Usability Evaluation Ecological Validity: Is More Always Better?

Romaric Marcilly, Helen Monkman, Sylvia Pelayo, Blake J Lesselroth

healthcare

MDPI

Article

Usability Evaluation Ecological Validity: Is More Always Better?

Romaric Marcilly [1,2,*], Helen Monkman [3,4], Sylvia Pelayo [1,2] and Blake J. Lesselroth [3,4]

[1] Univ. Lille, CHU Lille, ULR 2694—METRICS: Évaluation des Technologies de Santé et des Pratiques Médicales, F-59000 Lille, France; sylvia.pelayo@univ-lille.fr
[2] Inserm, CIC-IT 1403, F-59000 Lille, France
[3] School of Health Information Science, University of Victoria, Victoria, BC V8P 5C2, Canada; monkman@uvic.ca (H.M.); blake-lesselroth@ouhsc.edu (B.J.L.)
[4] School of Community Medicine, University of Oklahoma, Tulsa, OK 74135, USA
* Correspondence: romaric.marcilly@univ-lille.fr

Abstract: Background: The ecological validity associated with usability testing of health information technologies (HITs) can affect test results and the predictability of real-world performance. It is, therefore, necessary to identify conditions with the greatest effect on validity. Method: We conducted a comparative analysis of two usability testing conditions. We tested a HIT designed for anesthesiologists to detect pain signals and compared two fidelity levels of ecological validity. We measured the difference in the number and type of use errors identified between high and low-fidelity experimental conditions. Results: We identified the same error types in both test conditions, although the number of errors varied as a function of the condition. The difference in total error counts was relatively modest and not consistent across levels of severity. Conclusions: Increasing ecological validity does not invariably increase the ability to detect use errors. Our findings suggest that low-fidelity tests are an efficient way to identify and mitigate usability issues affecting ease of use, effectiveness, and safety. We believe early low-fidelity testing is an efficient but underused way to maximize the value of usability testing.

Keywords: usability; evaluation; user testing; ecological validity; nociception index

Citation: Marcilly, R.; Monkman, H.; Pelayo, S.; Lesselroth, B.J. Usability Evaluation Ecological Validity: Is More Always Better? *Healthcare* **2024**, *12*, 1417. https://doi.org/10.3390/healthcare12141417

Academic Editor: Victor R. Prybutok

Received: 29 May 2024
Revised: 1 July 2024
Accepted: 13 July 2024
Published: 16 July 2024

1. Introduction

Ecological validity (i.e., "test fidelity") is "*the extent to which behaviour in a test situation can be generalised to a natural setting*" [1]. Many testing conditions, such as the realism of the physical environment or the responsiveness of software functionality, can affect test validity and the ability to predict real-world performance. Ecological validity is high when each test condition closely mimics reality [2], whereas differences between the live and test environments may limit the accuracy of observations [3–6]. Borycki and Kushniruk's Cognitive Socio-Technical Framework [7] says ecological validity should increase as tests move from the bench to the bedside. However, this is not always possible because of organizational resource constraints, including time, money, and the availability of skilled testers. It is, therefore, important to identify which conditions have the greatest effect on validity and estimate the cost-effectiveness of replicating them with the highest possible fidelity.

Studies have investigated the effect of ecological validity on usability testing outcomes. Most of them evaluated only one ecological validity dimension at a time, usually comparing two levels of that dimension [8–10]. Only a few studies have looked at multiple dimensions concurrently [3,4]. Our goal was to study how manipulating several dimensions of ecological validity simultaneously can affect testing results, including detecting technology use errors [8]. We focused on summative testing of an acute care and intraoperative pain monitor. We compared two levels of ecological fidelity and asked the following questions: (1) were all use errors detected in both test settings? (2) did the level of fidelity influence

the number of errors made by participants? (3) were any novel errors (i.e., not identified in the risk analysis) detected?

2. Models of Ecological Validity

Many research teams have published models of ecological validity [3,11]. For example, Sauer's Four-Factor Framework of Contextual Validity categorizes conditions according to activity system components (i.e., user, technology, task, and environment). For our study, we used van Berkel's Seven Dimensions Framework (2020). This model lists seven dimensions of ecological validity that are important to consider when designing usability studies: (1) user roles, (2) the evaluation environment, (3) the presence of user training, (4) the clinical scenario, (5) whether patients are involved during testing, (6) attributes of the hardware, and (7) the software. We chose this model because it captured the largest number of variables in the most granular detail across the technology development lifecycle. In the next section, we describe each dimension in greater detail [11].

2.1. User Roles

It is important to recruit representative end-users for testing whenever possible rather than developers or professional testers. People involved in the design process—including clinician developers—cannot substitute for actual users [11,12]. The characteristics of the participant sample, such as professional role (e.g., physician, nurse, specialist), technical skill (e.g., computer literacy, years of experience), level of clinical training, usage habits, and preferences or values about the task or technology should match the target population to ensure the sample is representative of intended end-users [3,4,6,11]. There is some disagreement among experts on whether it is important to recruit only novices [13] or novices and experts [3,14,15]. We believe the decision to recruit multiple levels of user expertise depends on the objectives of the test (e.g., counting the number of usability issues, estimating learnability, or measuring error tolerance) [3]. Participants' conditions during testing (e.g., fatigue, beginning vs. end of shift, mood) may also affect test results and should match the end-user's context during actual use. To limit inequity or implicit bias, it may also be necessary to account for age, gender, and ethnicity.

While including representative end-users in testing is always preferable, it is often challenging to recruit healthcare professionals with the necessary subject matter expertise, knowledge of the target setting, and who have protected time to participate in testing. Usability professionals often must adapt and improvise to meet sponsor deadlines. If recruiting all representative users is cost or time-prohibitive, we suggest developing user personas. Personas are fictional but evidence-based representations of user groups that can guide recruitment strategies or testers [16]. We believe personas are most helpful for identifying technology requirements and guiding early design decisions but should not replace testing with representative users.

2.2. Environment

The environment dimension includes the physical (i.e., "*built*") and the social environment. The physical test environment refers to attributes of the facility or equipment that influence user behavior. This might include the configuration of the testing room or the presence of background noise. Test environment fidelity could range from an administrative office equipped with a desktop computer (i.e., low fidelity) to a simulation lab reproduction of a hospital room (i.e., high fidelity) to an actual unoccupied hospital room (i.e., naturalistic) [6]. The social environment refers to the presence of other humans or workflow interference (e.g., competing clinical requests, text messages, and pages) during the test [1]. Some tasks require multiple healthcare providers to work collaboratively. Teamwork and interprofessional interactions can be reproduced using scripts and actions performed by members of the research team (i.e., low fidelity), actors (i.e., high fidelity), or real-life colleagues (i.e., naturalistic).

