impact and the rest of the environment but as a test, it's perfect*" (P06). The design was also noted as a guide for participants with P12 sharing "*I quite like how simple it is because I think if there's loads of like, other strategies, could be quite distracting. When it's just like, a cardboard with a hole cut out. I know that that's what I have to do. Just put my hand through, so I quite like the setup.*"

Fig. 6 One of the researchers holding the Caput in place while the participant is examining the foetal position

To observe the participants' VE techniques, we opted to have an open birth canal, as discussed above. However, for the midwives, this was a point of frustration and some found it "*off putting*" (P09) and "*you've got a lot more space to move your hand all the way around*" (P03). The open space and lack of a cervix and vaginal wall were found to be initially disorientating for these users.

During the interviews, it became clear that the vaginal wall was something that some participants relied on to assist with their examinations. P09 shared "*there was none of the, like the vaginal wall that you'd be pushing against with your hand to get to the heads · So, you would be in but your hand would be more sturdy because you're being clamped by the woman. So, you're just at the cervix and you're trying to feel the head. That wasn't as easy to do which is why you have to hold on· Whereas normally, you've got that barrier the woman's vagina would be gripping· you feel a bit more guarded and [here] you just rotate your fingers. Whereas normally you can't rotate your whole hand or put more fingers in.*" Additionally, it was observed that P18 was using multiple fingers during their observations, when asked about this they responded "*I usually have two fingers in but when I was doing the manoeuvring on the skull, I was resting my other fingers on there, which I wouldn't usually do. But I guess because there's usually less space. It felt harder to almost manoeuvre my fingers because there was so much space. Whereas you kind of sometimes use like the vaginal wall I would use to kind of as like a, like a surface to be able to push against to be able to move your fingers a bit more so it felt more awkward on there on the models, which I would expect it to but I don't know whether that's because I'm just so used to examine in a normal situation*". Some of the midwives felt that the inclusion of having a cervix would help focus access to the skull with P15 sharing "*is that sometimes I got confused on where the end of the skulls was. Quite often when I was touching, [the clinical researcher] was like, 'Oh, you're going too far.' Whereas if there was a cervix there, I'd kind of know that that was my stopping point*".

Additionally, during the first few observations of midwives, the clinical researcher observed that the midwives were trying to first examine the cervix to determine the dilation of the phantom model. However, as our phantom birth canal was an abstract concept, there was no cervix. For future participants, we made them aware of the limitations of the model prior to commencing and put more emphasis on the state of the foetus for this scenario. Despite these limitations, all of the participants noted that without the presence of a vaginal wall and cervix it allowed us to observe and record their examinations.

7 Discussion

When conducting user evaluations, creating an environment that best represents that used in clinical practice yields better outcomes and can highlight areas of design that require further attention early in the development phase. Although there have been several studies that have conducted early user evaluations on medical devices (Candidori et al. 2024; Dimitrakakis et al. 2022; Goldman et al. 2013; Li et al. 2012; Ramalhinho et al. 2023; Kotecha et al. 2021; Schnittker et al. 2016), the methods by which the fidelity of environmental factors are considered are typically not specified. The seven dimensions of ecological validity can be used to provide researchers with an understanding as to what aspects of their study they can improve to increase the ecological validity of their study (van Berkel et al. 2020). Furthermore, when discussing '*Environment*' Van Berkel et al. discuss the need for attention to be given to practical, ethical, and study design to inform the realism of the study. Although they provide a high-level overview of what factors researchers should attend to to improve ecological validity, this dimension does not provide guidance for the level of fidelity of artefacts that can provide an appropriate level of realism. Therefore, we propose that using mixed-fidelity artefacts can establish adequate ecologically valid environments for evaluating early-stage medical devices.

By using PRETAR (Blandford et al. 2008) as a guide in our study, as done in Table 1, we were able to identify what was achievable given the contrasting elements to reach a desirable outcome. With this information, coupled with the feedback from the pilot study, we were able to identify where more effort was required to create a satisfactory use environment that would help to identify key areas of improvement for the suture glove.

Additionally, extending on the idea of mixed-fidelity by McCurdy et al. (2006), we were able to evaluate whether mixed-fidelity artefacts were able to provide the same level of realism for user evaluation. Although our use environment had a striking difference in relation to the fidelity of the apparatus, with one being extremely high-fidelity (the skulls) and the other being very low-fidelity (the birth canal), we were able to create an environment that was considered "*quite high fidelity in the sense that is the same struggle to try and identify what position the baby's in*" (P08).

