applied sciences

MDPI

Article

Investigating and Measuring Usability in Wearable Systems: A Structured Methodology and Related Protocol

Giuseppe Andreoni [1,2]

[1] Laboratory TeDH–Technology and Design for Healthcare, Department of Design, Politecnico di Milano, 20158 Milan, Italy; giuseppe.andreoni@polimi.it
[2] Bioengineering Laboratory, Scientific Institute IRCCS "E.Medea", 23842 Bosisio Parini, Italy

Featured Application: Assessment of usability of wearable systems in their design phase or in their application with users.

Abstract: Wearable systems are pervading our lives in several applications: from fitness to sport, from health monitoring to rehabilitation, up to prosthetics and empowering human functions through exoskeletons. If the technological requirements are mainly quantitative and easy to measure, their usability, acceptance, and user experience are generally poorly studied. There is a lack of a structured methodological approach to develop a comprehensive protocol. This paper aimed at providing these methodological bases and at defining some of the related tools. The first action was to clearly define the objectives of the study: (a) to identify design inconsistencies and usability problems or errors; (b) to validate the use of wearable systems under controlled test conditions with representative users; and (c) to establish a baseline in terms of user performance and user satisfaction levels. A five-step approach should be adopted: (1) define the target users; (2) conduct a task analysis for identifying the context, the parameters to be measured, and the methodology to collect data; (3) prepare a protocol and the investigation tools; (4) execute the usability experiments; and (5) analyze and report the data. This segmentation of the complex task of usability measurement into single steps can help in elaborating a proper protocol where users, usability factors and parameters, and their recording tools (questionnaires or measurement methods) are correctly identified and prepared for the experimental activity. The application of this methodology can support researchers, developers, and users in improving the deployment of these devices in our lives and the exploitation of these systems for increasing our quality of life.

Keywords: wearable systems; usability assessment; wearability ergonomics

Citation: Andreoni, G. Investigating and Measuring Usability in Wearable Systems: A Structured Methodology and Related Protocol. *Appl. Sci.* **2023**, *13*, 3595. https://doi.org/10.3390/app13063595

Received: 30 December 2022
Revised: 21 February 2023
Accepted: 9 March 2023
Published: 11 March 2023

1. Introduction

Wearable systems, or even more simply wearables, are a wide variety of body-worn objects and accessories that, in the last two decades, have been introduced to carry out several functions: to measure some human physical or physiological parameters, to support communication, such as earbuds, to support risky or demanding activities, or even to replace some human functions. Their main applications are in sport, fitness, and healthcare. Chan et al. [1] found several systems supporting complex healthcare applications and enabling low-cost, wearable, non-invasive alternatives for the continuous 24 h monitoring of health, activity, mobility, and mental status, both indoors and outdoors. Wearable systems consist of various components and devices, ranging from sensors and actuators to multimedia devices.

The main features of wearable systems are:

(a) They are "always on" and accessible by the user;
(b) They should be controllable and interactive;
(c) They augment human capabilities or senses;

(d) They should be unobtrusive, i.e., they have to operate in synergy with the body and not limit the user's functions and mobility;
(e) They can be used as a communications media.

In a macro-generalization, they can be categorized into four families:

1. Systems dedicated to empowering human functions; exoskeletons used in different workplaces are a good example of this family;
2. Systems integrating monitoring functions (for humans and/or the environment); smart garments, smart bracelets, smart watches, and smart shoes belong to this category and are applied in sport, fitness, and medicine;
3. System performing actions on the human body; orthoses and other systems for protection or rehabilitation can be included in these examples;
4. Systems able to replace human functions such as prostheses for the upper or lower limbs, and even insulin pumps replacing the physiological insulin management can be included in this group.

For all these typologies, usability becomes a mandatory requirement more than a simple feature. The ISO norms define "usability" as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [2,3]. In this regulatory framework, it is clear how usability is an outcome of use, but no specific processes or methods for taking account of usability in the design development or evaluation have been described. Firstly, this paper tried to address this topic in specific relation to usability assessment. Some general guidelines to define methods and usability assessment are provided in [4], in particular in clause 5. The methodologies proposed by the technical report are divided into methods either involving users or not. In the iterative design process, user participation is essential, so that the observation of users, performance-related measurements, critical incidents analysis, questionnaires, interviews, and thinking aloud are the recommended methods.

Despite this enormous importance in terms of their health-related applications, only a few studies have faced the usability aspects in a complete setting, mainly by adopting simple questionnaires to investigate them. In the next subsection, a survey of the usability assessment methods in wearable systems is presented.

Related Works

Keogh et al. chose the system usability scale (SUS) questionnaire to investigate the usability and acceptability of wearable sensors in the elderly. The SUS [5–8] measures the usability of a device/system/technology through a ten-item questionnaire with five response options for respondents from 1, strongly disagree, to 5, strongly agree, resulting in a potential minimum score of 0 and a maximum score of 100.

Keogh et al. also applied the intrinsic motivation inventory (IMI) to assess participants' experiences related to the target activity involved with wearing the device. IMI [9] is a multidimensional questionnaire containing twenty-two items on a seven-point Likert scale, ranging from 1, not at all true, to 7, very true. The measure assesses six subscales, namely interest/enjoyment, perceived competence, effort/importance, pressure/tension, value/usefulness, and perceived choice. It is indirectly related to usability but is otherwise interesting. Similarly, Domingos et al. [10] explored the user experience of wearing an activity tracker in a similar cohort of older adults; usability and acceptance were evaluated through the technology acceptance model (TAM) [11], SUS, and the user satisfaction evaluation questionnaire [12].

In a wider perspective, Keogh et al. [13] recently provided a systematic review about usability studies for wearables: they identified thirty-seven studies in which a substantial heterogeneity in the quality of reporting, the methods used to assess usability, the devices used, and the aims of the studies precluded any meaningful comparisons. Questionnaires were used in the majority of the studies (70.3%; n = 26), which was followed by those who used interviews (n = 17; 45.9%), while the methods of analysis were not reported in over a

third of the studies (n = 6; 35.3%). Their conclusion was that usability of wearable devices is a poorly measured and reported variable in the specific field of chronic health conditions.

Regarding wearables and their related applications for physical activity monitoring, Mc Callum et al. [14] found in their systematic review that usability was investigated by sixteen studies (14.4%, 16/111), out of which nine (56%, 9/16) used SUS, four (25%, 4/16) used interviews, two (13%, 2/16) used focus groups, and one (6%, 1/16) used observations of the participants completing timed tasks. Three dimensions were identified in relation to assessing the usability, namely (1) the burden of wearing and using the device, (2) interface complexity, and (3) perceived technical performance.