While replicating real-world conditions is desirable [2–4,6,11,17–21], especially during safety investigations [6,22], it is often not feasible given cost constraints. For example, testing a new surgical device a surgeon uses might call for a detailed simulation of the operating theatre, including medical equipment, an interactive mannequin, and actors to portray the interdisciplinary team. Unfortunately, many manufacturers and organizations cannot afford this level of ecological validity. In these circumstances, usability professionals must compromise between the realism of the test conditions and the built environment [23]. While it is important to replicate the most important attributes of the real environment, this is an unresolved area of active research [8,24–31].

2.3. Training

Van Berkel et al. [11] pointed out that researchers often train participants on a system before testing in a simulation. To avoid confounding, test proctors must offer the same training and materials actual users would receive. Including product training as part of the testing protocol can identify education gaps and ways to improve new user orientation [11].

2.4. Test Scenario

The test scenario provides context for the test participants; it describes the clinical use case, care setting, and task goals [32]. The fidelity of the scenario influences how seriously participants behave in a study (i.e., behavioral fidelity) [11]. In goal-based scenarios, usability professionals observe participants as they determine and execute the steps necessary to achieve the goal [32]. These are higher fidelity than simple tests of product features or user acceptance tests wherein proctors provide participants with step-by-step instructions to complete tasks. The breadth and depth of scenarios should, therefore, be representative of real-life activities [3]. Breadth is the extent to which activity system complexity is captured in the test (e.g., single task vs. parallel tasks) [3]. Depth is the level of detail and completeness with which a task is simulated (i.e., the proportion of real-life steps included in the test) [3,6]. The instructions researchers give testers for reporting findings can influence the results [18–20]. For example, with static prototypes (e.g., drawings, wireframes), participants may be asked to describe what they would do (e.g., click, type, scroll) to operate a product, whereas with interactive prototypes, participants may verbalize their thoughts while using the product.

Kushniruk and colleagues also suggested explicitly defining the urgency and typicality of tasks when designing the scenario [6]. Urgency indicates the level of immediacy and pressure associated with completing a task. Tasks can range from non-urgent (e.g., submitting a routine electronic order for acetaminophen) to urgent (e.g., submitting an order for a stat antihypertensive during a medical emergency). Typicality refers to how a task represents the usual, normal, or expected system use or workflow. Both extremes (i.e., typical and atypical) can be important during testing. For example, minor—yet common—usability issues may profoundly affect efficiency and user satisfaction. Rare usability issues may be more difficult to detect and cause catastrophic outcomes [12].

2.5. Patient Involvement

Van Berkel et al. (2020) [11] cautioned that including real patients in usability testing can generate valuable insights but at considerable risk. Patients may identify usability issues that are impossible to detect with actors. However, there are potential patient safety risks. There are both physical and psychological risks to consider. For example, a patient participating in an interview may re-experience painful events or memories. Usability testing, therefore, often includes a proxy for patients. For example, a study might instead use a mannequin or actor (i.e., a standardized patient) [8].

2.6. Software

Usability professionals sometimes conduct tests using early HIT prototypes (e.g., paper prototypes or wireframes). Generally, the prototyping method and degree of realism

can influence participants' reactions. The dimensions to consider include feature breadth (i.e., the proportion of finished features present), feature depth (i.e., level of feature detail) [3], physical similarity, interaction similarity, visual appearance [33], and data similarity [11].

While it is important to test as early as possible in the product design lifecycle—even with paper prototypes—there is a complex interaction between prototype fidelity and outcome measures [3,5,9,34,35]. Some studies suggest that prototype fidelity does not affect the number and type of usability problems detected [3,9]. However, we believe that when measuring participant behavior (e.g., clinically relevant performance), efficiency (e.g., task completion time), and effectiveness (e.g., task success rate), prototype fidelity is relevant [5,34,35]. Furthermore, it may be necessary to pre-populate the system with patient data. These data might be fabricated or real, anonymized patient data. If fabricated, it is important to include extreme values and test at the edges of input ranges (i.e., "boundary value analysis") to identify rare occurrences.

2.7. Hardware

Technology hardware can create usability issues or affect the goal success rate. A study by Andre et al. looked at the design and performance of four automatic external defibrillators [36]. The team found that participants could not use two machines to deliver a shock. The hardware design and packaging significantly influenced the ability of untrained caregivers to use the equipment properly. While hardware should be accounted for when designing tests or evaluating usability data, high-fidelity hardware prototypes are often expensive and time-consuming to produce [6].

3. Materials and Methods

3.1. Health Information Technology

We studied a novel pain monitor that uses calculations from an electrocardiogram (ECG) tracing to estimate the autonomic nervous system response to painful or stressful stimuli. The HIT measures the R-R interval between two QRS complexes [37]. An algorithm then calculates an analgesia nociception index (ANI): a unitless index ranging between 0 and 100, with higher values indicating more parasympathetic activity associated with analgesia and lower values indicating more sympathetic activity associated with pain (i.e., 0 = great pain or stress; 100 = adequate anesthesia). The goal is to keep the patient's ANI between 50–70.

The graphical user interface of the HIT (Figure 1) displays the instantaneous value of the ANI (ANIi) and its average over time (mean ANI, ANIm). The display includes numerical values, graphs, and information about the ECG and signal quality. The ANI monitor must be reset between patients to avoid errors (i.e., new patient data displayed with the previous patient's threshold calibration). The pain monitor can be used in intensive care units, operative theatres, and post-surgical care units. It is typically located at the head of the bed, close to other vital sign monitors.

3.2. Risk Analysis to Inform Scenario Development

We used a mixture of published articles and grey literature to conduct an a priori risk analysis. We studied safety incident reports databases and complaints files, published usability studies on earlier versions of the pain monitor, and conducted interviews with end-users using similar devices to identify potential usability errors and their consequences. We identified eight usability errors (Table 1) associated with physician and nursing tasks. We classified errors into three severity levels: mild: no injury, no patient discomfort (n = 1 error); moderate: light patient discomfort (n = 4 errors); severe: serious injury or death (n = 3 errors). We developed scenarios to test for each of the eight use errors.