Through the use of PRETAR and feedback from the pilot study, we were able to gain an understanding of which aspects of the use environment required high-fidelity and which did not. Additionally, our multidisciplinary research group allowed us to thoroughly discuss the feasibility of each design decision during the study's planning phase. This ensured that any changes to one artefact did not adversely affect the quality of other researchers' primary objectives, which all have competing interests (van Berkel et al. 2020). When researchers incorporate their known constraints within the development artefacts to create a realistic use environment, their medical devices can be assessed in a more ecologically valid environment that will help to uncover design considerations earlier in the product development lifecycle, leading to more user-friendly medical devices and better patient outcomes. Additionally, our findings support that of Sonderegger and Sauer (2010) in that low-fidelity objects can still support participants in their decision-making and allow them to envision how something can work in practice. Conversely, high-fidelity artefacts used to create a sense of realism may be regarded as the '*gold standard*', when assessing medical devices. However, during the early-stages of development, this may introduce unnecessary costs and time in their development. PRETAR has been a useful tool for assessing study requirements before developing any artefacts, and it can effectively determine the appropriate level of abstraction for each artefact. However, it became evident that not all elements of the PRETAR framework are necessary; rather, the first four phases-purpose, resources and constraints, ethics, and techniques for data gathering-play a valuable role in assessing design requirements.

The application of mixed-fidelity artefacts provides medical device developers with the ability to quickly test and evaluate medical devices during early-stage development without incurring high costs and unnecessary time to create realistic environments.

7.1 Environmental artefact fidelity: lessons and reflections

In this study, we employed the PRETAR framework to inform design decisions regarding the appropriate level of abstraction for each environmental artefact. In the following sections, we critically analyse each design choice, evaluating its benefits and limitations in the context of the study's objectives.

7.1.1 The suture glove

The suture glove in its current state was low-fidelity. Although the vision is to develop a wireless device, we felt it was appropriate to evaluate the device at this point to determine whether it was worth investing in further development. As discussed by Marcilly et al. (2024), a low-fidelity evaluation of a device does not decrease the ability to identify use errors and they encourage developers to evaluate their medical devices with low-fidelity prototypes to '*anticipate problems and guide the development and implementation*'. Furthermore, the capability of the device was at a level where the sensor was attached to a surgical glove and changes between surfaces could be detected.

Therefore, we concluded that this was the minimum viable design that could be used in a mock scenario.

Despite our attempts to ensure that the device was tested in an accurate representation of its ideal use, the glove still posed limitations that affected the study. The two biggest constraints were that we were double gloving to ensure a sterile environment. Doing so caused the wire from the sensor to become detached on multiple occasions and caused frustration to the participant as it took a few minutes to re-glove. Secondly, the wire between the sensor and the connection to the electrometer was purposely short to ensure that a good reading could be captured. This '*tethered*' the participant to the device and required that we made changes to the way they interacted with the device and how the sensor was connected to the electrometer, as it was causing discomfort. However, this also resulted in wires being pulled off the sensor if participants got excited or forgot they were tethered and yanking their hand out of the birth canal.

Furthermore, as discussed in Sect. 3.3, using this glove in its current state was only ethically viable if it was conducted in a mock environment, to remove unknown risks to any patients.

7.1.2 Foetal skulls and caput

The foetal skulls used in this study were high-fidelity. This was necessary to determine whether the gloves could work accurately and repeatably on a foetal skull, our first research objective. Fidelity was improved based on feedback from the pilot study participant. Because of the high-fidelity foetal skulls, we could determine that the sensors in their current state were not sensitive enough to distinguish the subtle changes of such a complex model compared to that of the pilot study skulls. Therefore, these skulls allowed us to determine this research objective.

A few participants still found the skulls to not be satisfactory: most notably that the skulls did not have ears. This level of fidelity had not occurred to us as being necessary for this study; further, producing ears would have been difficult and costly. However, overall the skulls were well received and allowed us to reach our objective.

The caput caused the largest disparity of perceptions regarding fidelity. Developing a realistic caput was a challenge, not only to design something that was a snug fit to the foetal skull but also with the same feeling of loose skin. This is partly due to the constraints with the materials available to produce the caput, but also the time and cost associated with producing something more life-like. Additionally, during the studies, multiple users interacted with the researchers' fingers holding the caput in place as no thought had been put into how to hold the caput in place: all the effort was put into creating them. In hindsight, a lower-fidelity caput, similar to that used in the pilot study with gel between plastic wrap, could have been sufficient with the addition of a fixing to hold the caput in place.

