Multimodal Technologies and Interaction

MDPI

Article

EUREKATAX: A Taxonomy for the Representation and Analysis of Qualitative Usability Test Data

Panagiotis Germanakos [1,*] and Ludwig Fichte [2]

[1] UX S/4HANA Product Engineering, Intelligent Enterprise Group, SAP SE, 69190 Walldorf, Germany
[2] UX Research, Shopify Inc., Ottawa, ON K2P 1L4, Canada; ludwig.fichte@shopify.com
* Correspondence: panagiotis.germanakos@sap.com

Received: 30 March 2020; Accepted: 18 May 2020; Published: 25 May 2020

Abstract: Usability tests serve as an insightful source of feedback for product teams that want to deliver user-centered solutions and enhance the User Experience (UX) of their products and services. However, in many cases, formative usability tests in particular may generate a large volume of qualitative and unstructured data that need to be analyzed for decision making and further actions. In this paper, we discuss a more formal method of analyzing empirical data, using a taxonomy, namely Engineering Usability Research Empirical Knowledge and Artifacts Taxonomy (EUREKATAX). We describe how it can provide guidance and openness for transforming fuzzy feedback statements into actionable items. The main aim of the proposed method is to facilitate a more holistic and standardized process to empirical data analysis while adapting on the solution or context. The main contributions of this work comprise the: (a) definition of the proposed taxonomy which represents an organization of information structured in a hierarchy of four main categories (discover, learn, act, and monitor), eight sub-categories, and 52 items (actions/operations with their respective properties); (b) description of a method, that is expressed through the taxonomy, and adheres to a systematic but modular approach for analyzing data collected from the usability studies for decision making and implementation; (c) formulation of the taxonomy's theoretical framework based on meticulously selected principles like experiential learning, activity theory: learning by expanding, and metacognition, and (d) extended evaluation into two phases, with 80 UX experts and business professionals, showing on the one hand the strong reliability of the taxonomy and high perceived fit of the items in the various classifications, and on the other hand the high perceived usability, usefulness and acceptability of the taxonomy when put into practice in real-life conditions. These findings are really encouraging, in an attempt to generate comparable, generalizable and replicable results of usability tests' qualitative data analysis, thereby improving the UX and impact of software solutions.

Keywords: user experience; user-centred design; taxonomy; qualitative data analysis; usability tests

1. Introduction

In today's highly competitive technological environment, the concepts of User Experience (UX) and usability are at the center of attention, with many organizations investing heavily on related activities each year [1]. The main objective is to enhance the quality of their products and services by designing and developing user-centered solutions so to stay ahead of their competition. *User Experience* is a broader notion that relates to a wide spectrum of concepts like usability, desirability, accessibility, usefulness, etc., of a prototype, component, or functional system. It might refer to "all aspects of the end-user's interaction with the company, its services, and its products" [2], or it might relate to "a momentary, primarily evaluative feeling (good-bad) while interacting with a product or service" [3]. In contrast, *usability* is a term that is more closely connected to the user interface of a product or a service, with many international organizations and researchers proposing various descriptions and alternatives over the years [4–8]. Currently, a commonly used definition of usability is the one of the International Organization of Standardization (ISO) in ISO 9241-11 as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [9]. These values could be regarded as Key Performance Indicators (also known as KPIs) for a business solution and can be evaluated through a series of usability tests that a product team may run during the validation phase of the User-Centered Design (UCD) and software development process [10]. Usability testing encompass various steps before and after the actual execution, that is taking place either on-site (i.e., customers' physical environment) or remotely. Such steps may include planning, recruitment of participants, preparation of validation scripts, observations and interviews, analysis of collected data, reporting, etc. The main aim of usability testing is to validate a solution by measuring effectiveness, efficiency and satisfaction of user interactions [6,11] with specific tasks via its user interface (the "touching" or connecting point with a technology or application). In recent years, there has been an extensive work on tools and methods (e.g., Reframer (by OptimalWorkshop), WhatUsersDo, Userfeel, Ovologger, Validately, UserZoom, etc.) that help researchers to collect data from users, either explicitly or implicitly (or using Mixed Methods Research [12,13]). In addition, there is a rich spectrum of solutions and models that facilitate the quantitative analysis of the captured feedback by applying computational techniques and algorithms that can produce statistical, mathematical, or numerical results liable to be interpreted and aligned with the research objectives. For the qualitative analysis of the data collected, a number of methods and workflows exist that employ either more rigorous or ad-hoc approaches to data analysis process, as well as tools that provide features and functionality to facilitate distinct data analysis actions, e.g., coding, clustering data (e.g., NVivo, Dedoose, F4analyse, Delve, etc.)

In this respect, a big challenge for researchers and practitioners nowadays, especially in the business sector, is to be able to create qualitative data analysis procedures and tools that will be able to capture the breadth and depth of the situation-specific scenarios and contexts, offering a more standardized and reproducible process and results. There is still the need for a more inclusive modeling of the activities that could guide and assist the transformation of the unstructured, uncertain and fuzzy seeds of information (e.g., opinions, suggestions, sentiments, experiences—a formative usability study of e.g., eight end-users (The term end-users refers to usability test session participants [also often referred to as "customers"] in their various business roles that interact with a product or a service considering a specific point of view that usually reflects their persona. A product team may conduct a number of usability tests with them so to observe and collect their feedback while accomplishing specific goals.) may produce an average of 150 feedback items of any nature) into a coherent representation of knowledge and actions. Such a method could complement existing quantitative approaches in order to generate more inclusive result of validation tests and thus multiplying the impact of an outcome or decision [14].

In this paper, we focus on the data analysis of empirical data gathered by various usability test sessions (especially formative usability tests). More specifically, we address the need of modeling the qualitative data analysis actions involved in this validation phase of a software's usability and we propose a taxonomy, namely Engineering Usability Research Empirical Knowledge and Artifacts Taxonomy (EUREKATAX). A taxonomy could be defined as a classification system that (a) organizes information based on predetermined categories ordered in a specific way, or (b) classifies multifaceted, complex phenomena according to common conceptual domains and dimensions [15,16]. It may be regarded as a language that communicates the understanding of a purposeful structured meaning of situation-specific matters, while leveraging the related expectations, knowledge and experiences. In business organizations, data analysis of usability studies is performed in teams composed of various members with different roles (such as UX interaction designers, product owners, development architects, business experts, etc.), educational backgrounds, mindsets, experiences, or synthesis and analysis skills. This results in diverse expectations regarding an outcome, the interpretations as well as the data analysis process itself. Providing a taxonomy that would bring awareness prior to execution would benefit in creating the appropriate (shared) mental models for the exact next step, saving time and effort, and offering semantically enriched knowledge for describing transformed goal-directed feedback items, measurable outcomes and, business value. More specifically, the main contributions of this work emphasize:

(a) The definition of a taxonomy, namely EUREKATAX, for usability tests data analysis—taking place during the validation phase of the UCD software development process. The proposed taxonomy consists of four main categories, eight sub-categories, and 52 items—operations with their respective properties.

(b) A qualitative data analysis method (and tool), that is expressed through the taxonomy, and adheres to a systematic but modular approach able to guide the transformation of empirical data into actionable items for decision making and implementation. It maintains the necessary openness and flexibility to be adopted in different scenarios and to the extent that a product team wants and needs.

(c) The formulation of EUREKATAX's theoretical framework, that supports the design of its structure and hierarchies, the rationale and learning during the data analysis process, and the actions towards the production of inclusive and meaningful results. We have carefully selected the underlying theoretical perspectives, directly related to the qualitative data analysis process executed by product teams, like Kolb's experiential learning, Engeström's activity theory: learning by expanding and metacognition.

(d) The extensive evaluation of the taxonomy on various levels of realization, aiming at triangulating the results and reaching to safe interpretations regarding its outcomes and validity. The first phase was conducted with 20 users (UX professionals and experts) investigating the reliability of the taxonomy, the proposed classifications of items and actions into the specified categories, and the theoretical considerations. The second phase was executed with 60 users—professionals in various business roles working into teams, for evaluating the implications and perceived usability of EUREKATAX once put into practice in a real-life scenario.

Preliminary results from Phase 1 reveal the high reliability of the proposed taxonomy and the internal consistency of its structure and hierarchies. In principle, experts show clear preference towards the various categories, sub-categories, operations, and properties that are realized in four different depths of detail and learning outcomes. Their responses to open-ended questions, include some interesting recommendations for enhancing EUREKATAX and valuable insights regarding items that would need further improvement and alignment with the current content of the taxonomy. The users and teams from Phase 2 express their positive behaviour towards the applicability and usefulness of the method (and tool) while exercising the various process steps and actions for analyzing qualitative data in a real-life setting. Their perceived usability and NPS scores show a central tendency towards the top values of the control factors' scales, with clear preference of recommending EUREKATAX to other prospective users. In addition, their explicit feedback that was collected in real-time during the data analysis process (e.g., strategies, opinions, feelings, wishes, experiences), in three distinctive phases, cross-verify the latter results, and highlight deeper meanings, strengths, and weaknesses for future consideration.

The remainder of this paper is structured as follows: Section 2, presents a literature review on related taxonomies and the method for constructing EUREKATAX. Section 3, deep dives on the theoretical dimensions that drive the logic behind its composition and influence the specific building blocks and hierarchies. In Section 4, we detail the composition of the theoretical model and the various classifications of EUREKATAX, i.e., discover, learn, act, and monitor. Section 5, presents the two evaluation phases, discussing the results, challenges, and limitations; and Section 6 draws the final conclusions.

2. Background Work and Method

This section makes an extensive reference to related work and taxonomies around the topic of investigation and details the methodological pillars of building EUREKATAX.

2.1. Related Work

It is quite sufficient that the proposed definitions around UX and usability, introduced earlier, embrace different concepts and dimensions, while at the same time the semantics that convey a consistent interpretation might alter depending the situation and context of application. This inherent reality motivated researchers and practitioners to invest a lot of effort to develop domain-specific usability models and constructs [17], methodologies, and guidelines [18–20] that could realize more prominently the UCD process and encapsulate more inclusively the entity user, i.e., understanding behaviors, humanistic characteristics, habits, daily routines, tasks, etc., in the environment that he functions and interacts. A typical example of the situation-specific implementation of the UCD process could be regarded the business sector for enterprise solutions. In this case, product teams most of the times need to comply with the short delivery cycles of products or services (e.g., four releases per year or even in shorter sprints), and at the same time to make sure that they incorporate UX essentials and the voice of the end-user throughout the project life-cycle. As such, the time constraints may not always allow for a full-fledged consideration of the traditional UCD process, either in terms of process steps or depth of requirements/data analysis. Now, the need shifts towards more agile methods such as Lean UX that are characterized by short iterations and the extraction of quick results without thorough documentation or detailed deliverables. Lean UX is all about multi-disciplinary teams, testing ideas early and often, prototyping, running UX sprints within a week, etc. Gothelf and Seiden provide a very practical guide on how to apply UX principles and methods into a lean/agile environment [21]. Nevertheless, although there are clear benefits from this approach to interaction design and development, there is also the risk of losing important insights during data collection and analysis or worse taking into consideration invalid assumptions, requirements and functionalities. EUREKATAX could be considered as a supplement to Lean UX practices, by providing a methodology-based structure to teams that have diverse backgrounds, limited time, but the willingness to document and learn from their own work and observations over time.

Moreover, a number of interesting research findings have been reported on information hierarchies, classification of actions, and taxonomies that concentrate either around the usability notion itself or its applications, bridging any arising gaps triggered by situation-specific scenarios or implementations. More specifically, Alonso-Rios et al. [22] presented an extensive review of the various concepts that describe the usability factor and proposed their organization through a detailed taxonomy accompanied by thorough descriptions of its attributes and relationships. Keenan et al. [23] highlighted the usability problems aspect and stressed the challenge of creating techniques that could suggest a meaningful and inclusive organization of them. In this respect, they suggested a taxonomic model (Usability Problem Taxonomy (UPT)) in which the detected usability problems are classified from both an artifact and a task perspective. Vilbergsdottir et al. [24] evaluated the Classification of Usability Problems (CUP) scheme aiming at further classifying the usability problems for providing a better guidance to developers regarding the understanding of the generated usability issues, how to tackle them with effective fixes or how to avoid having them in the future. Furthermore, Andre et al. [25] developed the User Action Framework, a structured knowledge base of usability concepts and issues, to guide the interaction development activities of usability engineering support tools and to facilitate high-quality usability problem reporting. From a different standpoint, Hermann et al. [26] introduced a general user taxonomy for Information Communication Technologies-systems and products, facilitating the development of intuitively usable design, especially devices and services tailored to different user groups (e.g., for the generation of different user models for model-based usability evaluation in early stages of product development), and Rajeshkumar et al. [27] collected and classified the various UX and usability evaluation methods used in different contexts and environments into different types of taxonomies, signifying also their correlations and disassociations. A taxonomy of definitions for usability studies in biometrics has been proposed by Michaels et al. [28] following a user-centred formulation of terminologies rather than system-oriented ones as traditionally used in the area. In his work, Gabbard [29] refers to a multi-dimensional taxonomy of usability characteristics for virtual environments, including usability suggestions and context-driven challenges; Singh et al. [30] discussed a taxonomy of usability requirements and design concepts for home telehealth systems which enable a more patient centric design; while Huang [31] outlines a taxonomy for measuring and improving the usability of E-Commerce systems. Lastly, Adamides et al. [32] present in their work a taxonomy of design guidelines for robot teleoperation after reviewing initially a set of 70 guidelines and eight categories ranging from architecture and scalability, error prevention and recovery to visual design and environmental information.

Most of the suggested classifications relate to the usability notion itself; the meaning, dimensions, problems, methods, measurable variables or its adoption in heterogeneous domains, or they express frameworks in the phase of the UCD validation process which relates to more rigorous examination of the empirical data collected. In this work, our main concern is to propose a comprehensive taxonomy that explores the possibility of defining a more systematic and rather flexible hierarchy of items so to guide the qualitative data analysis gathered from the numerous usability study sessions with the end-users. The aftermath would be the informed transformation of fuzzy and unstructured feedback into action items, producing rich, reusable, and comparable results ready for consumption. It is true that this is a big challenge, only by considering the transitional process step from data capture and cleansing to analysis, i.e., the data synthesis and consolidation across note takers and end-users (e.g., in a formative usability session one end-user, one moderator and two note takers might participate. If for each customer we conduct e.g., three sessions multiplied by three customers, this results in nine end-users by two sets of notes that need to be brought together so as to formulate a cross understanding of the information that will influence a solution). This stage is characterized by a certain degree of uncertainty during its implementation. It is situation-specific and highly contextual that it is not easy to decide what are the steps that can lead to an inclusive and well optimized result out of e.g., 300 comments generated from a formative usability study. As Spencer et al. mention in [33] "qualitative data are usually voluminous, messy, unwieldy, and discursive", or according to Miles [34] "an attractive nuisance". On the other hand, one of the main disadvantages of qualitative research methods (compared to e.g., quantitative), is the lack of standardization in data analytic techniques and procedures; meaning comparing, generalizing or replicating the results of the analyses may not be a feasible task [35]. We consider EUREKATAX as an inclusive framework towards addressing this gap.

2.2. Method for Constructing EUREKATAX

Building the taxonomy is an iterative process which might be based on a series of revisions and progressive refinements. For EUREKATAX, we first started by collecting and synthesizing information from the literature (including e.g., conferences, journals, books), using leading publications venues (e.g., the database of the ACM digital library) and targeted Web searches, for investigating published usability frameworks and classifications, and weighting their pros and cons. In parallel, we interviewed researchers, interaction designers and developers gathering their input and experiences. We also conducted numerous workshops with heterogenous project teams, performing qualitative analyses on data gathered from usability studies with customers regarding several business products. The follow-up one-to-one discussions and focus groups helped us to assemble valuable feedback, insights and needs from real-life scenarios and situation-specific events. Hence, the formation of EUREKATAX is based on specific empirical data consisting of statements, sentiments, suggestions, challenges, problems, lessons learned, findings, and facts of historical analyses, or more broadly previous representations of the body of investigation expressed from object-historical and theory-historical data [36]. The information, observations, and research outcomes were gradually cultivated and structured into a taxonomy identifying grouping characteristics with specific implications according to the various depths of data analysis. Below we present a synopsis of the main methodological milestones for the creation of the taxonomy:

1. Literature review of current frameworks, standards, classifications, and theoretical perspectives
2. Analysis of research findings from workshops, observations, interviews, and focus groups

3. Sorting useful related outcomes and classifications, and decision for re-usability, alignment or creation of new attributes
4. Creating the backbone—purposeful *categories* and association with theoretical considerations
5. Generating the multi-layer hierarchies and define *sub-categories*, *operations* and *properties*
6. Coin the exact definition for each attribute and populate the final structure and order of the taxonomy
7. Evaluate the proposed taxonomy using widely acknowledged and used methods as Delphi Card Sorting [37] and System Usability Scale (SUS) [38].

Especially, the very first three methodological milestones served as the groundwork for our own synthesis, for covering not only the requirements and limitations of a qualitative data analysis process but also to clarify more generic influential factors and theoretical aspects that would drive the reasoning behind the development of EUREKATAX and its adaptability to different domains and contexts of use. In this respect, for the construction of the taxonomy we also considered a number of complementary aspects, like:

(a) The data analysis process should facilitate creativity and openness to novelty, considering that it is inherently related to finding solutions to problems;
(b) The teams participating in the data analysis process are presenting different backgrounds, roles, skills, and communication styles. Hence, a presentation and functioning that would bring everyone on-board sharing the same mental models is deemed necessary;
(c) It should address the complexities of the feedback items and the generated problems and challenges. There is an essential need to decompose problems into sub-problems, considering abstraction, modularity and re-usage of previous solutions as necessary assets and abilities. Thus, our hierarchy applied the principles of deductive reasoning, it is structured and develops from the abstract to the concrete; starting with a general goal (category or activities) followed by a number of sub-categories (actions) and operations (conditions) with properties (knowledge units) in order to reach a specific logical conclusion or interpretation;
(d) There are specific elements e.g., conceptual, relationship, perspective, participant characteristics [16], that principally describe and generate the structure, themes (fundamental recurrent unifying concepts or statements about a subject matter) [39] and theory (a set of general, modifiable propositions that help explain, predict, and interpret events or phenomena of interest [15]) of a taxonomy. In this line, EUREKATAX assembles a set of categories and relationships that embrace theoretical perspectives like experiential learning [40], activity theory [36], and metacognitive strategies [41] (see Section 3) for increasing clarity, transparency, and control over the qualitative data analysis process; and
(e) It should demonstrate a strong and consistent association of its elements, relationships, and the produced knowledge. Our main concern for the proposed taxonomy was to define systematic and coherent classifications of information that would tightly couple its content with the theoretical principles employed. Accordingly, we apply on one hand a horizontal movement (or coverage) across its hierarchical structure, obtaining a holistic understanding, knowledge, and progressive skills acquisition while operating on the various goal-directed categories that present a continuous/gradual growth and development through a specific input (trigger) and a transformed valuable outcome (knowledge). On the other hand, we dive into each categorical classification following a vertical movement, identifying complementary potential and limitations during operation which are driven by situation-specific learning, highly contextual phenomena, and underlying forces of the organizational borders (relationships of concepts and data) for decision making and problem-solving.