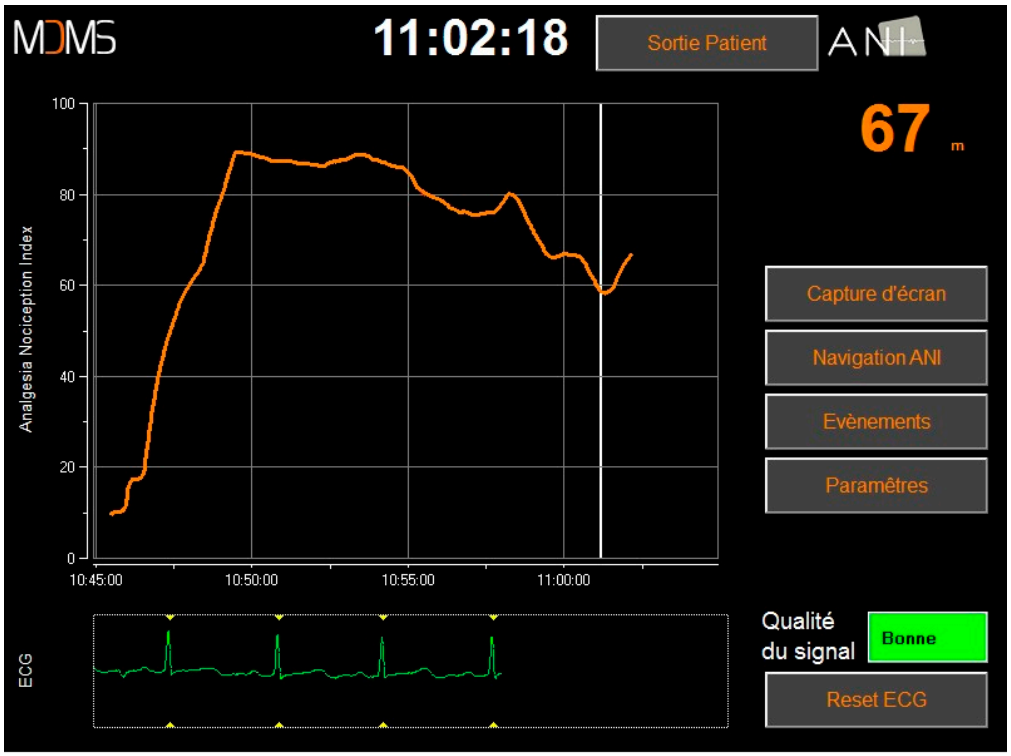

Figure 1. Screenshot of the analgesia nociception index (ANI) monitor.

Table 1. List of use error types and severity. Mild = no injury, no patient discomfort; moderate = light patient discomfort; severe = serious injury or death.

Type No.	Error Type	Description of the Error	Level of Severity
1	ANI-ECG * confusion	Participant confuses ECG with ANI	Mild
2	No detection of overdosage	The participant does not recognize when an ANIm value over 80 for an unconscious patient represents a medication overdose	Moderate
3	Focus on ANI	The participant only uses the ANI index and neglects other data sources to evaluate the patient's discomfort level	Moderate
4	Considering poor-quality data	The participant does not consider the quality of signal acquisition and bases her/his decision on poor-quality data	Moderate
5	Considering out-of-date data	The participant does not reset the ECG signal and bases her/his decision on out-of-date or erroneous data	Moderate
6	High ANI misunderstanding	The participant erroneously interprets the meaning of a high ANI on the screen	Severe
7	Low ANI misunderstanding	The participant erroneously interprets the meaning of a low ANI on the screen	Severe
8	Considering other patient data	The participant does not reset the values from the previous patient and bases her/his decisions on erroneous data	Severe

* ANI = analgesia nociception index; ECG = electrocardiogram.

3.3. Participants

In France, physicians and nurses with specialized training in anesthesiology typically manage pain in dedicated anesthesiology and resuscitation units. We conducted a preliminary context-of-use analysis in these two units and found no differences in how the devices were used. Therefore, to gather comprehensive data on the types of errors encountered with the ANI monitor, we recruited physicians and nurses for this study.

We included participants in this study if they were: (1) physicians (i.e., resuscitation clinicians or anesthesiologists) or nurses specialized in intensive care, (2) had at least two months of professional experience in an intensive care unit, (3) completed training with the ANI monitor, and (4) consented to be recorded. Participants were excluded if they had previously used this ANI pain monitor. Recruitment proceeded through convenience sampling. We recruited volunteers through announcements (i.e., newsletters and emails) in Lille Academic Hospital's units and through their professional networks.

3.4. Study Design and Test Conditions

We performed an in-lab experimental study using a one-factor within-subjects design (Figure 2). The within-participants variable was the level of fidelity; this included two conditions: low fidelity and high fidelity. While ecological validity can be conceptualized on a continuum, we designed two discrete levels for our study: low-fidelity and high-fidelity. We manipulated multiple ecological dimensions for each level (i.e., environment, scenario, patient involvement, software) (Table 2). All participants completed the same five scenarios twice—once for each test condition (i.e., low- and high-fidelity). We furnished each subject with two different but equivalent clinical cases to limit any carryover or priming effect. We counterbalanced the exposure order for each condition and clinical case. All test scenarios were developed by an anesthesiologist and modeled after real patient cases. A second anesthesiologist reviewed each case for face and content validity. During testing, a proctor observed participants either through one-way glass (high-fidelity) or while standing in the same room (low-fidelity). The proctor gave instructions through a loudspeaker (high-fidelity) or in person (low-fidelity). We audio- and video-recorded all usability tests.

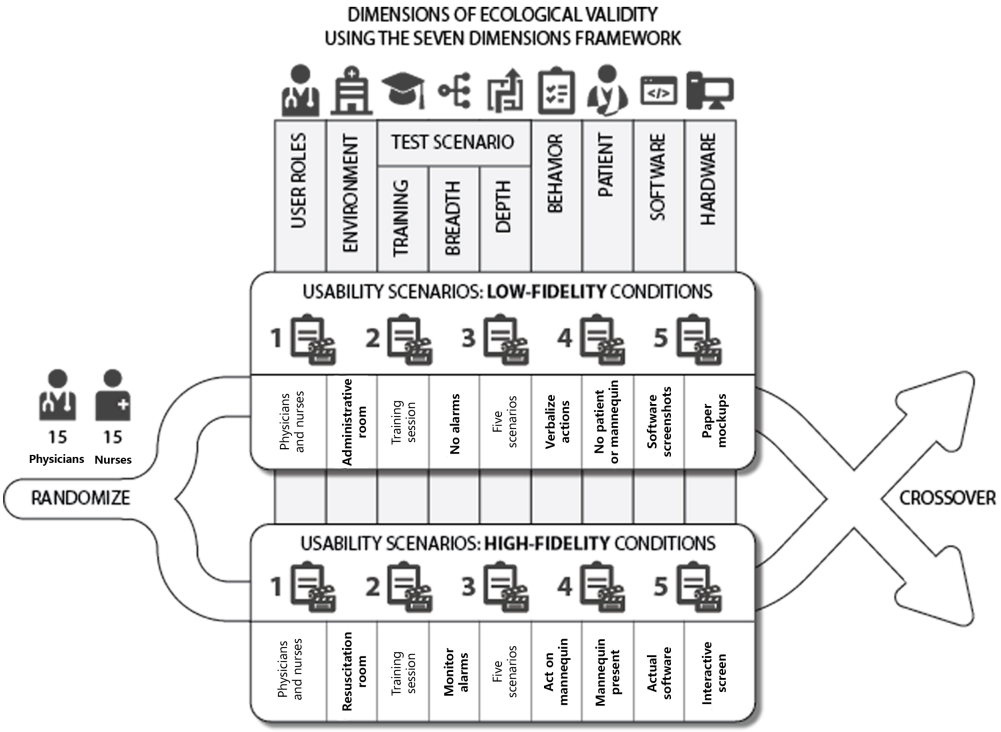

Figure 2. Diagram of the within-subjects study design. Thirty physicians and nurses completed five low-fidelity scenarios and five high-fidelity scenarios. We changed the fidelity in six of the seven ecological validity dimensions.

