Improvement of Usability in User Interfaces for Massive Data Analysis: An Empirical Study

Carlos Iñiguez-Jarrín[a,b], José Ignacio Panach[c], Oscar Pastor López[a]

[a]Research Center on Software Production Methods (PROS),
Universitat Politècnica de València. Camino Vera s/n. 46022, Valencia, Spain
[b]Departamento de Informática y Ciencias de la Computación,
Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito, Ecuador
[c]Escola Tècnica Superior d'Enginyeria,
Departament d'Informàtica, Universitat de València
Avenida de la Universidad, s/n, 46100 Burjassot, Valencia, Spain

{ciniguez, opastor}@pros.upv.es, joigpana@uv.es

Context: Big Data challenges the conventional way of analyzing massive data and creates the need to improve the usability of existing user interfaces (UI) in order to deal with massive amounts of data. How the UIs facilitate the search for information and helps in the end-user's decision-making depends on developers and designers, who have no guides for producing usable UIs. We have proposed a set of interaction patterns for designing massive data analysis UIs by studying 27 real case studies of massive data analysis. **Objective**: We evaluate if the proposed patterns improve the usability of the massive data analysis UIs in the context of literature search. **Method**: We conducted two replications of the same controlled experiment, one with 24 undergraduate students experienced in scientific literature search and the other with eight researchers who are experienced in biomedical literature search. The experiment, which was planned as a repeated measures design, compares UIs that have been enhanced with the proposed patterns versus original UIs in terms of three response variables: effectiveness, efficiency, and satisfaction. **Results**: The outcomes show that the use of interaction patterns in UIs for massive data analysis yields better and more significant effects for the three response variables, enhancing the discovery and visualization of the data. **Conclusions:** The use of the proposed interaction design patterns improves the usability of the UIs that deal with massive data. The patterns can be considered as guides for helping designers and developers to design usable UIs for massive data analysis web applications.

Keywords: user interface, Big Data, usability, interaction design patterns.

1 Introduction

Since the emergence of Big Data systems, the user interface (UI) of the tools has been adapting its approach in order to meet the user's data consumption needs. The concept of *fill in the form and submit* is being changed by the *pre-data collection and post-implementation analysis* concept to deal with large volumes of data and to enhance the limited human capacity of figuring out the relationships between the data. Examples of this last concept are the UIs of traditional business intelligence tools that are equipped with attractive and interactive visualizations to present and manipulate data from a wide range of relational data repositories.

In view of the massive and complex amount of data, the purpose of these UIs is to facilitate access to the data through easy-to-use interactive mechanisms for non-expert computer users. However, the facility with which these UIs allow the search and identification of relevant results depends to a large extent on the UI design performed by software developers who are non-experts in usability and designers who do not have enough guidelines to deal with the possible problems of interaction between the user and the massive data.

In software engineering, the *pattern* concept [1] is widely used to refer to solutions for repeated problems that appear in different contexts. Specifically, in the UI design context, the solutions for repeated problems are known as *interaction design patterns* and are referenced by the human-computer interaction (HCI) community under several names such as *user interface design patterns* [2], *interaction design patterns* [3], *HCI patterns* [4], or *design patterns* [5]. Tidwell [6] defines interaction design patterns as "possible good design solutions to a common interaction design problem within a certain context, by describing the invariant qualities of all those solutions".

The existing interaction design patterns solve important problems that are related to the interaction design in UIs such as how to organize the content in the UI, how to capture the input of user data, how to display or navigate the data, etc. These patterns provide solutions to problems of interaction with data that are available in fixed formats and stored in local data sources. However, these patterns do not consider the interaction problems when the data is *massive*, which is understood to be large amounts of data that can come from different data sources with different formats (i.e., structured, semi-structured, unstructured). In order to solve the design problems that are related to the interaction with massive data, we studied 27 Big Data case studies to look for solutions to problems related to consuming large amounts of data [7], specifically the problems related to how to visualize, explore, and manipulate the data. The resulting solutions were gathered as *interaction design patterns for dealing with massive data* (massiveData-ID patterns). These patterns bring together the common trends, best practices, and experiences of various designers to solve specific interaction design problems that arise when designing UIs that deal with massive data.