An effective taxonomy defines a sequence of instructions or levels of performance that might be expected for any given content element. EUREKATAX defines a system of guidance extending to four learning levels (depths) of detail regarding an outcome. It maintains an iterative, cyclic way of instruction across the zones of proximal development [42] of participants, as described in

Section 4. More specifically, our taxonomy consists of four main areas (categories) of qualitative data analysis: the discover, learn, act and monitor layers. Each of these layers are progressively disclosed and presented at various levels (depths) of detail, including specific examples and suggestions. The proposed taxonomy is an abstraction, construct, and enumeration of qualitative data analysis aspects and attributes that can be used by project teams for exploring the information captured from their usability studies, for assessing their product, application, functional prototype, or simply a design. At this point, we need to clarify that devising a qualitative data analysis taxonomy is not a trivial task, and even though EUREKATAX defines a unified hierarchical model, we could not claim that presents an exhaustive or complete archetype of empirical data analysis. However, it represents a comprehensive paradigm that supports (either as a guide or as standalone categories and classifications) the extraction of insightful learning outcomes and meaningful action items through several refinement cycles during the qualitative data analysis process.

3. Theoretical Considerations of the Taxonomy

In this section, we discuss the theoretical framework that EUREKATAX has been built upon. Theory provides the grounds and understanding of the use and the significance of a taxonomy, system, organization, interaction, or phenomenon. It provides insights regarding the interoperation, consistency, coherence, reliability, and interpretation of its elements and scope, while in parallel minimizes the probability of a result, learning outcome or finding to be the product of a chance. Moreover, it helps us formulate an awareness around potential causal links and confounding variables, or the nature of correlative or causal relationships, of a phenomenon and its context. Theory usually relates to the exploration of the systematic reasons for the events, experiences, and phenomena of inquiry, and could be used to predict and explain phenomena or it could provide a potential framework for guiding subsequent empirical research [16,43,44]. Thereupon, we elaborate on the theoretical perspectives and their contribution to the reasoning that drove the specific design of the taxonomy's hierarchical structure. These theoretical concepts have been carefully selected to act as the "glue" that holds together the various categories and elements of the taxonomy and enable a coordinated, consistent, and progressive flow of thinking and knowledge generation during the qualitative data analysis process. As discussed in Section 2.2, stepping stone principles of the taxonomy are: a flexible organization of information—to be used as a guide to qualitative data analysis or as distinct purposeful taxonomies with specific outcomes; multi-disciplinary teams that need to actively contribute, share, understand, and act upon the same feedback items; maintain overview, control, and critical assessment during the process with a horizontal and vertical reach; and move from the abstract to the concrete revealing highly linked layers in different depths of realization.

3.1. Experiential Learning

Henceforth, EUREKATAX should facilitate an active (or proactive) and not reactive (more traditional) learning experience during the analysis of empirical information. In the latter case, the learners (or team member, participant) follow a more "linear" learning process, responding solely to specific stimuli given by their instructor (or experts, people that are in charge, of an analysis, training, class, etc.), and having a dependent reactive role that limits their learning to the boundaries of the instructor. In the former case, learners engage in a collaborative mode of learning developing abilities like to scan quickly and analyze data to produce answers to potential questions or to formulate questions answerable from data, abilities to test data against criteria of reliability and validity, to formulate goals and assess current level of performance, etc. [45]. Therefore, they employ more experience-based approaches to learning as "the process whereby knowledge is created through the transformation of experience". [40]. For Kolb, experiential learning concerns a continuous internal cognitive process which involves the acquisition of abstract concepts generated by the new experiences of the individual and which can be applied flexibly in new scenarios, contexts, or situations. This learning theory could be represented as a four-stage cycle of the learning process with strong links amongst each other, differentiating from other cognitive or behavioral learning theories that overlook the subjective factor experience following more serial actions to knowledge extraction through acquisition, manipulation, and recall of functional operations regarding units of information or abstract symbols. The four stages of Kolb's model are: (a) the concrete experience (doing—having an experience, e.g., dealing with a specific challenge like evaluating the quality of a feedback item that might impact the implementation of the system), (b) reflective observation (reviewing—reflecting on the experience, e.g., answering related questions like who mentioned this feedback item? At which stage in the interaction process was mentioned? How was it experienced? What was the outcome?), (c) abstract conceptualization (concluding—learning from the experience, e.g., why is this happened in a certain way? If it refers to a usability issue, what did not work so well and why? How could this happen differently? What and how severe are the implications on the system?), and (d) active experimentation (planning—trying out what you have learned, e.g., how could similar situations be approached in the future? What could change? What could be done differently?). Each stage of the model builds upon the outcome of the previous one, and an individual could use each one separately, entering the cycle solely for one phase, but he will be able to receive the full benefit once he engages and executes all the stages as an effective learning procedure [46]. In this respect, an individual must be willing to be actively involved in the experience and be able to reflect on it, to use analytical skills to conceptualize the experience and to possess problem solving and decision-making skills so take advantage of the new ideas gained from the experience [47]. Central point to this perspective is that learning is conceived as a process that requires the resolution of conflicts between two or more dialectically opposing modes of dealing and adapting to the world, and not in terms of outcomes [40]. Such an approach is considered vital for the qualitative data analysis which could be regarded as a continuous, "structured", and expansive learning process to be effective as we will see later, that is based on knowledge and experiences "the foundation of, and the stimulus for, learning" [48], effective problem solving and decision making. Apart from the hard facts and supportive calculations, the active participation of the team members in this process with their different experiences, prior knowledge, senses, and feelings are deemed necessary for the more accurate transformation and interpretation of a feedback item. Such a mixed reflection will be able to assign more holistic and insightful semantic meaning to data under the current circumstances and context of use.

3.2. Learning by Expanding

On the other hand, as participants share and build-upon earlier experiences at the same time actively construct their own during the qualitative data analysis process, and while exchanging over a specific feedback item. These arising experiences, perspectives, feelings, thoughts, etc. In the cross boarders of the past and the future are blended with the object, goals, and history of the activity of qualitative data analysis at hand which "is not self-evident; it is typically at risk or in crisis, ambiguous, fragmented, and contested. The object is rediscovered as a result of historical and empirical work of data collection and analysis with the help of conceptual models by the participants, supported by the researcher-interventionist", as specified in Engeström's third generation of Activity Theory: Learning by Expanding [36] (p.xxxii). The main principles of expansive learning adopted in EUREKATAX lie on a: (i) horizontal movement towards learning while team members are collaborating during data analysis through three basic interacting types—coordination, cooperation, and reflective communication, enabling them to capture the dynamics of their teamwork in processes of problem solving and learning [49]. Together with a vertical movement which concerns the emerged negotiations of the participants for a feedback item given on contradicting motives, backgrounds and experiences for bringing change and development, building shared knowledge, concepts and meaning that could not be predicted or formulated in advance and outside this setting ("friction"); and (ii) performing actions (i.e., dialectics) ascending from the abstract to the concrete [50]—e.g., moving from fuzzy or multi-purpose feedback statements to single actionable items) depicting developmentally valuable learning as qualitatively modifying a challenging situation to discover and model its initial root-causes.

The main aim is to trace and reproduce theoretically the logic of its growth, of its past formation through the emergence and resolution of its inner contradictions [51]. An activity of data analysis is a dynamic process which entails various transformation phases before reaching a tangible learning outcome. It requires the collaboration of many subjects (team members) to exchange and debate over a common medium (i.e., EUREKATAX) for an object (i.e., feedback item). Inevitably, the various motives, angles of consideration and interpretations create a scenery of expansive learning that evolves through the zone of proximal development of the activity. According to Vygotsky's famous definition, the zone of proximal development (adjustments of the definition to the needs of this paper are presented in italic) "it is the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under *expert* guidance or in collaboration with more *skillful* peers" [42] (p. 86). The meaning that is assigned to a single datum collected during the usability testing sessions is the product of this contradiction that is not (and could not), be recognized at earlier stages or if produced by a single person. The big benefit, apart from the obvious personal development of the subjects, is that the initial seed of information has been re-conceptualized acquiring rich interpretations and becoming a purposeful action item. When the object of this activity, which has been refined with a variety of insights elicited from the subjects' backgrounds, expertise and points of views (e.g., design, development, architectural, scoping, contextual), embrace a radically wider horizon of possibilities than in a previous mode during the activity, then this expansive transformation could be considered as complete [52,53].

3.3. Metacognition

The theoretical framework that encapsulates the hierarchical formulation of our taxonomy would be problematic without the consideration of higher order meta-functionalities that would oversee, regulate, and appraise the qualitative data analysis process. Such meta-functionalities or activities include phases like planning, monitoring, evaluation and self-reflection, basic ingredients of the metacognitive strategies (i.e., sequential processes used to control cognitive activities and to ensure that a cognitive goal has been met [54]), that individuals exercise in their daily life enabling them to become successful learners. Metacognition is a form of cognition that refers to higher order thinking or simply "thinking about thinking", as it has been acknowledged from the area of Educational Psychology, emphasizing on the executive processes that orchestrate and actively control (or regulate) the cognitive procedure and cognitive processes [55,56]. There are many definitions that have been assigned to metacognition, but the first researcher that introduced and used the term was Flavell (1979) [41], referring to the awareness of learning and thinking; the knowledge someone has for his own cognitive processes, knowledge, and thinking; the ability to organize, monitor, adapt, and reflect on the thinking process, tasks, and cognitive strategies created for addressing a problem, handling a situation, or accomplishing a goal. Especially, for goals related to qualitative data analysis that entail information gathered from natural (highly contextual and uncertain) rather than controlled experimental settings (containing vague and unstructured input that is not easily reduced to numbers, such as ideas, statements, opinions and behaviors), there is an increased demand to maintain the control throughout the process and develop a self-awareness for recognizing more efficiently the cognitive strategies or methods that apply in each situation and phase of the analysis.

According to Flavell [41] metacognitive awareness consists of the metacognitive knowledge and metacognitive experiences or regulation. *Metacognitive knowledge* is "that segment of your stored knowledge that has to do with people as cognitive creatures and with their diverse cognitive tasks, goals, actions, and experiences" (p. 908), and is broken down into three categories: (a) knowledge of person variables (including knowledge of how someone learn and process information, individual knowledge and judgement of one's own learning abilities and the internal and external factors that might influence the success of the learning process); (b) task variables (the purpose, nature, and type of processing as well as the demand that is required for the execution of a specific learning task); and (c) strategy variables (knowledge about developed cognitive strategies and conditional knowledge as to which situations and how would be more appropriate to use them) [54,57]. *Metacognitive experience* includes the use of metacognitive strategies or regulation [56] and "are any conscious cognitive or affective experiences that accompany and pertain to any intellectual enterprise" ([41], p. 906). It mainly refers to the understanding and the feelings that occur during a thinking process or with respect to a cognitive goal, helping someone to evaluate the progress and expectations, to connect new information to old or to generate new knowledge and goals. Metacognitive experience affects metacognitive knowledge and vice versa, e.g., some metacognitive experiences are best described as metacognitive knowledge items (in this case the quality control or assessment happens as a conscious process, for example someone engages in a really frustrating situation trying to assign meaning to a complex feedback item, and then suddenly she recalls a similar challenge that she tackled successfully in the past regarding another feedback item, so she acts accordingly); or metacognitive experience might add, delete, or revise metacognitive knowledge. "Thus, metacognitive knowledge and metacognitive experiences form partially overlapping sets: Some experiences have such knowledge as their content and some do not; some knowledge may become conscious and comprise such experiences and some may never do so" ([41], p. 908). Finally, cognitive awareness could be distinguished into three types when referring to knowledge [58]: (a) declarative knowledge (include factual information and strategies that someone knows as well as knowledge about oneself as a learner and which factors can influence one's performance [59]), (b) procedural knowledge (includes knowledge about doing things; how can someone perform a task and which strategies are more effective to use for undertaking specific procedural steps of a task. It is expected that individuals with high degree of procedural knowledge can perform tasks more automatically (e.g., resolving certain types of problems) using a variety of strategies [60], and (c) conditional knowledge (refers to the knowledge of when and why to use a procedure or a cognitive strategy (or declarative and procedural knowledge [61]), employing skills and allocating resources for a more effective execution. For this decision, information for comparing two or more strategies is required, which strategy is better than another for a given situation and why, or under what circumstances a strategy might work or fail.)

Henceforth, it is quite apparent from the short analysis above, that a theoretical framework that would capture and utilize the knowledge as transformed experiences of participants, analyze and identify expansive learning outcomes for a feedback item from their blended backgrounds and expertise, and at the same time would maintain a meta-understanding through an active monitoring, action, and reflection over the whole process of qualitative data analysis would be considered vital for the proposed taxonomy.

4. EUREKATAX Description

EUREKATAX provides a comprehensive framework for analyzing qualitative data. The main aim is to provide a clear description of the hierarchies, attributes and relationships that frame the qualitative data analysis process during a system's evaluation phase in a structured, nonredundant, and nonconflicting way. Central points of reference for the proposed taxonomy are (a) the *tasks* (providing the minimum block of contextual information for a user and his interactions) that a user performs during a usability testing, which can be part of other tasks (receiving input or giving output) and might be directly related to an activity and a more generic process, or a part it; and (b) the *feedback items*, which represent the main source of information extracted as a reaction to the observed tasks (give participants the ability to witness the results (or outputs) of their tasks (or inputs)), and maintained throughout the lifecycle of the data analysis process until they have been transformed into action items. Those two concepts are fundamental qualities of the proposed taxonomy; every analysis and discussion is taking place with and for them as a combinatorial unit of evidence. The taxonomy that is designed around those two factors supports different phases in qualitative data analysis process ensuring the insightful production of actionable items based on the shared visions, objectives and priorities of a project team. This is achieved by the gradual horizontal and vertical refinement of the feedback items across four different depths of realization, transforming them from abstract concepts to specific units of information.

More specifically, Figure 1, depicts an overview of the taxonomy along with the influential theoretical dimensions discussed in Section 3. As we can observe, the idea behind EUREKATAX is to appreciate the qualitative data analysis as an iterative learning process extended across various cycles of development and learning outcomes, the so called *zones of proximal development* or cycles of expansive learning [42]. These cycles are monitored by a metacognitive awareness that coordinates and controls the qualitative analysis process through a goal-directed metacognitive regulation (metacognitive strategies for controlling cognitive activities that relate to e.g., the way that a feedback item will be analyzed or optimized—the specific strategies someone employs to improve the learning outcome regarding a target) and knowledge (have the knowledge of how someone learns), including (i) knowledge of person variables e.g., knowing that the team works more effectively in a design thinking room, (ii) task variables e.g., knowing that the data synthesis phase entails more complexity than the allocation of the feedback items into clusters, and (iii) strategy variables e.g., always iterate and cross-verify the clusters once generated), applied during the qualitative data analysis process. This process may include phases like (a) planning and organizing, whereby a team could skim over the collected data and decide on the goals of the qualitative data analysis at hand, how to approach and analyze the data set, how to divide big jobs into smaller manageable tasks, etc.; (b) monitoring, might involve checking progress with respect to the available timeframe, error discovery and handling, validation of the progress itself, e.g., if the specific cognitive strategies and activities provide the maximum expected benefit; (c) evaluating, includes assessment (in cases measurable) of the tasks executed, the outcome and effectiveness of the cognitive strategies followed for a specific job, initial reflection of what to keep and what to reject for the future and similar activities; and (d) self-reflecting, concerns the appraisal of team members and their actions during the data analysis process, skills acquisition through exchanging and practicing [62], things to consider or to avoid while working together for setting up an optimal experienced-based expansive learning environment.

The first level of a taxonomy usually describes its main *categories*, as the higher level of abstraction of the predetermined organization of information. Moving a team, during the data analysis process, horizontally—across the four different conceptual depths of the purposeful categories, obtains a holistic understanding and knowledge of the qualitative data analysis process while at the same time acquires or improves various related skills like prioritization and formulation of action plans for analyzing certain types of feedback items. In EUREKATAX, the main goal-directed categories, namely discover, learn, monitor, and act, embody basic concepts (as interrelated layers), which define its object, inner structure, and boundaries, suggesting a methodological guidance and conceptions how one can proceed in order to grasp the object adequately [63], i.e., the exploration and transformation of feedback items into meaningful and insightful action items. The *discover* category (or layer) concerns the creation of an understanding about the end-users participated in the usability study and the data using synthesis, consolidation and clearing techniques of feedback items that in the end will produce possible clusters. This layer is composed of two different classifications of data referring to the customers and guided exploration sub-categories. The main purpose of the *learn* category is to assign meaning to the extracted feedback items (or clusters) by identifying a semantic viewpoint of them, their relationships with other entities (e.g., interaction designs or use cases), and evaluate their type (e.g., usability issue or just a comment), significance and impact on the system under examination. This category consists of the data empathy and insightful recommendations sub-categories. The *act* category refers to the decisions and actions that someone takes regarding the feedback items based on his so far analysis and accumulated knowledge, the proposed solutions and other contextual information like priorities, strategic directions, resource availability, effort, capacity, skills, etc. This category is composed of three sub-categories, the informed decisions, wrap-up, and solutions area. Finally, the *monitor* category facilitates the continuous tracking, smart data inspection over the analyzed and semantically enriched data through meaningful data visualizations, and quick reporting. It contains the smart overview sub-category.

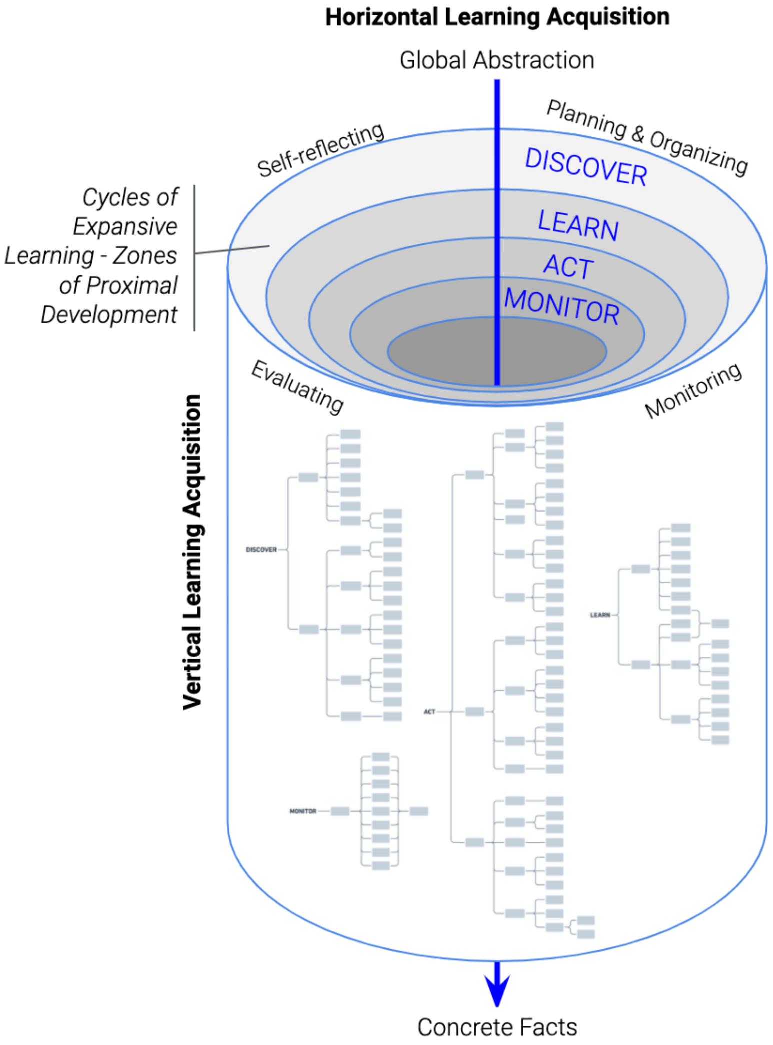


Figure 1. Engineering Usability Research Empirical Knowledge and Artifacts Taxonomy (EUREKATAX) Theoretical Model.

