International Journal of Human–Computer Interaction 39.1 (2023): 183-202

DOI: https://doi.org/10.1080/10447318.2022.2041885

A Supporting Tool for Enhancing User's Mental Model Elicitation and Decision-Making in User Experience Research

Marina Martín and José A. Macías (*)

Department of Computer Engineering, Universidad Autónoma de Madrid, Madrid, Spain.

(*) Corresponding author details:

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Tomás y Valiente 11 28049 Madrid. Spain

Tel. +34 91 497 6241

j.macias@uam.es

ORCID: 0000-0001-5071-0076

A Supporting Tool for Enhancing User's Mental Model Elicitation and Decision-Making in User Experience Research

User Experience (UX) research is intended to find insights and elicit applicable requirements to guide usable designs. Card Sorting is one of the most utilized methods. It is used to uncover the user's mental model and increase the usability of existing products. However, although Card Sorting has been widely utilized, most applications are based on spreadsheets. Furthermore, existing tools are principally intended to obtain qualitative information or customized quantitative outcomes to improve the information architecture. In this paper, a supporting tool based on the Card Sorting method is presented and detailed, including a comprehensive use case showing the main features. The tool implements predictive analysis of results through advanced statistics and machine learning techniques, providing comprehensive reports that enable evaluators and UX researchers to obtain high-level knowledge and important quantitative clues to enhance decision-making. The tool has been evaluated with participants and evaluators, obtaining relevant usability results and feedback.

Keywords: card sorting; usability evaluation; user-centered design; user experience, user research.

Subject classification codes: 43.040, 43.170.

1. Introduction

UX is rapidly gaining prominence today (Cayola & Macías, 2018; Veral & Macías, 2019). This paradigm is concerned with studying how people feel when they interact with a digital product or service. To this end, UX researchers use different methods (Farrell, 2017), being the Card Sorting (Spencer, 2009) one of the most utilized ones. Card Sorting represents a versatile method used in different development process phases for different purposes. Essentially, it requires users to sort cards into a set of provided categories or even enable them to create their new ones (Baxter et al., 2015). Card Sorting can be used in UX research and Design Thinking (Brown, 2008; Culén &

Gasparini, 2016) to explore the user's mental model (Farrell, 2017). In addition, it can be used as an evaluation method for improving the information architecture of interactive software or comparing different design solutions (Cayola & Macías, 2018; Macías et al., 2009; Rosenfeld et al., 2015; van Pinxteren et al., 2011a). In general, Card Sorting can be considered a reliable method to capture the user's mental model in conceptual research, allowing the detection of patterns and mental hierarchies of participants who perform the tests.

However, most Card Sorting tasks and analyses are still carried out manually, using homemade resources as predefined templates and custom spreadsheets (Macías, 2021; Spencer, 2009), resulting in poor analyses that limit effective decision-making. Also, a reduced set of commercial and academic tools exist. However, they provide specific data representations and run standard algorithms with customized parameters and settings, which intrinsically limits the utilization of alternative techniques and the gathering of enriched statistical outcomes beyond the information architecture. This reduces the expressivity of the quantitative analysis that can be improved by using predictive and advanced statistical facilities. Also, the mentioned tools do not usually support advanced help or shared knowledge for participants to effectively carry out sorting tasks and allow a more expressive elicitation of the user's mental model.

1.1. Research hypotheses

According to the arguments previously provided, the following research hypotheses are stated in order to carry out the research:

- H_1: The Card Sorting method and subsequent analysis are usually carried out manually. There are not many tools supporting advanced Card Sorting analysis and providing advanced knowledge to effectively guide evaluators (i.e., UX

researchers, usability engineers, and so on) in decision-making, as well as assisting participants to carry out sorting tasks in different ways to improve the elicitation of the user's mental model.

- H_2: It is possible to develop a supporting tool to help evaluators improve decision-making by providing enhanced reports through advanced statistical and predictive methods and assisting participants in the sorting tasks.
- H_3: The developed supporting tool reports acceptable values of usability for participants and evaluators.

1.2. Contribution

In order to verify the above hypotheses, this paper provides the contributions described below.

With the aim of corroborating H_1, a Systematic Literature Review (SLR) (Mengist et al., 2020) has been accomplished in order to study different academic approaches related to the advanced analysis of card-sorting data. Also, a competitive analysis of existing supporting tools has been accomplished. In general, the evidence found has helped corroborate that there are not many approaches related to the advanced acquisition of knowledge coming from Card Sorting, but only specific techniques, appearing on academic papers, that are used in isolation to solve specific problems on customized Card Sorting settings. As for commercial tools, most of them provide outcomes according to customized parameters and settings but not advanced statistical techniques, reducing the expressivity of the quantitative analysis for effective elicitation of the user's mental model. On the other hand, predictive techniques are rarely or not commonly used in existing supporting tools.

To corroborate H_2, a supporting tool, namely CAULDRON (interaCtive evAlUation tooL for aDvanced caRd sOrting aNalysis), has been developed. The tool

implements the Card Sorting method, allowing participants to sort cards and facilitating the interpretation of results by evaluators, providing advanced knowledge for effective decision-making. In addition to standard statistics and outcomes already existing in other approaches, such as general descriptive information, dendrograms, frequency analysis, classification matrices, and so on, CAULDRON includes the following advanced features that are inexistent or hard to find in existing supporting tools:

- A customized and quantitative version of the Delphi method (Reese et al., 2018b) has been implemented to share knowledge and help participants decide based on the classifications already made by other participants. This enriches the elicitation of the user's mental model in a group and sets up different experiments according to the evaluator criteria.
- An agreement measure, based on Fleiss' Kappa (Falotico & Quatto, 2015b), has been implemented to measure general and specific agreement to evaluate how participants have agreed to group the different cards into the corresponding categories.
- Correlation matrices have been implemented to study the correlation coefficients of both cards and categories.
- Multidimensional scaling (MDS), including the advanced Smacof algorithm (Borg et al., 2018), has been implemented to evaluate and visually represent similarities and differences of different groupings of cards and categories.
- Principal Component Analysis (PCA) (Jollife & Cadima, 2016b), together with K-means clustering (Paea & Baird, 2018a), have been combined and implemented in order to analyze, in a different way, groups of related cards and categories. The optimal number of clusters is automatically calculated by the tool.

- Heatmaps (Kassambara, 2018a; Macías, 2021) have been implemented to study differences and similarities among cards or categories. Also, dendrograms have been combined with heatmaps to show the classification matrix in an enriched way.
- Decision trees (Meyer-Baese & Schmid, 2014b) have been combined and implemented in order to analyze, in a different way, groups of related cards and categories. The optimal number of clusters is automatically calculated by the tool.
- An interactive consolidation board has been implemented in order for the researcher to create an ultimate Card Sorting. The tool automatically generates the board, considering the sorts carried out by participants. However, the researcher can conveniently modify this initial composition and export the results.

While existing approaches are more based on card analysis, the proposed approach reports on both card and category results, improving its analytical capability. Also, goodness indicators (Macías, 2021; Paea & Baird, 2018a), with corresponding benchmarks and tips to interpret them, have been included to assist the evaluator in interpreting results and carrying out effective decision-making. In addition, all the commented functionalities provide UX researchers with enhanced facilities to study the user's mental model and carry out custom evaluations.

In order to corroborate H_3, a user evaluation was carried out. To this end, evaluators and participants were recruited to evaluate the tool. The aim was twofold: evaluate the overall usability of the tool and identify specific problems to improve the tool further. In general, positive results were obtained from both evaluator and

participant assessments, indicating that the tool is usable according to the context of use stated.

This paper is structured as follows. Section 2 presents related work through an SLR and a comparative analysis of the existing tools. Section 3 reports on the tool, including design issues, main functionalities, a use case, and validation with experts. Section 4 addresses the usability evaluation of the tool, also showing the main results obtained. Finally, Section 5 reports on conclusions and future work.

2. Related work

User Experience is becoming popular in most domains, where the same paradigm can be found under different denominations (e.g., Client Experience, Driver Experience, etc.). In general, this paradigm is broader than the concept of usability itself (Borges & Macías, 2010; Castells & Macías, 2002; Veral & Macías, 2019), and it is closely related to the interaction of the user with a product or service. In this way, UX designers are devoted to investigating how to make the user's experience much more comfortable when using an interactive product or service (Lamprecht, 2020). To carry out this task, UX researchers utilize different methods to discover, explore, test, and listen to the user. Among them, exploratory methods, as is the case for Card Sorting, are the most utilized in user research (Foundation, 2020).

Card Sorting was initially used by psychologists (Wood & Wood, 2008) to study models and mental categorizations of their patients. Specifically, it appeared together with the Q methodology thanks to the physicist/psychologist William Stephenson around 1953 (Doubleday, 2013). Later on, it was used in the software domain, introduced in different books by experts such as Donna Spencer (Spencer, 2009) or Jacob Nielsen (Nielsen, 2004), who described its multiple variants and benefits for design and content categorization. In a nutshell, Card Sorting consists of sorting cards

labeled with meaningful terms and words into different categories. Card Sorting became popular to analyze information architecture of web applications, exploited by experts such as Rosenfeld and Morville (Rosenfeld et al., 2015), where three different variations coexist: open, closed, and hybrid Card Sorting. Open Card Sorting is mainly used at the beginning of a project, allowing participants to create their own categories and terminology to group contents. By contrast, in a Closed Card Sorting, the designer or evaluator provides the categories, and it is more beneficial to validate an existing set of categories and terms (Chaparro & Hinkle, 2008). Those variations provide different kinds of feedback in user research, being the hybrid approach advantageous when the information is partially incomplete, giving the user the freedom to create categories or select those created by the evaluator. In addition, the Delphi method can also be applied to Card Sorting. In this case, participants receive data and comments on results obtained from other participants (Doubleday, 2013), providing a different behavior and thus obtaining additional feedback in user research. As it can be seen, each of these variants can be useful in different phases of the development process, depending on the needs of both researchers and designers.

Once the Card Sorting tasks have finished, a second but essential step is analyzing the results for decision-making. In general, most of the Card Sorting analyses can be considered quantitative (Chaparro & Hinkle, 2008; Petrie et al., 2011; Righi et al., 2013; Wood & Wood, 2008). However, analyzing Card Sorting data is not an easy task, as it requires advanced statistical skills to interpret the results successfully (Macías, 2021). Most analyses are generally carried out manually or using basic statistics, which avoids obtaining advanced knowledge for effective decision-making. This way, specific statistic techniques, and supporting tools are needed.

The above concerns provide the primary motivation for this research and, in order to identify and analyze specific approaches and corroborate H1, two different activities were carried out. First, an SLR was performed in order to find academic-related works. Second, the most popular Card Sorting supporting tools were identified to carry out a comparative analysis and study existing strengths and weaknesses, classifying the statistics and data mining techniques they implement.

2.1. Systematic Literature Review

An SLR was carried out to find approaches dealing with advanced statistical and data mining techniques to analyze Card Sorting data. To carry out this task, the following research question was stated and used to help corroborate H_1:

RQ_1: What advanced statistical and data mining techniques are the most used in the context of Card Sorting?

To give an answer to the above research question, the following search string was composed:

((("card sorting" OR "card-sorting") AND ("analysis" OR "analytic" OR "analytical" OR "data" OR "mining" OR "outcome" OR "result" OR "presentation" OR "techniques" OR "categorization" OR "classification" OR "group")) AND ("information architecture" OR "usability" OR "software engineering"))

This search string was utilized in the following digital libraries to obtain related articles: Google Scholar, ACM Digital Library, IEEE Xplore, and SCOPUS.

Due to the large number of articles retrieved (see Table 1), a screening process was applied by considering inclusion and exclusion criteria. In this way, complete articles addressing the research issue stated, and written in the period 2010-2020 in the English language, were considered. By contrast, duplicated papers or those out of the inclusion criteria mentioned were excluded.

Table 1. The number of articles found and the final selection for each digital library.

Digital Library	**Articles Retrieved**	**Final Selection**
Google Scholar	253000	27
ACM Digital Library	2570811	15
IEEE Xplore	21	6
SCOPUS	746	32

Table 1 shows the initial number of papers obtained from each digital library using the aforementioned search string. Also, the final number of papers matching the inclusion criteria is included. As shown, the final number of selected articles has been largely reduced. This was mainly because most initial papers did not meet the specified requirements or were not helpful for the research. Those final papers were analyzed in detail in order to investigate the principal statistical and data mining technics used in the context of Card Sorting.

Once the papers were analyzed in detail, it was noticed that most of the existing literature is principally based on specific Card Sorting analysis using concrete datasets and statistical software (or even manual spreadsheets) in order to validate the results under concrete conditions. This implies that most of the techniques found have not been directly used in supporting tools but only as a proof of concept or application case. In addition, most of the techniques found are descriptive, whereas only a reduced number of approaches utilize predictive techniques to infer knowledge.