Furthermore, some participants stated that just including lubricant between the glove and the skull could have helped improve the realism of the study. However, this was not considered initially and we did not change the method once the study had begun to ensure all participants had a similar experience.

7.1.3 The phantom birth canal

The development of the phantom birth canal was in direct response to the different constraints between the purpose, resources and constraints, ethics, and techniques used for collecting the data of this study. As previously mentioned, it was necessary to observe the participant's hands whilst simultaneously reproducing the feeling of performing a VE, resulting in a low-fidelity model. Given more time and budget, we could have developed a high-fidelity model; however, it is unclear whether this would have significantly impacted the results obtained. Although the birth canal can be viewed as a crude representation of a birth canal, participants reported that it was able to recreate the same constraints as those they have during a real VE. Furthermore, due to the low-fidelity nature of the phantom birth canal, we consider there to be no risk to the participant.

7.1.4 User groups

Originally, we believed that this system would be useful in practice for both midwives and obstetricians, so we recruited participants from each of these groups as we had access to both. We also included student midwives and medical students with an interest in women's health to identify potential issues with vaginal examinations in general, as it is common for professionals to create workarounds to overcome issues that then become part of their routine (Lalley and Malloch 2010; Blandford et al. 2014). As neither group of students had ever conducted a VE before this study, it allowed us to identify other aspects of VE that could be improved. However, nothing was identified during the study.

The study highlighted that the suture glove would primarily be used in cases where the attendee was uncertain of the position of the fetus rather than in general examinations. During second stage labour, which is when the glove is designed to be used, these examinations are conducted by obstetricians. Thus, the user groups used for this study were not ideal. The time spent conducting studies with the incorrect user groups could have been better spent recruiting and observing more obstetricians. However, recent studies have explored the use of surrogates to perform user studies, that is students, engineers, etc., in place of intended users and found they can still provide valuable insight during evaluations (Yoo et al. 2024).

7.1.5 Study location

Although this study was able to reproduce the interaction of the clinician with the foetal skull, there were environmental factors that were ignored. Referring to the medical device usability regulatory standard IEC:62366, use environments should also take into consideration ‘chaotic versus calm, stress level and length of shift’ (International Electrotechnical Commission 2024), all of which are major contributors to birth management. This study, like others (Schnittker et al. 2016), was conducted in isolated rooms, whereas some participants expressed that in practice “*you’ve already got CTGs Going on what ’bop bop bop bop bomb’ and like other things, bleeping syringe drivers making other weird noises and alarm that have many alarms going off all the time.*” P01, and “*If she has an epidural she might just be making a couple of sounds but no epidural she might be screaming like there’s no tomorrow*” P16. As our study was an early assessment, we decided that this level of realism was unnecessary.

We considered using a room within the birthing suite, to which we had access, to conduct more realistic studies and have better access to our intended user groups. However, acquiring the necessary ethical clearance to conduct research studies within the hospital would have taken a great deal of time. Also, unlike surgical theatres which are typically pre-booked, a birthing suite may be required at a moment’s notice by the clinical team. Furthermore, conducting user studies in the hospital does not guarantee that participants will be able to join as they may also get called away at a moment’s notice.

7.1.6 Room setup

The constraint of the room setup was to achieve the purposes of two of our research objectives: to determine whether the suture glove is accurate and reproducible and to understand different participants’ vaginal examination methods. We decided to locate the monitor such that participants could not view it because the techniques they used to perform VEs might have changed with feedback from the monitor. Also, as some of the participants had never conducted VEs previously, it could have created habits that may not work in practice.

7.2 Application of PRETAR to determine level of fidelity

There are several standards and books written for medical device development that discuss how to conduct human factors research, most notably (Cassano-Piché et al. 2015; Food and Drug Administration (FDA) 2016; MHRA 2021; International Electrotechnical Commission 2024; Association for the Advancement of Medical Instrumentation 2009). Each of these describes a range of standard methods (i.e., contextual inquiry, task analysis, observations, usability studies, etc.) for evaluating medical devices. While not revolutionary, each of these methods can be used to assess specific traits of a medical device. However, none of these resources discuss what defines a satisfactory environment when conducting studies or the level of fidelity of environmental artefacts. What they do discuss, which is specific for medical devices, is that medical devices are applied to human beings, and therefore, researchers need to think of surrogate environments and surrogates for patients because it is not safe or ethical to deploy test prototypes in natural environments. van Berkel et al. (2020) begins to unpick this by highlighting that the study design should inform the realism of the study, but they do not provide a practical method to achieve this. Nagy and Haidegger (2022) define the level of environmental complexity (LoEC) in surgical practice, specifically for the evaluation of robotic-assisted minimally invasive surgery. However, these definitions only describe the characteristics of each level of environmental complexity and do not provide a means to determine the level of environmental complexity required to perform evaluations. For example, LoEC 1 - Training phantoms: described as a highly abstract representation of the surgical environment, and at the other end, LoEC 5 - Dynamic, realistic surgical environment: described as having realistic surgical phantoms conducted in in vivo environment with all relevant physiological motions. The contribution of our research is to provide a means by which researchers can determine the level of abstraction of environmental artefacts for early-stage medical device evaluation based on their study requirements.