The cyclic approach mirrors an iterative modular learning method that starts with the construction of basic or simple pieces of information and ends up as a more sophisticated amalgamation of information chunks regarding a specific topic. A team has the chance to enter any cycle of the taxonomy depending on the needs at hand, but it will fully benefit from a comprehensive learning outcome only if it explores the taxonomy at its full extent. The stages of EUREKATAX are mutually supportive and each one feeds the next with information, following a logical sequence in the qualitative data analysis process. A simple example: When a team has acquired the necessary knowledge (created a schema) about the end-users and how well they fit to the role under investigation then it can move to the next level of the taxonomy for discovering clusters of feedback items, that can accordingly classify as important or not depending on how many end-users have refer to them. So, the team uses previous findings and material as prerequisite for building upon new knowledge, which it becomes at the same time the new basic knowledge for further exploration in the next cycle, and so on and so forth. In the opposite scenario, whereby the team would jump directly to the clusters definition without integrating prior learnings about the end-users first, it would still have a result but it would not know how valuable it is; hence, each cycle (or isolated classifications in one cycle) in EUREKATAX, cannot represent a learning procedure on its own, but rather a necessary steppingstone towards the maximization of understanding and *learning experience* (i.e., the more a team deepens in its structures the more solid learning outcomes it receives).

Vertical exploration reflects a perpendicular movement to each category for assigning different and more inclusive semantic meaning to a feedback item with respect to a task. Someone can locate a complementary potential and restrictions during operations which are generating situation-specific learning e.g., considering platform or contextual factors (like strategy directions or delivery constraints) that influence the resolution of a specific usability issue. The extracted knowledge follows a progressive transformational route combining the experiences and knowledge of the team members converting a feedback item from a global abstraction—might have an unstructured format and convey a fuzzy message—to a concrete fact with a specific meaning and associations. Thus, each depth can produce a semantic transformation of a feedback item leading to the next depth of realization or a standalone interpretation with a more limited scope. Subsequently, in EUREKATAX, the various categories are decomposed in different classifications, as follows: Each *category* is decomposed into *sub-categories* (different, standalone, purposeful classifications, with specific input and output), and in turn into *operations* (describing someone's functions towards realizing the various sub-categories in a specific context, situation or location), which consequently may consist of several *properties* (specific characteristics or qualities). The particular hierarchical decomposition or notation could be loosely coupled with the one expressed by Engeström's *activity tfheory* whereby respectively: The 'categories' may relate to motivational 'activities', the 'sub-categories' to conscious with purpose 'actions', the 'operations' to the 'conditions' that declare the actions, and the 'properties' to the extracted 'knowledge units' from this process [36,64].

Hereafter, we provide a brief description of the various classifications of the taxonomy (constituting the main body of the theoretical model in Figure 1), elaborating on its four categories and the respective input and output information that facilitates a solid connection among them and its elements (the reader may also refer in [65] for a more detailed description of the categories and the potential realized through its implementation as a software application/tool).

4.1. Discover Category

The *discover* category (depth_1—Figure 2) is composed of two sub-categories as mentioned earlier. Those refer to "customers" and "guided exploration" classifications. A team should have a sufficient knowledge about the customers and end-users that participated in a particular usability study. An end-user represents a person that encapsulates the description and characteristics of a (business) role, or in more detail of a 'persona' as this has been defined by the team (a role might consist of more than one persona, depending on the viewpoints and specifics of a solution, e.g., the role of project manager might include the personas of project managers that are related to (a) an application for project planning generation and allocation of resources, and (b) to an application that monitors the execution of the project plan as well as the collaboration with the consultants). The "customers" sub-category receives as input the profiles of the end-users containing details like organization name, country, end-user name, actual role, end-user alias, and role description, and provides output related to end-users insights and the role-fit (may be general background check—across the business scenario and tasks, or task-based—measuring the degree of goodness of the role for a specific task). This allocation pre-supposes the verification of the initial (expected) end-user profile, prepared by the team before the end-users recruitment for the usability studies, with the profiles of each end-user collected during the execution of the session (actual). The team then measures to what extent there is a match among the two and where, generating a percentage e.g., 75% fit in tasks 1–3 and 100% fit in tasks 4–6. As soon as there is a clear understanding about the end-users that have participated in the study then the team can move on to the qualitative analysis by utilizing the "guided exploration" sub-category. This classification receives as input the raw data of the validation script notes collected during the usability study and through various iterative operations produce optimized and semantically enriched clusters containing feedback expressed with active wording for enhancing the clarity of a fact or event. Such operations include tasks identification and description, and allocation of the respective feedback items to each one of them after the synthesis, consolidation, and cleaning process of data. The end-users weighted references are assigned on each feedback item based on their fit in the respective tasks identified earlier, and specific aggregation calculations are applied obtaining e.g., the total references per feedback item by the end-users. Furthermore, the success with assistance for each end-user on a scale from 0–4 is recorded, indicating the degree of external influence (e.g., tips or hints by the moderator of the usability test) an end-user had for accomplishing a task.

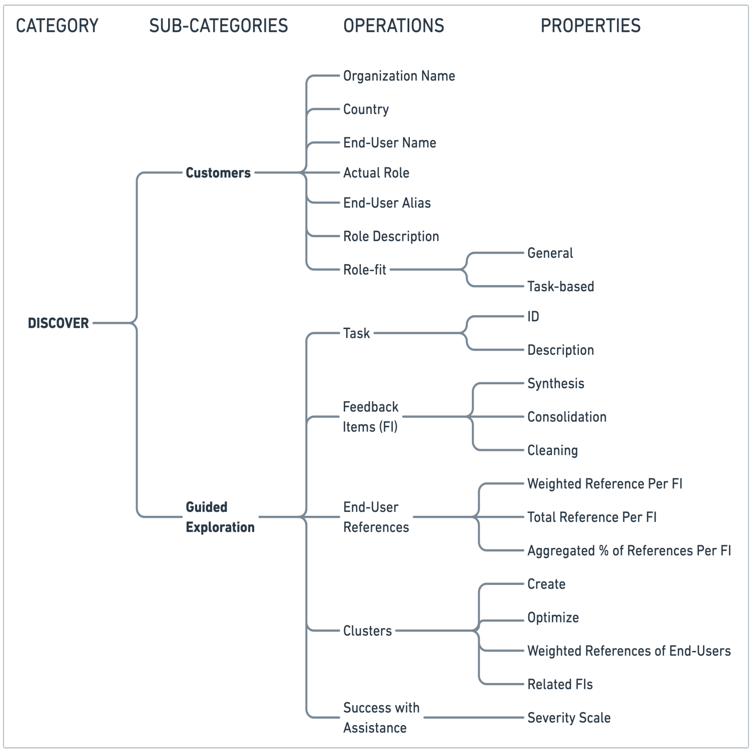


Figure 2. The discover category, sub-categories, operations, and properties.

4.2. Learn Category

The *learn* category (depth_2—Figure 3) also consists of two sub-categories: the "data empathy" and "insightful recommendations" classifications. Creating data empathy is the corner stone of the EUREKATAX taxonomy signifying the importance of facilitating the objective observation and unbiased interpretation of the data as well as the precise association of the various units of information in an attempt to reveal holistic and inclusive meanings of a feedback item (or cluster). "Data empathy" concentrates on a specific task each time and interaction requirements defined by the team prior to the usability study (each task is imported through the task_id from the "guided exploration" classification to the remaining classifications maintaining a semantic link, amongst other items, between them). It uses operations like cluster summary (optimize the content and insights of each cluster as those are dictated by the subsequent feedback items that is composed), semantic cluster name (the generation of a meaningful title conveying the essence of the clusters' content to be used for quick reference and input to other classifications for e.g., smart filtering, statistical analysis, and overview), association of a cluster with the related screen or design and use case defined before the usability test, an aggregated percentage of the weighted references made by the end-users, identification of the actual usability issue type (based on the Usability Problem Taxonomy [23]), and judgement of the given usability issue as positive, neutral, or negative.

The second sub-category facilitates a deeper understanding of the data under investigation allowing the identification of the (a) relevant importance for each cluster (to what extent a percentage of end-users' references for a feedback item need to be considered or not, e.g., a team might decide that given the specific sample size and user study specifics the 45% and above of the references should be regarded in the decision making process), (b) impact on the application (if a usability issue is severely influencing the functionality or perception of end-users for a specific product), and (c) priority (identify if a usability issue has a high, medium, low, or none priority to be addressed, e.g., in the next development sprint). In addition, recommendations how the team should proceed (e.g., if it needs to take an immediate action regarding a usability issue or not) could be calculated using the relevant importance and the impact on the application.

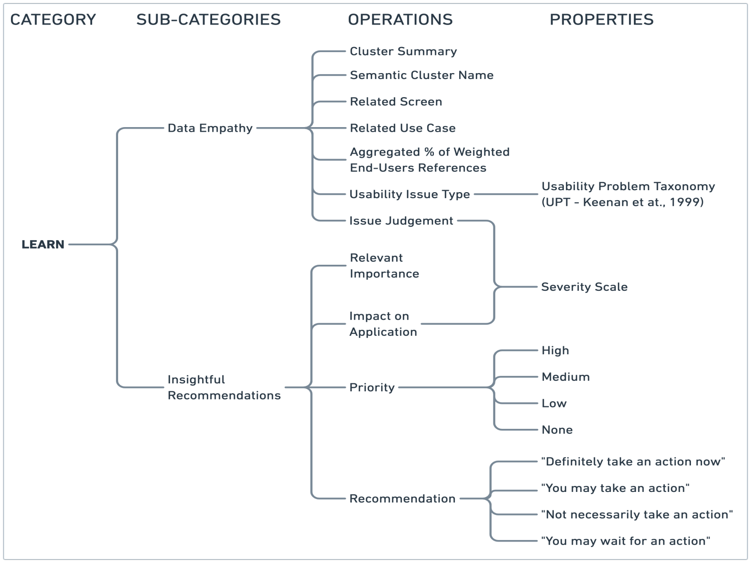


Figure 3. The learn category, sub-categories, operations, and properties.

4.3. Act Category

The *act* category (depth_3—Figure 4) entails three sub-categories, namely, "informed decisions", "wrap-up", and "solutions area". The "informed decisions" sub-category receives the insightful recommendations generated from the "learn" classification and yields possible high-level solutions and actions for the identified usability issues. The main operations that drive a team's decisions regarding the actions to be taken for each feedback item include the task_id, cluster summary, priority, and recommendation discussed earlier as well as possible solutions (discussion points and alternative suggestions that could be applied for tackling one or more usability issues), team decision (with properties like 'go', 'maybe' or 'no go' for a solution), and progress identification (as 'done', 'in progress' or 'not started'). The "wrap-up" classification handles all the information and analysis that takes place usually after the usability study's main tasks execution and includes operations like the analysis of post-questions (referring to impressions, improvement points or situation-specific comments), clustering of these feedback items (including optimization, allocation of weighted references of end-users and association with the related questions), weighted end-users' references per question and aggregation, and usability or UX test tools' responses (data collected by the use of any standardized supportive usability tool/questionnaire, mainly for cross-evaluation of the main tasks, like e.g., SUS [38] or UEQ [66]). The "solutions area" sub-category enables a deep dive into the alternative approaches for solving the discovered usability issues. The team has the chance to synergistically work towards detailing e.g., high priority usability issues, that have been assigned with a 'go' and are 'in progress', for identifying viable solutions. This sub-category contains operations like task_id, cluster summary, usability issues per task (with properties 'go' and 'in progress'), possible solutions (a list of alternative approaches discussed by the team for solving the usability issues), solutions effectiveness (indicating which usability issues are influenced (tackled) by which solutions and to what extent they are solved, e.g., one might be solved by 25% or another by 75% by one solution (i.e., partially to fully solved), and also the coverage a solution has across the usability issues, e.g., solves four of them (fully or partially) with a beneficial impact of 45%, once its total contribution is calculated). Lastly, an important operation is the viability which is realized through properties like estimated effort (how much time is needed for a specific solution to be implemented based on existing resources, expertise, know-how, difficulty, etc.), calculated risk (issues that might arise during the process and might hinder or delay the expected implementation, e.g., lack of existing guidelines, external collaborations), and likelihood of timely completion (a reconfigurable smart viability matrix allocates a solution, based on the assigned effort and risk, to a viability quadrant scale (i.e., 1 = high to 4 = low) indicating how probable it is to be successful).

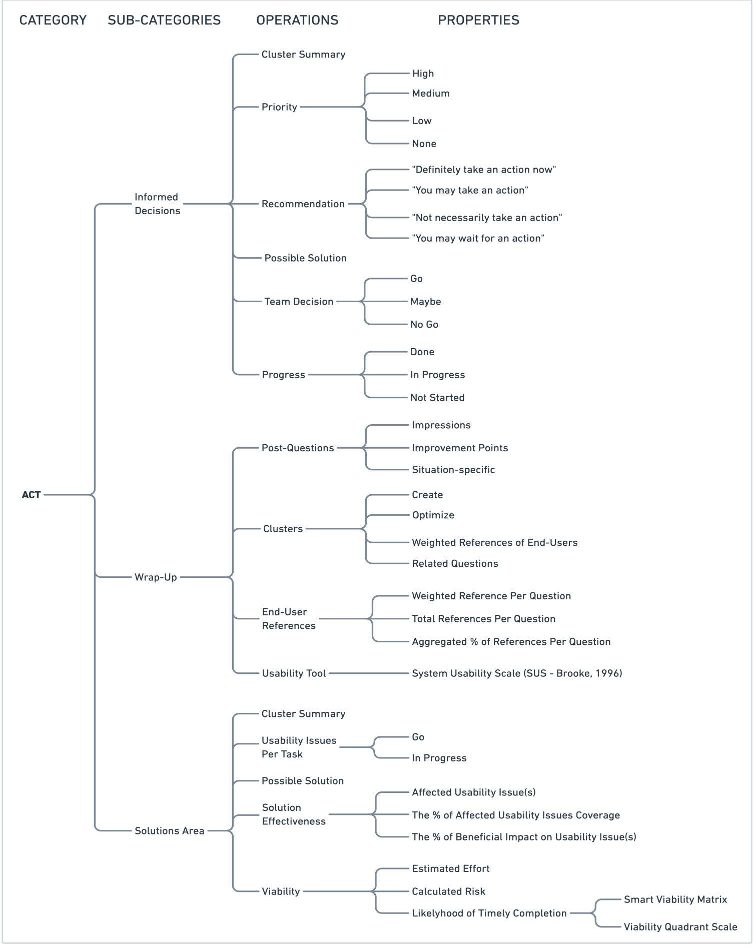


Figure 4. The act category, sub-categories, operations, and properties.

4.4. Monitor Category

The *monitor* category (depth_4—see Figure 5) consists of the "smart overview" sub-category which facilitates the continuous monitoring and exploration of the information that has been extracted from the previous classifications. It generates visually enhanced digital cards based on operations like persona creation, validation study, feedback overview, issues judegment, tasks assistance, feedback items' clusters, usability issue types, etc. Most of these operations receive input from the previous classifications and despite the presentation of a quick overview of the empirical research outcome, guides project teams to an informed drill-down on the reformulated semantic data (by applying e.g., filters) for prioritizing their actions and decisions. Furthermore, the visual appearance of data facilitates the quick transition from data analysis to documentation and fast reporting (e.g., by simply cropping and pasting the visual cards in perspective), a well-known time-consuming challenge in user research.

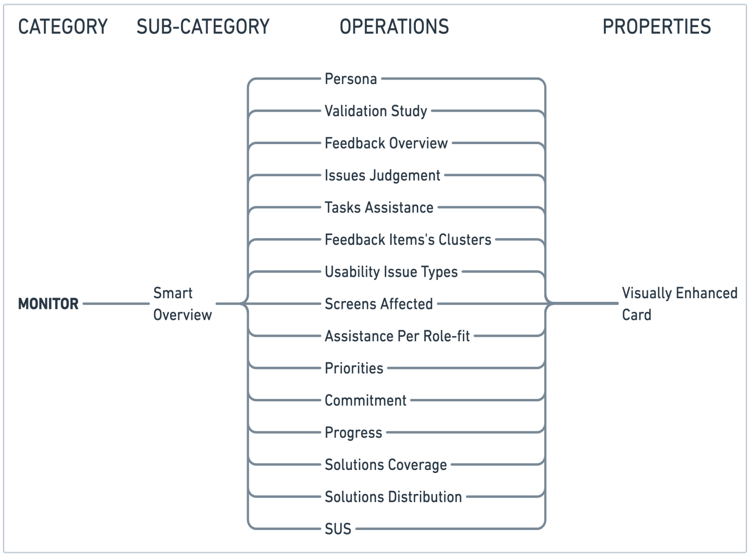


Figure 5. The monitor category, sub-category, operations, and properties.

5. Evaluation Method

Assessing the usefulness of a taxonomy in everyday application is a long and iterative process that involves various situation-specific perspectives and methods. Initially, our main concern was to validate EUREKATAX in the business sector with experts, regarding its more theoretical aspects, as well as with end-users testing its adoption in a real usability test data analysis scenario. More specifically, we designed a series of user studies executed into two distinctive phases: (a) In Phase 1, we were interested at verifying the theoretical considerations, as well as the design and development of the taxonomy's structure and elements, as a coherent model of information organization; and (b) Phase 2 emphasized on the practical implications, its usefulness, and acceptability once applied by teams in a real-life scenario; also, by taking advantage and comparing to their previous knowledge, methods, and experience. The user studies of the two evaluation phases are detailed below.

5.1. Phase 1—Evaluating the Structure and Elements of the Taxonomy

In the first phase we created a study by emphasizing on the classification and formulation of the terms under the four main categories, verifying to what extent they express a consistent methodology for the qualitative data analysis. Focal points included whether the: (a) generic categories were able to be progressively refined to obtain more concrete operators that would in turn generate a new subsequent organization of properties formulating consecutive taxonomic levels; (b) categories were clearly associated with the operators as permitted and the groups of properties avoiding any semantic misconceptions or restrictions of any nature (e.g., the operation "recommendations" is clearly associated with the "act" category, the "informed decisions" sub-category, and the related actions as properties; without hindering the use of more related actions or altering the understanding of what it is expressed); (c) input and outcomes of each category conveyed a unique but supplementary meaning across the various taxonomic levels generating an end-to-end understanding of the goal and the result (transformation or interpretation) of each activity (e.g., the generation of "clusters" as output from the "discover" category is on the one hand a distinctive outcome of the taxonomy from this classification that may not be confused with the nature of other outputs, and on the other hand it contributes to the accumulated knowledge produced by EUREKATAX since it might be used, e.g., as input to the "data empathy" category); (d) definitions were in adequate clarity expressing the exact purpose of their formation; and (e) there were no unnecessary overlaps or redundancies of the subsequent attributes avoiding any complications, confusions, or contradictions of the terms that could mislead or disrupt the process of the taxonomy's implementation.