On the one hand, most of the descriptive analyses are intended to present Card Sorting data using concrete settings (Bayram et al., 2016; Kelley, Lee, et al., 2017; Kelley, Wilcox, et al., 2017). Cluster analysis is one of the most used methods to analyze Card Sorting data (Adamides et al., 2015; Ali et al., 2019; Ballweg et al., 2018; Cho et al., 2018; De Lima Salgado et al., 2019; Doubleday, 2013; El Said, 2014; Erol,

2018; Gatsou et al., 2012; Gonzalez-Zuniga & Carrabina, 2016; Goodman-Deane et al., 2008; Guo & Yan, 2011; Huang & Ku, 2016; Lantz et al., 2019; Lucci & Paternò, 2015; Maat & Lentz, 2011; Mesgari et al., 2019, 2015; Nurcahyanti & Suhardi, 2014; Paea & Baird, 2018b; Palmer & O'Neill, 2010; Petrie et al., 2011; Reese et al., 2018a; Robles et al., 2019; Roth, 2013; Sampson, 2005; Santos & Boticario, 2015; Schmettow & Sommer, 2016; Shen & Prior, 2013; Slegers & Donoso, 2012; Thomas & Johnson, 2013; Urrutia et al., 2017; Vashitz et al., 2013; Verhoeven & Gemert-Pijnen, 2010; J. Wentzel et al., 2016; Jobke Wentzel et al., 2016; Zainuddin & Staples, 2016)(Gabe-Thomas et al., 2016; Katsanos et al., 2019; Pisanski & Žumer, 2010; van Pinxteren et al., 2011b), including specific unsupervised clustering algorithms such as K-means (Gabe-Thomas et al., 2016; Huang & Ku, 2016; Paea & Baird, 2018b; Urrutia et al., 2017) to create cluster of cards for further analysis. Also related, Hierarchical Cluster Analysis is often utilized in Card Sorting to analyze card clusters through dendrograms (Baxter et al., 2015; Kaufman & Rousseeuw, 2005). In addition, other advanced statistical techniques can be found in the literature to quantitatively analyze Card Sorting data, such as distance analysis (Katsanos et al., 2019) using similarity matrices (Maida & Obwegeser, 2012). Some of them are represented by heatmaps (Schmettow & Sommer, 2016) to analyze relationships among cards visually. Frequency analysis is also utilized to analyze the frequency of classification of cards into different categories (El Said, 2014; Mahmood et al., 2018; Olaverri-Monreal et al., 2013; Rehring et al., 2020; Robles et al., 2019). In addition, correlations are also helpful to study the classification relationship among different cards (Kaufman & Rousseeuw, 2005). Other agreement statistics can be found in the existing literature, as is the case for Fleiss' Kappa (Beyer & Pinzger, 2014; Dubois et al., 2011; Eli et al., 2011; Mesgari et al., 2015; Nawaz et al., 2011; Nayebi et al., 2018; Scapin et al., 2011), which is principally

used for measuring the level of agreement among user classifications according to specific benchmarks (Falotico & Quatto, 2015a). Finally, advanced multivariate statistical techniques found are Principal Component Analysis (Jollife & Cadima, 2016a) and Factor Analysis (Maat & Lentz, 2011; Mesgari et al., 2019; Reese et al., 2018a), used to reduce the number of initial classification categories in open Card Sorting settings. Also, Multidimensional Scaling (Balloo et al., 2016; Lantz et al., 2019; van Pinxteren et al., 2011b), (Young, 1997), an alternative to Factor Analysis, is used principally used for representing similarly classified cards on a two or three-dimensional map. This becomes useful for visually analyzing differences and similarities among cards (Kassambara, 2018b). Even so, we can appreciate different versions of the Multidimensional Scaling such as the WMDS (Inukai & Kamisasa, 1974), NMDS (Diniz-Filho et al., 2013), or Smacof (Borg et al., 2018), which is one of the most utilized implementations.

On the other hand, predictive techniques were found in a reduced number of articles, although they can help predict values and increase knowledge (Bou-Hamad & Jamali, 2020; Meyer-Baese & Schmid, 2014a) for effective decision-making. In this sense, unsupervised methods such as Binary Decision Trees (Hepting & Almestadi, 2013) and Group Decision Making (Morente-Molinera et al., 2019) are found, principally applied to specific problems to predict card categorizations.

These findings allow answering the RQ_1, concluding that, in general, there are not many approaches related to the advanced acquisition of knowledge from Card Sorting. Instead, only traditional and specific techniques can be found in isolation in academic papers to solve specific problems on customized Card Sorting settings, and principally for dealing with cards (and not with categories, commonly). Moreover,

predictive techniques are rarely used, although they can be considered interesting for producing advanced knowledge for decision-making.

Most of the methods found have been considered in the implementation of CAULDRON, including Decision Trees. Those have been configured to produce practical and advanced knowledge for the evaluator to carry out effective decision-making.

2.2. Competitive analysis

In order to complement the SLR, the competitive analysis of the most representative Card Sorting tools was carried out. The objective was to map the principal techniques found in the previous sub-section into the existing supporting tools to analyze the kind of advanced analysis provided by each one. To carry out this task, the following research question was stated and used together with RQ_1 to help corroborate H_1:

RQ_2: What advanced statistical and data mining techniques are implemented in the most representative Card Sorting tools?

An additional search was carried out using the Google engine to find the most representative Card Sorting tools. The aim was to find supporting tools related explicitly to Card Sorting, selecting those more used or better rated in the Card Sorting community. In this way, the following representative tools were found: Proven By Users (ProvenByUsers, 2020), Optimal Workshop (Optimal Workshop, 2020), usabiliTEST (usabiliTEST, 2020), UserZoom (UserZoom, n.d.), and xSort (XSort, 2020). In addition, other supporting tools, coming from the academic domain, were found, such as Casolysis (Casolysis, n.d.) and WeCaso (WeCaSo, 2016). While WeCaso is used to carry out sorting tasks and export data, Casolysis is used to process such data using statistical analysis; in general, all the analyzed tools include expressive user interfaces to effectively carry out sorts and check results.

Table 2 summarizes the principal advanced analysis techniques included in the analyzed tools so far, compared with those provided by CAULDRON. Most tools provide primary analyses such as general statistics by sort and participant, dendrograms, frequency and classification matrix analyses. Other advanced statistical and data mining techniques are hard to find in existing tools, such as advanced classification based on card or category matrices, Fleiss' Kappa agreement, and multivariate statistics such as MDS and PCA. Also, other clustering strategies, such as K-means are less used in existing tools. As for predictive techniques, such as Decision Tree, they are not commonly used. As shown in Table 2, CAULDRON utilizes most of all mentioned techniques to improve decision-making, including the Delphi method to guide participants and obtain additional user feedback. Only SLA and DBSCAN are the two techniques not considered in CAULDRON, since they are redundant considering the other techniques included. On the one hand, SLA (Service Level Agreement) (Endmann et al., 2015) is used as a measure of agreement, whereas DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Endmann et al., 2015) is another kind of cluster analysis that may be reductant when implementing MDS, K-means and Hierarchical Cluster Analysis (dendrograms) at the same time.

Table 2. Comparison of the different statistical and data mining techniques included in each evaluated tool. Also, the proposed supporting tool CAULDRON has been included to compare the featured analysis techniques.

Analysis Technique	**Proven By Users**	**Optimal Workshop**	**usabilitiTEST**	**UserZoom**	**Casolysis**	**xSort**	**CAULDRON**
Advanced Classification	X	X	X				X
PCA		X					X

MDS		3D Cluster View	X		X		X
Advanced Frequency Analysis using Delphi							X
DBSCAN					X		
Heatmap					X		X
Advanced Agreement using Fleiss' Kappa							X
Correlation							X
SLA					X		
K-means							X
Decision Tree							X

These findings allow answering the *RQ2*, concluding that most existing tools, even the commercial ones, include primary statistical analyses mainly focused on descriptive data mining (frequency analysis, basic general statistics, and Hierarchical Cluster Analysis). On the other hand, advanced statistical techniques are hard to find in most existing tools. Predictive techniques are usually not included, even when all these techniques provide enriched information to predict and detect groups or patterns in the data obtained and carry out advanced analysis to improve decision-making.

Answers to RQ_1 and RQ_2 help corroborate H_1, concluding that, in most cases, practical Card Sorting analysis is carried out manually, using statistical software or even custom spreadsheets. In addition, most advanced statistical techniques are used in the academic domain. Also, there are not many tools supporting advanced analysis. According to such findings, CAULDRON was implemented by including all the most important advanced statistical and data mining techniques commented to effectively guide evaluators (i.e., UX researchers, usability engineers, and so on) in decision-making, also allowing to improve the elicitation of the user's mental model in UX research.

3. The proposal

In order to address the drawbacks mentioned above, a supporting tool has been developed that implements the Card Sorting method and provides advanced analysis. In this way, design clues will be presented with a use case and a validation with experts to corroborate H_2.

3.1. Design issues

CAULDRON is a dual language (i.e., English and Spanish) Card Sorting supporting tool featuring a responsive design. It has been developed using web technology to be used in any platform (computer, mobile, tablet, etc.) and operating system. Figure 1 shows the architectural components in CAULDRON. The tool has been implemented in Python following a client-server architecture. Django has been used as a web development framework, thus following an MVC (Model, View, and Controller) design pattern to facilitate code reuse and maintenance. Also, a PostgreSQL database system has been used to persist all the information. While the client-side (front-end) contains all the interactive parts of the tool, implemented with specific JavaScript libraries such

as JQuery, Bootstrap, and Popper, the server-side (back-end) implements all the most essential functionality.

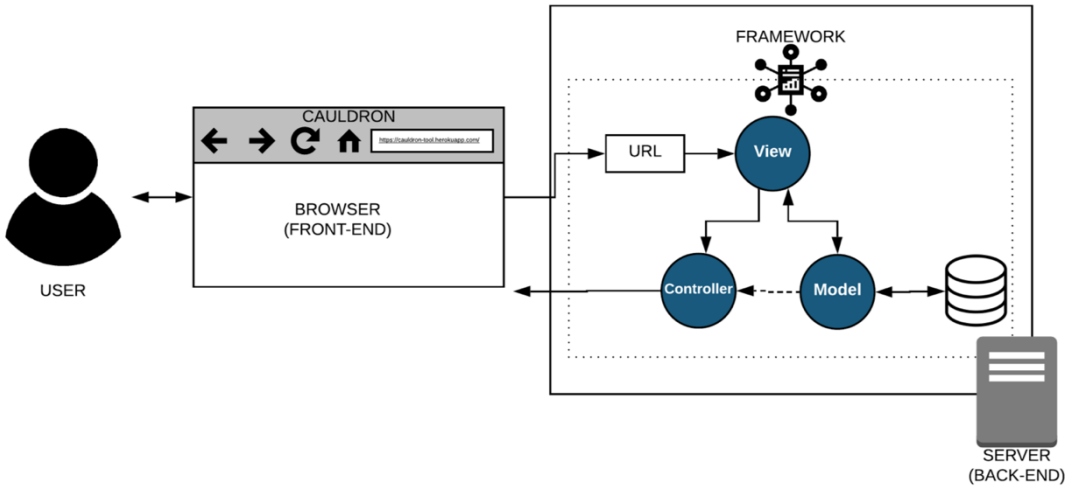

Figure 1. Architectural detail of CAULDRON, showing main components in a client-server deployment.

As for the main requirements, the implementation of CAULDRON has been divided into the following functional (Ruiz et al., 2021) subsystems:

- **User management subsystem**: It enables evaluators and participants to sign up and sign in. In addition, users can modify their personal information and recover the password.
- **Evaluator subsystem**: It enables evaluators to carry out main related tasks such as creating, modifying, removing, opening, and closing a Card Sorting evaluation, which can be open, closed, or hybrid, including the possibility to activate the Delphi facility for participants. Once a Card Sorting evaluation is created, a unique ID is generated in order for participants to participate in it. In addition, evaluators can track the participants' progress and have access to statistical reports generated by the tool. Also, a consolidation board has been implemented for evaluators to make a final decision (establish an ultimate Card

Sorting according to all participants' sorts). Likewise, the tool enables evaluators can save different information to disk.

- **Data analysis subsystem**: It is used by the evaluator subsystem in order to obtain statistical reports automatically generated by the tool. These reports are based on results obtained from applying advanced statistical and data mining techniques to a Card Sorting evaluation once it has been closed by the evaluator. Reports comprise diverse information to guide evaluators, including charts and goodness indicators that can be used for effective decision-making.
- **Participant subsystem**: It enables participants to carry out sorting tasks through an interactive sorting board. In this way, participants can only participate in an evaluation previously crated by the evaluator, using the provided ID. The Card Sorting board provides interactive functionalities for participants to carry out sorting tasks easily, including the possibility of having suggestions (obtained from other participants' sorts) as long as the evaluator has turned on the Delphi facility. Also, participants can save a current Card Sorting for being continued later.
- **Administrator subsystem**: This subsystem is used by the admin user, which is created by the framework and has access to all the information in the database, including evaluations and users.

In addition, non-functional requirements concerning usability have explicitly been included (Lee et al., 2021). To this end, a responsive aesthetical design has been considered to develop the tool, including other facilities to provide users with continuous help and guidance through tips, labels, and detailed information throughout the interaction.

3.2. Use Case

In order to illustrate the differential contribution of the tool -i.e., the statistical and data mining reports provided by CAULDRON, a use case, based on the results obtained through the user evaluation described in Section 4, is detailed. To this end, the different steps, together with the figures presenting the main results, are described below to indicate how the tool can be used to infer knowledge and carry out effective decision-making.

1. Let us suppose that a UX researcher wants to carry out user research to analyze how the users classify different internet applications by creating their own categories (open Card Sorting), thus eliciting information about the conceptual categories projected through the users' mental model, in addition to other relevant information about this. To carry out this task, the UX researcher signs up or signs in CAULDRON and creates a new open Card Sorting evaluation (obtaining the corresponding ID). This is shown in Figure 2, where the evaluation with ID "HKQQaaRG7b" has been created after configuring main evaluation characteristics, activating the Delphi facility, detailing help messages for participants, and finally creating the set of cards to use, which are the following (a total of 19): "Amazon Music", "Avast Antivirus", "Discord", "Dropbox", "Facebook", "Google Drive", "Google Play Music", "iCloud Drive", "Instagram", "McAfee", "Microsoft Teams", "Skype", "SoundCloud", "TeamViewer", "Telegram", "Twitter", "WhatsApp", "YouTube", and "Zoom".