We used PRETAR in this study to support reasoning about fidelity, and on reflection, we identified that the first four stages of PRETAR are necessary to help determine the level of fidelity, which Fig. 7 illustrates. Firstly, the *Purpose* of the study shapes what is important for fidelity. If what you are studying is the clinical accuracy of some aspect of a medical device, then you need to ensure that it can perform its task accurately. However, if you are evaluating the usability or experience of using an early-stage medical device then it is important that you are measuring the outputs of the medical device accurately and how people feel about it, but the level of fidelity of the environment maybe less important. Secondly, *Resources & Constraints*

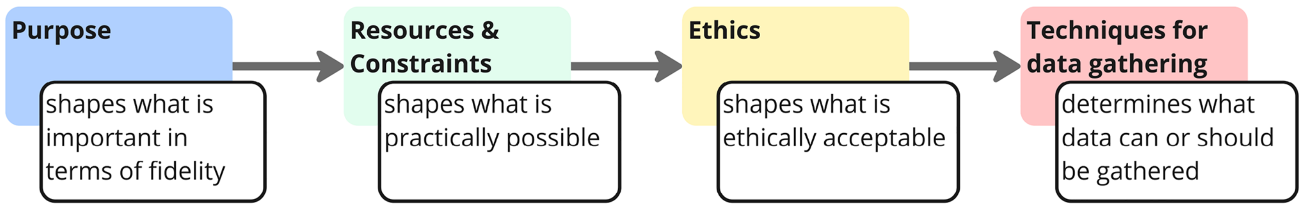

Fig. 7 Use of PRETAR in relation to determining the fidelity level of environmental artefacts of early-stage medical devices

shape what is practically possible and where you need to invest your efforts. The version of the suture glove evaluated in this study was not wireless, whereas the intention is that the final version of the suture glove will be. Although this affected how the participants were able to perform certain tasks, assessing the suture gloves as an early prototype allowed us to gain insight into how the participants reacted to it and what elements of the suture glove required more attention. Thirdly, *Ethics* shape what is ethically acceptable. For early prototypes, most usability studies would be considered unethical to evaluate with patients. Therefore, developing a surrogate patient that is still able to provide the necessary level of realism is important. Lastly, *Techniques for data gathering* determines what data can and/or should be collected. The limiting factor in this study was the need to observe the movements and techniques of the participants' hands, which required a birth canal that allowed us to view the hands of the participants. By uisng PRETAR as a guide and considering all of these aspects, researchers can determine the fidelity of each environmental artefact, which may result in low-fidelity, high-fidelity, or, as in our case, mixed-fidelity artefacts.

8 Conclusion

This paper explores the development of mixed-fidelity artefacts to create realistic use environments. Our findings indicate that mixed-fidelity environmental artefacts can provide adequate ecologically valid environments, particularly for early-stage user evaluations of medical devices. Additionally, they help reduce the cost and time required for artefact development, enabling researchers to allocate more resources toward refining and assessing their medical devices.

The first four phases of the PRETAR framework-purpose, resources and constraints, ethics, and techniques used for collecting data-play a valuable role in guiding these decisions. While high-fidelity environments are often considered ideal for medical device evaluation, mixed-fidelity artefacts can offer comparable insights at a fraction of the cost and complexity in formative studies.

Acknowledgements This research was funded in whole, or in part, by the Wellcome Trust [203145/Z/16/Z]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Author Contributions J.O. data curation, formal analysis, investigation, methodology, resources, validation, visualisation, and writing - original draft. S.J. conceptualization, data curation, formal analysis, investigation, methodology, resources, participant recruitment, validation, and writing - review and editing. C.S.F. conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, and writing - review and editing. L.M. data curation, formal analysis, investigation, participant recruitment, and writing - review and editing A.D. conceptualization, project administration, supervision D.S. conceptualization, project administration, supervision A.L.D. conceptualization, project administration, supervision M.K.T. conceptualization, project administration, supervision A.B. methodology, supervision, and writing - review and editing. All authors read and approved the final manuscript.