Table 2. Description of the test environment according to the level of fidelity.

Ecological Validity Dimension	The Low-Fidelity Condition	The High-Fidelity Condition
1. User roles	Physicians and nurses specialized in intensive care. Participants had a minimum of two months of experience in intensive care and received training on the pain monitor two days before the test.	
2. Environment	Administrative room without other types of medical equipment and devices.	Simulated resuscitation rooms were similar to actual resuscitation rooms. The room was equipped with furniture and real medical technology and devices (infusion pumps, ECG * monitor, etc.). The simulation space mimicked a real resuscitation room in terms of temperature, ambient sounds, interruptive alarms, and disinfectant smell.
3. User training	All participants attended a training session at least two days before the test. This matches current training protocols with new equipment.	
4a. Scenarios, breadth	We did not set off monitor alarms to interrupt the participants.	Interrupting alarms from the monitors interrupted the participants, just like in real-life resuscitation rooms.
4b. Scenarios, depth	Five goal-based scenarios to test all eight identified use errors. Each scenario was performed twice, once in each condition. We furnished a summary of the patient's case, including a description of the patient (e.g., age, gender, conditions), the clinical course, and a list of medications taken.	
4c. Scenarios, behavior	Participants were asked to verbalize how they would respond and what actions they would take.	Participants were asked to act on the mannequin as they would in real life.
5. Patient involvement	We did not include a patient or representation (i.e., a mannequin). Instead, the test moderator described the patient's status. We provided screenshots of the patient parameters required for medical decision-making.	We used a mannequin capable of reproducing physiologically realistic reactions of the human body.
6. Hardware	Participants are shown screenshots printed on paper and a video on a computer screen with no possibility of interaction with the computer.	The ANI * pain monitor is an interactive screen framed by a plastic shell. Users can interact by directly pressing the interactive buttons on the screen. Depending on which buttons are pressed, parameters are modified, or windows are opened on the interface.
7. Software	The test was primarily performed using screenshots of the pain monitor. Scenario 4 (testing error 8) required the participant to see blinking ANI curves; thus, a video of the blinking screen was shown instead of a screenshot. Participants could see screenshots of data typically rendered on ECG, respiratory, and ANI pain monitors.	The test was performed with an actual ANI pain monitor. Participants could see the patient's data typically rendered on ECG, respiratory, and ANI pain monitors in a live environment.

* ANI = analgesia nociception index; ECG = electrocardiogram.

3.5. Measurements

We recorded use errors during each test. We first interviewed two anesthesiologists and created a reference standard of acceptable behaviors and answers for each scenario. In the high-fidelity condition, we asked participants to behave in each scenario as they would in real life. In the low-fidelity condition, we instructed them to tell the observer what they would do for each task. During testing, a usability professional compared test subject behaviors and answers to the reference standard. For double-pass verification, a second usability professional reviewed all recordings and scored behaviors or answers.

3.6. Procedures

After determining the eligibility and consent of each participant, we explained the testing procedure. Then, participants completed all five scenarios in one condition (i.e., either high- or low-fidelity) and, after a short break, completed the next five scenarios in the other condition. We determined the order of scenarios in each condition using a randomization table. After the tests, we held debriefing sessions using a semi-structured interview guide to explore participants' perspectives on the technology (e.g., perceived usefulness and usability) and the root cause(s) of their errors. The total test duration was approximately 70 min.

3.7. Statistical Analyses

For each test and each participant, we counted the total number of errors and scored each by type and severity (i.e., mild, moderate, severe). We calculated descriptive statistics for errors at each level of fidelity. We calculated the overall frequency of error occurrence and the frequency by ecological level. Due to the small sample size and the rarity of the errors, we could not calculate inferential statistics. Statistical analyses were performed with Jamovi software (version 2.3.21, The Jamovi project).

3.8. Ethical Considerations

This study was conducted in France and is categorized as human and social science research. In accordance with French biomedical research law, our study protocol was exempt from ethical board approval or oversight [38,39]. This study was conducted in accordance with the Declaration of Helsinki.

We recruited all participants voluntarily and provided financial compensation of 150€ (approximately 160 U.S.$) for participation.

4. Results

Both groups were similar in age and sex (Table 3).

Table 3. Demographic characteristics of the participants.

Profile	Number (Females; Males)	Mean Age in Years (*SD*)
Anesthesiologists	15 (9; 6)	28.26 (2.54)
Nurses	15 (9; 6)	31.93 (6.14)
Total	30 (18; 12)	30.01 (5.05)

Across both conditions, we identified thirty-one errors (Figure 3); there were seven moderate and twenty-four severe errors. We saw users commit five of the eight possible error types (Figure 3 and Table 4) listed in our risk analysis (Table 1). Four error types appeared in both conditions (#2, #6, #7, and #8), and one only appeared in the low-fidelity condition (#4). We did not identify any unexpected errors (i.e., errors that defied categorization according to our risk analysis) in either condition.

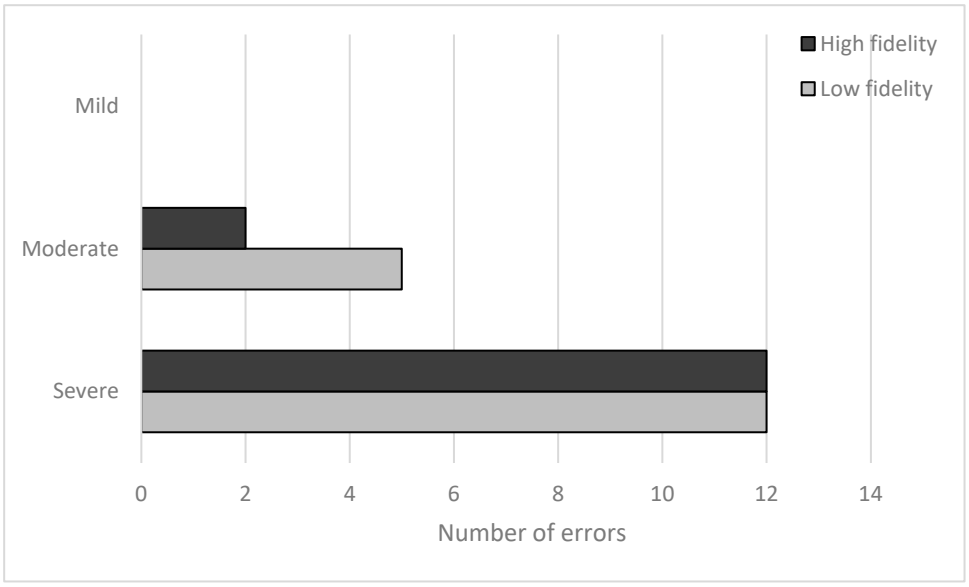

Figure 3. Number of errors committed in each condition according to the severity of the errors.