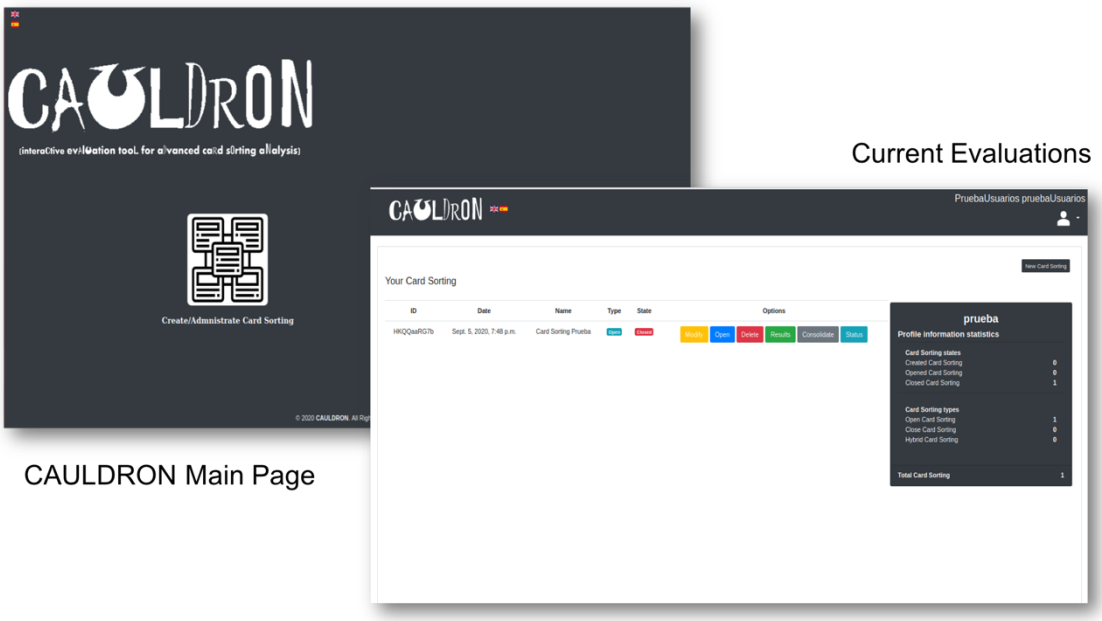

Figure 2. Main page with the access for evaluators (left) and the evaluations management screen (right) in CAULDRON, including the different options to deal with evaluations and showing information about participants and card sorts.

2. Once the Card Sorting evaluation has been created, participants can sign up or sign in CAULDRON to participate in a Card Sorting by introducing the ID provided by the UX researcher. Then, as shown in Figure 3, participants can carry out sorting tasks using the interactive sorting board, where the user can drag and drop the cards into categories or move them from one category into another. Categories can be created upon demand (open Card Sorting). In the meantime, the UX researcher can track the evolution of the evaluation using the "Status" button shown in Figure 2, obtaining the number of complete or remaining sorting tasks per participant. In addition, as the Delphi facility has been activated, a circled question mark appears in the cards that have reached at least 75% of classification in a given category. The idea is to suggest this category to other participants when they hover the mouse pointer over the mark. In the example shown in Figure 3, this mark appears in "Facebook", "Instagram", and

"Twitter" cards. This way, whenever participants hover the mouse over these marks, the category "REDES SOCIALES" (SOCIAL NETWORKS) is revealed in this case, denoting that those cards have been classified in the mentioned category by at least 75% of participants. This way, participants can use the suggested category if desired.

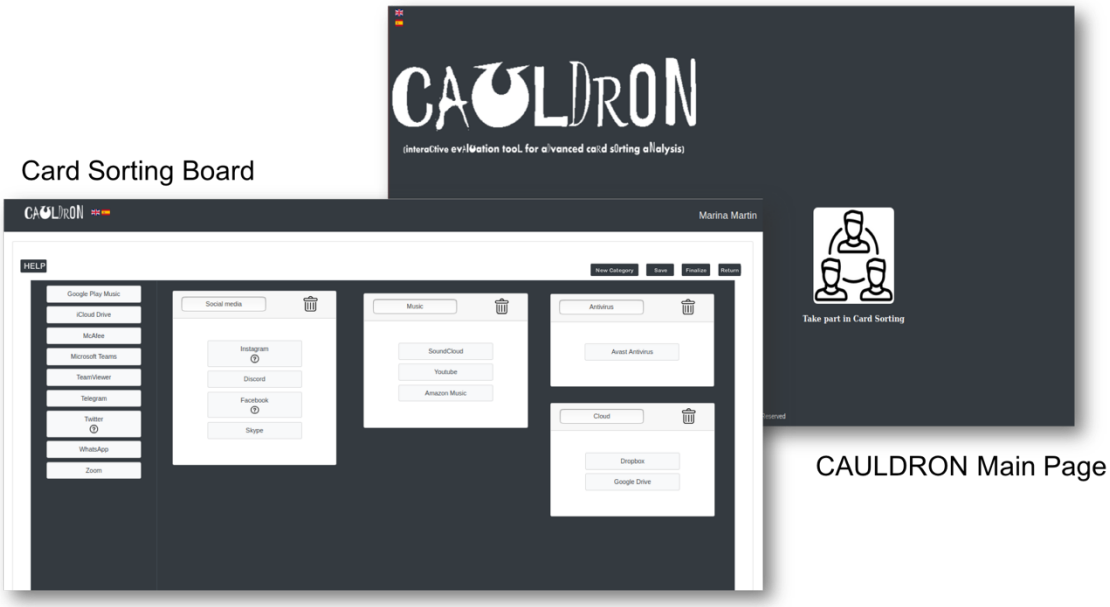

Figure 3. Main page with the access for participants (right) and the interactive Card Sorting board (left) for the participant's sorting tasks in CAULDRON.

3. Once all the Card Sorting participants have finished their sorting tasks, the UX researcher can close the Card Sorting and observe the results that are automatically generated by clicking on the "Results" button shown in Figure 2. This allows the UX researcher to observe basic statistical information such as the time consumed by each participant to carry out the sorting tasks, the number of categories created, etc. In this use case, the categories created by participants (a total of 17) were the following: "ALMACENAMIENTO EN NUBE" (CLOUD STORAGE), "Antivirus" (Antivirus), "ANTIVIRUS" (ANTIVIRUS), "APPS COMUNICACIÓN" (COMUNICATION APPS), "APPS DE MÚSICA" (MUSIC APPS), "ENTRETENIMIENTO"

(ENTERTAIMENT), “Mensajería” (Messaging), “Musica” (Music), “MÚSICA” (MUSIC), “Nube” (Cloud), “NUBE” (CLOUD), “Redes Sociales” (SOCIAL Networks), “REDES SOCIALES” (SOCIAL NETWORKS), “Reproduccion de contenido” (Content streaming), “Videoconferencias” (Videoconferences), “VIDEOCONFERENCIAS” (VIDEOCONFERENCES), and “Videos” (Videos). As can be seen, participants created different categories, but some of them represent the same concept, which is very common in open Card Sorting. In addition, the UX researcher has access to other advanced analyses through reports containing advanced statistics, goodness indicators and benchmarking, data mining, and interactive charts that can be exploited by using the mouse pointer to obtain further and detailed information, zoom in, or out and save them to disk. These analyses are the following:

3.1 A first analysis consists of observing the classification matrix and the agreement among participants. This information is shown in Figure 4, where the classification matrix also indicates the corresponding frequency, that is, the number of participants that have classified a card into a given category, where thick points indicate high frequencies. It is worth mentioning that this matrix is also used to implement the Delphi facility to suggest the categories that have obtained the highest rates equal to or over 75%. The consolidation facility is also based in this matrix to suggest a final classification to the UX researcher, as it will be explained later on. The Delphi facility is helpful to recommend categories created by other participants and thus reduce the total number of categories, reducing the dispersion in the classification matrix. This is seen in Figure 4, where cards such as “Facebook”, “Instagram”, and “Twitter” appearing in Figure 3 with the circled question mark, have been classified with less dispersion than others so far. This means that participants have used only two categories (“REDES SOCIALES” and “Redes Sociales”) to classify these cards, receiving the category

"REDES SOCIALES" a 75% of the classification. The Delphi facility is even more helpful in closed Card Sorting, where the evaluator proposes a fixed number of categories. This way, the UX researcher can configure different experiments to study the group's mental model (i.e., a group of experts in a particular domain) considering the collective feedback provided among participants. In addition, the Fleiss' Kappa is provided (see Figure 4) as a general value and benchmark and for each card (as it can be shown at the top of the matrix). This indicator is helpful to analyze the level of agreement among participants. This becomes much more useful in closed Card Sorting, as open ones produce more dispersion when participants create their own categories. In short, analyzing the information appearing in Figure 4, it is possible to affirm that there is a fair general agreement among participants, as they decided to create similar categories (e.g., "NUBE", "Nube", and "ALMACENAMIENTO EN NUBE") that conceptually mean the same thing (i.e., cloud) and thus classify the same cards.

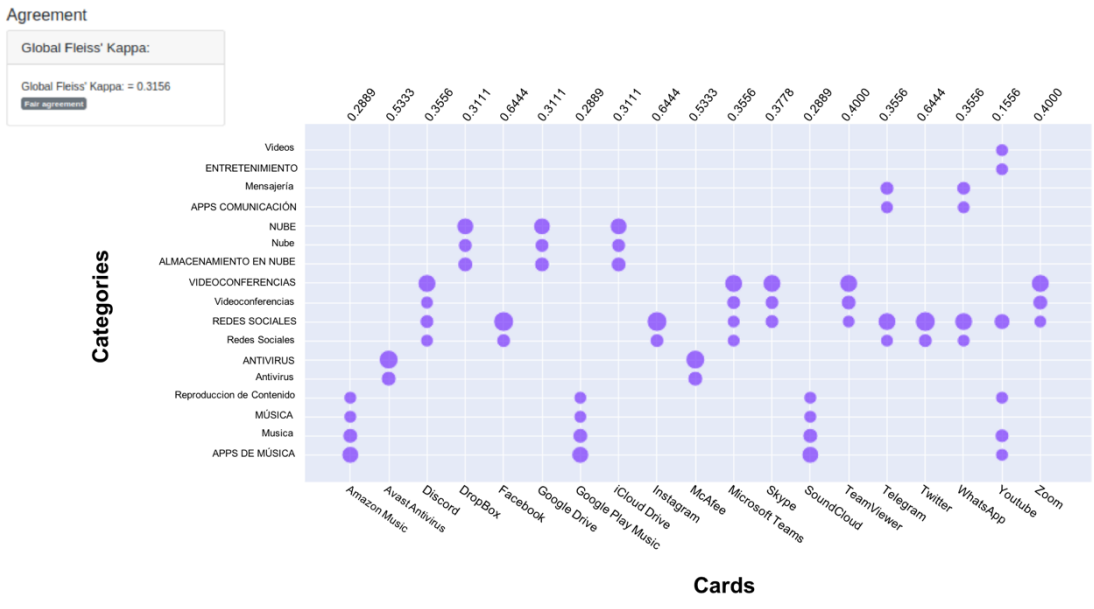

Figure 4. Frequency and agreement analysis.

3.2 Another analysis consists of studying the classification's big picture by combining heatmaps with dendrograms and showing the dissimilarity among items (cards or categories) also using heatmaps. Figure 5 shows this information. On the one hand, the heatmap at the left represents another view of the classification matrix denoting, where rectangles tend to be blue, a higher relationship (high frequency) between a card and the corresponding category. Also, the dendrograms provide a visual representation of the different groups that can be considered for cards and categories (i.e., predicted groups are represented in different colors). This visualization results more complete than dendrograms in isolation, as it is possible to observe relationships between cards and categories and, at the same time, cards and categories dendrograms, gathering all the information for a complete analysis. On the other hand, it is possible to also analyze the dissimilarity among cards or categories using heatmaps (see the two charts at the right in Figure 5). As for cards, a value closer to 0 (brown color) indicates that two cards can be considered as similar (less dissimilarity) as they have been classified into similar categories. Similarly, for categories, a value closer to 0 (brown color) indicates that two categories can be considered as similar (less dissimilarity) as they classify similar cards. These charts are helpful to quickly and visually study the relationship among items. For instance, analyzing the information provided in Figure 5 is possible to determine that some cards such as "Instagram", "Facebook", and "Twitter" are highly related, as they are classified as "REDES SOCIALES", and they are included in the same group as "YouTube", "Telegram" and "WhatsApp", according to the dendrogram of cards. Similarly, "APPS DE MÚSICA", and "Musica" can be considered as similar categories (they classify similar cards), and they are grouped in the dendrogram of categories. All this can be individually corroborated by the dissimilarity heatmaps as well.