5.1.1. Participants

We recruited a number of participants, using personal invitations, that had above the average knowledge of User Experience methods and tools and involved in the past in numerous usability tests and qualitative data analyses. A total of 20 experts (11 Female, 9 Male) participated in the study with average years of experience M = 7.83, SD = 5.91. Their actual roles were ranging from UX Designers, User Researchers, Usability Testing Experts to UX Managers and Consultants. All the users participated voluntarily and provided their consent that their interactions with the taxonomy's Web-forms would be recorded anonymously in the context of an experimental user research study.

5.1.2. Procedure

Among the various approaches (e.g., Delphi Card Sorting, Remote Card Sorting, Search Analysis—[37]), we decided to use the Closed Remote Card Sorting method, as it was best fitting to our purpose and research circumstances. We created an electronic environment composed of Web-forms (see Figure 6), each one outlining the different hierarchies of the EUREKATAX based on its four distinct categories described above. Participants could interact sequentially with the environment progressing from each classification to the next and providing their feedback. For each sub-category, operations, and properties we collected two types of participants' assessments regarding the validity of each classification. At first, users had to respond to what extent a classification of items fits together in terms of relevance (1 = low, 2 = medium, 3 = high fit), and second, they could provide in a free text format their additional comments, wishes or alternative views for each category. Given that the validation followed the unmoderated protocol we wanted to ensure that the various items would be understandable by the users who share different experiences and backgrounds. In this respect, we decided to use more descriptive phrases in cases that the meaning of the classification items was not self-explanatory, e.g., we re-phrased the operation "task" expressing it as "information about the tasks, that the study ran upon", or the sub-category "data empathy" described as "deep dive into your data and create links with existing user research artifacts". This was an iterative process step during the preparation phase of our taxonomy for validation which lasted for two weeks in total including parallel pilot testing of the taxonomy's semantics with targeted users, controlling this way the span of different roles and experiences. We ran the study remotely for two months (from 1 July 2018 until 1 September 2018), adhering to the ecological validity paradigm [67], and allowing the users to access the study on their preferred time and location, without also imposing any restrictions on the duration for providing an answer. To avoid any learnability effect during the process, each participant could execute only once the study which lasted approximately 15 min.

5.1.3. Hypotheses

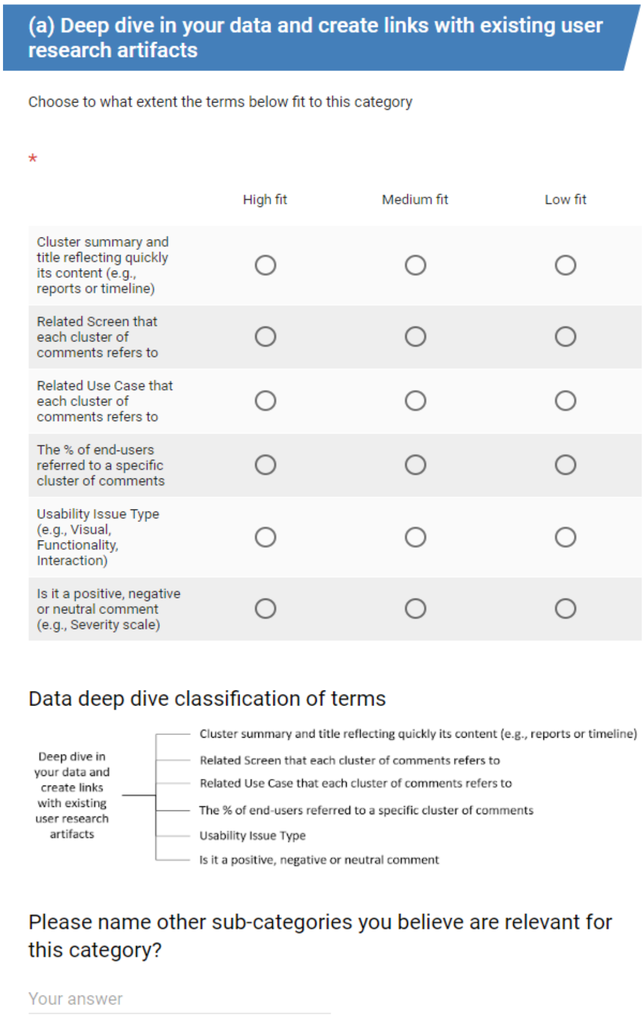
The following overarching hypotheses were formulated for the purpose of this evaluation phase:

Hypothesis 1 (H1). *There is a high consistency among the various classifications of the taxonomy.*

Hypothesis 2 (H2). *There is a significant preference of users towards the high/medium fit when compared to the low fit with regard to the relevance of the items to the various (sub-) categories.*

5.1.4. Analysis and Discussion of the Results

To evaluate the internal consistency of the EUREKATAX hierarchies (i.e., four main categories), a reliability analysis was carried out on the values comprising the eight consecutive sub-categories and 52 items of the taxonomy. Cronbach's alpha showed the classifications of terms to reach strong reliability, $\alpha = 0.945$. All items appeared to be worthy of retention, resulting in a decrease in the alpha if deleted. Such a result reveals the high internal consistency of the taxonomy and acceptance of our first hypothesis (H_1). For the sake of completeness, we have also calculated the reliability of each of the eight subsequent classifications of terms (see Table 1). From the results we can observe that even in the distinctive sub-categories of the taxonomy the internal consistency reached an acceptable reliability, with an exception the "customers" classification with α = 0.653. In this case, there were two items (item 2 and item 5) that by dismissing them could increase alpha close to 0.7 (i.e., α = 0.695), the minimum empirical value for acceptable validity. As such, removal of these items could be considered. On the other hand, for the "insightful recommendations", "informed decisions", and "smart overview" classifications, removing items 4, 6, and 5 could increase the reliability of the scales even more to α = 0.773, α = 0.796 and α = 0.888 respectively. As such, possible elimination of these items will be further investigated.



(a) Deep dive in your data and create links with existing user research artifacts

Choose to what extent the terms below fit to this category

*

	High fit	Medium fit	Low fit
Cluster summary and title reflecting quickly its content (e.g., reports or timeline)	○	○	○
Related Screen that each cluster of comments refers to	○	○	○
Related Use Case that each cluster of comments refers to	○	○	○
The % of end-users referred to a specific cluster of comments	○	○	○
Usability Issue Type (e.g., Visual, Functionality, Interaction)	○	○	○
Is it a positive, negative or neutral comment (e.g., Severity scale)	○	○	○

Data deep dive classification of terms

Deep dive in your data and create links with existing user research artifacts

- Cluster summary and title reflecting quickly its content (e.g., reports or timeline)
- Related Screen that each cluster of comments refers to
- Related Use Case that each cluster of comments refers to
- The % of end-users referred to a specific cluster of comments
- Usability Issue Type
- Is it a positive, negative or neutral comment

Please name other sub-categories you believe are relevant for this category?

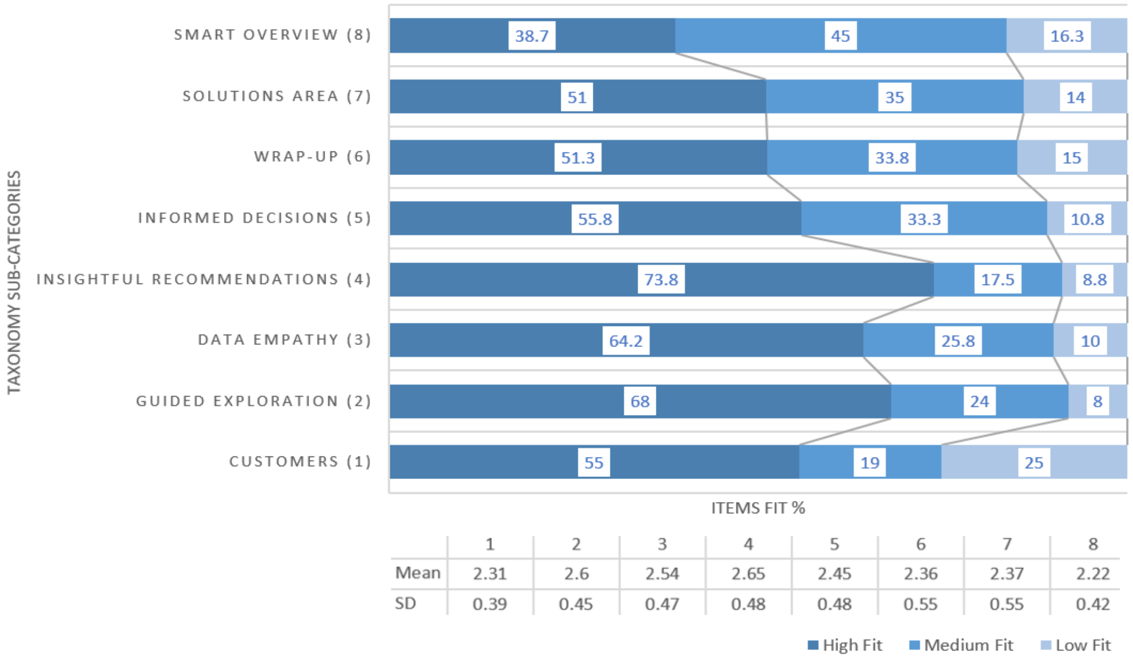
Your answer

Figure 6. The on-line environment consisting of the various hierarchies of EUREKATAX.

Table 1. Reliability alpha values of EUREKATAX sub-categories.

Sub-Category Name	No. of Items	Cronbach's α
Customers	7	**0.653**
Guided exploration	5	0.751
Data empathy	6	0.807
Insightful recommendations	4	0.764
Informed decisions	6	0.793
Wrap-up	4	0.767
Solutions area	5	0.811
Smart overview	15	0.885

Descriptive statistics showed a strong tendency of participants towards the "high fit" preference of items in relation to the various classifications they belonged to (M = 2.44, SD = 0.40). More specifically, multiple response analysis revealed that 53.6% of the terms were "high fit" in the various subsequent classifications of the taxonomy, 32% "medium fit", and 14.4% "low fit". This result could be interpreted as a total of 85.6% of "high and medium fit" of terms in the various hierarchies leading to confirm also our second hypothesis (H_2), since it showed a strong preference of the taxonomy's items in their respective categories by the participants. Figure 7, depicts the distribution of fit across the various sub-categories of the taxonomy. Given that our data were not normally distributed the non-parametric Friedman repeated measures test was carried out to compare the total perceived fit for the various classifications; whether there was an equal distribution of the preferred fit on the items in the eight classifications of the taxonomy (H_0). There was found to be a significant difference between the classifications, χ^2 (7) = 29.245, p = 0.000, i.e., significant variance on the preference of the experts regarding the fit of the terms, resulting in rejecting the H_0. As we can observe, the two sub-categories that scored higher were the "insightful recommendations" and "guided exploration" with M = 2.65, SD = 0.48 and M = 2.6, SD = 0.45 respectively, while those that scored lower were the "customers" (M = 2.31, SD = 0.39) and "smart overview" (M = 2.22, 0.42).



	1	2	3	4	5	6	7	8
Mean	2.31	2.6	2.54	2.65	2.45	2.36	2.37	2.22
SD	0.39	0.45	0.47	0.48	0.48	0.55	0.55	0.42

Figure 7. Distribution of items fit in the taxonomy (%).

Next, our main concern was also to investigate in more detail the distribution of the items' perceived "low fit", belonging in the lower scored sub-categories, by the experts across the various classifications of the taxonomy. Accordingly, we would be able to identify any discrepancies and locate specific "weak" items liable to be revisited in the future in an attempt to increase the consistency and quality of the usability tests data analysis process. Exploring Figure 6, descriptive statistics and frequency distributions, a targeted series of post hoc analyses with Wilcoxon signed-rank tests were conducted at the classifications of the four categories of the taxonomy that scored relatively high in "low fit" items, with a variable Bonferroni correction applied each time, depending the number of items in the respective sub-category (e.g., the "Guided Exploration" has five operations). More specifically, we analyzed those items that were below the mean of each sub-category by comparing them with the items scored high in the same sub-category. From the various combinations, the results revealed two items in the "customers" and one item in the "smart overview" classifications as the most problematic with a statistical significance difference. At first, for the "discover" category, the "customers" classification (M = 2.31, SD = 0.39) was investigated setting the significance level at $p < 0.0071$. The median (IQR) perceived fit for the items 3 (M = 1.30, SD = 0.57) and 5 (M = 1.90, SD = 0.91) when compared with item 4 (M = 2.80, SD = 0.52) were 1.00 (from 1.00 to 1.75) and 2.00 (from 1.00 to 3.00), respectively, with significance difference between them (for item 3 Z = −3.827, p = 0.000, while for item 5 Z = −2.973, p = 0.003). Items 3 and 5 refer to the operations "end-user name" and "end-user alias" respectively, showing a negative predisposition from the user experience experts regarding collecting and documenting the names or use any alias of their subjects participated in their usability studies, as such data would not add any added value to their empirical data analysis process. This is a really interesting finding, since it also complies with the EU General Data Protection Regulation (GDPR) put into practice end of May 2018, which relates to the data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA) [68]. On the other hand, for the "smart overview" classification (M = 2.22, SD = 0.42) which belonged to the "monitor" category, the significance level was set at $p < 0.003$. The median perceived fit for the item 15 (M = 1.85, SD = 0.74) when compared with the item 1 (M = 2.55, SD = 0.60) was 2.00 (from 1.00 to 2.00) with statistical significance difference (Z = −3.071, p = 0.002). This item refers to the "visualize the results of the standardized usability tool used" operation. A possible interpretation in this case would be that experts find it unnecessary to include tight together in the specific hierarchy the results of a third-party tool or process, since it cannot semantically co-exist with the other items in the same classification (or influence their actions), nor do its results derive from the analysis and interactions driven solely from the methodological approach represented by the EUREKATAX.

Finally, as mentioned earlier, we requested also from experts to provide additional feedback on the various classifications after the completion of the first task (i.e., allocation of their preferred fit of items to each sub-category), in a free text format. Applying a frequency of themes analysis [69] for synthesizing the collected open-ended responses, we managed to extract useful supportive insights regarding general impressions, challenges, new items, or alternative dimensions of the existing items useful for the future of this work. In this respect, we classified the comments into three categories, i.e., positive, negative and suggestions. Hereafter, we outline the main assessment points considering how often they reappeared in each category across the participants. On the positive side, many of these comments were fully aligned with the scope and rationale that the EUREKATAX has been built upon, indicating that they were in favor of items that e.g., bind the collected feedback items with the use cases, designs, and predefined usability issue types, or that there is a distinction of the end-users that match fully the expected user profile of a usability study for applying a weighted importance on their responses and reactions during a session. Other, more negative comments, emphasized on the lengthy size of the taxonomy, with an expert to mention that "we would need some time to familiarize ourselves with it", while another pinpointed a probable semantic confusion of the word "fit" and "I would like to see them (i.e., items) there (i.e., in a classification)"—he mentioned that "generally, I want to see less". Regarding the recommendations for improving the taxonomy, for the "customers" classification, it seems that participants were interested in more contextual information around the end-users that may interact with their solution, expressed through items like: their age, gender, and personality type; industry, department, and city that they are located; size of their organization (e.g., number of employees); hierarchical dependencies with other co-workers in the organizational plan; overview of responsibilities, number of years in the specific role and other peers in the same role; and their level of knowledge in the task, field, and business domain. With respect to, the "guided exploration" classification the experts commented that they were expecting to see the average time per role to fulfill the tasks as well as a comparison between power versus beginner end-users. They further indicated that they would prefer to have highlighted the part of the interaction design that needs improvement or commented in the "data empathy" sub-category. In the "informed decisions" hierarchy participants recommended to include items that clearly refer to the type of a solution discussed for a usability issue (e.g., bug fixing, new development, new guideline, etc.), and also the associated level of solution development (e.g., idea concept, UX design concept, proof of concept, prototype, released, etc.) This classification should also embrace the dependent tasks of the ones to which a solution is applied for improving the related usability issues, so as to avoid any additional indirect problems caused by this dependency. A final suggestion by the experts points out the importance of including an item in the "Smart Overview" sub-category for maintaining a bird's view across usability studies and controlling the improvement of usability issues over time based on the proposed solutions and action plans of the subsequent usability sessions.

5.2. Phase 2—Verifying the Usability, Acceptability and Usefulness of the Taxonomy in a Real-Life Scenario

From the outcomes of the first evaluation phase we could initially imply that EUREKATAX presents an internal consistency regarding its methodological phases and the modules of its application, as perceived by the UX experts. However, a necessary additional step in the process of constructing a qualitative taxonomy is to validate its usefulness, usability, and acceptability by the teams; when put into practice and tested in real-life situations. Accordingly, we would expect to facilitate a simple and guided transition from theory to practice through the underlined structured, yet flexible, iterative process to qualitative data analysis. More specifically, it would be essential to ensure that: (a) the information is organized for the user in the given context of use without hindering the analysis of qualitative data captured during the usability studies; (b) it justifies the effort invested and potential costs when implemented; (c) it reassures for the benefits of its existence and use; (d) it helps compensate on possible conflicts during exploration; and (e) it facilitates the objective and transparent extraction of findings and meanings.

5.2.1. Participants

We distributed personal invitations and advertised the user study in newsletters and mailing lists to recruit a number of professionals that would fit the expected sample characteristics. Aiming at increasing our study's internal validity, we targeted users that had already been taught and exercised knowledge found in common UX curricula, like usability testing preparation, scheduling and execution, moderation, note-taking, data synthesis and consolidation, qualitative data analysis fundamentals, etc.; they have been part of a product/project team; and they had already participated at least once in a formative usability testing or a user study of similar nature (e.g., remote validation). A total of 60 professionals (42 Female, 18 Male) eventually confirmed their participation in the study having different educational backgrounds, business goals and knowledge, and data analysis experience. After an initial screening we managed to classify the participants into three overarching categories based on their business role: 61.7% were UX professionals (e.g., user interaction designers/user researchers), 21.7% were product owners/managers and business experts, and 16.7% were architects/developers. This grouping would guide their allocation into teams, for maintaining a symmetrical distribution of roles across, e.g., avoiding one team having only product owners or UX professionals. Thus, we would be able to ensure a realistic composition of business groups of people blending different roles, responsibilities and specializations, and balancing the professional knowledge, rationale, viewpoints, and interpretations employed during the data analysis process. All the professionals participated voluntarily and provided their consent that all the data during the various activities of the workshops would be recorded anonymously in the context of an experimental user research study.