Sometimes, usability is intended as derived only from the technical reliability to support the adoption of the wearable device in clinical settings [15]. In their study, Martinato et al. demonstrated the accuracy of the measurement of physical activity in the elderly population by means of a smart watch so as to recommend its clinical use but without any usability assessment. The same approach was used by Hawthorne et al. [16], with an acceptance measure being collected through user interviews.

Bendig et al. [17] performed comprehensive usability assessments with 18 patients with Parkinson's disease using a mixed-methods usability battery containing the system usability scale, a rater-based evaluation of device-specific tasks, and qualitative interviews.

Usability is also essential to be considered since the design phase of these devices, because it is a basic requirement to match user acceptance of the product–service system, as well as its proper use, the reliability of the data measured by it, and the motivation and engagement of the overall approach or service that the system implements [18,19]. For example, in health-monitoring services, in some cases, it is crucial to wear a system in a specific body position to carry out the measurement with a dedicated procedure and receive feedback on the data quality to ensure that the entire process is complete and correctly executed. The semi-automatic monitoring of blood pressure with a body-worn cuff at the wrist represents a typical situation for this case.

These studies have reinforced the idea of the need for a standard, multifactorial, and integrated approach in usability assessment.

In light of these findings, in tackling the importance of these factors, the assessment of the usability of a wearable technology needs to be carefully planned and prepared though a structured approach. Users, the target application or function, and its specific requirements, as well as the morphological category, drives the development of a well-designed protocol.

This paper aims at providing a structured and standardized process to define a usability protocol for wearable systems. The aim of this paper is to propose a structured methodology to conduct usability analysis with wearables and to propose a set of reference questionnaires and charts to be used in experimental tests; this could lead to a common grid for usability assessments for further studies in this field.

In Section 2, the author presents a segmentation of usability assessments in five phases and a decision tree supporting the evolution of the process. This is the most relevant section from a methodological point of view. Section 3 introduces a set of questionnaires and methods supporting data collection in relation to the three phases of experimental activity: before the test in Section 3.1, during operation in Section 3.2, and after the test in Section 3.3. The data analysis methods are presented in Section 3.4. The final section discusses the relevance of such methodologies in the frame of the current research about wearables and their different fields of application.

2. Materials and Methods

This section presents the methodological approach and related guidelines for the definition of a structured usability assessment, starting from the identification of the objectives for the analysis and the process to designing a proper protocol (Section 2.1). In relation to the goals, Section 2.2 describes the typologies of usability metrics (quantitative parameters, errors, and subjective assessment). The assessment of the impact of the usability experience is presented in Section 2.3, while Section 2.4 discusses the reporting of the results.

2.1. Identify the Assessment Goals and Design the Corresponding Protocol

The goals of usability testing can target on different aspects. For example, they can include a definition of a baseline of user performance or establish and validate user performance measures and identify potential design concerns to be addressed in order to improve the efficiency, comfort, and end-user satisfaction of a system. Thus, in relation to wearables, usability tests should have three main objectives:

1. To determine the design inconsistencies and usability problem areas within a user's tasks. Regarding this, the potential sources of error may include different factors that can be divided into three main categories:
   (a) Wearability: dressing/undressing errors, such as the failure to locate some elements or components, the need for excessive movement or force to complete a certain function, failure in following the recommended flow of operation, or the wrong positioning of the system with respect to the underlying body area or anatomical point, thus producing discomfort, artifacts, or wrong measurements.
   (b) Efficiency: the functional performance to which the system is devoted. For example, the correct monitoring of some physiological parameters or, conversely, the failure to properly act in synergy with the user's movements and the expectation of orthotic systems.
   (c) Discomfort: comfort problems specifically related to biomechanical aspects, such as identifying zones with body interface issues (friction, limitation to movement, pressure, temperature, etc.) or eventual long-term postural problems produced by the wearable.
2. To validate the use of the wearable system under controlled test conditions with representative users. In this case, data will be used to access whether the usability goals regarding an effective, efficient, and well-received product have been achieved.
3. To establish a baseline in terms of user performance and user satisfaction levels and/or user interface for future product refinement or for a defined product category.

Undoubtedly, the identification of the exact assessment scenario drives the development of a proper investigation protocol and the related tools to evaluate the parameters of interest. At the same time, it is possible to define an integrated protocol aiming at evaluating all the above three aspects. Following this last choice, and in accordance with this framework, here, a methodology for preparing a well-designed protocol for the usability analysis of wearables is proposed considering all the three aspects.

The methodological approach to build a usability test protocol in the field of wearable technologies follows five steps (Figure 1):

1. Defining the target users of the product–service system; in several applications and settings, it is necessary to remember and consider that the wearer is the main user, but other secondary users are present and important, for example, in a clinical/medical setting, familiar caregivers or clinical operators interact with the wearable in some phases. The ISO norm is clear in defining the specified users, goals, and contexts of use as a particular combination of users, goals, and contexts of use for which usability is being considered for a product–service system.
2. Executing a preliminary task analysis for identifying the proper usage context and for understanding the usage scenario and the related parameters to define the different categories of involved actors, the different phases of use, and the usability factors to be investigated. Thanks to this analysis, it is possible to select the most suitable methodology to collect data (interviews, questionnaires, quantitative measurements, etc.) and their interpretation criteria.
3. Preparing and optimizing the tools for the usability analysis; suitable methods for collecting the data should be set up in relation to the parameters to be measured. Video recordings for user observation, questionnaires for subjective evaluations, tools to measure temporal or biomechanical parameters, or devices to carry out

other measures need to be identified and prepared in a proper configuration. At this stage, the inclusion criteria for participant recruitment needs to be identified and agreed among the researchers; this aspect is fundamental for the preparation of the documents for the tests (pre-test demographic and background information questionnaire, informed consent, audio/video recording consent, and experimental protocol and ethical committee acceptance, if applicable). About this last point, it must be remembered that in any usability study with users, participants have to sign an informed consent form that acknowledges that the participation is voluntary, that participation can cease at any time, and that in case the session is video-recorded, their identification privacy will be safeguarded.

Regarding recruitment, it is necessary to consider the dimension of the sample size of the users. Faulkner [20] demonstrated the risks of using only five participants and the benefits of using more. In the study, with a total population of sixty subjects, some of the randomly selected sets of five participants found 99% of the problems, but other sets found only 55%. Any panel of 10 users was able to reveal more than 80% of the usability problems. Any group of 20 users extracted by the total population was able to identify more than 95% of usability problems. Therefore, in usability testing, a recommended minimum sample size of ten users is recommended for each category (direct users, secondary users, and other users).