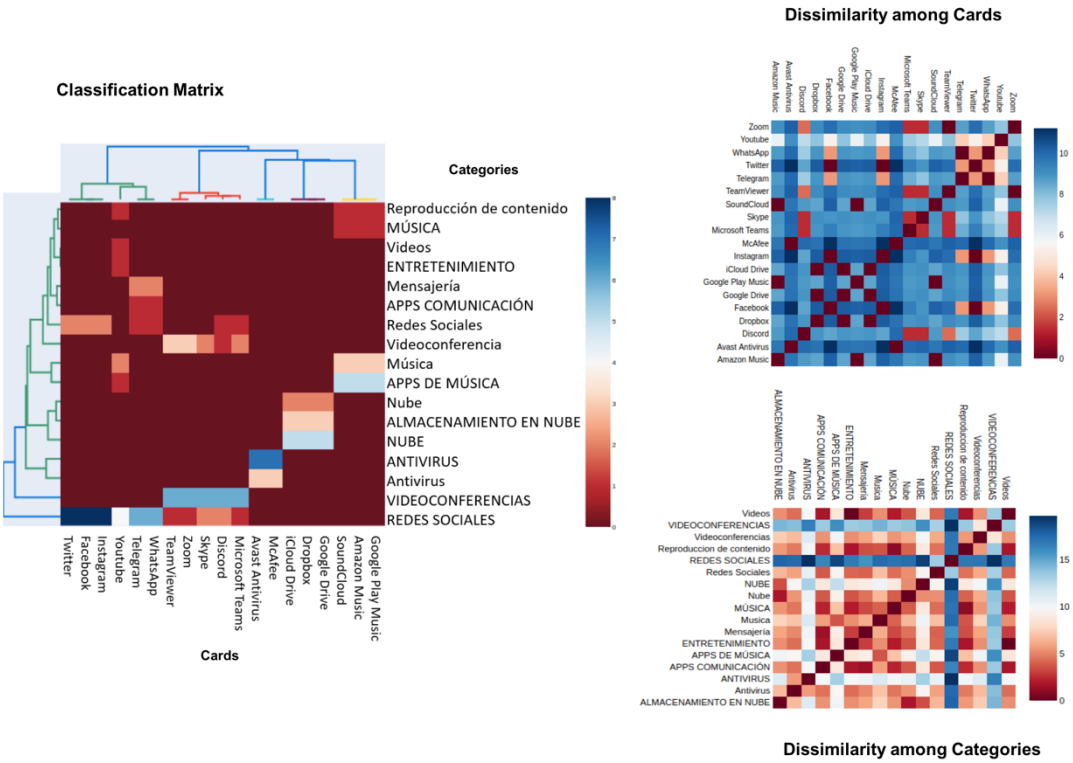

Figure 5. Heatmaps representing the classification matrix, including dendrograms to denote the different groups and the dissimilarity matrices for cards and categories.

3.3 Specific relationships among cards or categories can be found by using the correlation charts represented in Figure 6. This analysis enhances the information provided by the heatmaps, introducing the concept of linear relationship that can be useful to detect items that may be related (correlation closer to 1), no related (correlation closer to 0), or even inversely related (correlation closer to -1). This is the case for “MUSICA” and “REDES SOCIALES” categories, which have an opposite relationship. Similarly, “YouTube” and “Zoom” are cards with a close relationship, as shown in Figure 6.

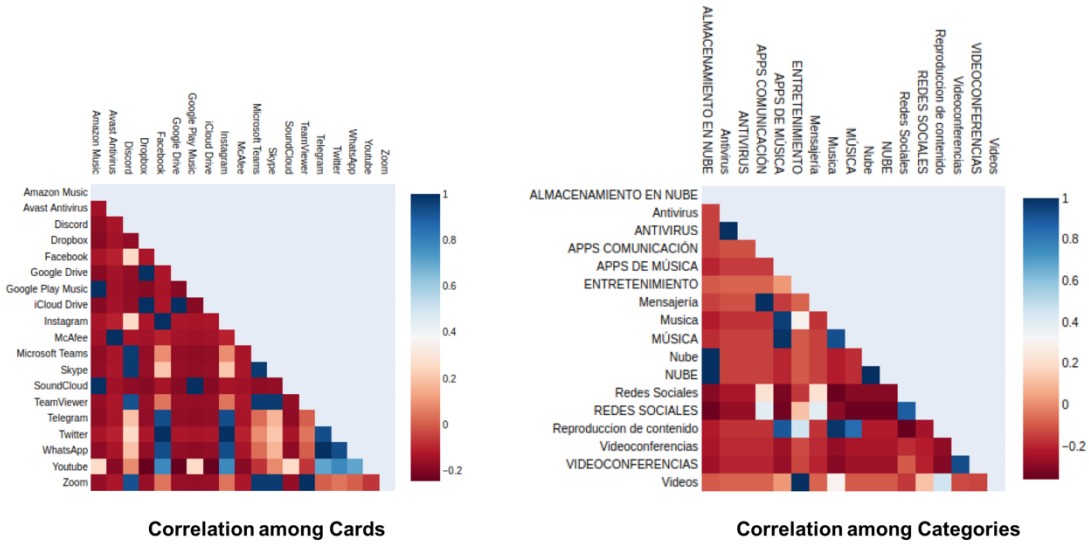

Figure 6. Charts representing correlations for cards (left) and categories (right).

3.4 Another analysis consists of a graphical representation of the items in a two-dimensional space to maximize the similarities and dissimilarities among cards or categories, allowing to visually identify groups of related items. A Multidimensional Scaling, precisely the Smacof approach, has been used to carry out this task. In Figure 7, the left scatterplot represents the groupings of cards. As shown, seven different groups of cards are clearly identified, which correspond to the expected groupings. Similarly, the right scatterplot represents groups of categories. In this case, the number of groups is not as precise as in the case of cards. However, some interesting groupings can be identified, such as the one corresponding to applications related to entertainment such as “MUSICA”, “Reproducción de Contenido”, “Vídeos”, and so on. Also, another group, which may be called “business software”, can be identified, including categories such as “Antivirus”, “Nube”, “Videoconferencia”, “ALMACENAMIENTO EN NUBE”, and so on. Finally, “REDES SOCIALES” appears apart, as it seems different from the rest. It is worth noting that categories such as “Videoconferencias”, and

"VIDEOCONFERENCIAS", which seem to be the same category, appear apart, as they have different distances according to the classifications carried out by participants, as it can also be appreciated in the dissimilarity heatmap. This is because although both categories classify the same cards, the rating from participants is different, as most participants have used "VIDEOCONFERENCIAS" instead of "Videoconferencias", giving a low rate and thus obtaining a higher distance between both categories. In general, the Multidimensional Scaling provides a strong visual tool that helps identify groups more rapidly than dendrograms, as it provides straight visual feedback and enables to identify near groups that could be related quickly. For instance, the group including cards related to social networks (such as "Facebook, "Instagram" and so on) is closer to the one containing cards related to videoconferencing (such as "Discord", "Microsoft Teams" and so on), as some participants decided to classify some of these cards into the former group, allowing the researcher to analyze the nature of the groups and the mental model of the participants related to the resulting classifications.

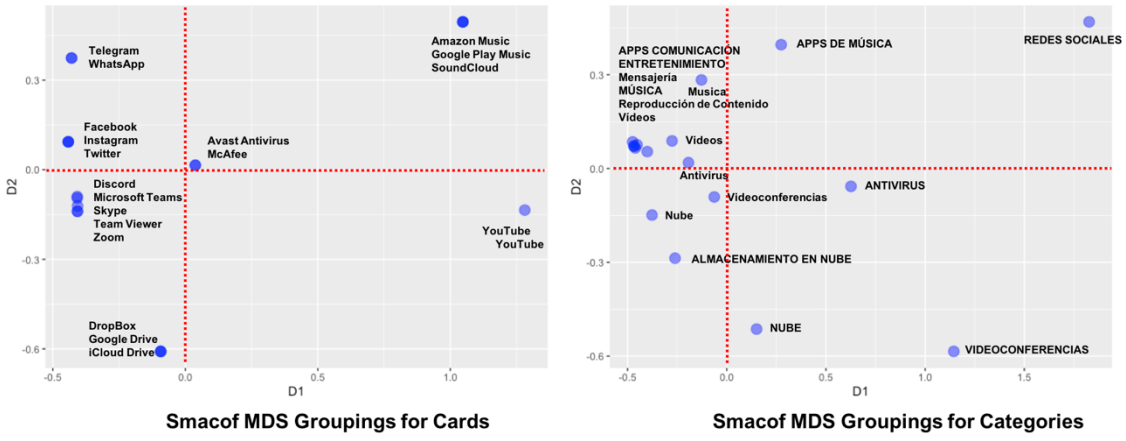

Figure 7. Scatterplots representing groups of cards (left) and categories (right) using the Smacof Multidimensional Scaling.

3.5 In addition to the Smacof Multidimensional Scaling, CAULDRON features different facilities to analyze clusters. One of them is the dendrogram, a tree diagram commonly based on an agglomerative clustering representation. However, dendrograms need to be analyzed in detail to study different clusters depending on one crucial parameter called *height*, which becomes cumbersome when the number of items is high. Another interesting strategy is the PCA (i.e., Principal Component Analysis) approach, where the clusters calculated with K-means are graphically depicted in two dimensions representing the two principal components that explain the majority of the variance. This has been implemented in CAULDRON. In this way, the tool automatically calculates the optimal number of clusters for cards or categories, although the UX researcher can change this value to study different cluster settings if desired (see the corresponding widgets on the top of Figure 8). The tool also computes specific goodness indicators to automatically carry out the best clustering possible. As shown in Figure 8, the optimal number of clusters for cards is 7, which corresponds to the number of groups already observed in Figure 7. As for categories, the optimal number of likely groups appears clearly than before with the Multidimensional Scaling. In this way, the optimal number of clusters is 7 for the case of categories. However, while the goodness indicator provides an optimal value for the case of cards, for the case of categories, this indicator is lower, indicating intermediate results, as it is difficult to separate the categories corresponding to cluster 7 (i.e., "APPS COMUNICACIÓN", "ENTRETENIMIENTO", "Mensajería", "MÚSICA", "Redes Sociales", "Reproduccion de contenido", and "Videos"), which includes categories of different nature. Other clusters appear clear, such as those representing music-related categories (“APPS DE MUSICA” and “MÚSICA”), or even those related to cloud technology (“Nube”, “NUBE” and “ALMACENAMIENTO EN NUBE”). This means that the categories

included in those clusters can be merged to reduce the total number. This clustering technique generally represents a more accurate approach than Multidimensional Scaling, principally based on dissimilarities.

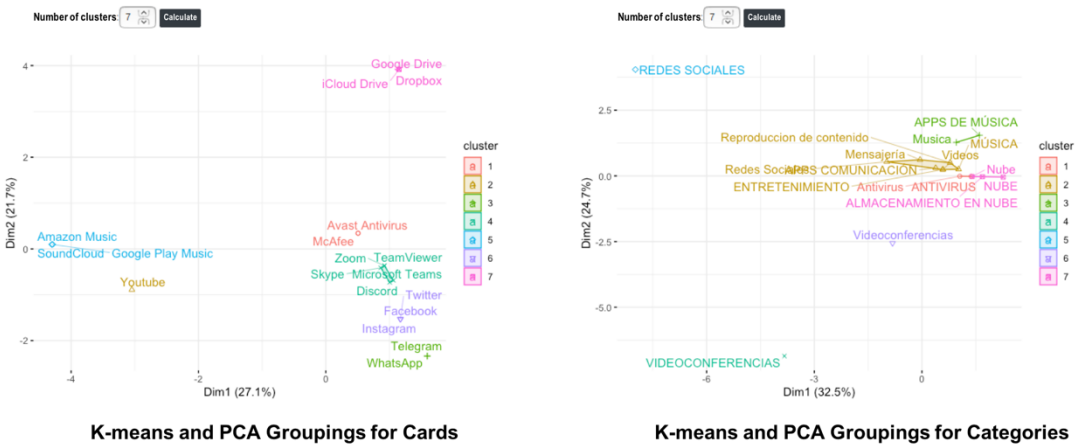

Figure 8. Clusters for cards (left) and categories (right) using K-means and PCA.

3.6 Another advanced analysis featured by CAULDRON is the predictive analysis of classifications using the participants' profiles. To this end, classification trees, which comprise a non-parametric supervised learning method, are used. The objective is to create a model that predicts the value of a target variable (the classification category) by learning simple decision rules inferred from the following variables related to the participants' profile obtained when they signed up: age, field (academic, professional, other), device used (mobile, tablet, computer), genre (female, male, other), years of experience with card sorting, years of experience with software development and evaluation, and the time spent to complete the sorting tasks. The idea is to obtain additional knowledge about how individual differences may affect the classification of the different cards. To carry out this task, CAULDRON generates a decision tree for each classified card, together with the corresponding help to guide

evaluators in interpreting results. Figure 9 shows two examples of decision trees generated, which correspond to “Instagram” and “Skype” cards. As shown for the case of “Instagram”, a classifier has been generated using two simple user variables: years of experience with software development and evaluation (experienceDevEv), and the time spent to complete the sorting tasks (time). The interpretation might be that regardless of experience, a percentage equal to or over 75% of participants have classified the “Instagram” card into the “REDES SOCIALES” category. In contrast, less than 25%, which spent more than 3.7 minutes in the classification, decided to classify the card into the “Redes Sociales” category. Both categories mean the same, but a distinction can be observed according to the use of capital letters, and this can may be related to the time spent by users. In the case of the “Skype” card, the years of experience (experienceDevEv), the gender (gender), and the time spent to carry out the sorting tasks (time) are the variables considered by the algorithm. This way, there is a first split related to the years of experience, where less experienced participants (60%) decided to classify the card as a videoconferencing application (“Videoconferencias” and “VIDEOCONFERENCIAS” categories). In contrast, more experienced participants (40%) considered “Skype” as a videoconferencing (those participants spending equal or less than 3.7 minutes in the classification) or a social network application (those participants spending more than 3.7 minutes in the classification). Also, for this case, it is observed that more experienced participants tended to use capital letters to codify categories (30% versus 10%), whereas fewer experienced users decided not to use capital letters (50% versus 10%). The gender variable is not relevant in this case. In general, the utilization of decision trees notably increases the analytical capability of the approach. The automatic creation of association rules allows finding temporal or causal relations, which help obtain further information from the sorting tasks and, in general,

help analyze specific patterns in the participant's mental model and behavior, helping carry out effective decision-making in UX research.