Table 4. List of use errors, severity, and the number (percentage) of participants who committed them.

Type No	Error Type	Description of the Error	Severity Level	Low-Fidelity	High-Fidelity
1	ANI *-ECG * confusion	Participant confuses ECG with ANI	Mild	0 (0%)	0 (0%)
2	No detection of overdosage	The participant does not recognize when an ANIm value over 80 for an unconscious patient represents a medication overdose	Moderate	4 (13%)	2 (7%)
3	Focus on ANI	The participant only uses the ANI index and neglects other data sources to evaluate the patient's discomfort level	Moderate	0 (0%)	0 (0%)
4	Considering poor-quality data	The participant does not consider the quality of signal acquisition and bases her/his decision on poor-quality data	Moderate	1 (3%)	0 (0%)
5	Considering out-of-date data	The participant does not reset the ECG signal and bases her/his decision on out-of-date or erroneous data	Moderate	0 (0%)	0 (0%)
6	High ANI misunderstanding	The participant erroneously interprets the meaning of a high ANI on the screen	Severe	3 (10%)	1 (3%)
7	Low ANI misunderstanding	The participant erroneously interprets the meaning of a low ANI on the screen	Severe	6 (20%)	7 (23%)
8	Considering other patient data	The participant does not reset the values from the previous patient and bases her/his decisions on erroneous data	Severe	3 (10%)	4 (13%)

* ANI = analgesia nociception index; ECG = electrocardiogram.

Across both conditions, participants made, on average, 1.03 errors (range = 0–7) out of 16 possible errors (8 errors × 2 conditions). They made 0.57 (range = 0–4) out of 8 possible errors in the low-fidelity tests and 0.47 (range = 0–3) in the high-fidelity tests. Sixteen participants (53%) did not commit any errors in either condition. Three participants did not commit any errors in the low-fidelity condition (10%), and one participant did not commit any errors in the high-fidelity condition (3%).

Overall, we observed 17 errors in low-fidelity conditions compared to 14 errors in the high-fidelity conditions. There were no differences between the number of mild or severe errors. By contrast, we identified five moderate errors in the low-fidelity condition and two in the high (Figure 4 and Table 5).

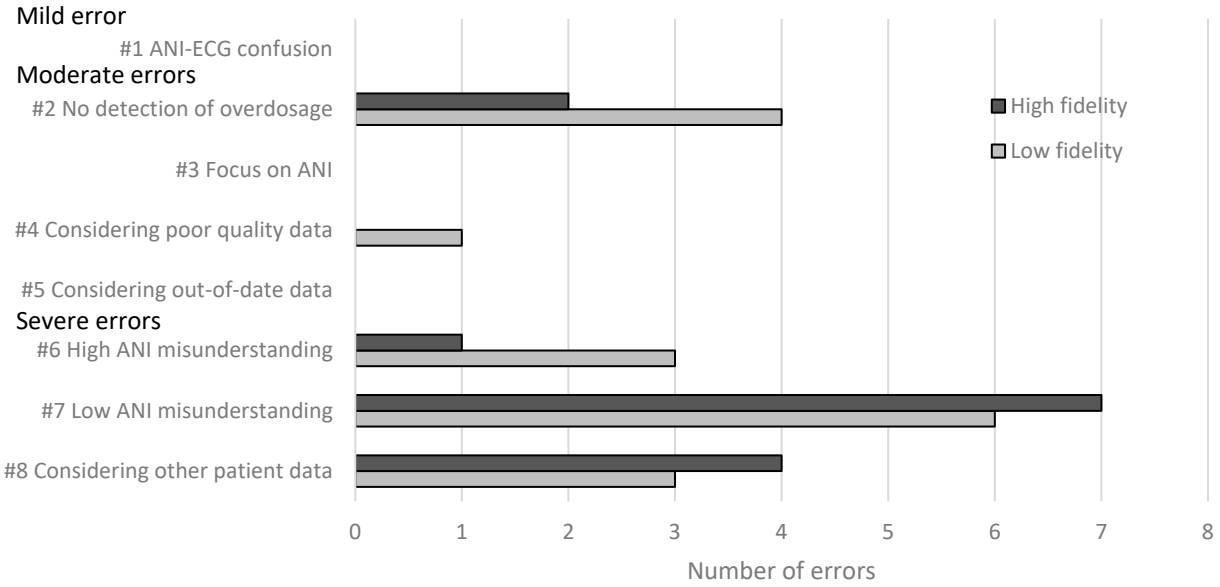

Figure 4. Number of errors made by the participants according to the type of error and the test condition.

Table 5. Number of errors committed in each condition according to their severity.

Error Severity	Low-Fidelity	High-Fidelity
Mild (1 possible error type)	0	0
Moderate (4 possible error types)	5	2
Severe (3 possible error types)	12	12
Total (8 possible error types)	17	14

5. Discussion

5.1. Principal Findings

While we saw some variation in the number and type of errors as a function of test fidelity and ecological validity, the difference was relatively modest (i.e., three additional errors in low-fidelity conditions) and inconsistent across severity levels. We observed more moderate errors in the low-fidelity condition but the same number of severe errors in both conditions. Overall, we observed the same error types in both conditions; increasing ecological validity did not improve our ability to detect specific types of usability issues. Nonetheless, given the small number of participants, scenarios, and observations, we hesitate to generalize to other technology evaluations or make broad testing recommendations. Instead, we believe there is a need for more comparative studies to identify subtle or specific effects of ecological validity on performance.

We propose several hypotheses to explain why participants made similar errors in both testing conditions. First, technology display screenshots (low-fidelity) looked like the live monitor (high-fidelity). Since the display did not include interactive features or affordances, the user experience may have been the same. Second, downstream participant behaviors were primarily cognitive (i.e., interpreting a display). There may have been insufficient task depth or breadth to see more cascading errors in workflow or ripple effects in complex adaptive systems. We might have seen more severe errors if we required participants to act on the data by adjusting medications or communicating with other clinicians. Finally, low-fidelity conditions may create a kind of "interference effect." The artificial conditions of low-fidelity mock-ups and the absence of contextual cues in a laboratory setting may create interface usability issues—for example, the inability to recognize system status—or limit a participant's situational awareness and response time.