Funding This research was funded in whole, or in part, by the Wellcome Trust [203145/Z/16/Z]; EPSRC [Grant number NS/A000050/1].

Data Availibility Statement As participants referred to identifying information in the studies, and did not explicitly consent to their pseudonymised data being made public for the research community, we are not able to provide open access to our data, but we will be able to provide pseudonymised data and notes for verification purposes if requested.

Declarations

Conflict of interest/Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

Anim-Somuah M, Smyth RM, Cyna AM, Cuthbert A (2018). Epidural versus non-epidural or no analgesia for pain management in labour. Cochrane Database Syst Rev, (5), 153. https://doi.org/10.1002/14651858.CD000331.pub4

Association for the Advancement of Medical Instrumentation (2009) ANSI/AAMI HE75-2009: human factors Engineering-design of medical devices. Association for the Advancement of Medical Instrumentation, Arlington, VA

Barbieri L, Angilica A, Bruno F, Muzzupappa M (2013) Mixed prototyping with configurable physical archetype for usability evaluation of product interfaces. Comput Indust 64(3):310–323. https://doi.org/10.1016/j.compind.2012.11.010

Blackler A (2008) Applications of high and low fidelity prototypes in researching intuitive interaction. Proc DRS2008, Design Res Soc Biennial Conf (pp. 1–17)

Blandford A, Adams A, Attfield S, Buchanan G, Gow J, Makri S, Warwick C (2008) The PRET A Rapporter framework: Evaluating digital libraries from the perspective of information work. Inf Process Manag 44(1):4–21. https://doi.org/10.1016/j.ipm.2007.01.021

Blandford A, Furniss D, Vincent C (2014) Patient safety and interactive medical devices: Realigning work as imagined and work as done. Clin Risk 20(5):107–110. https://doi.org/10.1177/1356262214556550

Borycki EM, Kushniruk AW (2010) Towards an integrative cognitive-socio-technical approach in health informatics: analyzing technology-induced error involving health information systems to improve patient safety. Open Med Inf J 4:181–187. https://doi.org/10.2174/1874431101004010181

Brooke J (1996) SUS: A 'Quick and Dirty' Usability Scale, Usability Evaluation In Industry. CRC Press, USA, p 6

Brooks J, Lopes P (2023). Smell & Paste: Low-Fidelity Prototyping for Olfactory Experiences. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery. (event-place: Hamburg, Germany)

Candidori S, Dozio N, Osouli K, Graziosi S, Zanini AA, Costantino ML, De Gaetano F (2024) Improving maternal safety: Usability and performance assessment of a new medical device for the treatment of postpartum haemorrhage. Appl Ergon 117:104223. https://doi.org/10.1016/j.apergo.2023.104223

Cassano-Piché A, Trbovich P, Griffin M, Lin YL, Easty T (2015) Human factors for health technology safety: evaluating and improving the use of health technology in the real world

Chandra D, Savoldelli G, Joo H, Weiss I, Naik V (2008) Fiberoptic Oral Intubation: The Effect of Model Fidelity on Training for Transfer to Patient Care. Anesthesiology 109(6):1007–1013. https://doi.org/10.1097/ALN.0b013e31818d6c3c

Chen J, Yoon D (2024) Exploring the Diminishing Allure of Paper and Low-Fidelity Prototyping Among Designers in the Software Industry: Impacts of Hybrid Work, Digital Tools, and Corporate Culture. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery. (event-place: Honolulu, HI, USA)

Deininger M, Daly SR, Lee JC, Seifert CM, Sienko KH (2019) Prototyping for context: exploring stakeholder feedback based on prototype type, stakeholder group and question type. Res Eng Design 30(4):453–471. https://doi.org/10.1007/s00163-019-00317-5

Derboven J, De Roeck D, Verstraete M, Geerts D, Schneider-Barnes J, Luyten K (2010). Comparing user interaction with low and high fidelity prototypes of tabletop surfaces. In: *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (pp. 148–157). New York, NY, USA: Association for Computing Machinery. (event-place: Reykjavik, Iceland)

DeWerd LA, Kissick M (2014) The phantoms of medical and health physics. Springer

Dieter G, Schmidt L (2021) Engineering Design, 6th edn. McGraw Hill Education, New york