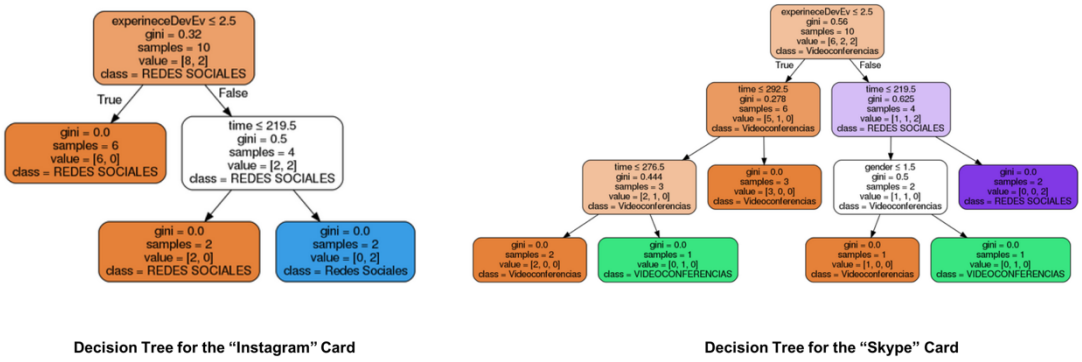

Figure 9. Decision trees generated for "Instagram" (left) and "Skype" (right) cards.

4. As shown, the variety of analyses provided enhances the elicitation of the user's mental model in this context, enabling effective decision-making. In addition, CAULDRON also provides other issues to facilitate the analysis to the researcher, such as replacing long item names with short synthetic labels or providing results related to categories (in addition to cards) to analyze a possible reduction. Finally, according to the results observed, the UX researcher can make a final decision by composing a consolidated and ultimate version of the Card Sorting. CAULDRON automatically provides the UX researcher with a preliminary version of the Card Sorting by considering the classification frequency previously described in Figure 4. Nevertheless, the researcher can modify this version and consolidate a final Card Sorting that can be exported to a CSV file. Figure 10 shows the consolidation boar, including an initial Card Sorting version where all the categories created by participants have been automatically included. Also, those cards having at least 75% of sorts into a single category have been automatically classified, as shown for the category "REDES SOCIALES", which includes the cards "Facebook", "Instagram", and "Twitter". The

researcher can use this board as a starting point to merge, remove and create new categories if desired. Also, cards can be moved from one category to another. This initial board provides additional knowledge of the users' mental model in order for the researcher to create the corresponding categories and carry through a consolidated version of the Card Sorting.

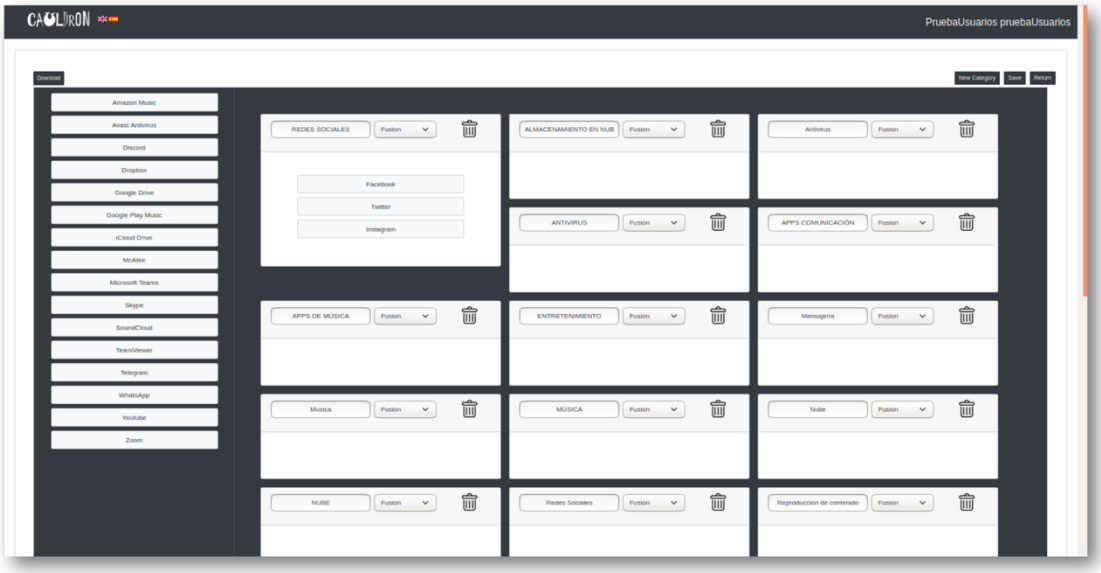

Figure 10. Initial consolidation board to help the UX researcher make a final decision on the Card Sorting.

All the clues reported in this section help corroborate H_2, concluding that it is possible to develop a supporting tool to help evaluators improve decision-making by providing enhanced reports through advanced statistical and predictive methods and assisting participants in the sorting tasks. This will be validated with experts down below.

3.3. Expert Validation

A validation with experts was carried out. The objective was to analyze the tool in-depth and obtain insights on which of the techniques implemented in the tool,

presented in the above use case, are better valued by experts, and compare CAULDRON with other tools that experts may have ever used to carry out Card Sorting analysis.

To carry out this task, we contacted 5 experts on Card Sorting. They are professionals in UX design and utilize Card Sorting in their daily problem-solving activities. In addition, we have utilized a publicly available dataset published on the Cardsorting.net web page (Cardsorting.net, 2014) as part of a tutorial (Blanchard et al., 2012). This dataset was directly loaded in our tool. It contains the results of a real Card Sorting involving different food items. In this open Card Sorting, 24 participants attempted to classify 40 cards into categories. The participants created 240 categories, but no normalization process was carried out. Thus, raw data represent the relationship between each card and the category into which it was classified. This setting becomes helpful for the purpose of our analysis, as experts must attempt to analyze the classifications and find relationships and knowledge using the techniques implemented in our tool.

The procedure carried out was the following:

a) A short introduction on the food Card Sorting and the CAULDRON tool was given to each expert (about 10 minutes).
b) Experts were asked to sign in as participants and carry out a classification to have a first contact with CAULDRON and the proposed Card Sorting dataset.
c) Experts were asked to sign in as evaluators and analyze the results of the proposed Card Sorting, investigating the different techniques provided by CAULDRON to understand the results and obtain insights about the sorting tasks performed.

d) Finally, we provided experts with a short survey in order to gather and analyze information about the following issues:

- Other Card-Sorting tools used by the experts.
- Strengths and weaknesses observed in CAULDRON, compared to other Card Sorting tools that experts may have ever used.
- Most valuable techniques in CAULDRON, also describing why they consider them as such.

Once the survey was analyzed, we found out that experts usually utilize other Card Sorting tools such as Proven by Users, Optimal Workshop, and UserZoom, analyzed in Section 2.2.

As for the strengths and weaknesses, on the one hand main strengths are related to the facility to observe the results using different advanced statistics and graphics of different kinds. Experts argued that although some of the visualizations can be seen as reductant, they are complementary and appreciated to observe the information from different perspectives and thus infer more precise knowledge about the user's mental model. On the other hand, weaknesses are related to the attractiveness of the user interface and the interaction when carrying out sorting tasks. In general, the user interface has been perceived as less elaborated than that provided by commercial tools. In addition, the analytics reported by commercial tools (showing the history of evaluations) also overcome the output reported by CAULDRON, as it only shows numerical information about the most essential concerns related to historical information about evaluations.

As for The CAULDRON's most valuable techniques, all the experts agreed that heatmaps combined with dendrograms, correlations, k-means, and decision trees, which are not implemented in the other tools that they use, report relevant information in order

to analyze and have advanced knowledge about the user's mental model. On the one hand, experts indicated that heatmaps help obtain a big picture of the classification to have a quick first analysis of the results. Also, combined with dendrograms, this visualization results more complete than dendrograms in isolation, as it is possible to observe relationships between cards and categories and, at the same time, have cards and categories dendrograms, gathering all the information for a more complete analysis. Also, correlations were valued by experts to detect items that may be related, not related, or even inversely related, complementing the information provided by the heatmaps. In addition, the visualization based on k-means combined with PCA, has been greatly valued by experts as this visualization provides more accurate information about clusters than dendrograms when the number of categories is high (as is the case for the Card Sorting analyzed by experts); experts also valued the goodness indicators in order to analyze the quality of the clusters. Finally, the decision tree is probably the most unexpected and valued visualization. All experts agree that decision trees provide valuable predictive information, obtaining additional knowledge about how individual differences may affect the classification of the different cards. This feature, which is not found in other Card Sorting tools, helps obtain further information from the sorting tasks and, in general, helps analyze specific patterns in the participant's mental model and behavior, helping carry out effective decision-making in UX research. Also, some experts reported on other interesting techniques not found in the tools that they use, such as the advanced agreement based on Fleiss' Kappa and the implementation of an assistance-based mechanism based on the Delphi method. By contrast, MDS representation was less valued due to the high number of cards and categories. In fact, MDS representation hinders the cluster visualization in this Card Sorting, being other

representations, such as k-means or heatmaps combined with dendrograms, perceived as more helpful for this specific case

In general, experts considered CAULDRON as a valuable tool, including advanced statistical and predictive techniques that are hard to find in most existing tools, even the commercial ones that include primary statistical analyses mainly focused on descriptive data mining (frequency analysis, basic general statistics, and Hierarchical Cluster Analysis). Also, experts confirmed that CAULDRON's techniques provide enriched information to predict and detect groups or patterns in the data obtained and carry out advanced analysis to improve decision-making.

4. Evaluation

In order to evaluate the proposal and thus corroborate H_3, a controlled evaluation with real users was carried out. The aim was to assess the usability of CAULDRON and discover principal flaws, also obtaining the perception of the users concerning the tool.

4.1. Method

Due to the pandemic conditions, the evaluation was carried out remotely. In this way, two different kinds of users were considered: evaluators and participants. Evaluators were in charge of creating a specific Card Sorting and checking the results, while participants were requested to participate in the Card Sorting created by the evaluators and thus carry out sorting tasks. Each evaluation session consisted of remotely contacting the users and providing them with a brief introduction on the purpose of the evaluation and the tasks to achieve (about 10 minutes). The Thinking Aloud protocol was used (Veral & Macías, 2019), asking the users to speak aloud to register all the comments and behavior to be analyzed later on. The time spent by each user to carry out the tasks was also registered. After completing the tasks, users were requested to fill in

the SUS questionnaire (de Castro & Macías, 2016) consisting of 10 questions that are evaluated on a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). In addition, users were requested to indicate both positive and negative issues about CAULDRON.

To perform the evaluation, evaluators were provided with the Card Sorting described in the previous section, whose specification is the following:

- Type: open Card Sorting.
- Delphi facility: activated.
- Cards (a total of 19): “Amazon Music”, “Avast Antivirus”, “Discord”, “Dropbox”, “Facebook”, “Google Drive”, “Google Play Music”, “iCloud Drive”, “Instagram”, “McAfee”, “Microsoft Teams”, “Skype”, “SoundCloud”, “TeamViewer”, “Telegram”, “Twitter”, “WhatsApp”, “YouTube”, and “Zoom”.

This way, the evaluation was carried out in three phases:

(1) All evaluators were requested to create the expected Card Sorting, which was the same in all cases, having (for this specific case), the same ID.
(2) The ID was distributed to all participants in order to complete the sorting tasks and the questionnaires.
(3) Evaluators were again contacted in order to check the results and complete the questionnaires.

The specific tasks will be described later on.

4.2. Variables

In order to conduct the evaluation, the following variables were considered:

- Quantitative variable:

    - Effectiveness: number of tasks successfully accomplished by users without support or help.
    - Efficiency: time spent by users to complete each task, represented in seconds.
    - Perceived usability: value between 0 and 100, representing the score obtained from the SUS questionnaire.
- Qualitative variables:
    - User behavior and observations obtained from the Thinking Aloud sessions.
    - Positive and negative issues obtained from the user's opinion (open questionnaire).

It was established a value of 80 as acceptance level for the perceived usability, as it represents a score rated as A (percentile from 90 onwards) in the SUS benchmarking (Lewis & Sauro, 2018).

4.3. Participants

A total of 20 users were recruited for the evaluation. They were different from those who participated in Section 3.3. More specifically, they were 10 evaluators and 10 participants. On the one hand, evaluators were software professionals having experience in evaluation and software development. They were 10 men aged between 21 and 24 (M= 22.0 SD = 0.817). On the other hand, participants were advanced computer science students and graduates with testing and software engineering skills. They were 6 men and 4 women aged between 20 and 22 (M= 21.6, SD= 0.699).