4. Executing the usability experiments, taking care to collect accurate and reliable data. The possible presence of a facilitator could be considered. A facilitator could have the role of briefing the participants, eventually reminding them some tasks (without "leading" the experimentation, but rather leaving the maximum freedom to the users so as to allow natural behavior) and/or helping in collecting data (interviews and questionnaires) or carrying out specific measurements but without the effect (effective or perceived) of a supervised test that would not produce true and reliable data due to the effects of the facilitator's presence. A facilitator could also instruct the participant to 'think aloud' so that a verbal record exists of their interaction with the system and its application. A facilitator could also observe and note user behaviors, user comments, and system actions and eventually enter them into a proper case report form (CRF) or data-logging application. In case of the non-participation of a facilitator, all the tools (such as video recording or other systems for self- or automated filling of the CRF) should be prepared. From all these recommendations, the general rule for the conduction of the experiments is that the facilitator should not evaluate the participant but simply record or support the user's assessment of the system.
5. Analyzing and extracting the outcomes and providing the statistical analysis and its interpretation; this task should be preferably conducted by a third researcher who is different to the facilitator so to avoid polarization in interpreting the data and drawing conclusions. From a statistical point of view, specific attention should be given to the nature and characteristics of each variable to properly select the test (normal distribution, paired test for pre–post assessments, etc.).

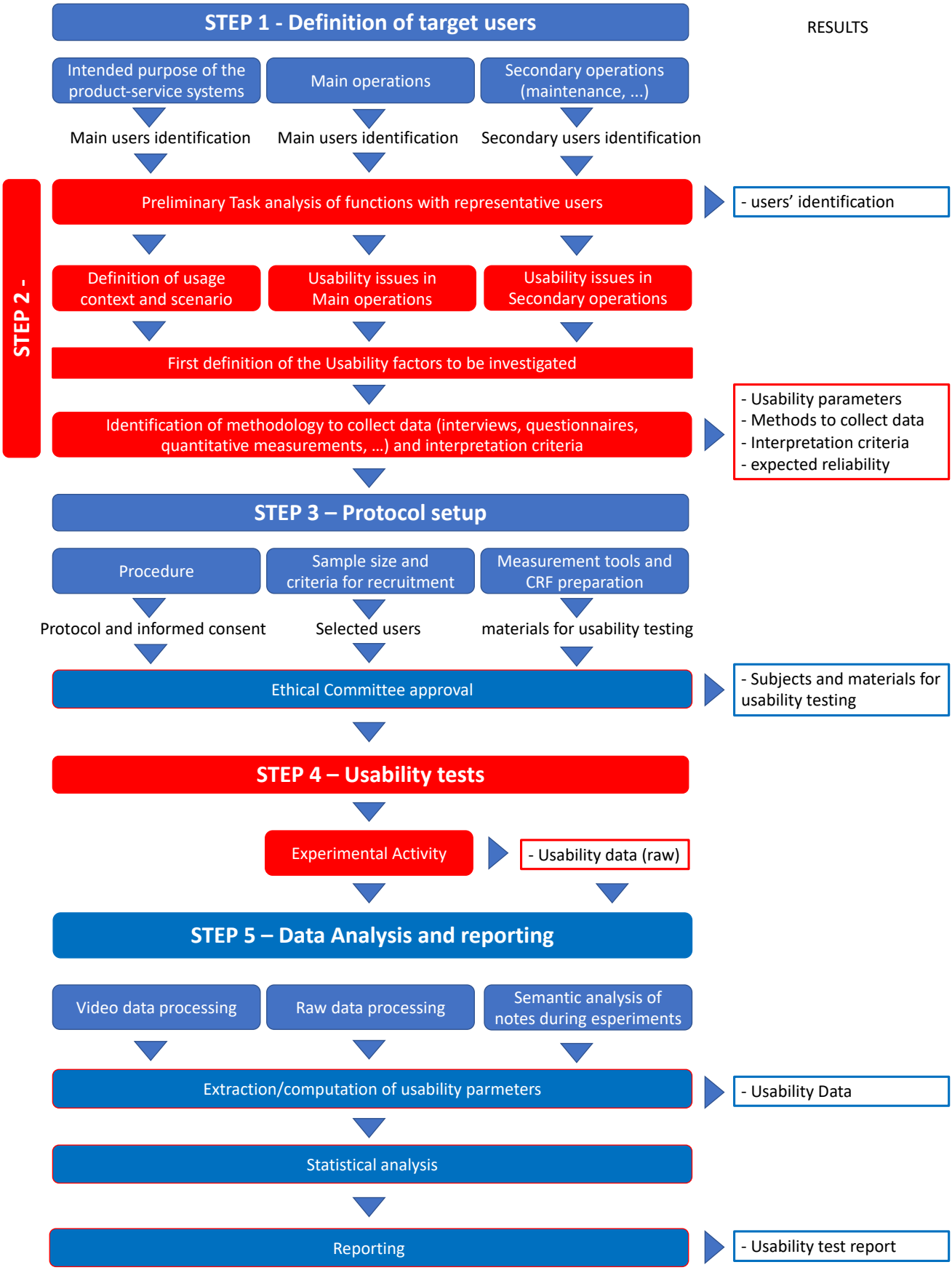


Figure 1. A decision tree describing the proposed methodological approach to prepare a comprehensive usability assessment and the results of each step.

2.2. Define the Usability Metrics

In relation to the different typologies of assessment, proper usability criteria and metrics should be defined. Usability metrics refer to a user's performance measured against the specific performance goals necessary to satisfy the usability requirements. Different aspects could be addressed: some of them are quantitative such as (a) time-to-completion of scenarios, (b) scenario completion success rates, (c) adherence to operational sequences, and (d) error number and rates; other factors are more qualitative and could be analyzed through subjective evaluations. Typically, objective parameters could be measured by suitable tools during the experimental test, while subjective measurements regarding ease of use and satisfaction can be collected via questionnaires and during debriefing at the conclusion of the session. The main advantage of quantitative methods is that they can collect quantifiable data so that the results are easy to compare. The questionnaires can be standardized, such as SUS or TAM, or can be customized using rating scales based on visual analog scales (VAS) [21,22] and the Likert scale [23]. A common mistake is to build questionnaires with a fixed rating scale, and it is worth noting that the Likert scoring system has a variable rating scale from three-point evaluation up to nine-point evaluation according to the semantic factors to be considered. The Likert scales measure agreement. In a Likert scale, respondents are asked how much they agree or disagree with a set of statements. An overall position is derived after analyzing all the responses to the related questions. In addition, usability assessment questionnaires, such as SUS and the standardized user experience percentile rank questionnaire (SUPR-Q) [24], use a Likert scale.