In this paper, we evaluate whether the use of the proposed massiveData-ID patterns produces positive or negative changes in the usability of UIs from the end-user's perspective. The main contribution of this paper is the design and execution of an experiment to analyse differences in terms of usability depending on whether massiveData-ID patterns are used. We evaluate effectiveness, efficiency, and the satisfaction of the end-user during the use of UIs that support massive data analysis. The experiment is a repeated-measures design that compares UIs that incorporate massiveData-ID patterns with UIs that do not incorporate them using two experimental problems. Five patterns were analysed in the experiment. The experiment was conducted in two replications: Replica 1 (R1), with 24 students from the undergraduate degree program in Information and Documentation of the University of Valencia (UV) and Replica 2 (R2), with eight researchers from the PROS[1] Research Centre's Genome group of the Polytechnic University of Valencia (UPV). The subjects of both replicas work every day with UIs to analyse massive data of literature: The subjects in R1 are proficient in bibliographic database management and information search, especially with the Scopus[2] literature database, and the subjects in R2 are researchers who are involved in the biomedical field and have experience in the search and the analysis of biomedical literature using the PubMed[3] search engine as a source of information. Our experiment consists of solving two literature search problems, one for each subject profile, using literature search UIs to find relevant documents on a topic of interest taking into account literature attributes such as authors, subject, bibliographical references, etc. The large number of documents that are interrelated through literature attributes and the interaction mechanisms required to analyse the document contents make this domain a suitable environment for evaluating the use of UIs that are designed with the proposed massiveData-ID patterns.

The results show that, independently of the experimental problem to be solved with the UIs, the UIs designed with massiveData-ID patterns achieved better usability scores for effectiveness, efficiency, and satisfaction in both replicas (R1 and R2), with the R1subjects achieving higher scores than the R2 subjects.

2 Related Works

This work focuses on improving the usability of UIs for massive data analysis through the use of interaction design patterns. In this section, we review works dealing with: a) the design of data analysis UIs, especially those for analyzing huge amounts of literature since it is the domain of our experiment; b) the use of interaction design patterns for designing UIs; and c) the evaluation of interaction design patterns.

2.1 Design of Data Analysis User Interfaces

The interfaces for analyzing data have made a great leap from command-line UIs to web-based UIs by incorporating interactive data visualizations to show the trends in the data. Examples of such interfaces are those implemented by Business Intelligence tools such as Tableau (www.tableau.com), Qlik (www.qlik.com), and TIBCO (www.tibco.com), which are designed to connect to a variety of structured data sources that are often locally stored. These UIs are designed following a typical structure consisting of "selecting data source-visualizing-showing analysis". Indeed, once the data sources are selected, the user can create several data visualizations by dragging and dropping the fields from the connected data source and then he/she can incorporate the visualizations on a dashboard to show the data analysis.

Other examples of data analysis UIs are those that are specifically developed to meet the needs of interaction in a specific domain. These UIs are developed taking into account the specific filters and

[1] http://www.pros.webs.upv.es/
[2] www.scopus.com
[3] www.ncbi.nlm.nih.gov/pubmed/

visualizations required to exploit the data. A clear example of these are the UIs of the literature search systems that people use to find and analyse a large number of publications that are related to a topic of interest.

The design and redesign of UIs to improve the access and visibility of the content of the literature is a constant concern, especially for the search tools of biomedical and scientific literature. In the biomedical field, the UIs of relevant literature search tools are continually being improved to help researchers, physicians, and clinicians interact with the large amount of clinical literature; this is the case of the UIs of PubMed [8][9] or MedlinePlus [10]. Thus, the design of a single view to present the relationship of the literature with the content of different silos of biomedical data is one of the main challenges of designers and developers [11][12][13]. In the field of scientific literature, the UIs that are used for search engines (e.g., Web of Science, Scopus, Springer, IEEE explorer, and Discovery [14]) incorporate UI controls to search for articles, books, and journals and to filter them by author, field of research (e.g., genetics, the clinical sciences, immunology), source of titles (e.g., Plos One), or year of publication. On the other hand, the interfaces that are used by bibliometric systems apply powerful visualizations that are accompanied by statistical and mathematical methods to analyse publications that are based on the relationships between references, for both scientific literature [15] and biomedical literature[15][16].

In this paper, we focus on improving the design of UIs for searching literature by incorporating the best ideas and design solutions found in existing data analysis UIs.