In general, this kind of evaluation works well with the proposed sample size, as long as users are selected according to the objective of the evaluation and the kind of problems expected found (Cayola & Macías, 2018; Veral & Macías, 2019). Therefore,

the sample (20 participants) is representative enough (Nielsen & Landauer, 1993) to find the most critical usability problems (over 97%) that will be taken into consideration to study the user perception and improve the approach further.

The evaluation was carried out in accordance with the recommendations of national and international ethics guidelines, i.e., the Código Deontológico del Psicólogo and the American Psychological Association. The study does not entail any invasive procedure, and it does not carry any risk to the participants' mental or physical health, thus not requiring ethics approval according to the Spanish law BOE 14/2007. All subjects participated voluntarily and gave written informed consent in accordance with the Declaration of Helsinki. They were free to leave the evaluation at any time.

4.4. Apparatus

In order to carry out the evaluation, users were provided with the URL of the tool. As CAULDRON is a responsive tool, it automatically adapts to any platform. However, the recommended device was a laptop or a desktop computer to consider a standard environment and easy access to the developed functionality. More specifically, 55% of the users utilized a laptop for remote evaluation, and 45% utilized a desktop computer. As for the remote applications used to contact users, the most used were Microsoft Teams (35%), Skype (30%), and Discord (25%). Furthermore, the most used web navigator was Google Chrome, used by 75% of users.

4.5. Tasks

Different tasks for evaluators and participants were proposed to stablish the specific context to assess the usability of the tool.

Evaluators were requested to achieve the following tasks in sequential order:

- Sign up (ET_1). To perform this task, evaluators had to access the tool through the link provided, select the registration option, and introduce the required data.
- Log in (ET_2): Once registered, evaluators had to log into the tool by introducing the requested information (login and password) to verify their identity.
- Create a closed Card Sorting (ET_3): Once logged in, evaluators had to create a new Card Sorting by filling in all the required forms related to general settings, involved cards and categories, and the different help messages for participants. The set of cards was provided to evaluators, but the categories, in this case, were not specified, so evaluators were asked to introduce the categories that they desired.
- Modify a created Card Sorting (ET_4): Once the Card Sorting was created, evaluators had to modify it to transform it into an open Card Sorting (so, the categories created remain unused).
- Open the Card Sorting (ET_5): Evaluators had to open the created Card Sorting in order for participants to carry out the sorting tasks.
- Log out (ET_6): Evaluators had to log out from the tool.
- Log in (ET_7): Evaluators had to log in using the credentials created in task ET_1. This task was carried out once the participants had finished their sorting tasks.
- Check results (ET_8): Evaluators had to access the Card Sorting results and check the different statistics and data mining charts generated.
- Consolidate a final Card Sorting (ET_9): Evaluators had to consolidate a final Card Sorting using the interactive board and download the final version.
- Log out (ET_{10}): Finally, evaluators had to log out from the tool.

As previously commented, the evaluator assessment was carried out in two phases, first creating the Card Sorting (task ET_1 to ET_6) and then checking the results (tasks ET_7 to ET_{10}) obtained from the participants' sorting tasks accomplished.

As for participants, they were requested to achieve the following tasks in sequential order:

- Sign up (PT_1): Participants had to register in the tool, using the same procedure as for the case of evaluators.
- Log in (PT_2): Once registered, participants had to log also into the tool by introducing the requested information (login and password) to verify their identity.
- Participate in a Car Sorting (PT_3): To carry out this task, participants were provided with the ID of the Card Sorting previously created by the evaluators. This way, participants had to add this Card Sorting to their main panel.
- Carry out a Card Sorting (PT_4): Once the Card Sorting was added, participants had to carry out the sorting tasks using the interactive sorting board. As previously commented, the 19 cards were created by the evaluator, but participants had the freedom to create the desired categories (open Card Sorting) according to their own mental model.
- Finish and log out (PT_5): Participants had to close the Card sorting and log out from the tool once they finished their sorting tasks.

4.6. Analysis of Results

In general, the results obtained can be considered good enough to provide evidence about the usability of the tool.

As for effectiveness, average values of 89% (SD= 24.24%) and 90% (SD=25%) were obtained from the evaluator and the participant assessments, respectively. In the case of evaluators, the most problematic tasks were ET_5 and ET_8, which are related to opening the Card Sorting and checking the final results, where some evaluators needed some advice to find the right button to carry on. As for participants, task PT_4 was the most complex one. This was because participants were requested to carry out the (whole) Card Sorting, and some help was needed in some cases to guide the participants about the steps to follow to have all the sorting tasks finished.

As for efficiency, Table 3 shows a summary of the principal results obtained for the case of the evaluator assessment. In general, there were no significant problems with the tasks to accomplish. As shown, confidence interval values are lower than one minute in most cases, denoting that average time values are quite representative of what it would take an evaluator to complete the tasks. As expected, PT_1, PT_8, and PT_9 took longer as these tasks are related to the initial creation of the Card Sorting, the visualization and checking of results, and the consolidation of the final Card Sorting, respectively. However, it is worth mentioning some specific problems that arose in tasks PT_4 and PT_8, as some evaluators were confused about the right button to press to carry through the tasks so that these buttons will be revised.

Table 3. Efficiency results in seconds obtained from the evaluator assessment, where mean, min, max, standard deviation, median, and 95% confidence interval values for each task are shown.

	ET_1	**ET_2**	**ET_3**	**ET_4**	**ET_5**	**ET_6**	**ET_7**	**ET_8**	**ET_9**	**ET_{10}**
Mean	110.33	6.40	7.11	13.84	9.99	5.72	4.97	717.52	136.50	4.66
Min	70.80	1.00	5.21	4.04	1.19	2.30	3.21	440.40	6.04	3.10
Max	660.00	17.86	10.44	59.02	49.53	10.20	12.94	1141.80	321.60	7.20

SD	180.82	4.93	1.86	17.25	15.13	2.57	2.98	333.37	123.33	1.23
Median	97.50	7.20	7.43	13.98	12.13	6.05	4.22	756.90	261.00	4.53
CI (95%)	125.30	3.42	1.29	11.95	10.48	1.78	2.07	231.01	85.49	0.85

Table 4 shows the efficiency results obtained for the case of the participant assessment. No significant problems were found in this case either. Also, confidence interval values were narrow, denoting a precise estimation of the mean values. As expected, PT_4 took the longest, as this task is related to the main sorting activities that participants had to accomplish to carry out a Card Sorting. This was the most complex task, and some misunderstandings related to the tool's buttons arose. In particular, users did not fully understand the difference between the buttons "Save" and "Finalize" when finishing and submitting a Card Sorting. Therefore, these buttons will be revised.

Table 4. Efficiency results in seconds obtained from the participant assessment, where mean, min, max, standard deviation, median, and 95% confidence interval values for each task are shown.

	PT_1	**PT_2**	**PT_3**	**PT_4**	**PT_5**
Mean	84.00	12.77	66.52	286.46	6.92
Min	80.40	7.02	45.30	195.80	3.00
Max	87.60	31.72	90.00	390.00	22.03
SD	3.80	7.21	14.59	55.57	5.57
Median	84.00	10.49	66.60	276.90	5.40
CI (95%)	2.35	4.47	9.04	35.06	3.45

As for perceived usability, results were also promising. Figure 11 shows the SUS score obtained from evaluator (left) and participant (right) assessments. As shown,

participant scores depict higher values, mainly distributed in a narrow interval over 90, thus denoting high perceived usability. By contrast, evaluator scores are largely distributed, even with a relatively low min value. This denotes a more variated perceived usability from evaluators, although the median values were also high to be considered acceptable. In general, this difference can be because evaluators had to accomplish a higher number of tasks, so the interaction with the tool was more intensive and more critical according to the number of functionalities addressed. On the other hand, mean values obtained for the evaluator and participant assessments were 93.75 (SD=6.15) and 84.5 (SD=11.47), respectively, denoting high rates of perceived usability. Indeed, these values largely overcome the acceptance level stated, which was established in 80.

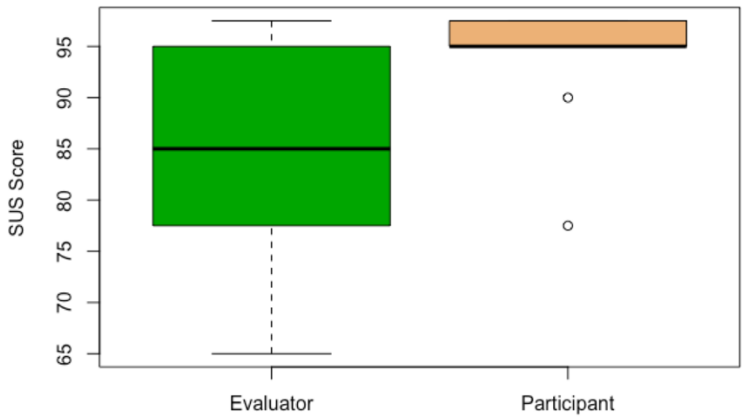

Figure 11. SUS scores obtained from the evaluator and participant assessments.

As for the qualitative information drawn from the evaluations (Thinking Aloud, positive and negative issues), users generally perceived the tool as comfortable, intuitive, and easy to use. Participants appreciated the Delphi facility in order to have suggestions about classifications. Also, evaluators denoted the quality of both graphics

and information provided in the statistical analysis, being the statistical calculations and the consolidation board perceived as helpful overall. On the other hand, some users suggested to speed up the statistical calculations, clarify the meaning of some buttons (such as “Save” and “Finalize”), improve the dialogs in some cases, increase the quantity of information about the statistical metrics, allow customizing the user interface’s colors, improve the color of the logo and improve the drag-and-drop mechanism in the interactive sorting board. All these issues will be revised in the future.

All in all, and according to the evidence obtained, it is possible to corroborate H3, concluding that the developed supporting tool reports acceptable values of general usability for both roles: participant and evaluator.

5. Conclusion

UX is becoming commonplace in most software developments today. Most companies are interested in designing high-quality interactive products to gain market share and remain competitive (Macías & Castells, 2002, 2001; Quintal & Macías, 2021). In this sense, specific methods should be utilized to make UX effective and, more specifically, accomplish convenient UX research to obtain initial clues for a successful design (Sánchez & Macías, 2019). Card Sorting has proved to be a suitable method to carry out user research in a UX strategy.

In general, there is a need for functional supporting tools (Macías, 2008) to automate the methods related to UX research. As demonstrated by the bibliographical study, although some supporting tools exist, most Card Sorting analyses are still carried out manually. In addition, existing tools primarily provide basic information using typical visualizations mainly intended to improve the information architecture, which restricts the capacity of further exploring the user’s mental model in detail during the user research.

In this paper, CALUDRON (interaCtive evAlUation tooL for aDvanced caRd sOrting aNalysis) has been presented. It comprises a supporting tool for advanced Card Sorting analysis. The tool features predictive analysis of results through advanced statistics and data mining, providing comprehensive reports that enable evaluators, UX researchers, and usability engineers to obtain high-level knowledge and important quantitative clues to enhance user's mental model elicitation and decision-making. Information about design, a detailed use case, a validation with experts, and an evaluation with real users have been comprehensively presented to detail the work. This helped corroborate the main hypotheses stated, reporting also good results in usability.

In general, the approach presented is intended to be an alternative to other existing tools, utilizing Card Sorting as an advanced method to carry out user research. This helps to provide not only basic information about sorting tasks and information architecture but advanced details to go beyond and analyze high-level information. This can be obtained thanks to the advanced visualization, statistical, and machine learning techniques implemented.

As for future work is expected to improve the tool with the issues found during the user evaluation. Also, it is expected to include new analyses principally based on machine learning techniques to obtain helpful information for user research. In this way, new user evaluations will be performed in the future, replicating the evaluation carried out.

Acknowledgements

This work was partially supported by the Spanish Government under grant number RTI2018-095255-B-I00]; and the Madrid Research Council under Grant number P2018/TCS-4314].

Disclosure Statement: No potential competing interest was reported by the authors.

References

Adamides, G., Christou, G., Katsanos, C., Xenos, M., & Hadzilacos, T. (2015).

Usability guidelines for the design of robot teleoperation: A taxonomy. *IEEE Transactions on Human-Machine Systems*, *45*(2), 256–262. https://doi.org/10.1109/THMS.2014.2371048

Ali, A. El, Ashby, L., Webb, A. M., Zwitser, R., & Cesar, P. (2019). Uncovering perceived identification accuracy of in-vehicle biometric sensing. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, 327–334. https://doi.org/10.1145/3349263.3351506

Balloo, K., Pauli, R., & Worrell, M. (2016). Individual Differences in Psychology Undergraduates' Development of Research Methods Knowledge and Skills. *Procedia - Social and Behavioral Sciences*, *217*, 790–800. https://doi.org/10.1016/j.sbspro.2016.02.147

Ballweg, K., Pohl, M., Wallner, G., & von Landesberger, T. (2018). Visual similarity perception of directed acyclic graphs: A study on influencing factors. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10692*, 241–255. https://doi.org/10.1007/978-3-319-73915-1_20

Baxter, K., Courage, C., & Caine, K. (2015). Card Sorting. *Understanding Your Users*, 309–322. https://doi.org/10.1016/b978-0-12-800232-2.00011-0

Bayram, K., Yıldızlı, H., & Saban, A. (2016). Determining Preservice Teachers' Goal Orientations for Learning through Card Sorting Activity. *International Journal of Learning and Teaching*, *8*(1), 40. https://doi.org/10.18844/ijlt.v8i1.522

Beyer, S., & Pinzger, M. (2014). A manual categorization of android app development issues on stack overflow. *Proceedings - 30th International Conference on Software Maintenance and Evolution, ICSME 2014*, 531–535. https://doi.org/10.1109/ICSME.2014.88

Blanchard, S. J., Aloise, D., & DeSarbo, W. S. (2012). The Heterogeneous P-Median Problem for Categorization Based Clustering. *Psychometrika*, *77*(4), 741–762. https://doi.org/10.1007/s11336-012-9283-3

Borg, I., Groenen, P. J. F., & Mair, P. (2018). *Applied Multidimensional Scaling and Unfolding*. Springer International Publishing. http://www.springer.com/series/8921

Borges, C. R., & Macías, J. A.-E. U. A. de M. (2010). Feasible database querying using a visual end-user approach. *Proceedings of the 2nd ACM SIGCHI Symposium on Engineering Interactive Computing Systems - EICS '10*, 187–192. https://doi.org/10.1145/1822018.1822047

Bou-Hamad, I., & Jamali, I. (2020). Forecasting financial time-series using data mining models: A simulation study. *Research in International Business and Finance*, *51*, 101072. https://doi.org/10.1016/j.ribaf.2019.101072

Brown, T. (2008). Design thinking. *Harvard Business Review*, *86*(6).