2.2.1. Quantitative Metrics

In relation to quantitative parameters, for example, an assessment of a scenario's completion, they require that the participant achieves a pre-defined goal or inputs specific data that would be used in the course of a typical task. The scenario is completed when the participant indicates that the scenario's goal has been obtained (whether successfully or unsuccessfully). Thus, the facilitator or the participant her/himself indicates a score that is binary (yes/no, 0/1) or numerical (e.g., duration of the given task in seconds with decimals, length of a movement in cm). Examples of quantitative parameters in wearable usability can include the following items:

- Success in self-wearability of the system: binary value of 0/1 or no/yes;
- Self-wearability time: time in seconds with two decimals;
- Self-taking-off time: time in seconds with two decimals;
- Success in self-removing the system: binary value of 0/1 or no/yes.

Other quantitative and derived parameters can be computed at the end of the experiments to complete the usability assessment. Among the most relevant ones, the completion rate is the percentage of test participants who successfully complete the task without critical errors. A completion rate of 100% is the goal for each task in usability tests.

Another key performance indicator is the error-free rate, that is, the percentage of test participants who complete the task without any errors (critical or non-critical errors). Typically, an error-free rate of 80% is the goal for each task in usability tests.

The ISO-TR 16,982 identifies several performance-related measurements that are also called task-related measurements (time spent to complete a task, number of tasks that can be completed within a predefined duration, amount of idle time (it is important to distinguish between system-induced delays, thinking time, and delays caused by external factors), and number of total key strokes).

2.2.2. Errors

During the activities of the usability test, it is also important to record the eventual errors in the execution. We can distinguish between critical and non-critical errors. *Critical errors* are deviations at completion from the targets of the scenario. A critical error is defined as an error that results in an incorrect or incomplete outcome. An example of a critical

error in wearable usability is mounting the device in a wrong position so as to prevent the system from collecting the correct data (e.g., a fall detection sensor). The recording of this kind of error is difficult and generally requires the presence of a facilitator because the participants may or may not be aware that the task goal is incorrect or incomplete. Special attention should be paid to the independent completion of the scenario, which is a universal goal. If help is obtained from the other usability test roles, then this is something that indicates the presence of a critical error. A specific category in critical errors is when the participant starts (or attempts to start) an action in a way that causes its final goal state to become unobtainable. In general, critical errors are unresolved errors during the process of completing the task or errors that produce an incorrect outcome. Instead, *non-critical errors* are errors that are recovered from by the participants, or ones that do not result in processing problems or unexpected results. In other words, a non-critical error is an error that would not have an impact on the final output of the task but would result in the task being completed less efficiently. For example, errors in some procedures that would cause the participant to not complete a given task according to the standard or most efficient procedure (e.g., excessive steps and keystrokes) are coded as non-critical errors. These errors may also be errors of confusion (ex., initially selecting the wrong function or using a user interface control incorrectly, such as attempting to edit an un-editable field). In addition, the detection and recording of these kinds of errors is difficult, and generally the participant is not aware of their occurrence. For this reason, video recordings of usability tests or the presence of a facilitator during the session is recommended. Although non-critical errors can go undetected by the participant, they are generally frustrating to the participant herself/himself. A specific analysis of exploratory behaviors can be carried out, such as identifying the opening of a wrong latch or button while searching for a component to close/open a wearable system while wearing/removing it. These "wrong" actions can be coded as a non-critical error and be reported in the proper section of the CRF or in the free notes, i.e., the blank space where free annotations by the participant or the facilitator can be written. Again, the ISO-TR identifies some typical quantitative parameters: number of errors, time spent recovering from errors, time spent locating and interpreting information in the user's guide, number of commands utilized, number of systems and features that can be recalled, the frequency of use of support materials (documentation, help system, etc.), the number of times that the user task was abandoned, and the number of digressions.

2.2.3. Subjective Assessments

Subjective opinions about specific tasks, the time to perform each task, features, and functionality must be collected. They are essential not only to assess the usability but also to highlight elements and their priority for system improvement. For this reason, it is crucial to focus on the intended purpose of the wearable system and to build a proper set of judgements with their significance and a scoring scale (over three points, five points, seven points, nine points, or even ten points according to the semantic value, as suggested by the Likert methodology) in relation to the functions and/or characteristics under evaluation. For example, functionality and acceptance should follow a seven-point scoring scale, while wearability can be assessed through a five-point scoring scale. For discomfort perception, a non-linear Borg scale is recommended. Other examples and the proposed items for usability assessments are included in the next section.

At the end of the usability test, participants can also rate their satisfaction with the overall system. This questionnaire provides the most relevant set of data collected during usability tests. Combined with the interview/debriefing session, these data are used to assess the attitudes of the participants and the impact of the system.

2.3. Assessment of Impact of the Usability Experience

The collected data about usability have a positive (for example, comfort perception) or negative (for example, discomfort perception) valence. In both cases, it is necessary to evaluate their impact on the system and their relationship with the user, specifically

with respect to usability. Impact can be also interpreted as the influence of the system on personal behavior, social activity, or economic value.

In a positive experience, we can define "impact" as an assessment of the level of satisfaction in the utilization and experience of a system by the user. In the case of a positive experience, the participant experiences no or minimal problems for the successful completion of the task. Typical examples of these impacts could be the participant's willingness to buy the system, their intention to increase their use of the system, the possible adoption of new behaviors in daily life to accommodate the system, or the eventual promotion of the system in a group of friends/relatives. These factors could be rated over a five-point Likert scale, indicating the different levels of impact: high, moderate, low, or null when there is a neutral evaluation or no influence on user experience.

In a negative experience, we can define "impact" as the assessment of the severity and frequency of the problems encountered during the use and experience of a system by the user and the level of influence of this problem on successful task completion. To identify and prioritize the eventual recommendations for the improvement of a system, it is important to adopt a method for the classification of the impact of a problem emerging from the analysis of the data collected during the evaluation activities. This method must consider the combination of two main factors: the severity (S) of the problem and the frequency (F) of occurrence or the number of users experiencing critical or non-critical problems during the evaluation ($N_c + N_{nc}$). In the first approach, the frequency of occurrence of problems or critical errors or non-critical errors during the execution of a given task is simply computed as the number of these errors (N_c and N_{nc}) or the ratio between the number of errors and the time to task completion (ttc). In this case, the impact I is computed as:

$$I = S \times F = \frac{S_c \times N_c + S_{nc} \times N_{nc}}{ttc} \quad (1)$$

where S is the severity, F is the frequency, S_c is the severity of critical errors, N_c is the number of critical errors, S_{nc} is the severity of non-critical errors, N_{nc} is the number of non-critical errors, and ttc is the time to task completion.