5.2.2. Procedure and Hypotheses

For the second evaluation phase, we decided to run a number of interactive sessions with product teams in the form of on-site workshops. The main concern was to create a realistic environment where participants would have the chance to put EUREKATAX into use. To facilitate the process we developed an Excel functional prototype tool (see Figure 8), composed of five distinctive modules that allow a smooth and consistent application of the taxonomy's theoretical perspectives, hierarchies and classification of items. The tool offered functionalities like: (i) create understanding of customers (or end-users), data empathy and clustering through guided exploration (module 1—the discover category); (ii) assign meaning and get insightful recommendations (module 2—the learn category); (iii) meet the issues and expand on challenges by making informed decisions and inclusive wrap up (module 3—the act category); (iv) deep dive into the solutions and spot the coverage and viability (*module 4—the act category*); and (v) keep track continuously and ease reporting with a smart overview (*module 5—the monitor category*). The teams had the chance to apply EUREKATAX through the prototype deep diving into the specifics of each data analysis phase. In fact, we simulated a real-life business scenario of data analysis, setting up the starting scene at the stage where the teams return to their base from their customer visits—after they have executed a number of formative usability tests with their end-users, and had as a next step to analyze the collected feedback.

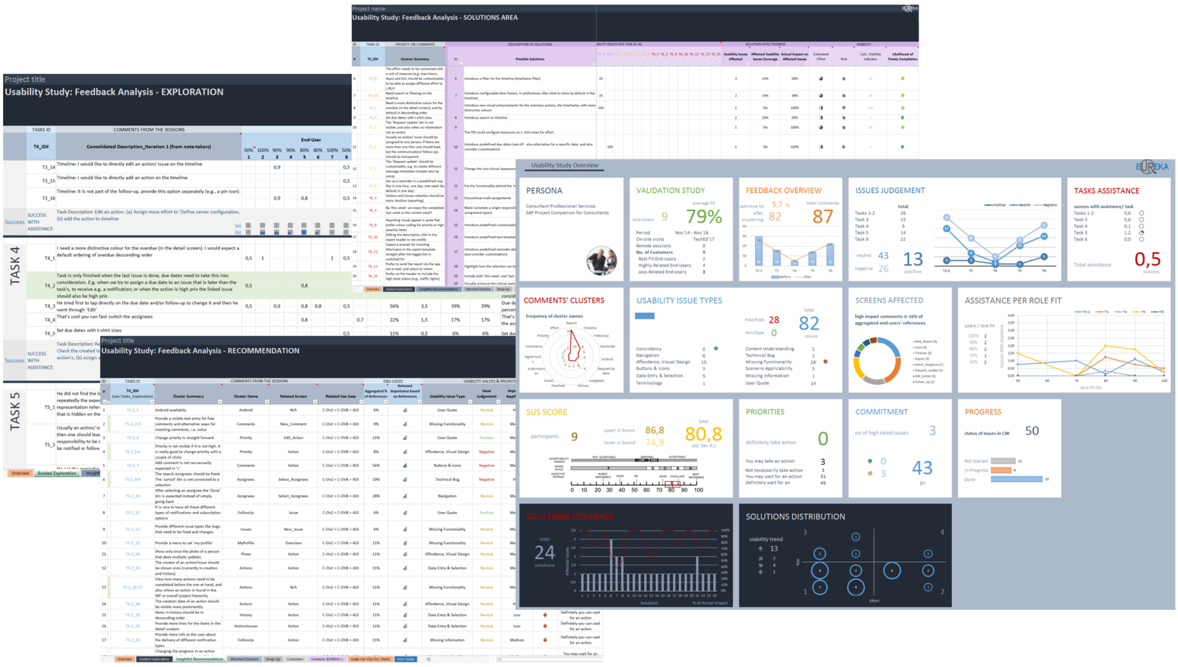
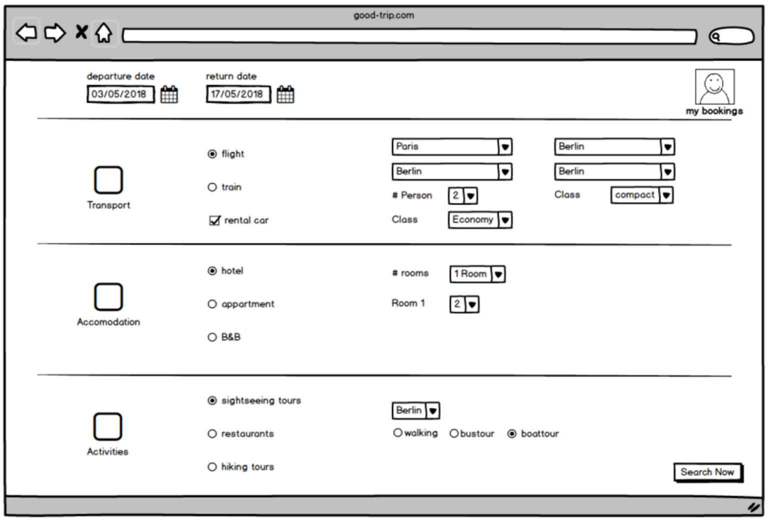


Figure 8. Snapshots of the EUREKATAX functional prototype tool.

For the scope of this study, and in order to control the content that the teams would use (i.e., the empirical data-set as one of the controlled conditions of our experiment), we recruited four end-users beforehand and ran a remote usability test with them. They had to interacted with a low-fidelity prototype regarding a travel agent Web-site (see Figure 9), devised explicitly for the scope of this study. The four screens of the prototype were referring to "make a booking", "search results", "my bookings", and a "notification message", showing the respective details, attributes, and fields. The end-users executed a number of predetermined tasks over those screens during four different recorded 90 min slots. A moderator was administering and monitoring the sessions, while two note-takers were capturing the feedback during the process. In total, we generated four different consolidated sets of approximately 120 feedback items each one, containing unstructured data like observations, ideas, comments, wishes, emotions, opinions, etc. To increase the familiarity of the participants (product teams) with the simulated scenario, and to make sure that they all had a common baseline understanding of the facts, we dedicated the time at the beginning to show the recordings of the usability tests captured earlier, we explained the prototypes and clarified any questions regarding the notes and the end-users profiles. Once everybody was feeling comfortable with the setting we started the analysis of the qualitative data by applying the EUREKATAX method (and the functional prototype tool). The teams were in the same physical environment with the instructors, interacting in a controlled environment with the prototype after a small introduction of each data analysis phase (and module). At first, we introduced each phase (reflecting the distinctive categories of the taxonomy) and then we allocated a specific time slot for the teams to put it into practice using the generated data-sets. Hence, we followed the moderated protocol during the execution of the sessions progressing gradually from the first step of the method to the last together with the participants, but without interfering with their work. Apart from the necessary guidance and time restrictions for executing each phase (aligning with the constraints of a real-life business case), the participants were free to apply the analysis according to their discussions and decision in their groups and as they believed was better fitting the circumstances, without any external influence from the instructors.



departure date 03/05/2018
return date 17/05/2018

● flight
○ train
☑ rental car

Paris ▾ Berlin ▾
Berlin ▾ Berlin ▾
Person 2 ▾ Class compact ▾
Class Economy ▾

● hotel
○ appartment
○ B&B

rooms 1 Room ▾
Room 1 2 ▾

● sightseeing tours
○ restaurants
○ hiking tours

Berlin ▾
○ walking ○ bustour ● boattour

Figure 9. Low-fidelity prototype screen snapshot.

In summary, four interactive workshops took place during the period September 2019–January 2020 with an average duration time of 8 h per workshop. In total 16 teams participated in this study, four teams in each workshop with 3–4 team-members on average each one. Each team applied the analysis once by joining one of the available slots in the period of 4 months. During these sessions, we had the chance to make observations regarding the utilization of EUREKATAX in the real-life business settings, extracting important insights during the application of its various phases, collecting feedback for improvement of its conceptual viewpoints, or suggestions for further optimization of its hierarchy and structural elements. As a consequence, we gathered three different types of feedback:

(a) We distributed to participants a SUS-like questionnaire after the completion of the data analysis process (at the end of each workshop) for measuring the perceived usability (including effort vs outcome) when comparing EUREKATAX to their previous experiences—methods and tools that they have used in the past for analyzing usability testing data. Given the lack of a standardized tool, to our knowledge, to systematically evaluate the perceived usability of a UX process, taxonomy, workflow, or methodology, we aligned the widely acknowledged System Usability Scale (SUS) [38,70,71] to our goals. SUS could be considered as an industry standard, technology independent tool used to determine the usability of consumer software, Web-sites, hardware, etc. Originally, it was composed of 10 items (questions) which measure, on a Likert-scale from 1 to 5, factors like satisfaction, memorability, ease-of-use, efficiency, learnability, etc. [72]. These theoretical dimensions could be considered sufficient for the purpose of this study, so we adjusted the corresponding items to best express the needs and requirements of the taxonomy (as a method and tool) under investigation. E.g., we rephrased item 3 to "I thought EUREKATAX was easy to use and I felt well guided throughout the data analysis", or item 5 to "I found the various data analysis phases/functions in EUREKATAX were well integrated". Moreover, we were also interested to bring the data analysis outcome in the center of attention, clarifying whether the process of EUREKATAX (e.g., are all phases, mechanics, materials, information, and presentations suitable for the teams? Are all the activities being implemented as intended?) meets the goals (measure if the objectives have been achieved and specific success criteria have been met at the specific time) by offering the expected outcome (e.g., how well has the method achieved its objectives (and sub-objectives)? Is there the right (accurate) information in place to make decisions?). Such factors could be considered of high importance when evaluating a method or workflow process [73] as to information systems or products. Hence, we extended the construct by adding two more items, i.e., item 11 "I feel that with EUREKATAX I have more structured and richer outcome to explore", and item 12 "I believe that the outcome from EUREKATAX does not justify the effort invensted". Participants could take as much time as they desired to complete this task. To measure the effectiveness of the current feedback we formulate the following two hypotheses:

Hypotheses 3 (H3). *The perceived usability score of all participants is higher than the standard average SUS score of 68 [74].*

Hypotheses 4 (H4). *Participants' scores distribution show a central tendency towards the top values (strong positive attitudes) of the adjusted SUS subsequent Likert scales.*

(b) In the same line, we also asked participants to allocate a score on the Likert-scale of 1 to 10 to the question "how likely is that you would recommend EUREKATAX to a colleague", identifying the Net Promoter Score (NPS) for the EUREKATAX (method and tool). NPS is a key simple success metric used extensively in the business sector as a standard for measuring the loyalty that exists between a provider and a consumer. It is associated with revenue growth [75] and UX qualities, e.g., it has been shown strong correlations between NPS and SUS [76]—attitudes toward usability explains the likelohood of users' to make a recommendation, e.g., if users rate usability as high, they are much more likely to recommend it to others [77]. In this respect, we formulate the following hypothesis:

Hypotheses 5 (H5). *Participants' NPS score is over 70, showing a strong tendency towards recommending EUREKATAX [78].*

(c) In each workshop, we used the Think-Aloud protocol in combination with the systematic observation [79] as a method of capturing the metacognition (and motivation) of the teams in three distinct time intervals during the execution of the data analysis tasks. Specifically, we simulated the three metacognition phases (as discussed in Section 3.3) through three sets of questions provided before-during-after the completion of the main data analysis tasks. The teams answered nine questions in total; three before they start the data analysis process (e.g., is this similar to a previous task? What do I want to achieve? What should I do first?), three during the learning process (e.g., am I on the right track? What can I do different? Who can I ask for help?), and three after finishing applying all the phases of the method (e.g., what worked well? What could I have done better? Can I apply this knowledge and skills to another situation?). The participants spent 5 min on brainstorming and 5 min on exchanging into their teams, in an attempt to share awareness for the same basic knowledge and standpoint to data analysis. Then, they wrote their thoughts as a group into post-its and pinned them on a shared wall. This way, we were able to collect, in real-time, information about the participants' thoughts, expectations, strategies, and decisions during the EUREKATAX data analysis process. In other words, clarifying the 'what', 'why', 'when' and 'wow' during the data analysis process. All this information would be associated and analyzed at a later stage for cross-verifying and enriching the outcomes of the study in relation to the participants scoring behaviour in the first two types of feedback using the SUS and NPS tools. The quality of this feedback type was evaluated using the following hypothesis:

Hypotheses 6 (H6). *The feedback items and actions expressed by over than 20% of the teams in each of the three time intervals were in their majority addressed by the classifications and functionalities of EUREKATAX.*

5.2.3. Analysis and Discussion of the Results

Evaluating the Perceived Usability and Usefulness of the Taxonomy

Initially, we re-assessed the reliability of the adjusted SUS tool. Given the required alignment and enhancement of its items, so to comply with the scope of EUREKATAX within the context of the current evaluation phase, the confirmation of its validity deemed necessary at this stage. Cronbach's alpha for the overall SUS revealed that the scale had sufficient reliability of $\alpha = 0.701$. Although this value might not be consistent to other stronger reported findings of e.g., 0.91 [80], we could argue for its acceptance since it meets the typical minimum standard of .70 for this type of measurement [81]. Furthermore, a closer look to its items declares that by dismissing two of them, item 10 ("I needed to learn a lot of things before I could get going with EUREKATAX") and 4 ("I think that I would need the support of a technical person to be able to use EUREKATAX"), could increase coefficient alphas to $\alpha = 0.705$ and $\alpha = 0.713$, respectively. Thereupon, adhering to the analysis recommendations of the SUS we normalized the scores of the scales to produce a percentile ranking as suggested in [74]. After the necessary calculations the SUS score was 68.6, corresponding to a percentile rank of 50% (participants had higher perceived usability than 50% of all products tested), with associated adjective rating B-. This value is above the standard average of 68 leading us to accept H_3. A possible interpretation of this score could refer to a positive marginal result, since it lies just above the standard average, however we should recognize that at this abstraction level it is in line to our expectations given the type and characteristics of the object of investigation. The usability evaluation of a taxonomy or methodological process entails an inherent rationale and complexity which may significantly vary to the one of e.g., a product, that the SUS tool had been originally designed to measure. Latter interactions may typically be based on well defined use cases (that include interchangeably the user and the system), with simpler structure and purpose of its interrelated entities, as well as a more clear reasoning and transparent mental models that draw a more direct line to the causation of a result as opposed to the former.

Triggered by the aforementioned outcome, our secondary goal was to elaborate on the internal structure of SUS for two reasons: First, we needed to understand how the two additional items (11 and 12) fit into the formation of SUS, and second if we should consider it as uni-variate or multivariate construct, guiding the next steps of our analysis for extracting the central tendency of the participants to the subsequent scales. The situation-specific behaviour of items 4 and 10 (as also confirmed in our preliminary investigation) is the main reason for the long lasting dispute in the research community regarding the nature of SUS. Although, it has been fundamentally assumed that the psychometric properties of SUS dictates an assessment of the single construct of usability [80], there are other studies with bigger samples that confirm its use as a two-factor (or more) solution [82,83]. For example in the latter case, Lewis and Sauro [83] named the first factor based on its content (composed of items 1, 2, 3, 5, 6, 7, 8 and 9), as 'usability' and the second factor (composed of items 4 and 10), as 'learnability'. Hence, our objective was to perform a factor analysis to examine the inner structure of the SUS in our study. To explore the factorial structure of SUS on our sample, all 10 + 2 items of the instrument were subjected to an exploratory factor analysis with Varimax (orthogonal) rotation. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = 0.708. Barlett's test of sphericity $\chi^2(66) = 146.757$, $p < 0.001$, indicating that the correlation structure is adequate for factor analysis. The principal component analysis and the Kaiser's criterion of eigenvalues greater than 1 [84] yielded a three-factor solution as the best fit for the data, accounting for 51.841% of the variance. The results of this factor analysis are presented in Table 2.

Table 2. Factor analysis table for the adjusted system usability scale (SUS).

	Loadings			
Adjusted SUS Items	**Factor 1: Utility**	**Factor 2: Usability**	**Factor 3: Learnability**	**Communality**
Q8	**0.784**			0.659
Q2	**0.726**		0.369	0.694
Q5	**0.710**	0.438		0.746
Q6	**0.694**			0.499
Q7	**0.362**	0.359		0.345
Q3		**0.674**		0.468
Q9	0.303	**0.667**		0.557
Q1	0.324	**0.633**		0.546
Q11		**0.555**		0.378
Q10			**0.707**	0.515
Q4			**0.700**	0.492
Q12			**0.427**	0.322
Eigenvalue	3.292	1.552	1.377	
% of Total Variance	27.434	12.932	11.475	
Total Variance			**51.841%**	

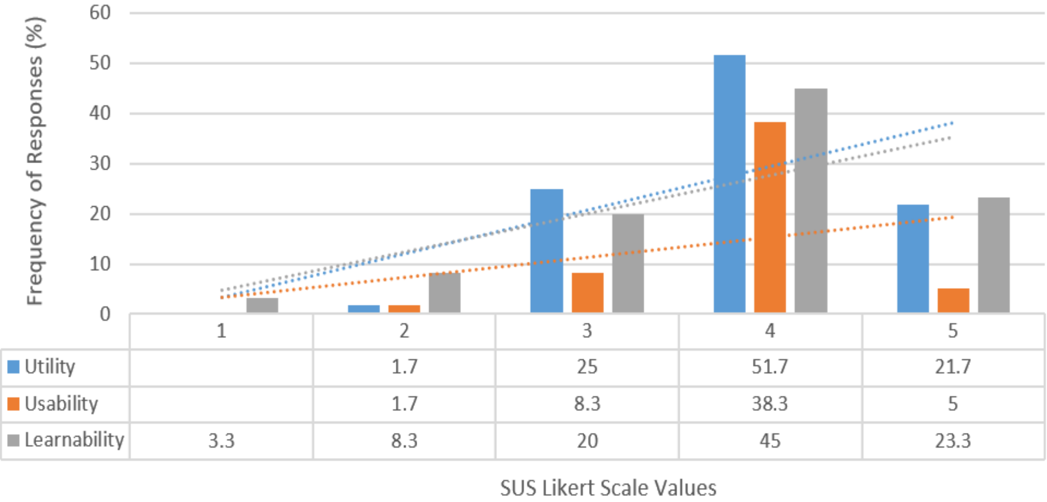
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. [a,b,c]

[a]. Rotation converged in six iterations. [b]. Boldface values refer to the loadings of the items of each factor.
[c]. Underlined communality values indicate small amount of variance

Likely, two of the three factors obtained with a similar structure as those found previously in [83], namely the usability and learnability. Analyzing in more detail the semantic meaning of the items' content of each factor, we decided for consistency purposes to keep the latter name for factor 3 and build upon the distinction of the former with a "twin" term called utility for factors 2 and 1 respectively, as more suitable names that best characterize the related items in each factor. Usability and utility theoretically are similar since they are both highly influencing the design and development of qualitative user-centred solutions [85,86], however there are some indirect complementary differences justified (also) by our analysis. On one hand, utility refers to how the information flows from one phase to the other, the practicality and usefulness of tasks, and how beneficial the functionality and the mechanics to accomplish a given task are; whereby on the other hand, usability concerns in addition the efficiency, safety, memorability, learnability, and satisfaction during interaction [87], issues closely related to the user. Thus, the three factors are: (a) *utility* with five sub-scales: 8, 2, 5, 6, and 7. This factor had an eigenvalue of 3.292 and accounted for 27.434% of the variance. As mentioned, it was labeled as such due to the high loadings by items that relate to the usefulness for accomplishing successfully the goals of data analysis, process complexity, consistency of tasks, and interrelation of phases and functionality; (b) *usability* with four sub-scales: 3, 9, 1, and 11. This factor had an eigenvalue of 1.552 and accounted for 12.932% of the variance. Its label derived from the high loading of its items that referred to aspects closer to the user per se, like safety and guidance in relation to a structured outcome (produced by the new item 11), ease-of-use, confidence, engagement, desire, and satisfaction; and (c) *learnability* with three sub-scales: 10, 4, and 12. This factor had an eigenvalue of 1.377 and accounted for 11.475% of the variance. Interestingly, this factor had by two thirds exactly the same structure as in [83], allowing for the assumption that the new item 12 (referring to the effort vs outcome) was perceived by the participants as 'learning effort', and hence the assigned label could be considered as most appropriate in this case. Hence, we identified high loadings in items that concern the familiarity and easiness to accomplish a task for the first time, efficiency on tasks not exposed in the past, support or expert assistance, and prior training.