5.2. Comparison to the Literature and Implications of Findings

While we did not see more errors or usability issues in the high-fidelity test conditions, we believe the interaction between test fidelity and error detection is complex. Increasing the fidelity of each dimension may not invariably increase test sensitivity or the ability to identify potential errors [40]. However, we cannot conclude that low-fidelity tests are always better or more sensitive than high-fidelity tests. On the contrary, it is still possible that high-fidelity simulations may identify significant and potentially severe usability issues that are complex, context-dependent, and otherwise hidden during tests with low ecological validity. Thus, low-fidelity usability tests should not be used in lieu of high-fidelity evaluations. Instead, they should be used earlier and potentially more frequently throughout the development and testing lifecycle. When selecting test fidelity, we believe the deciding factors to consider include the technology features of interest, the anticipated interaction between technology and the environment characteristics [40], the most important user goals, and the context of use. Nevertheless, this study's findings suggest that low-fidelity tests are an efficient and cost-effective way to identify and mitigate many issues impacting ease of use, effectiveness, and safety. These findings also align with recommendations from leading usability experts to test early and often [41–43]. High-fidelity tests may offer deep insights into future implementation challenges, but we should also embrace "discount" testing to operate and innovate at the pace of healthcare.

There are four unresolved issues that demand further study. First, if ecological validity can influence the sensitivity of tests to detect usability errors, we must know where to apply our efforts. In resource-constrained settings, how do usability professionals predict what level of fidelity is needed to identify all important errors? Perhaps only some test dimensions must be high-fidelity to meet testing goals. Second, when there are so many dimensions of ecological validity—and within each dimension, so many levels of fidelity—professionals need a framework to know (1) which dimensions are associated with specific error types and (2) what fidelity level is sufficient to test a product's performance thresholds. A starting point for developing this framework could be Kushniruk and Turner's User–Task–Context matrix, which lists three dimensions relevant to HIT design and example attributes for each dimension [44]. Attempts to build a similar model for each of van Berkel's seven dimensions would be more challenging. An ideal product would list the relevant attributes for each dimension and suggest strategies for creating low and high-fidelity versions. Third, we must know if there are certain dimensions that should always be high-fidelity. We presume testing participants should always possess knowledge of the clinical subject matter and use context. However, there may be other dimensions critical to safety or other key performance measures. Fourth, when designing tests with fidelity in mind, how low can you go? Jensen and colleagues proposed that extremely low-fidelity tests without a priori scenarios, functional prototypes, or patient data still provide valuable information [45]. These tests can foster discussion when identifying technology requirements, edge cases, or implicit user knowledge about the use context.

To address these issues, usability professionals should work towards a consensus on the minimum number of dimensions and attributes to consider when designing tests. One strategy that usability professionals can use when designing and reporting usability studies is to leverage—and expand—existing frameworks and guidelines [46–48]. Standardized reporting would enable researchers to strategically close gaps in our understanding of ecological validity and the effect specific dimensions have on the accuracy of findings.

5.3. Strengths and Limitations

There are several strengths of this work that deserve mention. First, we believe this is one of the first studies to use van Berkel's theoretical framework as a scaffolding to build a usability testing protocol. In doing so, we are building the empirical database to explain how decisions of ecological validity influence usability testing findings, technology design decisions, and implementation outcomes. We also provide an extensible model to guide future testing in this arena. Second, this is one of the first studies to compare the effect of test

fidelity in multiple dimensions (i.e., environment, scenario breadth, user behavior, patient involvement, and software). We compared two modalities of ecological validity in the lab, whereas most published reports compare the laboratory to the real world [24–26,29,30]. Third, we incorporated a HIT risk analysis into our protocol to develop a pre-identified list of errors. This increased the precision of our "testing forecast" and the instrumentation to search for these issues. At the same time, we could still identify and classify new error types.

There are also important study limitations affecting the explanatory power and generalizability of our results. First, we did not conduct comparative tests across all seven dimensions of van Berkel's model. We could have included levels for user roles, user training, scenario urgency, and scenario typicality. Second, we configured fidelity at only two levels when, in fact, every dimension of fidelity exists on a continuum. For example, for patient involvement, our high-fidelity condition included a mannequin. This could have been a "mid-fidelity" condition, and we could have also included a standardized patient or an actual patient for the "high-fidelity" condition. As we noted above, the static screenshots of the interface look very similar to the actual device. The similarity in exposures may have caused a Type II (i.e., false negative) error. Third, we did not conduct naturalistic testing. Healthcare systems are complex adaptive systems with emergent properties, changing actors, and widely distributed workflow and cognition. This makes it extremely difficult to know what other usability issues went undetected. Fourth, we only tested one HIT. It would be informative to know how the technology, target user, and context-of-use interact with fidelity and testing outcomes. Fifth, we recruited 30 users—and only 15 per role. We do not know if this was the correct number to surface all usability issues. Deciding on the correct power for summative usability testing is a hotly debated topic in the literature [49]. While it has been argued that as few as 5 participants can identify over 80% of usability issues, research has shown that many more participants may be needed depending on the heterogeneity of users, the complexity of the product, and the goals of testing (e.g., formative testing for iterative re-design or summative testing for user acceptance) [50–52]. However, our sample size is in line with recommendations for summative evaluations of medical devices [53]. Sixth, there was a risk of contamination between dimensions or the order of exposure. For example, the onscreen data may have improved situational awareness, promoted new behaviors like cross-checking signals, and thereby decreased "high ANI misunderstanding" errors (#6).

6. Conclusions

More ecological validity does not always seem to be better when evaluating the usability of HIT. Our results suggest that high ecological validity does not consistently provide more information about the quality and defects of HITs. Low-fidelity testing can be an effective and cost-effective way of identifying and mitigating many problems associated with ease of use, efficiency, and safety. However, we do not recommend replacing all high-fidelity testing with low-fidelity testing. Instead, low-fidelity testing can be used early and more often so that usability researchers can better anticipate problems and guide development and implementation teams at the pace of healthcare.

Author Contributions: Conceptualization, R.M. and S.P.; methodology, R.M. and S.P.; formal analysis, R.M., H.M. and B.J.L.; investigation, R.M. and S.P.; data curation, R.M.; writing—original draft preparation, R.M., H.M. and B.J.L.; writing—review and editing, R.M., H.M., S.P. and B.J.L.; project administration, S.P.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the French Agence Nationale de la Recherche (grant number: ANR-15-CE36–0007).

Institutional Review Board Statement: This study was conducted in France and is categorized as human and social science research. In accordance with French biomedical research law, our study protocol was exempt from ethical board approval or oversight [38,39]. This study was conducted in accordance with the Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all participants involved in this study.

Data Availability Statement: Data is available on demand.