Instead, considering users experiencing errors, the frequency of users is the ratio (or percentage) between the number of subjects experiencing critical (n_c) or non-critical errors (n_{nc}) and the total number of participants (n_{tot}). So, following this approach, we can define the impact I as:

$$I = \frac{S_c \times n_c + S_{nc} \times n_{nc}}{n_{tot}} \quad (2)$$

where S_c is the severity of critical errors, n_c is the number of participants experiencing critical errors, S_{nc} is the severity of non-critical errors, n_{nc} is the number of participants experiencing non-critical errors, and n_{tot} is the total number of participants.

According to these two factors (severity and frequency), we can distinguish four levels of negative impact:

- High, i.e., the presence of problems that prevent the user from completing the task (critical error) and their number and/or frequency;
- Moderate, i.e., the presence and high frequency of situations that cause difficulty to the user, but she/he succeeds in completing the task (non-critical error with medium impact);
- Low, i.e., the occurrence and low frequency of repetition of minor problems that do not significantly affect the completion of the task (non-critical error with low impact);
- Null, i.e., the absence of any negative conditions.

A typical example of the impact of some wearable systems is the limitation of mobility while wearing them.

2.4. Data Reporting

Usability data reporting is the main activity of the test and is achieved by providing assessments and suggestions to solve the eventually encountered problems. Usability assessments are also tasks that measure the first experience with a system, because expectations can affect the overall score. For this reason, it is recommendable to include a section measuring expectations before the system's use (pre-test questionnaire).

The outcome of this comprehensive approach is a usability CRF that is described in the following section.

3. Results

The aim of this study is to provide a set of standard and common metrics to assess usability in wearables. In this way, further studies could gather general data with a common setup for sharing knowledge and experience, and they could have the possibility for comparison. The methodological approach presented here combines a set of quantitative metrics and subjective evaluations. The quantitative metrics are related to well-defined temporal parameters (for example, time to task completion such as wearing time, removal time, or time to activate specific functions using the wearable system), binary logical values (for example, successful or unsuccessful task completion), and numerical values (for example, the number of errors performed by the user in given tasks). Instead, subjective assessments follow different approaches both for similar and different applications. Related works have demonstrated that if some common tools such as SUS or TAM focus on a general assessment, the methodologies lack in the assessment of specific but common tasks such as wearing a system, using it, and taking off it. These steps are only rarely and partially investigated, but they represent the basic and common user experience for most systems and applications. The following subsections present a reference set of questionnaires to evaluate the usability of wearable systems. These questionnaires implement a standard scoring method based on the semantic value of items (according to the Likert Scale) for tasks, performance, and appearance/aesthetics. In fact, in accordance with the approach described previously, the main results of steps 1–3 are the sets of questionnaires and methods supporting the data collection, and these are presented in the following Sections 3.1–3.3. The recommendations for analyzing the measured data from a statistical point of view and producing the final report are presented in Section 3.4.

3.1. Pre-Test Questionnaire

In this phase, the user her/himself takes the system, or the facilitator introduces system with a general description. At this step, the users can touch the system but not wear and/or use it.

Then, the user fills in a short questionnaire for describing their product expectations. It consists of five items whose reporting is conducted by means of a seven-point or five-point Likert scale according to the semantic value of the investigated factor, as shown in Tables 1 and 2.

Table 1. Scoring table for the pre-test questionnaire for the factors with their semantic meaning and the corresponding value of 1–7 in the Likert scale system.

Item Score	Expected Functionality	First Look Aesthetics	First Look Acceptance
1	very negative	totally unacceptable	totally unacceptable
2	negative	unacceptable	unacceptable
3	slightly negative	slightly unacceptable	slightly unacceptable
4	neutral	neutral	neutral
5	slightly positive	slightly acceptable	slightly acceptable
6	positive	acceptable	acceptable
7	very positive	totally acceptable	totally acceptable

Table 2. Scoring table for the pre-test questionnaire for the factors with their semantic meaning and the corresponding value of 1–5 in the Likert scale system.

Item Score	Expected Wearability (Considering both Wearing and Removing the System)	Expected Comfort
1	very difficult	very bad
2	difficult	bad
3	neutral	neutral
4	easy	good
5	very easy	very good

3.2. Test Questionnaires

The test questionnaires are organized according to the different phases of the investigated experience. In the case of wearables, three main events can be distinguished: wearing the system, using the system, and, finally, removing the system. The following subsections are dedicated to the presentation of specific tools (scales, charts, and questionnaires) to assess operation during these activities.

3.2.1. Wearing the System

The first operation is generally to wear the system, so the user is asked to wear the system. Her/his ability to put the system on themselves and the related time is recorded.

- Successful self-wearability is assessed through a binary score: 0/1 (N/Y);
- Wearing time is measured in seconds with two decimals.

In case of assistance by the facilitator/other, the level of the received assistance can be rated through a VAS score of 0–10, whose given value should be recorded with one decimal.

The observed non-critical errors are noted by the facilitator or by the user herself/himself.

Wearability and perceived comfort are assessed through the scheme already reported in Table 2.

To gain a better specification of the perceived comfort or related problems, the level of the perceived discomfort both globally and in specific zones of the body could be measured using a 0–100 VAS score (as shown in Figure 2a) and/or a body part discomfort (BPD) scoring system [25,26], as shown in Figure 2b.

Some free observations can be reported by the user as needed in a dedicated blank space.