The communalities of the variables included are rather low overall with three variables (7, 11, and 12) having a small amount of variance (34.5%, 37.8%, and 32.2% respectively) in common with the other variables in the analysis. This may indicate that the variables chosen for this analysis are only weakly related with each other. However the KMO and Bartlett's Test of Sphericity both indicate that the set of variables are at least adequately related for factor analysis. Consequently, this means that we have identified three clear patterns of response among the participants regarding perceived usability using the SUS—one pattern labeled as utility, the other as usability, and the third as learnability. These three tendencies are independent of one another (i.e., they are not correlated).

Next, our aim was to discover the central tendencies of the participants in these three factors—patterns of responses. This would help us to clarify in more detail the scoring behaviour in relation to the actual influence that EUREKATAX had to the teams, when they put the method into practice, as well as expose any potential biases e.g., respondents avoiding using extreme response categories, agree with statements as presented, or scoring for seeking (social) acceptance. After normalizing all scales to be positive, we computed the three new factors (variables) consisting of the pertinent items for each one as revealed earlier from the factorial analysis. Descriptive analyses such as frequency distributions and median were obtained to characterize the collected data. Thereupon, we observe for the three factors that the central tendency of participants' scoring preference is concentrated towards the top (positive) values of the scales (i.e., 4 and 5), see Figure 10 (implying the acceptance of H_4). More specifically, for the utility factor (Mdn = 4, IQR = 1) professionals (N = 60, 73.4%) recognize EUREKATAX as a useful taxonomy and method that offers beneficial functional capabilities for achieving the data analysis goals, coherent process steps, and consistent information flow between and within the various tasks and phases. Regarding the usability factor (Mdn = 4, IQR = 0.88), participants (N = 60, 75%) found the taxonomy to be a method that offers the necessary empathy to the end-user, since it entails vital human-centred characteristics during the data analysis process like guidance, confidence, satisfaction, etc. Finally, for the learnability factor (Mdn = 4, IQR = 1), again users (N = 60, 68.3%) were consistent with the idea that EUREKATAX is a method that provides high familiarity to the data analysis tasks, easy to execute them for the first time, without necessarily an expert's support or prior training.



	1	2	3	4	5
■ Utility		1.7	25	51.7	21.7
■ Usability		1.7	8.3	38.3	5
■ Learnability	3.3	8.3	20	45	23.3

Figure 10. Frequency of perceived preferences regarding the three SUS factors.

We analyzed the NPS score according to [88]. The result after the calculations revealed a strong NPS score of 80 (Mdn = 10, IQR = 1), labeled as 'excellent' (indicating the acceptance of H_5). In the respective categories of the scale [75] EUREKATAX received the following allocation of scores for N = 60: For *promoters*, 32 (53.3%) of the professionals responded with a score of 10 and 16 (26.7%) with 9. Nobody from the participants lied to the category of *Detractors*—there were no scores between 0 and 6. For the participants labeled as *passives* (their behavior falls between promoters and detractors), 11 (18.3%) responded with 8 and only 1 (1.75) with 7. This result suggests that professionals were enthusiastic with EUREKATAX and will recommend it (generate a lot of positive word-of-mouth) to other colleagues or third-parties, since they are motivated to influence other prospective users of the taxonomy with their feedback [78]. In relation to SUS (with which NPS has been found to have a strong positive correlation of 0.61 [89]), although at a first sight we could acknowledge that the current overall score of the perceived usability in our case does not justify the high NPS score [77,89], if we elaborate on the subsequent factors of utility, usability, and learnability and appreciate the high rating behaviour of the participants (i.e., distribution of scores around the top values of the scales), we could argue for a strong association that explains the high likelihood of professionals (especially the promoters) to speak in favor and endorse EUREKATAX.

Evaluating the Participants' Thoughts, Knowledge and Strategies When Applying EUREKATAX during the Data Analysis Process

We captured participants' feedback on the three distinctive phases of metacognition during the data analysis tasks' execution, in a free text format. Applying a thematic analysis [69,90] we managed to identify patterns (or themes) within the collected qualitative data that were interesting and underlined even more the impact of EUREKATAX on the participatory teams during the data analysis process. We explored the collected data following a rather flexible methodological approach (as opposed to more rigid methodologies for qualitative data analysis [91]) since we were interested at allowing the concepts and themes to emerge naturally from the contents of the feedback, avoiding any bias or constraints from the beginning (e.g., following a classification technique based on predetermined conditions and research variables). All the feedback items (collected originally in post-its) created a quite rich and diversified data corpus which was composed of 83 unstructured items (e.g., statements, opinions, feelings, actions, assumptions, etc., at each one of the three phases). Accordingly, in order to manage the initial complexity, we present a minimum viable perspective of the analysis, focusing on comprehensively addressing the respective research challenge and hypothesis. More specifically, we used an adapted version, to our case, of the Braun and Clarke's (2006) six-step framework [69,92], as described below:

Step 1: At first, a bottom-up or inductive thematic analysis applied driven by the data themselves. We started by synthesizing and consolidating the feedback across all 16 teams. Then, we cleaned and sorted the data, correcting any grammatical or syntactical errors, and filled in any gaps of incomplete statements by recalling insights from memory and the workshop sessions.

Step 2: We organized our data in a meaningful and systematic way creating clear summaries of feedback. We structured them under each phase, taking care of not altering the message that is communicated. Main concern was to make ourselves familiar with the semantic meaning of the entire body of the data-set and how it is related.

Step 3: We went through the transcripts and generated initial codes reducing the data volume into small purposeful chunks if information. We did not follow an exhaustive coding process (e.g., code every piece of text or line separately), but we rather created codes that had a distinctive semantic meaning and would be manageable for the next stages of the analysis. For this exercise we used *open coding*; developing and modifying the codes as we worked through the coding process rather than having pre-set codes. This was an iterative process that included many discussions and improvement of the initial versions.

Step 4: Next, we decided to refine even more the initial codes by creating a corresponding high-level code for each one of them. We followed the same process as in Step 3, and we came up with meaningful one-to-three words that best describe each case, e.g., for the code "discuss with note takers in groups and share experiences" we composed the "note-takers experiences" high-level code. This would help us to engage more effectively with the data-set, e.g., screening faster the various items, and identify any similarities or repetitions, especially across the three metacognition phases.

Step 5: In the following step, two independent coders reviewed all the coded feedback items and extracted separately relevant themes, considering the semantic meaning of each one. Then, they repeated this exercise and compared their notes; cross validating each outcome until they reached an agreed consolidated version of themes for each code. The themes referred to extracts that were connected to the purpose of the feedback items. Although, given the size of our sample, there were codes with considerable overlapping, we managed after several rehearsals to produce six themes covering all the range of the codes and related them as they best fit together. For example we had several codes that dealt with data cleansing, comparison, data synthesis, and consolidation, and sharing experiences. We assemble these into a theme called "prepare and organize your data". Other codes concerned the creation of clusters and summaries, the organization of weighted information, and the consideration of the end-users roles in the data analysis; so, we collated them into a theme called "data reduction and specification". Another example, some codes were linked to topics as creating priorities, identify solutions, effort and risks, and making decisions. Respectively, we formulated a theme called "priorities and actions". The rest of the themes referred to "review and explore the data", "coding and classification", and "presentation and explanation of findings", to complete the list. After that, we grouped the codes under the corresponding themes identifying and optimizing any potential overlaps. For example, a code in the 'before' phase that referred to "organize findings and bring structure" would be associated with the theme "review and explore the data", while another in the 'during' phase that points out "role fit should be considered", would contribute to the theme "data reduction and specification". As a result, all the codes had an interrelated high-level code and all of them were associated with one theme, which in turn (one theme) might pertain to more than one code. (Table A1 contains a full list of the themes, high-level codes, and codes extracted).

Step 6: Lastly, we applied theoretical thematic analysis (i.e., top-down—driven by our focus and specific research questions [69]) since we were interested to associate the derived themes and coded data to the various phases and classifications of EUREKATAX for verification. Hence, we used the taxonomy's categories discover, learn, act, and monitor (and the respective sub-categories and operations) to organize the derived themes (and the subsequent coded items) into hierarchies determining their relationship based on their recurrence (for at least 20% of the teams, N = 16—see Figure 11).

Applying descriptive analysis like frequencies and importance we obtained an understanding how teams perceived the whole data analysis process in the three distinctive phases of metacognition, while executing the data analysis tasks using EUREKATAX and where do they place emphasis most in relation to the derived codes. In total, we generated 65 codes, allocated into the respective phases as follows: 21 codes in 'before', 22 codes in 'during', and 22 codes in the 'after' phase. Hereafter, we outline the codes and frequencies of over than 20% of the participating teams (N = 16—see Figure 12 for all the codes' frequencies across the teams). We build upon the assumption that at least four out of 16 teams (approximately 12–16 participants) would serve in our case as the critical mass of the sample to justify the importance and validity of the collected feedback items (i.e., codes) [93].

Regarding the 'before' phase (where participants had a small introduction but not yet become familiar with data analysis stages or tasks), teams would categorize/cluster findings into groups after the usability testing sessions (e.g., use colour coding) (69%); go through the feedback and clean data (56%); digitize feedback and organize findings by bringing structure to data (44%); prioritize comments and find issues (31%); compare notes of different users, finding out similarities/differences, and would involve all stakeholders in the data analysis process (25%). The majority of the teams in the 'during' phase (participants found in the middle of the data analysis process, having partial knowledge of the method), mentioned that role fit should be taken into consideration during the data analysis process (56%); while many of them would create clusters of the points in the notes (44%), by also considering the task, role fit and the usability issues (50%). A significant number would calculate weighted references, instead of testing users with the same (31%); where others were wondering if time would always be sufficient (given the business constraints) to apply EUREKATAX (31%). In addition, 31% of the teams would prioritize the points and summarize them based on users' weights. Finally, 25% expressed the wish to be able to trace back the comments to the original notes. For the 'after' phase (participants concluded at this stage the data analysis process), teams felt that they had learned a new powerful method and tool and that they will definitely apply it in their daily work (44%). The 38% of the teams mentioned that they found the tool very useful, rational, and they liked that everything was traceable; they learned to analyze feedback in a more constructive manner; they valued the use of weights during the data analysis process; they emphasized that data visualization will help them to better compare and create an end-to-end understanding of the feedback; and they liked the systematic and consistent way to analyze the data increasing precision of the outcome. Moreover, 25% believed that they learned how to extract what to do next and how to create priorities, how to analyze the effort to implement the usability issues identified, as well as how to summarize the analysis results and create a variety of effective reports. In conclusion, they found EUREKATAX a great tool that was worth the effort; made analysis less subjective, and definitely they would like to see it as a fully functional application as it would be more powerful and useful. As we can observe in Figure 12, there are also some codes that are on the borderline of the set threshold, and with 19% recurrence may also contribute considerably in the cumulative appraisal of the taxonomy. For the sake of completeness these relate to beliefs, opinions, and standpoints of participants with regard to actions that relate to keeping the original notes intact for future reference, comparing notes of different users (maybe using the storytelling method), and capturing similarities and differences ('before' phase). They would summarize and label the end-users' points, and appreciate recommendations based on the calculated priorities ('during' phase). As a highlight learning outcome, they mentioned that they can organize the end-users' feedback better now ('after' phase).

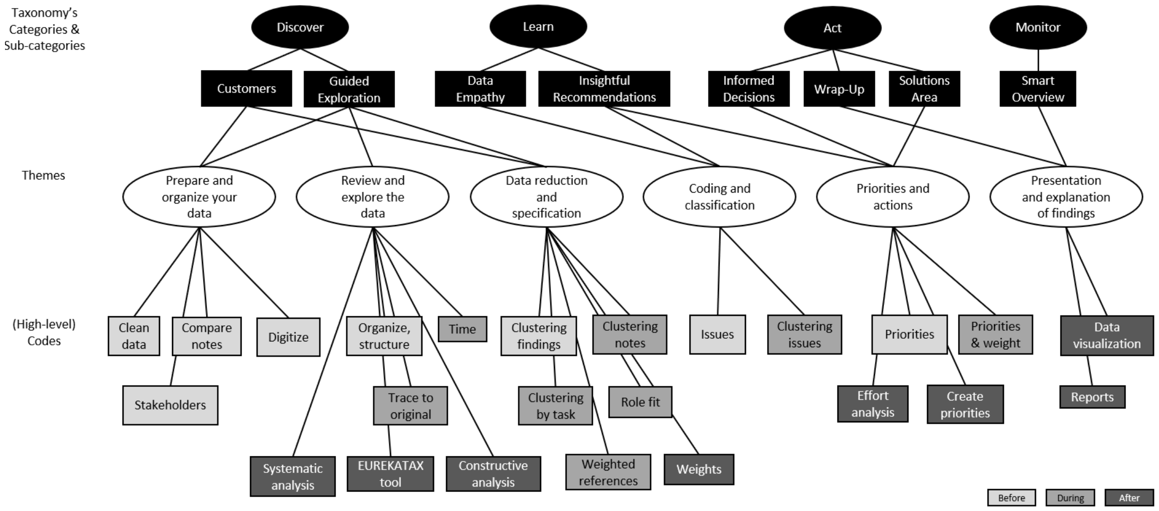


Figure 11. Thematic map—illustrating the relationship between themes, codes and taxonomy classifications.

The feedback collected during the three phases, emphasizes on cognitive meta-functionalities of the participants that regulate their goal-directed strategies and knowledge during the usability tests data analysis process. It shows that the participants' perception and attitudes, as influenced by EUREKATAX data analysis tasks, explained to a large extent the scoring and central tendency towards the top values of the SUS (e.g., the code "like the tool is very useful. It is rational. Everything is traceable") and NPS (e.g., the code "I will also recommend it to my colleagues") tools. In such a way, they scored high because they find in EUREKATAX what they would aim for during the data analysis process, in terms of expectations regarding the actions (and operations), and the effort invested in relation to the outcome. In particular, when participants "thinking about their thinking", evaluating continuously their knowledge, strategies, and learning around the related tasks, we can observe that the proposed taxonomy addresses significantly their expectations, experiences, needs, and wishes at each phase of the process (see Figure 11). Again, with reference to at least the 20% of our sample, at the beginning, when participants had little knowledge about our taxonomy and method their feedback items were fully satisfied by the hierarchies and operations of the taxonomy.

For example, the "stakeholders" and "digitize" codes were covered by the "customers" sub-category, the "organize, structure" by the "guided exploration", the "clustering findings" by the "data empathy", the "priorities" by the "insightful recommendations", "informed decisions" and "solutions area", etc. During the analysis, where participants came across with the first two classifications (and the respective functionalities) of the taxonomy, expressed through their feedback their satisfaction, indicating that in practice would continue to work and act in line with the offerings of EUREKATAX. They were feeling that they were on track without missing or wishing to change something substantially in the data analysis process. For example, the code "trace to original" fulfilled by the sub-category "guided exploration", the "role fit", "clustering notes" and "weighted references", by the "customers" and "guided exploration" respectively, while the "clustering issues" by the "data empathy" and "insightful Recommendations". After the completion of the data analysis, when participants asked to evaluate overall their knowledge, comparing their past experiences and the new lessons that they learned, in relation to what they would use in the future, they were keen to employ most of the operations of the taxonomy, placing emphasis in the systematic and coherent perspective that the method brings to the analysis of their data-set. Indicatively, their feedback at this stage relates to codes like "systematic analysis", "constructive analysis", and the method or tool *per se* "EUREKATAX tool", which are realized by the "guided exploration" sub-category; or the codes "effort analysis" and "reports" by the "solutions area" and "smart overview" respectively.

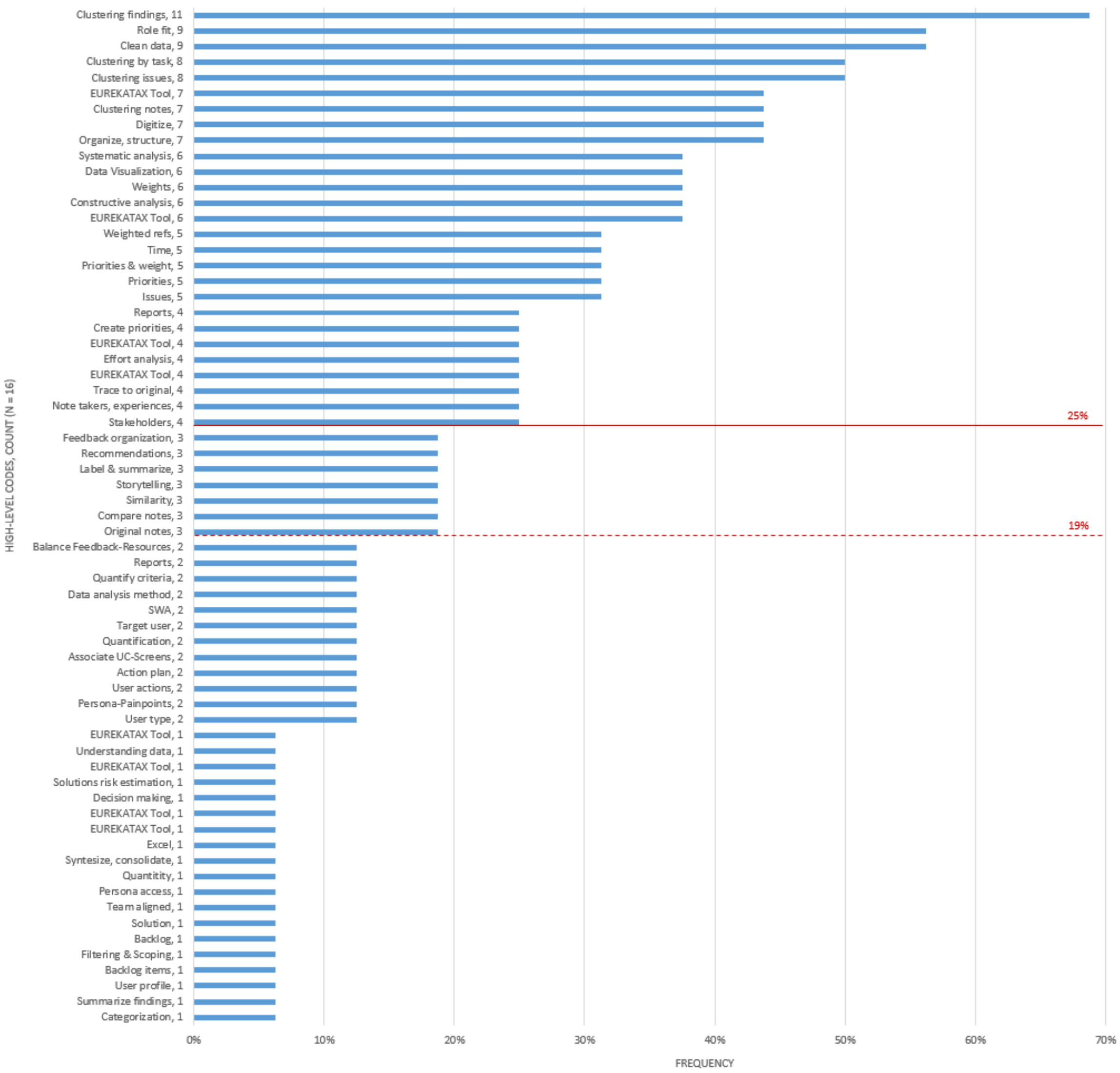


Figure 12. Distribution of codes frequencies across the participating teams (N = 16).