Dimitrakakis E, Aylmore H, Lindenroth L, Dwyer G, Carmichael J, Khan DZ, Stoyanov D (2022) Robotic Handle Prototypes for Endoscopic Endonasal Skull Base Surgery: Pre-clinical Randomised Controlled Trial of Performance and Ergonomics. Annal Biomed Eng 50(5):549–563. https://doi.org/10.1007/s10439-022-02942-z

Dumas JS, Fox JE (2012) Usability testing. The human-computer interaction handbook: fundamentals, evolving technologies, and emerging applications, p. 1221–1241. CRC Press Boca Raton, FL, Publisher

Edwards HM, McDonald S, Zhao T, Humphries L (2013) Reflecting on the PRET A Rapporter Framework Via a Field Study of Adolescents' Perceptions of Technology and Exercise. Human Technology: An Interdisciplinary Journal on Humans in ICT Environments. University of Jyväskylä, Agora Center, Publisher

Ericsson A (2003) Valid and non-reactive verbalization of thoughts during performance of tasks towards a solution to the central problems of introspection as a source of scientific data. J Conscious Stud 10(9–10):1–18

Food and Drug Administration (FDA) (2016) Applying human factors and usability engineering to medical devices: guidance for industry and Food and Drug Administration staff. Food and Drug Administration, Rockville, MD, USA

Ghi T, Eggebø T, Lees C, Kalache K, Rozenberg P, Youssef A, Tutschek B (2018) ISUOG Practice Guidelines: intrapartum ultrasound. Ultrasound Obstetr & Gynecol 52(1):128–139. https://doi.org/10.1002/uog.19072

Goldman RE, Bajo A, MacLachlan LS, Pickens R, Herrell SD, Simaan N (2013) Design and Performance Evaluation of a Minimally Invasive Telerobotic Platform for Transurethral Surveillance and Intervention. IEEE Trans Biomed Eng 60(4):918–925. https://doi.org/10.1109/TBME.2012.2226031

Gosselin M-C, Neufeld E, Moser H, Huber E, Farcito S, Gerber L, Kuster N (2014) Development of a new generation of high-resolution anatomical models for medical device evaluation: the Virtual Population 3.0. Phys Med Biol 59(18):5287. https://doi.org/10.1088/0031-9155/59/18/5287

Hess M (2015) Online survey about current use of 3D imaging and its user requirements in cultural heritage institutions. 2015 Digital Heritage (Vol. 2, pp. 333–338)

Hotton E, O'Brien S, Draycott TJ (2019) Skills training for operative vaginal birth. Best Pract & Res Clin Obstetr & Gynaecol 56:11–22. https://doi.org/10.1016/j.bpobgyn.2018.10.001

International Electrotechnical Commission (2024) Medical devices-Part 1: application of usability engineering to medical devices (ISO Standard No.: 62366-1: 2015

Jaufuraully S, Salvadores Fernandez C, Bagchi B, Gupta P, Desjardins A, Siassakos D, Tiwari MK (2023) Preventing stillbirth from obstructed labor: A sensorized, low-cost device to train in safer operative birth. Front Global Women's Health. https://doi.org/10.3389/fgwh.2022.1039477

Jensen LS, Nissen L, Bilde N, Özkil AG (2018) Prototyping in Mechatronic Product Development: How Prototype Fidelity Levels Affect User Design Input. DS 92: Proceedings of the DESIGN 2018 15th International Design Conference (pp. 1173–1184)

Jensen MB, Elverum CW, Steinert M (2017) Eliciting unknown unknowns with prototypes: Introducing prototrials and prototrial-driven cultures. Design Stud 49:1–31. https://doi.org/10.1016/j.destud.2016.12.002

Jerome B, Kazman R (2005) Surveying the Solitudes: An Investigation into the Relationships between Human Computer Interaction and Software Engineering in Practice. In: Seffah A, Gulliksen J, Desmarais MC (eds) Human-Centered Software Engineering - Integrating Usability in the Software Development Lifecycle. Springer, Netherlands, Dordrecht, pp 59–70

Joyner JS, Kong A, Angelo J, He W, Vaughn-Cooke M (2022) Development of Low-Fidelity Virtual Replicas of Products for Usability Testing. Appl Sci. https://doi.org/10.3390/app12146937

Keenan KE, Jordanova KV, Ogier SE, Tamada D, Bruhwiler N, Starekova J, Hernando D (2024) Phantoms for Quantitative Body MRI: a review and discussion of the phantom value. Magn Reson Mater Phys Biol Med. https://doi.org/10.1007/s10334-024-01181-8