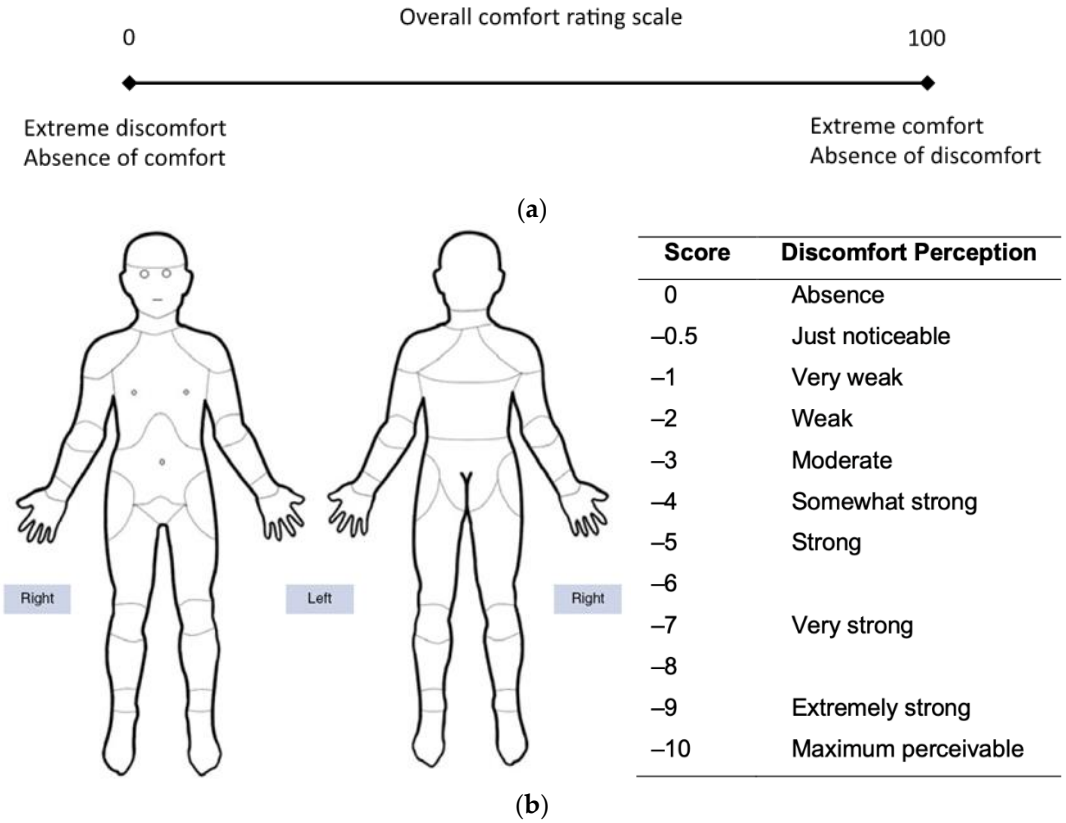
3.2.2. Use

Then, the user can start using the system. Usually, its operation is driven by a piece of software, so the user needs to switch on the device and eventually a related application to pair the system, personalize some settings, and activate some functions. For this, several tasks/subtasks related to global operation could be identified. The most frequently adopted usability scoring method is the SUS tool that fits well with the usability assessment of wearable systems and is a questionnaire composed of 10 items in order to assess the perceived usability of technological systems. In any case, a proper amount of usage time must be considered before administering the test in order to have a consistent usability assessment. Otherwise, and in some cases this should be considered, the usability analysis should be split into two phases: (a) after the first usage (to investigate the intuitiveness of the interface and its functions) and (b) after 2/3 days/1 week (to better explore the overall usability).

Specific and quantitative measures of the system's efficiency, such as successful task completion and the related time, could also be taken. For example, the wearable system must carry out or support some functions; in this case, simple, logical, and immediate scores can be introduced, such as:

- Successful completion of the operation: binary value 0/1 (N/Y);
- Rate of the level of received assistance: VAS 0-10 (value with one decimal);
- Time in seconds to complete the action, measured with two decimals;

- Number of actions in the period.



Score	Discomfort Perception
0	Absence
–0.5	Just noticeable
–1	Very weak
–2	Weak
–3	Moderate
–4	Somewhat strong
–5	Strong
–6	
–7	Very strong
–8	
–9	Extremely strong
–10	Maximum perceivable

Figure 2. The comfort scoring charts system: (**a**) the 0–100 VAS scale to report a general comfort assessment and (**b**) the body part discomfort chart to report the eventually perceived discomfort in specific body areas and its intensity level.

In addition, some subjective evaluations could be recorded on these aspects through some dedicated questionnaires using a Likert-scale-based scoring method in accordance with the proper semantic meaning (Table 3).

Table 3. Scoring table for the pre-test questionnaire for the factors with their semantic meaning and the corresponding value of 1–7 in the Likert scale system.

Item Score	Appropriateness of the Received Support	Importance of the Received Support
1	very negative	not important at all
2	negative	not important
3	slightly negative	slightly not important
4	neutral	neutral
5	slightly positive	slightly important
6	positive	important
7	very positive	very important

At the end of this phase, the assessment of the overall comfort/discomfort though the methods described in the previous step (the 0–100 VAS scale and the BPD chart) should be repeated. Again, a blank space for free text and notes could be given to report free observations or suggestions on how to improve usability.

3.2.3. Removing the System

After completing the use, the subject may or may not have to remove the system. This operation is also important in the usability analysis. Its evaluation follows a similar scheme to that presented in the above Section 3.2.1, with modifications related to taking off the system.

This operation is not simply the reverse of wearing the system: according to the different types of technology, there can be very different situations. Taking off an accessory such as a smartwatch or removing a data logger from a worn support (such as a sensorized t-shirt), placing it on its support for recharging, and then removing the smart garment that is very closely fitted to the body, or even taking off a lower-limb prosthesis and the liner that forms the interface to the body, require very particular sub-task with a high complexity: the method of grasp the device, the force required, the movements required, the number of elements to be removed, and the final operation (charging, cleaning, or mounting or placing onto a support) must be considered during the preliminary task analysis to best prepare the usability assessment of this last phase. From these considerations, we can conclude that it is preferable—but not obvious—that the same assessment scheme used for the analysis of wearing the system could be adopted, but proper adjustments and customization should be introduced.

3.2.4. Other Operations

For some specific categories of wearable systems, other functions could be investigated in terms of usability because they strongly align with it.

Recharge, maintenance, software/firmware updates, and interaction with other devices are example of these operations. If they are apparently secondary tasks, they could indeed be crucial in usability: a short battery duration with the need of recharging the system once a day could be one of the main reasons not to use the wearable device. For this reason, great attention and the highest priority should also be given to these aspects.

The assessment preparation should follow the same approach presented here, with a well-conducted task analysis and the identification of all the factors of usability that are to be evaluated. A proper test protocol and a related CRF can be then prepared.

3.3. Post-Test Questionnaire

To verify the alignment with the expectations recorded at the beginning of the test and to obtain a final overall assessment of the perceived usability, a set of subjective measures could be prepared as a final questionnaire to be administered to the users (Tables 4 and 5).

Finally, the assessment should include the level of satisfaction, support, and importance of the system in supporting the task/functions to which it is devoted to (Table 6).

Table 4. Scoring table for the post-test questionnaire for the factors with their semantic meaning and the corresponding value of 1–7 in the Likert scale system.

Item Score	Perceived Functionality	Perceived Aesthetics	Overall Acceptance
1	very negative	totally unacceptable	totally unacceptable
2	negative	unacceptable	unacceptable
3	slightly negative	slightly unacceptable	slightly unacceptable
4	neutral	neutral	neutral
5	slightly positive	slightly acceptable	slightly acceptable
6	positive	acceptable	acceptable
7	very positive	totally acceptable	totally acceptable

Table 5. Scoring table for the post-test questionnaire for the factors with their semantic meaning and the corresponding value of 1–5 in the Likert scale system.