If we would take a more distant stance for interpreting (on a meta-level) how the knowledge of participants has been transformed in time, although there are distinctive variations in the codes across the three phases of metacognition, we can recognize some feedback items (or codes), that play a central role for the teams in time. Such as, structuring and clustering the information collected from the usability tests sessions, considering the role-fit of the end-users in the data analysis, generating weighted references to comments, set priorities, sharing experiences and reporting, etc. The latter facts are visible, either directly or indirectly, and are being progressively intensified (given their repetition from the critical mass of the sample) towards the completion of the data analysis using EUREKATAX. Apart from the tangible added value for the taxonomy, this shows how the team builds upon the previous findings and lessons learned each time, exploring the new knowledge, which in turn becomes the new experience, so to continue delve into the next learning units, and so on and so forth. In essence, they engage in an experiential iterative modular learning process, where they learn by expanding in their zone of proximal development. Participants, in principle, were not confused or lost, or they did not feel that what they put into practice in this real life scenario was out of reach given their past expertise or their roles. At the same time, no big gaps or strong diversification from their expectations according to their formulated mental models around usability tests data analysis have been identified during the workshops, that would hinder the execution of the data analysis tasks or negatively affect their UX perception and attitude towards EUREKATAX. The aftermath of the aforementioned analysis leads us at accepting also H_6.

5.3. Validity and Limitations of the Studies in the Two Evaluation Phases

The validity of a study is primarily affected by its internal, external, and ecological validity. Internal validity reflects the accuracy of data and the conclusions drawn based on this data; external validity indicates whether the data and the conclusions drawn can be generalized to a wider extent [94], and ecological validity requires that the experimental design, procedure and setting of the study approximate the real-life context that is under investigation [67]. With the aim to increase the internal validity of all our studies during the evaluation process of EUREKATAX, in Phase 1 we recruited a sample of participants with a significant expertise on UX methods, techniques, and tools. Each expert had already led or contributed extensively in various phases of usability studies, prior to conducting the taxonomy study. Thus, the research design was set up in order to avoid inference errors since the participants involved were more experienced than novice users, with respect to usability studies execution and analysis of empirical data. In Phase 2, we recruited professionals that had different backgrounds and business roles. They already had some experience with related activities to usability testing and they had all participated at least once in the qualitative data analysis process with their product teams. Our main concern for these series of studies was to simulate a real-life scenario to data analysis in the business sector, therefore the balanced formulation of the teams would increase the accuracy of the information flow, interpretations and decisions during the various data analysis tasks. Regarding external validity, given that future studies will contribute to the external validity of the reported research, we suggest that using the EUREKATAX as a qualitative data analysis method and tool in different contexts and technological settings could improve the overall user experience regarding products' usability and quality. The latter argument can be supported by the results and feedback items of the participants especially in Phase 2, where teams had the opportunity to embrace the various classifications of the proposed taxonomy by putting them into practice through the recommended operations. Their perception, beliefs, and opinions suggest the benefits of EUREKATAX, also if generalized and applied to other contexts of use. For example, one code from our analysis mentions "great tool, makes analysis less subjective, worth the effort", while another "this method and tool could be also applied to other situations, e.g., interviews" (see Table A1). Finally, there was also an effort to increase ecological validity of the research since in Phase 1, the taxonomy's tasks were integrated in Web-based forms and the participants were involved at their own physical environments without the intervention of any experimental equipment or person. With respect to Phase 2, we simulated a business scenario around the object of investigation in the physical organizational environment of the participants, that were familiar and exercised their daily business responsibilities. We scheduled more than one time-slots for the workshops, so that they have the chance to join one of them at their convenience. In addition, during the execution of the tasks, we avoided employing any new methods, material or equipment, that would make them feel uneasy or uncomfortable.

Regarding limitations, in Phase 1, the study was evaluated with a rather "closed system" testing method, a variation of Delphi Card Sorting. This means, that all the elements of the taxonomy (like categories, sub-categories, etc.) were provided in a specific structure requesting from the participants to assess their alignment and suitability. Although the specific approach has been carefully selected among others aiming at a wide and diversified reach of experts, someone could argue that there is an increased bias during its execution, since it already creates a basic understanding through the predefined semantic association of the elements and their interactions. Alternatively, more inductive evaluation approaches could he used allowing experts to construct the taxonomy's hierarchies from the elements themselves. Another limitation would relate to the remote validation protocol we used, and the inability to assess whether the meaning of the items has been equally realized by the participants. In this respect, they did not have the opportunity to ask for any clarification questions or to engage into discussions regarding blurred points. As discussed earlier, we tried to minimize the likelihood of falsified interpretations of the given items with the pilot testing we ran during the preparation phase, but still it is rather improbable to be certain about a common base of understanding across the participants. A similar validation method could consider inviting experts in the same physical space with the testers giving them the option to exchange whenever necessary. In Phase 2, the use of questionnaires, like SUS and NPS, to collect the perceived usability and satisfaction of participants refer to self-reported data, which may be unreliable. They measure subjective user perception, not objective performance (based on metrics like task completion rates, time on task, or errors). These metrics reveal what the user's satisfaction level was, but do not pinpoint any weaknesses or strengths in their experience (or what you can change to improve it). Subjective metric of preferences (e.g., satisfaction) usually tell a clearer story when combined with performance metrics (given their imminent correlation [95]). To partially compensate on this weakness and cross-triangulate the captured data from the questionnaires, we designed the second part of this evaluation phase, collecting in real-time (in three distinctive time intervals) the experience of the teams regarding the 'what', 'why', 'when' and 'how'. Guided by their metacognitive awareness and functionalities, we managed to create a deeper understanding when they apply a higher order thinking during qualitative data analysis, and discover causation or associations between the data that led us to more safe and informed conclusions. Moreover, another concern raised by the participants during the data analysis process was with respect to time. Given the project that they might be assigned to, potential organizational constraints, as well as the pressure to deliver in short periods of time the next updates of their products (usually four times per year), they were somehow skeptical if they would always have the time to implement the method (i.e., "time" code). Indeed, we could consider this as a loose point to EUREKATAX and highly situation-specific, which however could be managed in most situations by the cyclic (or modular) approach of the taxonomy (see Section 4). Another limitation could be the use of Excel as a prototype tool to realize the classifications and operations of the taxonomy as functionalities, and support teams to receive the real experience of the process while applying it. Although, Excel qualified as the most suitable tool for us at this stage, users found it challenging to use, as expressed by some of the findings: "I need to prepare the .xls tool in advance, this might be cumbersome", or "Excel is not my friend for descriptive tasks". We should note that Excel is just the medium to serve our purpose for the current study and it does not represent a recommended implementation platform nor an evaluation object (or factor) of consideration in our study. This would be rather improbable given the nature and goals of the usability tests data analysis tasks. In this respect, we are currently exploiting opportunities for developing a fully-functional application based on the proposed taxonomy, triggered always from the feedback of the end-users and the respective codes like "I learned a new powerful method and tool—I will definitely apply it in daily work" or "If it is developed as an app it will be much more useful and powerful". Lastly, because of the time limit, the two evaluation phases were conducted with a sample size of 20 UX experts and 60 business professionals respectively, that at this research stage we would safely assume that served its purpose. However, in relation also to the enhancement of taxonomy's external validity, a larger and more varied (in terms of roles and experience) sample should be recruited as a representative size of population that would be able to justify the generalization of the taxonomy to larger groups and diverse contexts of use.

6. Conclusions and Discussion

In this paper we presented a comprehensive method towards the analysis of qualitative data collected from the usability studies. This method is expressed through the definition of the EUREKATAX taxonomy, that is composed of four main categories, eight sub-categories, and 52 items—operations with their respective properties. Main aim of this taxonomy is to provide the grounds for a more systematic and rather flexible hierarchy of information that would guide the transformation of fuzzy empirical data (like opinions, behaviors, sentiments, ideas, experiences, etc.) into concrete actionable items ready to be injected to the implementation plans of a software development team. Such actionable items, may be a set of concrete solutions (and alternatives) that could be employed for tackling specific usability issues, for improving the UX and quality of software designs and products. Although, there are various research works around the composition of taxonomies that refer to the definition of the term "usability" and the surrounding influential aspects (e.g., errors, usability problems), or to the diverse application domains (like virtual reality environments, biometrics, telehealth systems, robotics, etc.), to our knowledge only a few works have rigorously approached the actual process of qualitative data analysis proposing a systematic taxonomic model that would elaborate on the consolidation, synthesis and imminent conversion of unstructured data into meaningful seeds of executable tasks. In this line of thought, our taxonomy describes the data analysis stages with the expected flexibility while business experts progressively operate on a certain feedback item, moving from general to more detailed specification of information organization depending on the desired level of understanding and the expected outcome. For example, the members of a team might be constraint in terms of time, since they need to deliver quickly and they do not have the capacity for a full-fledged data analysis. In this respect, they might decide to move rigorously through the first two levels of the taxonomy and then utilize the attributes that are most important to them in order to have supportive argumentation, given the circumstances, about critical feedback items. They could then progress in the future to the next levels so to obtain a more solid outcome.

For the construction of EUREKATAX we considered the characteristics and mechanics of the validation phase that iteratively takes place during the UCD and software development process, e.g., flexibility of tasks execution, knowledge transferability and integration, experience-based interpretations, transparent workflows, manageable iterative process modules with tangible outcomes, heterogeneous team work environments, diversity of roles and educational backgrounds of professionals, etc. In this respect, our main concern was to base our taxonomy on solid theoretical grounds that would benefit the participants throughout the whole analysis process and would optimize the subsequent steps and effort invested. The theoretical concepts have been meticulously blended into a framework that maintain a coherent organization of the categories and elements of EUREKATAX that facilitates a coordinated, consistent, and progressive flow of thinking and knowledge generation. Such theoretical aspects refer to well received experience-based models for learning, like Kolb's experiential learning, applied horizontally and vertically on the various interaction activities of data analysis. This takes place in a dynamic collaborative mode for the participants, moving from the abstract to the concrete, and engaging into beneficial negotiations for problem solving and decision making (putting into practice Engeström's learning by expanding paradigm of activity theory). The capstone of the theoretical framework is a "bird's eye view" for regulation and quality assessment of the data analysis process, which is expressed by a metacognitive perspective with meta-functionalities like planning, monitoring, evaluation, and self-reflection with respect to specific tasks. In other words, a high order thinking and awareness of cognitive processes, strategies and knowledge that ensures the successful accomplishment of the goals set.

Henceforth, the purpose of our taxonomy's evaluation extends to various levels of realization. In this respect, we decided to design two main evaluation phases considering the related objectives, requirements, methodological challenges and constraints: Phase 1 was concerned with the verification of the taxonomy's theoretical considerations; structure and hierarchies of its categories, sub-categories, operations and properties; and the respective classifications as a coherent model of information organization. Phase 2 was focused on the more practical implications of the taxonomy's utilization in a real-life business setting, validating its perceived usability, usefulness, and acceptability by the professionals when they engage with the suggested usability tests data analysis tasks (or operations). During Phase 1, we wanted primarily to validate the internal consistency of EUREKATAX and measure its reliability, without neglecting its subsequent classifications. Cronbach's was a strong indicator of the taxonomy's reliability. On the other hand, it revealed items that might be redundant or repetitive, which should be reconsidered in future iterations. Second, we were interested to assess the perceived fit of the items into the various classifications ensuring their semantic association with the categories and sub-categories and ensuring the harmonious information flow in the whole organization of information. This was not an easy task, considering that the items' meanings may differ (not identical) across categories, operations may relate to different sub-categories or the sub-categories themselves might have different level of granularity with different properties assigned to their operations. Third, we emphasized on items that scored relatively low, and by applying a series of post hoc analysis tests we managed to locate those that exhibit a particular problematic case for EUREKATAX, isolating them for future investigation and transformation. Lastly, we analyzed the open ended feedback we received from the UX experts finding really valuable insights and recommendations for the improvement of the taxonomy's content and structure.

Regarding Phase 2, we wanted the measure the perceived usability and UX of teams when they put the EUREKATAX into practice. In this respect, we developed an Excel prototype realizing all operations of the taxonomy as functionalities; enabling participants to follow the suggested data analysis method and tasks based on a real-life scenario and conditions. We used an enriched version of the SUS and NPS questionnaires to collect their subjective preferences regarding the subsequent factors. Initial data analysis with respect to the former showed the reliability of the tool as well as the above the standard threshold perceived usability of the participants for the EUREKATAX. This was a positive result given the peculiarities and complexity that the evaluation of a taxonomy (or method) entails as opposed to a typical product or software tool that SUS is primarily designed to measure. A deeper exploratory factor analysis at this stage revealed the multivariate nature of the tool in our case (in contrast to its uni-variate original development), that produced three factors, namely utility, usability and learnability. This finding could be perceived as well as positive regarding the structure and explainability of its scales given the situation-specific adjustments and the incorporation of two additional items. Descriptive analysis and frequencies distribution showed the central tendency of the sample towards the top values of the factors' scales. This was also true for the NPS tool where participants showed their strong preference towards recommending further, to other prospective users, EUREKATAX. Furthermore, during this evaluation phase we wanted to support and cross-verify the subjective positive preferences of the teams by extracting also their explicit feedback during the data analysis process. Using a systematic approach which was driven by three distinctive metacognition phases we collected their strategies, beliefs, opinions, feelings, wishes, experiences, etc. In real-time during execution, formulating a basis of valuable empirical insights. Following thematic and descriptive analyses we managed to create a consistent and interrelated data set which produced codes and themes that helped us to understand causation and association of data that otherwise would be intractable. In other words, we were able to triangulate the results of the subjective preferences of participants collected with the use of questionnaires, and apart from the overall positive user experience, perceived usefulness and acceptability of EUREKATAX, we could highlight deeper meanings, strengths, and weaknesses. For example, the results revealed that the classifications of the taxonomy fulfill the majority of the teams' expectations regarding the data analysis process of usability test data since there were identified association and patterns of the produces codes, high-level codes and themes with the respective categories and hierarchies of EUREKATAX. This is very encouraging for the future of our work since participants' feedback items are inline with the method's rationale, structure, elements and sequence of actions. On the other hand, it has been pinpointed that time might be a challenge considering the characteristics of a business ecosystem and the projects' management. However, this finding at the same time justifies the validity of the taxonomy's design as a cyclic and expansive iterative learning process where teams can progress and benefit according to their current needs and constraints.

In summary, the evaluation results showed the validity and acceptance of the taxonomy, but also revealed many points for future consideration and improvement. The ambition is to produce an organization of information that will assign a more inclusive meaning of what it represents (i.e., a guided method for usability tests' qualitative data analysis), as a definition, and what should be expected once measured to the benefit of the end-users and the UX quality of their software solutions.

Author Contributions: Conceptualization, P.G. and L.F.; methodology, P.G. and L.F.; software, P.G. and L.F.; validation, P.G. and L.F.; formal analysis, P.G.; investigation, L.F.; resources, P.G. and L.F.; data curation, P.G. and L.F.; writing—original draft preparation, P.G. and L.F.; writing—review and editing, P.G. and L.F.; visualization, P.G. and L.F.; supervision, P.G. and L.F.; project administration, P.G. and L.F.; funding acquisition, P.G. and L.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We would like to thank the professionals and individuals who have participated in the user studies, usability tests and workshops and provided their constructive feedback and valuable suggestions. Furthermore, our special thanks goes to our managers Matthias Berger and Thomas Uhl for their encouragement and support and to our colleagues UX Researchers Alexandra Matz and Lwam Leber for their insights, thoughts and continuous assistance throughout the process, the preparation and execution of the workshops and evaluation sessions.

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

Appendix A. Themes

Table A1. Themes identified along with the associated codes and respective high-level codes.