Kelley T (2001) Prototyping is the shorthand of innovation. Design Manag J 12(3):35–42. https://doi.org/10.1111/j.1948-7169.2001.tb00551.x

Kotecha B, Wong PY, Zhang H, Hassaan A (2021) A novel intraoral neuromuscular stimulation device for treating sleep-disordered breathing. Sleep & Breath Schlaf & Atmung 25(4):2083–2090. https://doi.org/10.1007/s11325-021-02355-7

Lalley C, Malloch K (2010) Workarounds: The Hidden Pathway to Excellence. Nurse Leader 8(4):29–32. https://doi.org/10.1016/j.mnl.2010.05.009

Li AC, Kannry JL, Kushniruk A, Chrimes D, McGinn TG, Edonyabo D, Mann DM (2012) Integrating usability testing and think-aloud protocol analysis with "near-live'' clinical simulations in evaluating clinical decision support. Int J Med Inf 81(11):761–772. https://doi.org/10.1016/j.ijmedinf.2012.02.009

Lim Y-K, Pangam A, Periyasami S, Aneja S (2006) Comparative analysis of high- and low-fidelity prototypes for more valid usability evaluations of mobile devices. *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles* (pp. 291–300). New York, NY, USA: Association for Computing Machinery. (event-place: Oslo, Norway)

Lipsett BJ, Reddy V, Steanson K (2023). *Anatomy, Head and Neck: Fontanelles*. StatPearls Publishing, Treasure Island (FL). Retrieved from http://europepmc.org/books/NBK542197

Makri S, Blandford A, Cox AL (2011) This is what I'm doing and why: Methodological reflections on a naturalistic think-aloud study of interactive information behaviour. Inf Process & Manag 47(3):336–348. https://doi.org/10.1016/j.ipm.2010.08.001

Marcilly R, Monkman H, Pelayo S, Lesselroth BJ (2024) Usability Evaluation Ecological Validity: Is More Always Better? Healthcare. https://doi.org/10.3390/healthcare12141417

Massoth C, Röder H, Ohlenburg H, Hessler M, Zarbock A, Pöpping DM, Wenk M (2019) High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students. BMC Med Educ 19(1):29. https://doi.org/10.1186/s12909-019-1464-7

McCurdy M, Connors C, Pyrzak G, Kanefsky B, Vera A (2006) Breaking the fidelity barrier: an examination of our current characterization of prototypes and an example of a mixed-fidelity success. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1233–1242). New York, NY, USA: Association for Computing Machinery. (event-place: Montréal, Québec, Canada)

Meck A-M, Draxler C, Vogt T (2022) A Question of Fidelity: Comparing Different User Testing Methods for Evaluating In-Car Prompts. *Proceedings of the 4th Conference on Conversational User Interfaces.* New York, NY, USA: Association for Computing Machinery. (event-place: Glasgow, United Kingdom)

Melvin MT, Siddiqui NT, Wild E, Parotto M, Perelman VS, You-Ten KE (2023) Achieving Competency in Fiber-Optic Intubation Among Resident Physicians After Higher- Versus Lower-Fidelity Task Training: A Randomized Controlled Study. Anesth Analg 137(1):200–208. https://doi.org/10.1213/ANE.0000000000006283

MHRA (2021) Guidance on applying human factors and usability engineering to medical devices including drug-device combination products in Great Britain

Nagy TD, Haidegger T (2022) Performance and Capability Assessment in Surgical Subtask Automation. Sensors 22(7):2501. https://doi.org/10.3390/s22072501 (Retrieved 2025-03-10, from https://www.mdpi.com/1424-8220/22/7/2501 (Number: 7 Publisher: Multidisciplinary Digital Publishing Institute))

NENC Preterm Birth Group (2021) Management of preterm birth (Tech. Rep.). NHS North East and North Cumbria. Retrieved from https://nornet.org.uk/files/pdf/nenc-preterm-birth-guideline-v2-15-feb-2022.pdf

NHS (2022). NHS Maternity Statistics, England - 2021–22 (Tech. Rep.). NHS maternity Statistics

Opie J, McCann E, Qureshi H, Brookshaw T, Erro E, Selden C, Blandford A (2024) An academician's approach to the application of human factors standards: a case study on a liver support system. Hum Factors Healthcare. https://doi.org/10.1016/j.hfh.2024.100070