Item Score	Wearability (Considering Both Wearing and Taking Off the System)	Perceived Comfort
1	very difficult	very bad
2	difficult	bad
3	neutral	neutral
4	easy	good
5	very easy	very good

Table 6. Scoring table for the post-test questionnaire for the factors with their semantic meaning and the corresponding value of 1–7 in the Likert scale system.

Item Score	Perceived Satisfaction	Perceived Support in the Function	Overall Importance
1	very negative	not important at all	not important at all
2	negative	not important	not important
3	slightly negative	slightly not important	slightly not important
4	neutral	neutral	neutral
5	slightly positive	slightly important	slightly important
6	positive	important	important
7	very positive	very important	very important

3.4. Final Usability Test Report

A usability test report should be provided at the conclusion of the usability test. It should consist of a report and/or a presentation of the results, including an evaluation the usability metrics and their comparison against the pre-defined goals, subjective evaluations, and specific usability problems and possible recommendations for their resolution. A nonparametric statistical analysis could be used for the proper parameters and for the pre/post-test analysis. A nonparametric statistical analysis should be applied when the distribution of the measures is not normal (i.e., following a Gaussian curve). Therefore, in the case when the researcher needs to compare two independent means, instead of a two-sample *t*-test that is used in the case of a normal distribution, it is recommended to use the Mann–Whitney U test. The same applies for comparing two dependent means: the paired *t*-test should be changed into the Wilcoxon signed rank test. For the analysis of correlations, the Spearman rank should be used instead of the Pearson correlation. In the case when the statistical analysis is related to more than two conditions, in the case of independent means, the Kruskal–Wallis test should be adopted instead of the ANOVA (one way) test, while for dependent means, the Friedman test applies instead of the ANOVA test for repeated measures.

4. Discussion

This paper presented an integrated methodology for the development and execution of usability tests for wearable systems. Wearable systems are new and miniaturized technologies that are pervading our lives in several applications: measuring performances in fitness or sport activities, e-health and m-health systems for monitoring physiological signs in chronicity management at home or when monitoring patients in hospital, as well as during rehabilitation processes, and even a prosthesis for an upper or lower limb can be considered a wearable device. New applications at the workplace have recently been introduced with wearable exoskeletons empowering human functions in completing demanding or complex tasks. Usually, two main factors are analyzed: the technological characteristics and performance, and the usability and acceptance of these systems by the users. Technological requirements are mainly quantitative and easy to directly measure. Instead, their usability, acceptance, and the user experience is poorly studied: Keogh et al. [13] confirmed this finding. Furthermore, in studies found in the scientific literature, the research methodology

is usually simplified or considers only limited factors. From these findings, it emerged that there is a lack of a structured methodological approach to develop a comprehensive protocol. This paper aimed at providing a methodological basis and the related tools for developing a structured protocol. As defined in the methodological section, the first action is to clearly define the objectives of the study. Three main aspects could be addressed: (a) identify design inconsistencies and usability problems or errors; (b) validate the use of the wearable system under controlled test conditions with representative users; and (c) establish a baseline in terms of user performance and user satisfaction levels. This aspect seems to be underestimated, and most of the related studies are reactive and target the second goal, i.e., they involve an evaluation of the usability of a certain device. Indeed, in the design phase, the first goal is crucial and can provide significant elements for product optimization and refinement before the final development and delivery. The third goal can be related to the creation of a reference or gold-standard score to compare systems belonging to the same category of systems. The three goals are different and are also related to different phases of the lifecycle of a product. For this reason, the experimental protocol should be designed accordingly and with proper tools, considering adequate measurement methodologies and supporting proper data analysis techniques, for example, from a statistical point of view (for example, a pre–post comparison to assess the evolution of the product during the design phase for goal 1, a usability measurement for a commercial system for goal 2, and an analysis of usability for a panel of devices for goal 3).

Leveraging on the many research activities carried out in the field of wearable sensors design and development, this paper tried to provide a complete and integrated perspective of the usability assessment of wearable systems. The paper contributes to highlighting the different aspects of usability, dividing the three main tasks (wearing, utilization, removing) and the different parameters to be considered in the evaluation of the operations carried out by the user in each phase. The generalization in the main tasks drove the proposal of a set of quantitative metrics to be adopted as standard indexes and the identification of the most relevant qualitative factors to be investigated by means of questionnaires, namely functionality, aesthetics, acceptance, wearability, comfort, satisfaction, support to the user, and importance. The previous approaches usually adopt a common interpretation grid (in general, a five-point Likert scale) for all factors, ignoring the fact that the Likert methodology uses different validated scoring scales in function of the semantic meaning of the factors.

A parameter that is usually ignored by all studies is the expectation: indeed, this factor influences the satisfaction and acceptance that are so relevant for a complete usability evaluation.

The proposed five-step method ((1) define the target users; (2) execute a preliminary task analysis for identifying the context, the parameters to be measured, and the methodology to collect data; (3) prepare the protocol and the investigation tools; (4) execute the usability experiments; and (5) analyze and report the data) provides a step-by step guideline to researchers to avoid deviations from the objective together with a set of usability factors and parameters, their definitions, and the questionnaires or measurement methods.

The availability of clear and reliable usability data is a fundamental step for supporting researchers and developers in improving the quality and usability of these systems. Due to their increasing diffusion and application in our lives, their good acceptance and usability is a necessary condition for the full exploitation of these systems that aim at increasing our health and quality of life.

The methodological limitation of this paper consists of its non-radical innovation: it applies known indexes and methods to measure usability parameters, such as SUS, TAM, and other indexes. These methods were chosen whilst reflecting on the identified usability factors for wearables and as a balance between specificity and easy measurability. So, the five-step protocol design guideline, the reference charts, and the questionnaires to measure usability are proposed to setup a reference standard framework for usability studies. Maybe this is ambitious, but it is necessary for implementing the third goal in order to gain an

integrated and coherent comparison among a set of homogeneous systems in terms of function for typologies and functions. A second limitation of this study lies is in the different categories of wearables that can be identified with very specific purposes and applications: for this reason, the proposed generalization can appear superficial. The methodology here presented can be integrated by a set of specific assessments (measures, quantitative metrics, and subjective evaluations) in terms of the function of the parameters of interest that the preliminary task analysis can identify. The last limitation of this study is that no application of the method is presented here. This would have required a too extended version of the text, and a future paper will demonstrate the reliability of the method in some applications. This is also the direction of future research that will explore the usability of different kinds of systems (smart garments, body-worn accessories, exoskeletons, and prosthetics) in selected and representative applications.