(Phases of Metacognition) Codes	High-Level Codes
Prepare and organize your data	
(Before) Go through feedback and clean data	Clean data
(Before) Digitize feedback	Digitize
(Before) Involve all stakeholders in the analysis	Stakeholders
(Before) Discuss with note takers in groups and share experiences	Note takers, experiences
(Before) Compare notes of different users, find out similarities/differences	Compare notes
(Before) Storytelling	Storytelling
(Before) Create action plan	Action plan
(During) Quantification of qualitative information (check for repititive feedback)	Quantification
(During) Helped team to be on the same page	Team aligned
(During) How to make sure that we always have enough feedback from users?	Quantitity
(During) It is difficult to pick-up the feedback items from the notes	Syntesize, consolidate
(During) Excel is not my friend for descriptive tasks	Excel
(After) I need to prepare the .xls tool in advance, this might be cumbersome	EUREKATAX Tool

Table A1. *Cont.*

(Phases of Metacognition) Codes	High-Level Codes
Review and explore the data	
(Before) Organize findings and bring structure	Organize, structure
(After) I learned a new powerful method and tool—I will definitely apply it in daily work	EUREKATAX Tool
(After) Like the tool is very useful. It is rational. Everything is traceable	EUREKATAX Tool
(After) To analyze the feedback in a more constructive way	Constructive analysis
(After) I liked the systematic way to analyze data—more consistency and precision	Systematic analysis
(During) Would we always have enough time to implement the methodology	Time
(During) Trace the points with original notes	Trace to original
(After) Great tool, makes analysis less subjective, it worth the effort	EUREKATAX Tool
(After) If it is developed as an app will be much more useful and powerful	EUREKATAX Tool
(Before) Save original notes	Original notes
(After) I can organize users' feedback better now	Feedback organization
(Before) User type analysis	User type
(Before) Write down the comments in relation to user-step (action)	User actions
(During) New useful way of data analysis	Data analysis method
(Before) Categorize customers	Categorization
(Before) Define user profile	User profile
(Before) Filtering information and scoping analysis	Filtering and Scoping
(During) Not every piece of feedback has to be resolved	Solution
(During) Need continuous access to the persona for a better understanding	Persona access
(After) This tool is mindblown	EUREKATAX Tool
(After) I will also recommend it to my colleagues	EUREKATAX Tool
(After) The better I understand the data, the better I can help to improve the product	Understanding data
(After) This method and tool could be also applied to other situations, e.g., interviews	EUREKATAX Tool
Data reduction and specification	
(Before) Categorize/cluster findings into groups (e.g., use colour coding)	Clustering findings
(During) Role fit should be considered	Role fit
(During) Cluster comments by task considering end-users' role fit	Clustering by task
(During) Cluster and document the notes and points	Clustering notes
(After) User weights is important	Weights
(During) Calculate weighted references, instead of testing users with the same	Weighted refs
(During) Identify user type—target group scope	Target user
Coding and classification	
(During) Cluster by usability issue type—classify	Clustering issues
(Before) Find usability issue types	Issues
(Before) Capture the frequency of similar comments	Similarity
(During) Label the points—summarize (find similarities)	Label and summarize
(Before) Match painpoints to the persona	Persona-Painpoints
(During) Relate points to use cases and screens	Associate UC-Screens
(During) Assistance times is a good way to understand the usability test result	Success with Assistance
Priorities and Actions	
(Before) Prioritize comments	Priorities
(During) Prioritize the points—summarize (and based the users' weight)	Priorities and weight
(After) Analyze the effort to implement the issues identified	Effort analysis
(After) I learned how to extract what to do next and how to create priorities	Create priorities
(During) Recommendation based on priorities is good	Recommendations
(After) The criteria for judgement may be more quantified (e.g., KPIs)	Quantify criteria
(After) How to balance customer feedback and development resources Balance	Feedback-Resources
(Before) Create backlog	Backlog items
(During) Create backlogs based on the output	Backlog
(After) Learned how to deal, measure and make decision based on the data I have	Decision making
(After) I learned how to capture solutions on specific issues and estimate risk of the changes	Solutions risk estimation
Presentation and explanation of findings	
(After) Visualize data analysis results for comparison and better end-to-end understanding	Data Visualization
(After) I learned how to summarize the analysis results and create a variety of reports	Reports
(After) Could use this tool to generate high level reports with more explanations	Reports
(Before) Summarize findings in slides to share knowledge with others	Summarize findings

References

1. UserTesting. 2017 UX and User Research Industry Survey Report. Available online: https://info.usertesting.com/ux-industry-survey-2017.html (accessed on 20 November 2018).
2. Nielsen, J. Nielsen Norman Group. 2016 Available online: https://www.nngroup.com (accessed on 20 November 2018).
3. Hassenzahl, M. User experience (UX) towards an experiential perspective on product quality. In Proceedings of the 20th Conference on l'Interaction Homme-Machine, Metz, France, 2–5 September 2008; pp. 11–15.
4. Hertzum, M. Images of usability. *Int. J. Hum. Comput. Interact.* **2010**, *26*, 567–600. [CrossRef]
5. Tractinsky, N. The usability construct: A dead end? *Hum. Comput. Interact.* **2018**, *33*, 131–177. [CrossRef]
6. Nielsen, J. *Usability Engineering*; Elsevier: Amsterdam, The Netherlands, 1994.
7. Preece, J.; Benyon, D.; Davies, G.; Keller, L.; Rogers, Y. *A Guide to Usability: Human Factors in Computing*; Addison-Wesley: Reading, MA, USA, 1993; Volume 183.
8. Preece, J.; Rogers, Y.; Sharp, H.; Benyon, D.; Holland, S.; Carey, T. *Human-Computer Interaction Reading*; Addison-Wesley: Reading, MA, USA, 1994.
9. Iso, W. *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)*; 9241-11; The International Organization for Standardization: Geneva, Switzerland, 1998; Volume 45, p. 9.
10. Preece, J.; Rogers, Y.; Sharp, H. *Interaction Design: Beyond Human-Computer Interaction*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
11. Schneiderman, B. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*; Addison-Wesley: Reading, MA, USA, 1998;Chapter 15; pp. 510–549.
12. Creswell, J.W.; Creswell, J.D. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*; Sage Publications: Thousand Oaks, CA, USA, 2017.
13. Johnson, R.B.; Onwuegbuzie, A.J. Mixed methods research: A research paradigm whose time has come. *Educ. Res.* **2004**, *33*, 14–26. [CrossRef]
14. Creswell, J.W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*; SAGE Publications Inc.: Thousand Oaks, CA, USA, 2008.
15. Patton, M.Q. Qualitative interviewing. *Qual. Res. Eval. Methods* **2002**, *3*, 344–347.
16. Bradley, E.H.; Curry, L.A.; Devers, K.J. Qualitative data analysis for health services research: Developing taxonomy, themes, and theory. *Health Serv. Res.* **2007**, *42*, 1758–1772. [CrossRef]
17. Krug, S. *Don't Make Me Think!: A Common Sense Approach to Web Usability*; Pearson Education India: New Delhi, India, 2000.
18. Barnum, C.M. *Usability Testing Essentials: Ready, Set... Test!*; Elsevier: Amsterdam, The Netherlands, 2010.
19. Rubin, J.; Chisnell, D. *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
20. Dumas, J.S.; Dumas, J.S.; Redish, J. *A Practical Guide to Usability Testing*; Intellect Books: Bristol, UK, 1999.
21. Gothelf, J.; Seiden, J. *Lean UX: Designing Great Products with Agile Teams*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2016.
22. Alonso-Ríos, D.; Vázquez-García, A.; Mosqueira-Rey, E.; Moret-Bonillo, V. Usability: A critical analysis and a taxonomy. *Int. J. Hum. Comput. Interact.* **2009**, *26*, 53–74. [CrossRef]
23. Keenan, S.L.; Hartson, H.R.; Kafura, D.G.; Schulman, R.S. The usability problem taxonomy: A framework for classification and analysis. *Empir. Softw. Eng.* **1999**, *4*, 71–104. [CrossRef]
24. Vilbergsdottir, S.G.; Hvannberg, E.T.; Law, E.L.C. Assessing the reliability, validity and acceptance of a classification scheme of usability problems (CUP). *J. Syst. Softw.* **2014**, *87*, 18–37. [CrossRef]
25. Andre, T.S.; Hartson, H.R.; Belz, S.M.; McCreary, F.A. The user action framework: A reliable foundation for usability engineering support tools. *Int. J. Hum. Comput. Stud.* **2001**, *54*, 107–136. [CrossRef]
26. Hermann, F.; Niedermann, I.; Peissner, M.; Henke, K.; Naumann, A. Users interact differently: Towards a usability-oriented user taxonomy. In Proceedings of the International Conference on Human-Computer Interaction, Beijing, China, 22–27 July 2007; pp. 812–817.
27. Rajeshkumar, S.; Omar, R.; Mahmud, M. Taxonomies of user experience (UX) evaluation methods. In Proceedings of the 2013 International Conference on Research and Innovation in Information Systems (ICRIIS), Kuala Lumpur, Malaysia 27–28 November 2013; pp. 533–538.

28. Micheals, R.J.; Stanton, B.; Theofanos, M.F.; Orandi, S. *A Taxonomy of Definitions for Usability Studies in Biometrics*; US Department of Commerce, National Institute of Standards and Technology: Gaithersburg, MD, USA, 2006.
29. Gabbard, J.L. A taxonomy of Usability Characteristics in Virtual Environments. Ph.D. Thesis, Virginia Tech, Blacksburg, VA, USA, 1997.
30. Singh, J.; Lutteroth, C.; Wünsche, B.C. Taxonomy of usability requirements for home telehealth systems. In Proceedings of the 11th International Conference of the NZ Chapter of the ACM Special Interest Group on Human-Computer Interaction, Auckland, New Zealand, 8–9 July 2010; pp. 29–32.
31. Huang, A. A research taxonomy for e-commerce system usability. *AMCIS 2002 Proceedings*; 2002; p. 94. Available online: https://aisel.aisnet.org/amcis2002/94/ (accessed on 23 May 2020).
32. Adamides, G.; Christou, G.; Katsanos, C.; Xenos, M.; Hadzilacos, T. Usability guidelines for the design of robot teleoperation: A taxonomy. *IEEE Trans. Hum. Mach. Syst.* **2015**, *45*, 256–262. [CrossRef]
33. Ritchie, J.; Lewis, J.; Nicholls, C.M.; Ormston, R. *Qualitative Research Practice: A Guide for Social Science Students and Researchers*; Sage: Thousand Oaks, CA, USA, 2013.
34. Miles, M.B. Qualitative data as an attractive nuisance: The problem of analysis. *Admin. Sci. Q.* **1979**, *24*, 590–601. [CrossRef]
35. Cohen, D.J.; Crabtree, B.F. Evaluative criteria for qualitative research in health care: controversies and recommendations. *Ann. Fam. Med.* **2008**, *6*, 331–339. [CrossRef] [PubMed]
36. Engeström, Y. *Learning by Expanding*; Cambridge University Press: Cambridge, UK, 2014.
37. Soranzo, A.; Cooksey, D. Testing taxonomies: Beyond card sorting. *Bull. Assoc. Inf. Sci. Technol.* **2015**, *41*, 34–39. [CrossRef]
38. Brooke, J. SUS-A quick and dirty usability scale. *Usability Eval. Ind.* **1996**, *189*, 4–7.
39. Boyatzis, R.E. *Transforming Qualitative Information: Thematic Analysis and Code Development*; Sage: Thousand Oaks, CA, USA, 1998.
40. Kolb, D. *Experiential Learning as the Science of Learning and Development*; Prentice Hall: Englewood Cliffs, NJ, USA, 1984; Available online: https://www.researchgate.net/publication/235701029_Experiential_Learning_Experience_As_The_Source_Of_Learning_And_Development (accessed on 23 May 2020).
41. Flavell, J.H. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *Am. Psychol.* **1979**, *34*, 906. [CrossRef]
42. Vygotsky, L.S. *Mind in Society: The Development of Higher Psychological Processes*; Harvard University Press: Cambridge, MA, USA, 1980.
43. Kaplan, A. *The Conduct of Inquiry: Methodology for Behavioural Science*; Routledge: Abingdon, UK, 2017.
44. Merton, R.K. *On Theoretical Sociology: Five Essays, Old and New*; Technical Report, Simon and Schuster: New York, NY, USA; 1967.
45. Knowles, M. *The Adult Learner: A Neglected Species*; Gulf Publishing: Houston, TX, USA, 1990.
46. Kolb, D.A.; Fry, R.E. *Toward an Applied Theory of Experiential Learning*; MIT Alfred P. Sloan School of Management: Cambridge, MA, USA, 1974.
47. Kolb, D.A.; Boyatzis, R.E.; Mainemelis, C. Experiential learning theory: Previous research and new directions in perspectives on thinking, learning, and cognitive styles (educational psychology series). *Perspect. Think. Learn. Cogn. Styles* **2001**, *1*, 227–247.
48. Boud, D.; Cohen, R.; Walker, D. *Using Experience for Learning*; McGraw-Hill Education (UK): New York, NY, USA, 1993.
49. Engeström, Y. *From Teams to Knots: Activity-Theoretical Studies of Collaboration and Learning at Work*; Cambridge University Press: Cambridge, UK, 2008.
50. Sannino, A. Activity theory as an activist and interventionist theory. *Theory Psychol.* **2011**, *21*, 571–597. [CrossRef]
51. Engeström, Y.; Sannino, A. Studies of expansive learning: Foundations, findings and future challenges. *Educ. Res. Rev.* **2010**, *5*, 1–24. [CrossRef]
52. Engeström, Y. Expansive learning at work: Toward an activity theoretical reconceptualization. *J. Educ. Work* **2001**, *14*, 133–156. [CrossRef]
53. Engestrom, Y. Innovative learning in work teams: Analyzing cycles of knowledge creation in practice. *Perspect. Act. Theory* **1999**, *377*, 404.

54. Livingston, J.A. Metacognition: An Overview. 2003. Available online: https://files.eric.ed.gov/fulltext/ED474273.pdf (accessed on 23 May 2020).
55. Devine, J. The role of metacognition in second language reading and writing. In *Reading in the Composition Classroom: Second Language Perspectives*; Heinle & Heinle Publishers: London, UK, 1993; pp. 105–127.
56. Brown, A. Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In *Metacognition, Motivation, and Understanding*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1987 .
57. Vandergrift, L.; Goh, C.C.; Mareschal, C.J.; Tafaghodtari, M.H. The metacognitive awareness listening questionnaire: Development and validation. *Lang. Learn.* **2006**, *56*, 431–462. [CrossRef]
58. Jacobs, J.E.; Paris, S.G. Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educ. Psychol.* **1987**, *22*, 255–278. [CrossRef]
59. Schraw, G. Promoting general metacognitive awareness. *Instruct. Sci.* **1998**, *26*, 113–125. [CrossRef]
60. Pressley, M.; Borkowski, J.G.; Schneider, W. Cognitive Strategies: Good Strategy Users Coordinate Metacognition and Knowledge. 1987. Available online: https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/4401/file/Schneider_W_Cognitive_strategies_Kopie.pdf (accessed on 23 May 2020).
61. Garner, R. When children and adults do not use learning strategies: Toward a theory of settings. *Rev. Educ. Res.* **1990**, *60*, 517–529. [CrossRef]
62. Dreyfus, S.E.; Dreyfus, H.L. *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*; Technical Report; California Univ. Berkeley Operations Research Center: Berkeley, CA, USA, 1980.
63. Holzkamp, K. We don't need no education. *Forum Kritische Psychol.* **1983**, *11*, 113–125.
64. Leont'ev, A. Activity, Consciousness, and Personality. 1978. Available online: http://lchc.ucsd.edu/mca/Paper/leontev/ (accessed on 23 May 2020).
65. Germanakos, P.; Fichte, L. EUREKA: Engineering usability research empirical knowledge and artifacts. In Proceedings of the International Conference on Learning and Collaboration Technologies, Las Vegas, NV, USA, 15–20 July 2018; pp. 85–103.
66. Laugwitz, B.; Held, T.; Schrepp, M. Construction and evaluation of a user experience questionnaire. In Proceedings of the Symposium of the Austrian HCI and Usability Engineering Group, Graz, Austria, 20–21 November 2008; pp. 63–76.
67. Brewer, M.B.; Crano, W.D. Research design and issues of validity. In *Handbook of Research Methods in Social and Personality Psychology*; Cambridge University Press: Cambridge, UK, 2000; pp. 3–16.
68. Commission, E. *Data Protection—Rules for the Protection of Personal Data Inside and Outside the EU*; European Commission Data Protection: Brussels, Belgium, 2018.
69. Braun, V.; Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [CrossRef]
70. Sauro, J.; Lewis, J.R. Correlations among prototypical usability metrics: Evidence for the construct of usability. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 4–9 April 2009; pp. 1609–1618.
71. Brooke, J. SUS: A retrospective. *J. Usability Stud.* **2013**, *8*, 29–40.
72. Lewis, J.J.R.; Sauro, J. Revisiting the factor structure of the System Usability Scale. *J. Usability Stud.* **2017**, *12*, 183–192.
73. Stufflebeam, D.L.; Shinkfield, A.J. An analysis of alternative approaches to evaluation. In *Systematic Evaluation*; Springer: Berlin, Germany, 1985; pp. 45–68.
74. Sauro, J. *Measuring Usability with the System Usability Scale (SUS)*; MeasuringU: Denver, CO, USA, 2011. Available online: https://measuringu.com/sus/ (accessed on 23 May 2020).
75. Reichheld, F.F. The one number you need to grow. *Harvard Bus. Rev.* **2003**, *81*, 46–55.
76. Lewis, J. *Predicting Net Promoter scores from System Usability Scale scores;* MeasuringU: Denver, CO, USA, 2011. Available online: https://measuringu.com/nps-sus/ (accessed on 23 May 2020).
77. Sauro, J. The challenges and opportunities of measuring the user experience. *J. Usability Stud.* **2016**, *12*, 1–7.
78. Schneider, D.; Berent, M.; Thomas, R.; Krosnick, J. Measuring customer satisfaction and loyalty: Improving the 'Net-Promoter'score. In Proceedings of the Annual Meeting of the American Association for Public Opinion Research, New Orleans, LO, USA, 15–18 May 2008.
79. Winne, P.H.; Perry, N.E. Measuring self-regulated learning. In *Handbook of Self-Regulation*; Elsevier: Amsterdam, The Netherlands, 2000; pp. 531–566.

80. Bangor, A.; Kortum, P.T.; Miller, J.T. An empirical evaluation of the system usability scale. *Int. J. Hum. Comput. Interact.* **2008**, *24*, 574–594. [CrossRef]
81. Landauer, T.K. Behavioral research methods in human-computer interaction. In *Handbook of Human-Computer Interaction*; Elsevier: Amsterdam, The Netherlands, 1997; pp. 203–227.
82. Borsci, S.; Federici, S.; Lauriola, M. On the dimensionality of the System Usability Scale: A test of alternative measurement models. *Cognit. Process.* **2009**, *10*, 193–197. [CrossRef] [PubMed]
83. Lewis, J.R.; Sauro, J. The factor structure of the system usability scale. In Proceedings of the International Conference on Human Centered Design, San Diego, CA, USA, 19–24 July 2009; pp. 94–103.
84. Stevens, J.P. *Applied Multivariate Statistics for the Social Sciences*; Routledge: Abingdon, UK, 2012.
85. Norman, D.A.; Draper, S.W. *User Centered System Design: New Perspectives on Human-Computer Interaction*; CRC Press: Boca Raton, FL, USA, 1986.
86. Ames, A.L. Users first! An introduction to usability and user-centered design and development for technical information and products. In Proceedings of the IEEE International Professional Communication Conference on Communication Dimensions, Sante Fe, NM, USA, 24–27 October 2001; pp. 135–140.
87. Nielsen, J. *Usability 101: Introduction to Usability*; Nielsen Norman Group: Fremont, CA, USA, 2012. Available online: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ (accessed on 23 May 2020).
88. Owen, R.; Brooks, L.L. *Answering the Ultimate Question: How Net Promoter Can Transform Your Business*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
89. Sauro, J. *Does Better Usability Increase Customer Loyalty?* MeasuringU: Denver, CO, USA, 2010. Available online: https://measuringu.com/usability-loyalty/ (accessed on 23 May 2020).
90. Javadi, M.; Zarea, K. Understanding thematic analysis and its pitfall. *Demo* **2016**, *1*, 33–39. [CrossRef]
91. Vaismoradi, M.; Turunen, H.; Bondas, T. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nurs. Health Sci.* **2013**, *15*, 398–405. [CrossRef] [PubMed]
92. Maguire, M.; Delahunt, B. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *AISHE-J* **2017**, *9*.
93. Pareto, V. *Cours D'économie Politique*; Librairie Droz: Geneva, Switzerland, 1964; Volume 1.
94. Cook, T.D.; Campbell, D.T. *Quasi-Experimentation: Design and Analysis for Field Settings*; Rand McNally: Chicago, IL, USA, 1979; Volume 3.
95. Nielsen, J. *User Satisfaction vs. Performance Metrics*; Nielsen Norman Group: Fremont, CA, USA, 2012. Available online: https://www.nngroup.com/articles/satisfaction-vs-performance-metrics/ (accessed on 23 May 2020).