2.2 Patterns for Designing User Interfaces

Designing the UI means designing the interaction (the dialogue between the user and the interface), and the use of patterns is a widely accepted strategy to do that [17][18][19]. Patterns for designing the UI, commonly named *interaction design patterns*, capture the knowledge about successful solutions to recurring UI design problems in an easily understood way. Their popularity can be seen in the plethora of published pattern collections for designing UIs in several platforms (e.g., desktop [6][20], mobile [20][21], social media [20], and web [2][5][18][20][22][2][23][24]) and domains (e.g., cultural heritage [25], games [26], augmented work environments [27], information retrieval [28], and recommender systems [29]).

Patterns for supporting the design of interfaces that handle large amounts of data are scarce and specifically address interaction aspects such as visualization, filtering, and searching. Some of them can be found as a category within the published patterns such as Tidwell's book chapter entitled "Show Complex Data: Trees, Charts and Other Information Graphics" [20], which describes a category of patterns for dealing with the complex aspects of data presentation. The existing patterns cover some of the important problems of designing UIs that have massive amounts of data; however, as new problems appear, the list of patterns must grow to solve them. We propose patterns to address some interaction design problems and use them to improve the design of the existing UIs of a massive data domain: bibliographic search.

2.3 Evaluation of Patterns

The literature contains papers that focus on evaluating patterns that discuss *what* to evaluate about a pattern and *how* to do it. By *what* we mean the purpose for evaluating the pattern(s). Below, we summarize the related works that we found.

Seidel [30] mentions that the evaluation of an individual pattern or a set of patterns can be directed towards two high-level purposes: *writing* (the written representation of the pattern) and *applicability* (the application and use of the pattern). These high-level purposes contain low-level purposes: Completeness, Briefness, and Validity for writing; and Usability, Viability, and Impact for applicability. Guerra et al. [31] also evaluate the patterns taking into account two purposes, but considering two pattern characteristics: a) being repeatable (a recurrent solution used in existent software development community and b) being a reference (a solution used as a model for the development of new software).

By *how*, we mean the variables, instruments, people, or research methods that are used by authors to validate the patterns. Below, we summarize the related works that we found.

In [32], the authors evaluate 14 new patterns for designing the UIs of recommender systems through a workshop structured in the form of *Writer`s Workshop* [33]. This is commonly used in PLoP conferences, where a panel of experts in recommender systems and interface design study each pattern before the session. Once in the workshop, the experts discuss and examine the strengths and weaknesses of each pattern and propose improvements in the content and style of the pattern. In [31], the authors describe the evaluation of eight patterns conducted by both experienced and inexperienced developers in the pattern language field. While the inexperienced developers evaluate the patterns by designing UIs prototypes with and without patterns, the experienced developers evaluate the patterns by comparing them with their previous knowledge. Post-test questionnaires are given to determine the perception of the participants about the usage of individual

patterns and the whole set of patterns. In [34], an experiment evaluates the user satisfaction of six UI design patterns of search boxes and autocomplete. Forty-six undergraduate students performed a specific task for each design pattern using six UI prototypes (one for each pattern) developed and hosted on the Amazon cloud server. Sitting in front of a PC, the students used all of the prototypes and evaluated the ease of use of each pattern on an online questionnaire using a 7-point scale. The hypotheses were statistically analysed using repeated measures ANOVA.

As mentioned above, the evaluation of patterns is directed towards different purposes and is performed by considering different research methods. In this paper, we empirically evaluate the patterns through an experiment. The purpose is to evaluate the applicability of the pattern, specifically the impact caused by the patterns in terms of usability (i.e., effectiveness, efficiency, and satisfaction), when they are incorporated in UIs for massive-data analysis.

3 Design patterns for interacting with massive data (massiveData-IDP)

The literature contains interaction patterns to solve problems that are related to the interaction with large amounts of data. However, as the volume of data grows, new data consumption challenges arise, and therefore, new interaction problems dealing with massive data must be identified, studied, solved, and documented as interaction patterns. Therefore, we have identified and defined five patterns [7] for massive-data analysis environments through a systematic five-step process. Step 1 is to identify data consumption challenges by studying 27 real Big Data use cases from several domains (e.g., online stores, financial services, security). Four challenges were identified: enhancing data discovery, enlarging visualization, performing data analysis operations, and contextualizing data. Step 2 is to identify the existing tools facing these consumption challenges. Step 3 is to study the UIs implemented by the tools. Step 4 is to identify the design solutions used repeatedly by designers in the UIs. Step 5 is to document the design solutions in the pattern's format using a template with 6 sections (name, context, problem, solution, why, and example).