Funding: This study is part of the research that was funded by INAIL, grant number PDT 3/1–TUTA-Multimodal Wearable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Chan, M.; Estève, D.; Fourniols, J.Y.; Escriba, C.; Campo, E. Smart wearable systems: Current status and future challenges. *Artif. Intell. Med.* **2012**, *56*, 137–156. [CrossRef] [PubMed]
2. *ISO 9241-210:2019*; Ergonomics of Human-System Interaction—Part 210: Human-Centred Design for Interactive Systems. ISO: Geneva, Switzerland, 2019.
3. *ISO 9241-11:2018*; Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts. ISO: Geneva, Switzerland, 2018.
4. *ISO/TR REPORT 16982:2022*; Ergonomics of Human-System Interaction—Usability Methods Supporting Human-Centred Design. ISO: Geneva, Switzerland, 2022.
5. Keogh, A.; Dorn, J.; Walsh, L.; Calvo, F.; Caulfield, B. Comparing the Usability and Acceptability of Wearable Sensors among Older Irish Adults in a Real-World Context: Observational Study. *JMIR Mhealth Uhealth* **2020**, *8*, e15704. Available online: https://mhealth.jmir.org/2020/4/e15704 (accessed on 25 November 2022). [CrossRef]
6. System Usability Scale (SUS). Available online: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html (accessed on 30 November 2022).
7. Brooke, J. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* **1996**, *189*, 4–7.
8. Liang, J.; Xian, D.; Liu, X.; Fu, J.; Zhang, X.; Tang, B.; Lei, J. Usability Study of Mainstream Wearable Fitness Devices: Feature Analysis and System Usability Scale Evaluation. *JMIR Mhealth Uhealth* **2018**, *6*, e11066. [CrossRef]
9. Markland, D.; Hardy, L. On the Factorial and construct validity of the Intrinsic Motivation Inventory. *Res. Q. Exerc. Sport* **1997**, *68*, 20–32. [CrossRef]
10. Domingos, C.; Costa, P.; Santos, N.; Pêgo, J. Usability, Acceptability, and Satisfaction of a Wearable Activity Tracker in Older Adults: Observational Study in a Real-Life Context in Northern Portugal. *J. Med. Internet Res.* **2022**, *24*, e26652. Available online: https://www.jmir.org/2022/1/e26652 (accessed on 12 February 2023). [CrossRef]
11. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef]
12. Gil-Gómez, J.A.; Manzano-Hernández, P.; Albiol-Pérez, S.; Aula-Valero, C.; Gil-Gómez, H.; Lozano-Quilis, J. USEQ: A Short Questionnaire for Satisfaction Evaluation of Virtual Rehabilitation Systems. *Sensors* **2017**, *17*, 1589. [CrossRef]
13. Keogh, A.; Argent, R.; Anderson, A.; Caulfield, B.; Johnston, W. Assessing the usability of wearable devices to measure gait and physical activity in chronic conditions: A systematic review. *J. Neuroeng. Rehabil.* **2021**, *18*, 138. [CrossRef] [PubMed]
14. McCallum, C.; Rooksby, J.; Gray, C. Evaluating the Impact of Physical Activity Apps and Wearables: Interdisciplinary Review. *JMIR Mhealth Uhealth* **2018**, *6*, e58. Available online: https://mhealth.jmir.org/2018/3/e58 (accessed on 12 February 2023). [CrossRef]
15. Martinato, M.; Lorenzoni, G.; Zanchi, T.; Bergamin, A.; Buratin, A.; Azzolina, D.; Gregori, D. Usability and Accuracy of a Smartwatch for the Assessment of Physical Activity in the Elderly Population: Observational Study. *JMIR Mhealth Uhealth* **2021**, *9*, e20966. Available online: https://mhealth.jmir.org/2021/5/e20966 (accessed on 12 February 2023). [CrossRef] [PubMed]
16. Hawthorne, G.; Greening, N.; Esliger, D.; Briggs-Price, S.; Richardson, M.; Chaplin, E.; Clinch, L.; Steiner, M.; Singh, S.; Orme, M. Usability of Wearable Multiparameter Technology to Continuously Monitor Free-Living Vital Signs in People Living with

Chronic Obstructive Pulmonary Disease: Prospective Observational Study. *JMIR Hum. Factors* **2022**, *9*, e30091. Available online: https://humanfactors.jmir.org/2022/1/e30091 (accessed on 12 February 2023). [CrossRef]
17. Bendig, J.; Spanz, A.; Leidig, J.; Frank, A.; Stahr, M.; Reichmann, H.; Loewenbrück, K.; Falkenburger, B. Measuring the Usability of eHealth Solutions for Patients with Parkinson Disease: Observational Study. *JMIR Form. Res.* **2022**, *6*, e39954. Available online: https://formative.jmir.org/2022/10/e39954 (accessed on 12 February 2023). [CrossRef] [PubMed]
18. Moon, N.W.; Baker, P.M.; Goughnour, K. Designing wearable technologies for users with disabilities: Accessibility, usability, and connectivity factors. *J. Rehabil. Assist. Technol. Eng.* **2019**, *6*, 1–12. [CrossRef]
19. Andreoni, G.; Standoli, C.E.; Perego, P. Defining Requirements and Related Methods for Designing Sensorized Garments. *Sensors* **2016**, *16*, 769. [CrossRef] [PubMed]
20. Faulkner, L. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behav. Res. Methods Instrum. Comput.* **2003**, *35*, 379–383. [CrossRef]
21. Gift, A.G. Visual Analogue Scales. Measurement of Subjective Phenomena. *Nurs. Res.* **1989**, *38*, 286–287. [CrossRef]
22. Vagias, W.M. *Likert-Type Scale Response Anchors*; Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management, Clemson University: Clemson, SC, USA, 2006; Available online: http://media.clemson.edu/cbshs/prtm/research/resources-for-research-page-2/Vagias-Likert-Type-Scale-Response-Anchors.pdf (accessed on 25 November 2022).
23. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.* **1932**, *22*, 55.
24. Sauro, J. SUPR-Q: A Comprehensive Measure of the Quality of the Website User Experience. *J. Usability Stud.* **2015**, *10*, 68–86.
25. Corlett, E.N.; Bishop, R.P. A Technique for Assessing Postural Discomfort. *Ergonomics* **1976**, *19*, 175–182. [CrossRef]
26. Drury, C.G.; Coury, B.G. A methodology for chair evaluation. *Appl. Ergon.* **1982**, *13*, 195–202. [CrossRef] [PubMed]