Ramalhinho J, Yoo S, Dowrick T, Koo B, Somasundaram M, Gurusamy K, Clarkson MJ (2023) The value of Augmented Reality in surgery - A usability study on laparoscopic liver surgery. Med Image Anal 90:102943. https://doi.org/10.1016/j.media.2023.102943

Ramphul M, Kennelly M, Burke G, Murphy D (2015) Risk factors and morbidity associated with suboptimal instrument placement at instrumental delivery: observational study nested within the Instrumental Delivery & Ultrasound randomised controlled trial ISRCTN. BJOG: Int J Obstetr & Gynaecol 122(4):558–563. https://doi.org/10.1111/1471-0528.13186

Richter LL, Ting J, Muraca GM, Synnes A, Lim KI, Lisonkova S (2019) Temporal trends in neonatal mortality and morbidity following spontaneous and clinician-initiated preterm birth in Washington State, USA: a population-based study. BMJ Open 9(1):e023004. https://doi.org/10.1136/bmjopen-2018-023004

Rudd J, Stern K, Isensee S (1996) Low vs. high-fidelity prototyping debate. Interactions 3(1):76–85. https://doi.org/10.1145/223500.223514

Rueda D, Hoto R, Conejero A (2013). Study of the Influence of Prototype Aesthetic Fidelity (A Realism Factor) in Usability Tests. Holzinger A, Ziefle M, Hitz M, Debevc M (Eds.), Human Factors in Computing and Informatics (pp. 122–136). Berlin, Heidelberg: Springer Berlin Heidelberg

Schembri P, Anthony R, Pelc M (2017) Detection of Electroencephalography Artefacts using Low Fidelity Equipment

Schnittker R, Schmettow M, Verhoeven F, Schraagen J (2016) Combining situated Cognitive Engineering with a novel testing method in a case study comparing two infusion pump interfaces. Appl Ergon 55:16–26. https://doi.org/10.1016/j.apergo.2016.01.004

Sim G, Horton M, McKnight L (2016). iPad vs Paper Prototypes: Does Form Factor Affect Children's Ratings of a Game Concept? In: *Proceedings of the The 15th International Conference on Interaction Design and Children* (pp. 190–195). New York, NY, USA: Association for Computing Machinery. (event-place: Manchester, United Kingdom)

Sonderegger A, Sauer J (2010) The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. Appl Ergon 41(3):403–410. https://doi.org/10.1016/j.apergo.2009.09.002

Speicher M, Lewis K, Nebeling M (2021). Designers, the Stage Is Yours! Medium-Fidelity Prototyping of Augmented & Virtual Reality Interfaces with 360theater. Proc. ACM Hum.-Comput. Interact., *5*(EICS). https://doi.org/10.1145/3461727

van Berkel N, Clarkson MJ, Xiao G, Dursun E, Allam M, Davidson BR, Blandford A (2020) Dimensions of ecological validity for usability evaluations in clinical settings. J Biomed Inf 110:103553. https://doi.org/10.1016/j.jbi.2020.103553

Van Berkel N, Opie J, Ahmad OF, Lovat L, Stoyanov D, Blandford A (2022) Initial Responses to False Positives in AI-Supported Continuous Interactions: A Colonoscopy Case Study. ACM Trans Interact Intell Syst 12(1):2:1-2:18. https://doi.org/10.1145/3480247

Virzi RA (1989) What can you Learn from a Low-Fidelity Prototype? Proc Hum Fact Soc Annual Meet 33(4):224–228. https://doi.org/10.1177/154193128903300405

Viswanathan V, Linsey J (2011) Design Fixation in Physical Modeling: An Investigation on the Role of Sunk Cost. In: IDETC-CIE2011 (pp. 119–130). Volume 9: 23rd International Conference on Design Theory and Methodology; 16th Design for Manufacturing and the Life Cycle Conference

Yew J, Convertino G, Hamilton A, Churchill E (2020) Design Systems: A Community Case Study. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–8). New York, NY, USA: Association for Computing Machinery. (event-place: Honolulu, HI, USA)

Yoo S, Ramalhinho J, Dowrick T, Somasundaram M, Gurusamy K, Davidson B, Blandford A (2024) Can engineers represent surgeons in usability studies? Comparison of results from evaluating augmented reality guidance for laparoscopic surgery. Comput & Graph 119:103881. https://doi.org/10.1016/j.cag.2024.01.008

Zhou X, Rau P-LP (2019) Determining fidelity of mixed prototypes: Effect of media and physical interaction. Appl Ergon 80:111–118. https://doi.org/10.1016/j.apergo.2019.05.007