Cardsorting.net. (2014). *Cardsorting.net*. http://cardsorting.net

Casolysis. (n.d.). *Mensch-Computer-Interaktion, Softwaretechnologie - Universität Paderborn*. Retrieved July 11, 2021, from https://cs.uni-paderborn.de/mci/

Castells, P., & Macías, J. A. (2002). Un sistema de presentación dinámica hipermedia para representaciones personalizadas del conocimiento. *INTELIGENCIA ARTIFICIAL*, *6*(16). https://doi.org/10.4114/ia.v6i16.738

Cayola, L., & Macías, J. A. (2018). Systematic guidance on usability methods in user-centered software development. *Information and Software Technology*, *97*, 163–175. https://doi.org/10.1016/j.infsof.2018.01.010

Chaparro, B. S., & Hinkle, V. D. (2008). Card-Sorting: What You Need to Know about Analyzing and Interpreting Card Sorting Results. *Usability News*, *10*(2), 1–6.

Cho, H., Yen, P. Y., Dowding, D., Merrill, J. A., & Schnall, R. (2018). A multi-level usability evaluation of mobile health applications: A case study. *Journal of Biomedical Informatics*, *86*, 79–89. https://doi.org/10.1016/j.jbi.2018.08.012

Culén, A. L., & Gasparini, A. A. (2016). Design Thinking Processes: Card Methodologies for Non-designerse. In P. Minaříková & L. Z. Suchá (Eds.), *Librarians as Designers. Case studies on the improvment of library services* (pp. 73–85). Masarykova Univerzita.

de Castro, A., & Macías, J. A. (2016). SUSApp. *Proceedings of the XVII International Conference on Human Computer Interaction*, 1–8. https://doi.org/10.1145/2998626.2998667

De Lima Salgado, A., Dias, F. S., Mattos, J. P. R., De Mattos Fortes, R. P., & Hung, P.

C. K. (2019). Smart toys and children's privacy: Usable privacy policy insights from a card sorting experiment. *SIGDOC 2019 - Proceedings of the 37th ACM International Conference on the Design of Communication*, 1–8. https://doi.org/10.1145/3328020.3353951

Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., Telles, M. P. de C., Rangel, T. F., & Bini, L. M. (2013). Mantel test in population genetics. *Genetics and Molecular Biology*, *36*(4), 475–485. https://doi.org/10.1590/S1415-47572013000400002

Doubleday, A. (2013). Use of card sorting for online course site organization within an integrated science curriculum. *Journal of Usability Studies*, *8*(2), 41–54.

Dubois, E., Bortolaso, C., Bach, C., Duranthon, F., & Maumont, A. B. (2011). Design and evaluation of mixed interactive museographic exhibits. *International Journal of Arts and Technology*, *4*(4), 408. https://doi.org/10.1504/IJART.2011.043441

El Said, G. R. (2014). Card sorting assessing user attitude in E-learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8523 LNCS*(PART 1), 261–272. https://doi.org/10.1007/978-3-319-07482-5_25

Eli, J. A., Mohr-Schroeder, M. J., & Lee, C. W. (2011). Exploring mathematical connections of prospective middle-grades teachers through card-sorting tasks. *Mathematics Education Research Journal*, *23*(3), 297–319. https://doi.org/10.1007/s13394-011-0017-0

Endmann, A., Fischer, H., & Krökel, M. (2015). Mensch und Computer 2015 - Usability Professionals. In A. Endmann, H. Fischer, & M. Krökel (Eds.), *Mensch und Computer 2015 - Usability Professionals*. DE GRUYTER. https://doi.org/10.1515/9783110443882

Erol, T. (2018). Dimensions of Holistic Automotive Seat Comfort Experience: A Card Sorting Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1007–1011. https://doi.org/10.1177/1541931218621232

Falotico, R., & Quatto, P. (2015a). Fleiss' kappa statistic without paradoxes. *Quality and Quantity*, *49*(2), 463–470. https://doi.org/10.1007/s11135-014-0003-1

Falotico, R., & Quatto, P. (2015b). Fleiss' kappa statistic without paradoxes. *Quality*

and Quantity, *49*(2), 463–470. https://doi.org/10.1007/s11135-014-0003-1

Farrell, S. (2017). UX Research Cheat Sheet. In *Nielsen Norman Group*. https://www.nngroup.com/articles/ux-research-cheat-sheet/

Foundation, I. D. (2020). *What is User Centered Design? | Interaction Design Foundation*. Www.Interaction-Design.Org. https://www.interaction-design.org/literature/topics/user-centered-design

Gabe-Thomas, E., Walker, I., Verplanken, B., & Shaddick, G. (2016). Householders' mental models of domestic energy consumption: Using a sort-and-cluster method to identify shared concepts of appliance similarity. *PLoS ONE*, *11*(7). https://doi.org/10.1371/journal.pone.0158949

Gatsou, C., Politis, A., & Zevgolis, D. (2012). Novice user involvement in information architecture for a mobile tablet application through card sorting. *2012 Federated Conference on Computer Science and Information Systems, FedCSIS 2012*, 711–718.

Gonzalez-Zuniga, D., & Carrabina, J. (2016). Clustering to categorize desirability in software: Exploring cluster analysis of Product Reaction Cards in a stereoscopic retail application. *2016 Digital Media Industry and Academic Forum, DMIAF 2016 - Proceedings*, 193–197. https://doi.org/10.1109/DMIAF.2016.7574931

Goodman-Deane, J., Langdon, P., Clarke, S., & Clarkson, P. J. (2008). Categorising design methods: How designers view the roles of user methods in design. *Contemporary Ergonomics 2008*, 273–278.

Guo, J., & Yan, P. (2011). User-centered information architecture of university library website. *ICCRD2011 - 2011 3rd International Conference on Computer Research and Development*, *2*, 370–372. https://doi.org/10.1109/ICCRD.2011.5764153

Hepting, D. H., & Almestadi, E. H. (2013). Discernibility in the analysis of binary card sort data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8170 LNAI*, 380–387. https://doi.org/10.1007/978-3-642-41218-9_41

Huang, S. L., & Ku, H. H. (2016). Brand image management for nonprofit organizations: Exploring the relationships between websites, brand images and donations. *Journal of Electronic Commerce Research*, *17*(1), 80.

Inukai, Y., & Kamisasa, H. (1974). Multidimensional Scaling. *Japanese Psychological Review*, *17*(1), 79–105. https://doi.org/10.24602/sjpr.17.1_79

Jollife, I. T., & Cadima, J. (2016a). Principal component analysis: A review and recent developments. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 374, Issue 2065). Royal Society of London. https://doi.org/10.1098/rsta.2015.0202

Jollife, I. T., & Cadima, J. (2016b). Principal component analysis: A review and recent developments. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 374, Issue 2065). Royal Society of London. https://doi.org/10.1098/rsta.2015.0202

Kassambara, A. (2018a). *Heatmap in R: Static and Interactive Visualization*. Datanovia. https://www.datanovia.com/en/lessons/heatmap-in-r-static-and-interactive-visualization/

Kassambara, A. (2018b). *Heatmap in R: Static and Interactive Visualization*. Datanovia.

Katsanos, C., Avouris, N., Stamelos, I., Tselios, N., Demetriadis, S., & Angelis, L. (2019). Cross-study Reliability of the Open Card Sorting Method. *Conference on Human Factors in Computing Systems - Proceedings*, 1–6. https://doi.org/10.1145/3290607.3312999

Kaufman, L., & Rousseeuw, P. J. (2005). Finding Groups in Data. In L. Kaufman & P. J. Rousseeuw (Eds.), *Wiley Series in Probability and Statistics* (Vol. 603). Wiley.

Kelley, C., Lee, B., & Wilcox, L. (2017). Self-tracking for mental wellness: Understanding expert perspectives and student experiences. *Conference on Human Factors in Computing Systems*, 629–641. https://doi.org/10.1145/3025453.3025750

Kelley, C., Wilcox, L., Ng, W., Schiffer, J., & Hammer, J. (2017). Design features in games for health: Disciplinary and interdisciplinary expert perspectives. *DIS 2017 - Proceedings of the 2017 ACM Conference on Designing Interactive Systems*, 69–81. https://doi.org/10.1145/3064663.3064721

Lamprecht, E. (2020). *The Difference Between UX & UI Design - A Layman's Guide (2021 Guide)*. Career Foundry.

Lantz, E., Keeley, J. W., Roberts, M. C., Medina-Mora, M. E., Sharan, P., & Reed, G.

M. (2019). Card Sorting Data Collection Methodology: How Many Participants Is Most Efficient? *Journal of Classification*, *36*(3), 649–658. https://doi.org/10.1007/s00357-018-9292-8

Lee, S. C., Nadri, C., Sanghavi, H., & Jeon, M. (2021). Eliciting User Needs and Design Requirements for User Experience in Fully Automated Vehicles. *International Journal of Human–Computer Interaction*, 1–13. https://doi.org/10.1080/10447318.2021.1937875

Lewis, J. R., & Sauro, J. (2018). Item Benchmarks for the System Usability Scale. *Journal of Usability Studies*, *13*(3), 158–167. https://www.researchgate.net/profile/James-Lewis-8/publication/330225055_Item_Benchmarks_for_the_System_Usability_Scale/links/5c34fb41a6fdccd6b59ce868/Item-Benchmarks-for-the-System-Usability-Scale.pdf

Lucci, G., & Paternò, F. (2015). Analysing how users prefer to model contextual event-action behaviours in their smartphones. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9083*, 186–191. https://doi.org/10.1007/978-3-319-18425-8_14

Maat, H. P., & Lentz, L. (2011). Using sorting data to evaluate text structure: An evidence-based proposal for restructuring patient information leaflets. *Technical Communication*, *58*(3), 197–216.

Macías, José Antonio. (2008). Intelligent assistance in authoring dynamically generated Web interfaces. *World Wide Web*, *11*(2), 253–286. https://doi.org/10.1007/s11280-008-0043-3

Macías, José Antonio. (2021). Enhancing Card Sorting Dendrograms through the Holistic Analysis of Distance Methods and Linkage Criteria. *Journal of Usability Studies*, *16*(2), 73–90. https://uxpajournal.org/wp-content/uploads/sites/8/pdf/JUS_Macias_Feb2021.pdf

Macías, José Antonio, & Castells, P. (2001). A generic presentation modeling system for adaptive web-based instructional applications. *Conference on Human Factors in Computing Systems - Proceedings*, 349–350. https://doi.org/10.1145/634067.634273

Macías, José Antonio, & Castells, P. (2002). Tailoring dynamic ontology-driven web documents by demonstration. *Proceedings of the International Conference on Information Visualisation*, *2002-Janua*, 535–540. https://doi.org/10.1109/IV.2002.1028826

Macías, José Antonio, Granollers, T., & Latorre, P. (2009). New Trends on Human-Computer Interaction. In José A. Macías, A. Granollers Saltiveri, & P. M. Latorre (Eds.), *New Trends on Human-Computer Interaction: Research, Development, New Tools and Methods*. Springer London. https://doi.org/10.1007/978-1-84882-352-5

Mahmood, F., Wan Adnan, W. A., Md Noor, N. L., & Mohd Saman, F. (2018). Emotional response towards cultural-based e-government portal design using card sorting method. *Communications in Computer and Information Science*, *886*, 12–22. https://doi.org/10.1007/978-981-13-1628-9_2

Maida, M., & Obwegeser, N. (2012). Evaluation of Techniques for Structuring Multi-Criteria Decision Problems. *INTERNATIONAL CONFERENCE ON INFORMATION RESOURCES MANAGEMENT (CONF-IRM)*.

Mengist, W., Soromessa, T., & Legese, G. (2020). Method for conducting systematic literature review and meta-analysis for environmental science research. *MethodsX*, *7*, 100777. https://doi.org/10.1016/j.mex.2019.100777

Mesgari, M., Okoli, C., & De Guinea, A. O. (2019). Creating Rich and Representative Personas by Discovering Affordances. *IEEE Transactions on Software Engineering*, *45*(10), 967–983. https://doi.org/10.1109/TSE.2018.2826537

Mesgari, M., Okoli, C., & Ortiz De Guinea, A. (2015). Affordance-based user personas: A mixedmethod approach to persona development. *2015 Americas Conference on Information Systems, AMCIS 2015*, 1–17.