Acknowledgments: The authors would like to thank all the participants, as well as Pierre-François Gautier for his help in carrying out this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sonderegger, A.; Sauer, J. The Influence of Laboratory Set-up in Usability Tests: Effects on User Performance, Subjective Ratings and Physiological Measures. *Ergonomics* **2009**, *52*, 1350–1361. [CrossRef] [PubMed]
2. Morten, H. User Testing in Industry: A Case Study of Laboratory, Workshop, and Field Tests. In *Proceedings of the User Interfaces for All: Proceedings of the 5th ERCIM Workshop*; GMD: Sankt Augustin, Germany, 1999; pp. 59–72.
3. Sauer, J.; Seibel, K.; Rüttinger, B. The Influence of User Expertise and Prototype Fidelity in Usability Tests. *Appl. Ergon.* **2010**, *41*, 130–140. [CrossRef] [PubMed]
4. Sauer, J.; Sonderegger, A. Methodological Issues in Product Evaluation: The Influence of Testing Environment and Task Scenario. *Appl. Ergon.* **2011**, *42*, 487–494. [CrossRef] [PubMed]
5. Sauer, J.; Sonderegger, A. The Influence of Prototype Fidelity and Aesthetics of Design in Usability Tests: Effects on User Behaviour, Subjective Evaluation and Emotion. *Appl. Ergon.* **2009**, *40*, 670–677. [CrossRef] [PubMed]
6. Kushniruk, A.; Nohr, C.; Jensen, S.; Borycki, E.M. From Usability Testing to Clinical Simulations: Bringing Context into the Design and Evaluation of Usable and Safe Health Information Technologies. Contribution of the IMIA Human Factors Engineering for Healthcare Informatics Working Group. *Yearb. Med. Inform.* **2013**, *8*, 78–85. [PubMed]
7. Borycki, E.M.; Kushniruk, A.W. Towards an Integrative Cognitive-Socio-Technical Approach in Health Informatics: Analyzing Technology-Induced Error Involving Health Information Systems to Improve Patient Safety. *Open Med. Inform. J.* **2010**, *4*, 181–187. [CrossRef] [PubMed]
8. Marcilly, R.; Schiro, J.; Genin, M.; Somers, S.; Migaud, M.-C.; Mabile, F.; Pelayo, S.; Del Zotto, M.; Rochat, J. Detectability of Use Errors in Summative Usability Tests of Medical Devices: Impact of the Test Environment. *Appl. Ergon.* **2024**, *118*, 104266. [CrossRef] [PubMed]
9. Boothe, C.; Strawderman, L.; Hosea, E. The Effects of Prototype Medium on Usability Testing. *Appl. Ergon.* **2013**, *44*, 1033–1038. [CrossRef]
10. Uebelbacher, A. The Fidelity of Prototype and Testing Environment in Usability Tests. Doctoral Thesis, University of Fribourg, Fribourg, Switzerland, 2014.
11. van Berkel, N.; Clarkson, M.J.; Xiao, G.; Dursun, E.; Allam, M.; Davidson, B.R.; Blandford, A. Dimensions of Ecological Validity for Usability Evaluations in Clinical Settings. *J. Biomed. Inform.* **2020**, *110*, 103553. [CrossRef]
12. Thomas, J.C.; Kellogg, W.A. Minimizing Ecological Gaps in Interface Design. *IEEE Softw.* **1989**, *6*, 78–86. [CrossRef]
13. Nielsen, J. *Usability Engineering*; Academic Press: Boston, MA, USA, 1993; ISBN 978-0-12-518405-2.
14. Kjeldskov, J.; Skov, M.B.; Stage, J. Does Time Heal? A Longitudinal Study of Usability. In Proceedings of the Australian Computer-Human Interaction Conference 2005 (OzCHI'05), Canberra, Australia, 21–25 November 2005.
15. Rubin, J. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*; Wiley Technical Communication Library; Wiley: New York, NY, USA, 1994; ISBN 978-0-471-59403-1.
16. Park, H.; Monkman, H.; Wenger, A.; Lesselroth, B. Portrait of Ms. Diaz: Empirical Study of Patient Journey Mapping Instruction for Medical Professional Students. *Knowl. Manag. E-Learn. Int. J.* **2020**, *12*, 469–487. [CrossRef]
17. Dahl, Y.; Alsos, O.A.; Svanæs, D. Fidelity Considerations for Simulation-Based Usability Assessments of Mobile ICT for Hospitals. *Int. J. Hum.-Comput. Interact.* **2010**, *26*, 445–476. [CrossRef]
18. Schmuckler, M.A. What Is Ecological Validity? A Dimensional Analysis. *Infancy* **2001**, *2*, 419–436. [CrossRef] [PubMed]
19. Kieffer, S.; Sangiorgi, U.B.; Vanderdonckt, J. ECOVAL: A Framework for Increasing the Ecological Validity in Usability Testing. In Proceedings of the 2015 48th Hawaii International Conference on System Sciences, Kauai, HI, USA, 5–8 January 2015; pp. 452–461.
20. Wang, Y.; Mehler, B.; Reimer, B.; Lammers, V.; D'Ambrosio, L.A.; Coughlin, J.F. The Validity of Driving Simulation for Assessing Differences between In-Vehicle Informational Interfaces: A Comparison with Field Testing. *Ergonomics* **2010**, *53*, 404–420. [CrossRef] [PubMed]
21. Dahl, Y.; Alsos, O.A.; Svanæs, D. Evaluating Mobile Usability: The Role of Fidelity in Full-Scale Laboratory Simulations with Mobile ICT for Hospitals. In *Human-Computer Interaction. New Trends*; Jacko, J.A., Ed.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5610, pp. 232–241. ISBN 978-3-642-02573-0.
22. Ben-Menahem, S.M.; Nistor-Gallo, R.; Macia, G.; Von Krogh, G.; Goldhahn, J. How the New European Regulation on Medical Devices Will Affect Innovation. *Nat. Biomed. Eng.* **2020**, *4*, 585–590. [CrossRef]
23. Kjeldskov, J.; Skov, M.B. Studying Usability In Sitro: Simulating Real World Phenomena in Controlled Environments. *Int. J. Hum.-Comput. Interact.* **2007**, *22*, 7–36. [CrossRef]