Meyer-Baese, A., & Schmid, V. (2014a). Statistical and Syntactic Pattern Recognition. In *Pattern Recognition and Signal Analysis in Medical Imaging* (pp. 151–196). Elsevier. https://doi.org/10.1016/b978-0-12-409545-8.00006-6

Meyer-Baese, A., & Schmid, V. (2014b). Statistical and Syntactic Pattern Recognition. In *Pattern Recognition and Signal Analysis in Medical Imaging* (pp. 151–196). Elsevier. https://doi.org/10.1016/b978-0-12-409545-8.00006-6

Morente-Molinera, J. A., Ríos-Aguilar, S., González-Crespo, R., & Herrera-Viedma, E. (2019). Dealing with group decision-making environments that have a high amount of alternatives using card-sorting techniques. *Expert Systems with Applications*, *127*, 187–198. https://doi.org/10.1016/j.eswa.2019.03.023

Nawaz, A., Clemmensen, T., & Hertzum, M. (2011). Information Classification on University Websites : A Cross-Country Card Sort Study. *Information Systems Research Seminar in Scandinavia (IRIS)*, *34*(15), 528–542.

Nayebi, M., Kuznetsov, K., Chen, P., Zeller, A., & Ruhe, G. (2018). Anatomy of functionality deletion: An exploratory study on mobile apps. *Proceedings - International Conference on Software Engineering*, 243–253. https://doi.org/10.1145/3196398.3196410

Nielsen, J. (2004). *Card Sorting: How Many Users to Test*. Jakob Nielsens Alertbox.

Nielsen, J., & Landauer, T. K. (1993). Mathematical model of the finding of usability problems. *Conference on Human Factors in Computing Systems - Proceedings*, 206–213. https://doi.org/10.1145/169059.169166

Nurcahyanti, W. E., & Suhardi. (2014). Information architecture assessment of BPS headquarter official website. *2014 International Conference on Information Technology Systems and Innovation, ICITSI 2014 - Proceedings*, 177–182. https://doi.org/10.1109/ICITSI.2014.7048260

Olaverri-Monreal, C., Lehsing, C., Trubswetter, N., Schepp, C. A., & Bengler, K. (2013). In-vehicle displays: Driving information prioritization and visualization. *IEEE Intelligent Vehicles Symposium, Proceedings*, 660–665. https://doi.org/10.1109/IVS.2013.6629542

Optimal Workshop. (2020). *Optimal Workshop*. Optimal Workshop. https://www.optimalworkshop.com/

Paea, S., & Baird, R. (2018a). Information Architecture (IA): Using multidimensional scaling (MDS) and k-means clustering algorithm for analysis of card sorting data. *Journal of Usability Studies*, *13*(3), 138–157. https://uxpajournal.org/information-architecture-card-sort-analysis/

Paea, S., & Baird, R. (2018b). Information Architecture (IA): Using multidimensional scaling (MDS) and k-means clustering algorithm for analysis of card sorting data.

Journal of Usability Studies, *13*(3), 138–157.

Palmer, F., & O'Neill, E. (2010). Interpreting technology-mediated identity: Perception of social intention and meaning in Bluetooth names. *ACM International Conference Proceeding Series*, 232–239. https://doi.org/10.1145/1952222.1952273

Petrie, H., Power, C., Cairns, P., & Seneler, C. (2011). Using Card Sorts for Understanding Website Information Architectures: Technological. *Methodological and Cultural Issues. Human-Computer Interaction – INTERACT 2011*, *6949 LNCS*(PART 4), 309–322. https://doi.org/10.1007/978-3-642-23768-3_26

Pisanski, J., & Žumer, M. (2010). Mental models of the bibliographic universe. Part 1: Mental models of descriptions. *Journal of Documentation*, *66*(5), 643–667. https://doi.org/10.1108/00220411011066772

ProvenByUsers. (2020). *Online Card Sorting from Proven By Users*. https://www.provenbyusers.com/

Quintal, C., & Macías, J. A. (2021). Measuring and improving the quality of development processes based on usability and accessibility. *Universal Access in the Information Society*, *20*(2), 3. https://doi.org/10.1007/s10209-020-00726-7

Reese, T., Segall, N., Nesbitt, P., Del Fiol, G., Waller, R., MacPherson, B. C., Tonna, J. E., & Wright, M. C. (2018a). Patient information organization in the intensive care setting: Expert knowledge elicitation with card sorting methods. *Journal of the American Medical Informatics Association*, *25*(8), 1026–1035. https://doi.org/10.1093/jamia/ocy045

Reese, T., Segall, N., Nesbitt, P., Del Fiol, G., Waller, R., MacPherson, B. C., Tonna, J. E., & Wright, M. C. (2018b). Patient information organization in the intensive care setting: Expert knowledge elicitation with card sorting methods. *Journal of the American Medical Informatics Association*, *25*(8), 1026–1035. https://doi.org/10.1093/jamia/ocy045

Rehring, K., Brée, T., Gulden, J., & Bredenfeld, L. (2020). Conceptualizing EA cities: Towards visualizing enterprise architectures as cities. *27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019*.

Righi, C., James, J., Beasley, M., Day, D. L., Fox, J. E., Gieber, J., Howe, C., & Ruby, L. (2013). Card Sort Analysis Best Practices. *Journal of Usability Studies*, *8*(3),

69–89.

Robles, T. de J. Á., Rodríguez, F. J. Á., Benítez-Guerrero, E., & Rusu, C. (2019). Adapting card sorting for blind people: Evaluation of the interaction design in TalkBack. *Computer Standards and Interfaces*, *66*. https://doi.org/10.1016/j.csi.2019.103356

Rosenfeld, L., Morville, P., & Arango, J. (2015). *Information Architecture for the World Wide Web* (4th Editio). O'Reilly Media, Inc.

Roth, R. E. (2013). An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2356–2365. https://doi.org/10.1109/TVCG.2013.130

Ruiz, J., Serral, E., & Snoeck, M. (2021). Unifying Functional User Interface Design Principles. *International Journal of Human–Computer Interaction*, *37*(1), 47–67. https://doi.org/10.1080/10447318.2020.1805876

Sampson, F. (2005). Taking UX offshore. *Interactions*, *12*(6), 8–9. https://doi.org/10.1145/1096554.1096569

Sánchez, E., & Macías, J. A. (2019). A set of prescribed activities for enhancing requirements engineering in the development of usable e-Government applications. *Requirements Engineering*, *24*(2), 181–203. https://doi.org/10.1007/s00766-017-0282-x

Santos, O. C., & Boticario, J. G. (2015). Practical guidelines for designing and evaluating educationally oriented recommendations. *Computers and Education*, *81*, 354–374. https://doi.org/10.1016/j.compedu.2014.10.008

Scapin, D. L., Marie-dessoude, P., Winckler, M. A., & Detraux, C. (2011). *Personal Information Systems: User Views and Information Categorization*. *c*, 40–47.

Schmettow, M., & Sommer, J. (2016). Linking card sorting to browsing performance – are congruent municipal websites more efficient to use? *Behaviour and Information Technology*, *35*(6), 452–470. https://doi.org/10.1080/0144929X.2016.1157207

Shen, S. T., & Prior, S. D. (2013). My favorites (bookmarks) schema - One solution to online information storage and retrieval. *ACM International Conference*

Proceeding Series, 33–40. https://doi.org/10.1145/2503859.2503865

Slegers, K., & Donoso, V. (2012). The impact of paper prototyping on card sorting: A case study. *Interacting with Computers*, *24*(5), 351–357. https://doi.org/10.1016/j.intcom.2012.05.005

Spencer, D. (2009). *Card Sorting: Designing Usable Categories* (Marta Justak (ed.)). Louis Rosenfeld. www.rosenfeldmedia.com

Thomas, R. L., & Johnson, I. (2013). Merging methodologies: Combining individual and group card sorting. *User Experience, and Usability. Design Philosophy, Methods, and Tools*, *8012 LNCS*(PART 1), 417–426. https://doi.org/10.1007/978-3-642-39229-0_45

Urrutia, J. I. G., Brangier, E., & Cessat, L. (2017). Is a holistic criteria-based approach possible in user experience?: Study of the classification of 58 criteria linked to UX. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10288*, 395–409. https://doi.org/10.1007/978-3-319-58634-2_29

usabiliTEST. (2020). *usabiliTEST: Usability Testing Tools for Everyone*. http://www.usabilitest.com/

UserZoom. (n.d.). *Actionable UX Insights for Better Digital Experiences*. Retrieved July 11, 2021, from https://www.userzoom.com/es/

van Pinxteren, Y., Geleijnse, G., & Kamsteeg, P. (2011a). Deriving a recipe similarity measure for recommending healthful meals. *Proceedings of the 15th International Conference on Intelligent User Interfaces - IUI '11*, 105–114. https://doi.org/10.1145/1943403.1943422

van Pinxteren, Y., Geleijnse, G., & Kamsteeg, P. (2011b). Deriving a recipe similarity measure for recommending healthful meals. *Proceedings of the 15th International Conference on Intelligent User Interfaces - IUI '11*, 105–114. https://doi.org/10.1145/1943403.1943422

Vashitz, G., Nunnally, M. E., Parmet, Y., Bitan, Y., O'Connor, M. F., & Cook, R. I. (2013). How do clinicians reconcile conditions and medications? The cognitive context of medication reconciliation. *Cognition, Technology and Work*, *15*(1), 109–116. https://doi.org/10.1007/s10111-011-0189-0

Veral, R., & Macías, J. A. (2019). Supporting user-perceived usability benchmarking through a developed quantitative metric. *International Journal of Human Computer Studies*, *122*, 184–195. https://doi.org/10.1016/j.ijhcs.2018.09.012

Verhoeven, F., & Gemert-Pijnen, J. E. W. C. van. (2010). Discount User-Centered e-Health Design: A Quick-but-not-Dirty Method. *Lecture Notes in Artificial Intelligence*, *6389*, 101–123.

WeCaSo. (2016). *Web-based card sorting tool*. https://cs.uni-paderborn.de/en/mci/news-single/neue-version-des-web-basierten-card-sorting-tools-wecaso-new-version-of-the-web-based-card-sorting-tool-wecaso

Wentzel, J., Müller, F., Beerlage-de Jong, N., & van Gemert-Pijnen, J. (2016). Card sorting to evaluate the robustness of the information architecture of a protocol website. *International Journal of Medical Informatics*, *86*, 71–81. https://doi.org/10.1016/j.ijmedinf.2015.12.003

Wentzel, Jobke, de Jong, N. B., & van der Geest, T. (2016). Redesign based on card sorting: How universally applicable are card sort results? *HCI*, *9745*, 381–388. https://doi.org/10.1007/978-3-319-40247-5_38

Wood, J. R., & Wood, L. E. (2008). Card Sorting: Current Practices and BeyondJUS. *Journal of Usability Studies Archive, 4*(1), 1–6.

XSort. (2020). *Free card sorting application for Mac*. https://xsortapp.com

Young, F. W. (1997). *Multidimensional Scaling*. University of North Carolina. https://doi.org/10.4324/9780203767719

Zainuddin, E., & Staples, S. (2016). Developing a shared taxonomy of workaround behaviors for the information systems field. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 5278–5287. https://doi.org/10.1109/HICSS.2016.652

Biographies

Marina Martín received her B.Cs. in 2020 from Universidad Autónoma de Madrid. She is currently a student in the Big Data & Data Science Master at the same university. Marina

Martín works as an R&D engineer focused on mobility and Big Data. Her interests are Human-Computer Interaction and Data Science.

José A. Macías received his Ph.D. in Computer Science in 2003 from Universidad Autónoma de Madrid, where he currently works as Associate Professor. He has been working on Human-Computer Interaction for more than 20 years, serving as President in the Spanish main HCI association and Co-Chair in the Spanish SIGCHI.

Table 1. The number of articles found and the final selection for each digital library.

Table 2. Comparison of the different statistical and data mining techniques included in each evaluated tool. Also, the proposed supporting tool CAULDRON has been included to compare the featured analysis techniques.

Table 3. Efficiency results in seconds obtained from the evaluator assessment, where mean, min, max, standard deviation, median, and 95% confidence interval values for each task are shown.

Table 4. Efficiency results in seconds obtained from the participant assessment, where mean, min, max, standard deviation, median, and 95% confidence interval values for each task are shown.

Figure 1. Architectural detail of CAULDRON, showing main components in a client-server deployment.

Figure 2. Main page with the access for evaluators (left) and the evaluations management screen (right) in CAULDRON, including the different options to deal with evaluations and showing information about participants and card sorts.

Figure 3. Main page with the access for participants (right) and the interactive Card Sorting board (left) for the participant's sorting tasks in CAULDRON.

Figure 4. Frequency and agreement analysis.

Figure 5. Heatmaps representing the classification matrix, including dendrograms to denote the different groups and the dissimilarity matrices for cards and categories.

Figure 6. Charts representing correlations for cards (left) and categories (right).

Figure 7. Scatterplots representing groups of cards (left) and categories (right) using the Smacof Multidimensional Scaling.

Figure 8. Clusters for cards (left) and categories (right) using K-means and PCA.

Figure 9. Decision trees generated for "Instagram" (left) and "Skype" (right) cards.

Figure 10. Initial consolidation board to help the UX researcher make a final decision on the Card Sorting.

Figure 11. SUS scores obtained from the evaluator and participant assessments.