24. Sun, X.; May, A. A Comparison of Field-Based and Lab-Based Experiments to Evaluate User Experience of Personalised Mobile Devices. *Adv. Hum.-Comput. Interact.* **2013**, *2013*, 619767. [CrossRef]
25. Kaikkonen, A.; Kallio, T.; Kekäläinen, A.; Kankainen, A.; Cankar, M. Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing. *J. Usability Stud.* **2005**, *1*, 4–16.
26. Kjeldskov, J.; Graham, C.; Pedell, S.; Vetere, F.; Howard, S.; Balbo, S.; Davies, J. Evaluating the Usability of a Mobile Guide: The Influence of Location, Participants and Resources. *Behav. Inf. Technol.* **2005**, *24*, 51–65. [CrossRef]
27. Sauer, J.; Sonderegger, A.; Heyden, K.; Biller, J.; Klotz, J.; Uebelbacher, A. Extra-Laboratorial Usability Tests: An Empirical Comparison of Remote and Classical Field Testing with Lab Testing. *Appl. Ergon.* **2019**, *74*, 85–96. [CrossRef]
28. Nielsen, C.M.; Overgaard, M.; Pedersen, M.B.; Stage, J.; Stenild, S. It's Worth the Hassle!: The Added Value of Evaluating the Usability of Mobile Systems in the Field. In Proceedings of the 4th Nordic Conference on Human-Computer Interaction Changing Roles—NordiCHI '06, Oslo, Norway, 14–18 October 2006; pp. 272–280.
29. Duh, H.B.-L.; Tan, G.C.B.; Chen, V.H. Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. In Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services—MobileHCI '06, Helsinki, Finland, 12–15 September 2006; p. 181.
30. Kjeldskov, J.; Skov, M.B.; Als, B.S.; Høegh, R.T. Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Mobile Human-Computer Interaction—MobileHCI 2004*; Brewster, S., Dunlop, M., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3160, pp. 61–73. ISBN 978-3-540-23086-1.
31. Baillie, L.; Schatz, R. Exploring Multimodality in the Laboratory and the Field. In Proceedings of the 7th International Conference on Multimodal Interfaces—ICMI '05, Torento, Italy, 4–6 October 2005; p. 100.
32. Usability.gov, Scenarios, Usability.Gov, Improving the User Experience. 2013. Available online: https://www.usability.gov/how-to-and-tools/methods/scenarios.html (accessed on 16 February 2024).
33. Virzi, R.A.; Sokolov, J.L.; Karis, D. Usability Problem Identification Using Both Low- and High-Fidelity Prototypes. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Common Ground—CHI '96, Vancouver, BC, Canada, 13–18 April 1996; pp. 236–243.
34. Säde, S.; Nieminen, M.; Riihiaho, S. Testing Usability with 3D Paper Prototypes—Case Halton System. *Appl. Ergon.* **1998**, *29*, 67–73. [CrossRef] [PubMed]
35. Sauer, J.; Franke, H.; Ruettinger, B. Designing Interactive Consumer Products: Utility of Paper Prototypes and Effectiveness of Enhanced Control Labelling. *Appl. Ergon.* **2008**, *39*, 71–85. [CrossRef] [PubMed]
36. Andre, A. Automated External Defibrillator Use by Untrained Bystanders: Can the Public-Use Model Work? *Prehospital Emerg. Care* **2004**, *8*, 284–291. [CrossRef] [PubMed]
37. Logier, R.; Jeanne, M.; De Jonckheere, J.; Dassonneville, A.; Delecroix, M.; Tavernier, B. PhysioDoloris: A Monitoring Device for Analgesia/Nociception Balance Evaluation Using Heart Rate Variability Analysis. In Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, Argentina, 31 August–4 September 2010; pp. 1194–1197.
38. Peute, L.W.; Lichtner, V.; Baysari, M.T.; Hägglund, M.; Homco, J.; Jansen-Kosterink, S.; Jauregui, I.; Kaipio, J.; Kuziemsky, C.E.; Lehnbom, E.C.; et al. Challenges and Best Practices in Ethical Review of Human and Organizational Factors Studies in Health Technology: A Synthesis of Testimonies: A Joint Contribution from the International Medical Informatics Association's Human Factors Engineering and the European Federation for Medicatl Informatics' Human and Organizational Factors of Medical Informatics Working Groups. *Yearb. Med. Inform.* **2020**, *29*, 58–70. [CrossRef] [PubMed]
39. Toulouse, E.; Masseguin, C.; Lafont, B.; McGurk, G.; Harbonn, A.; Roberts, J.A.; Granier, S.; Dupeyron, A.; Bazin, J.E. French Legal Approach to Clinical Research. *Anaesth. Crit. Care Pain Med.* **2018**, *37*, 607–614. [CrossRef] [PubMed]
40. Wiklund, M.E.; Kendler, J.; Strochlic, A.Y. *Usability Testing of Medical Devices*; CRC Press: Boca Raton, FL, USA, 2011; ISBN 978-1-4398-1183-2.
41. Nielsen, J. Applying Discount Usability Engineering. *IEEE Softw.* **1995**, *12*, 98–100. [CrossRef]
42. Krug, S. *Rocket Surgery Made Easy: The Do-It-Yourself Guide to Finding and Fixing Usability Problems*; Voices That Matter; New Riders: Berkeley, CA, USA, 2010; ISBN 978-0-321-65729-9.
43. Kushniruk, A.W.; Patel, V.L. Cognitive and Usability Engineering Methods for the Evaluation of Clinical Information Systems. *J. Biomed. Inform.* **2004**, *37*, 56–76. [CrossRef] [PubMed]
44. Kushniruk, A.; Turner, P. A Framework for User Involvement and Context in the Design and Development of Safe E-Health Systems. *Stud. Health Technol. Inform.* **2012**, *180*, 353–357.
45. Jensen, S.; Nøhr, C.; Rasmussen, S.L. Fidelity in Clinical Simulation: How Low Can You Go? *Stud. Health Technol. Inform.* **2013**, *194*, 147–153.
46. Schulz, K.F.; Altman, D.G.; Moher, D.; CONSORT Group. CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials. *BMJ* **2010**, *340*, c332. [CrossRef]
47. Ogrinc, G.; Davies, L.; Goodman, D.; Batalden, P.; Davidoff, F.; Stevens, D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): Revised Publication Guidelines from a Detailed Consensus Process. *BMJ Qual. Saf.* **2016**, *25*, 986–992. [CrossRef] [PubMed]

48. Peute, L.W.; Driest, K.F.; Marcilly, R.; Bras Da Costa, S.; Beuscart-Zephir, M.-C.; Jaspers, M.W.M. A Framework for Reporting on Human Factor/Usability Studies of Health Information Technologies. *Stud. Health Technol. Inform.* **2013**, *194*, 54–60. [PubMed]
49. Bevan, N.; Barnum, C.; Cockton, G.; Nielsen, J.; Spool, J.; Wixon, D. The "Magic Number 5": Is It Enough for Web Testing? In Proceedings of the CHI '03 Extended Abstracts on Human Factors in Computing Systems—CHI '03, Ft. Lauderdale, FL, USA, 5–10 April 2003; p. 698.
50. Barnum, C.M. *Usability Testing Essentials: Ready, Set . . . Test!* 2nd ed.; Morgan Kaufmann: Amsterdam, The Netherlands, 2021; ISBN 978-0-12-816942-1.
51. Lewis, J.R. Usability: Lessons Learned . . . and Yet to Be Learned. *Int. J. Hum.-Comput. Interact.* **2014**, *30*, 663–684. [CrossRef]
52. Caron, A.; Vandewalle, V.; Marcilly, R.; Rochat, J.; Dervaux, B. The Optimal Sample Size for Usability Testing, From the Manufacturer's Perspective: A Value-of-Information Approach. *Value Health* **2022**, *25*, 116–124. [CrossRef]
53. Food and Drug Administration. *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, 2016.