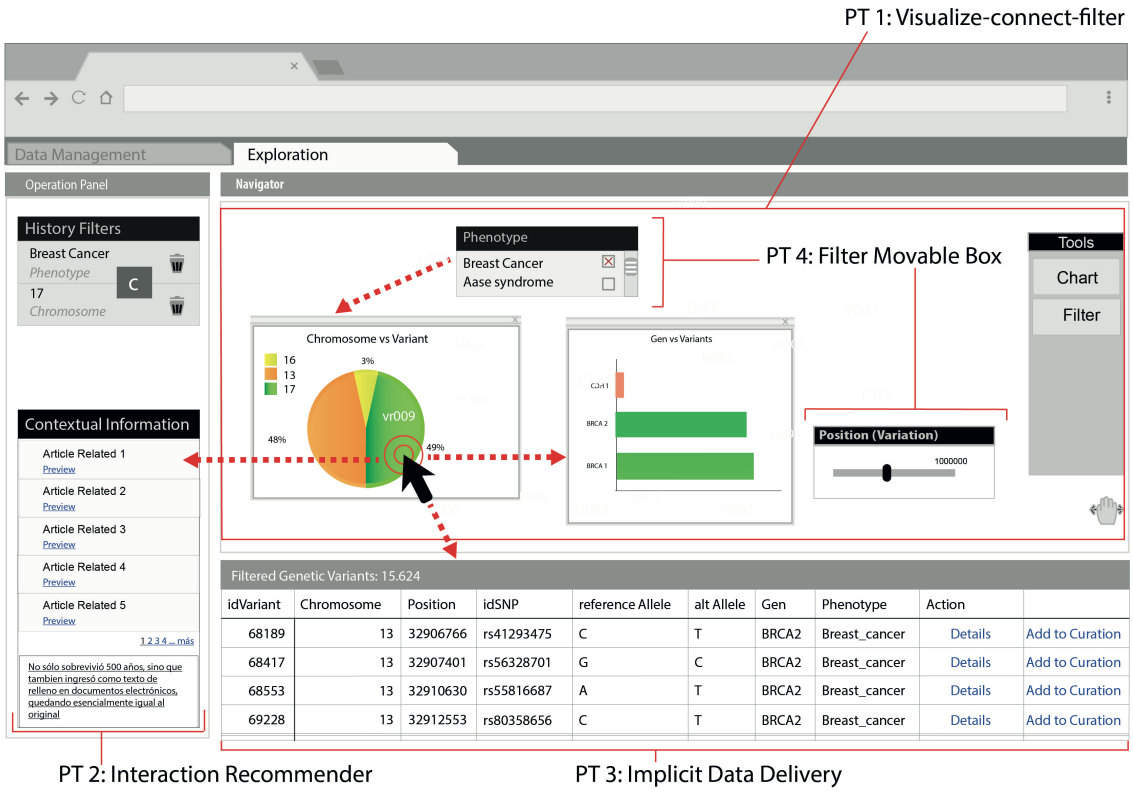
We call this set of resulting patters "Interaction Design Patterns for dealing with massive data", which we refer to in this work as *massiveData-ID* patterns. Here, we synthesize the description of four massiveData-ID patterns that we use in this experiment by describing the problem tackled and the proposed solution for each pattern. We also illustrate the solution in an example of a UI that is designed with them (**Fig. 1**).

PT 1: Visualize, Connect, Filter
Problem: How can the user visualize the multiple changes in the data behavior caused by a simple data filter?
Solution: Provide a canvas where the user is free to create interactive data charts and to place and organize them across the canvas according to his/her needs. When the user applies a filter condition by interacting with a data chart, the remaining charts respond automatically by updating their state according to the filter condition that affects the entire data set.

PT 2: Interaction Recommender
Problem: How can the user be provided with knowledge obtained from his/her interactions?
Solution: Show emerging system recommendations, including information about context, potential explorable items, alerts, or notifications. The system produces recommendations based on the history of performed interactions. Each interaction provides information about the element of the data schema involved in the interaction, so the system can suggest the exploration of new paths according to the data schema.

PT 3: Implicit Data Delivery
Problem: How can the delay effect be reduced when showing and navigating through large data volumes?
Solution: Request and deliver the data every time the user implicitly expresses the intent to navigate through them.

PT 4: Filter Movable Box
Problem: How can the user define filter conditions and apply them in a fluid way within the analysis space?
Solution: Allow the user to create filter conditions from any existing attribute in the underlying data schema. This minimizes the unnecessary movements of the user in the analysis space by allowing the user to place the filters anywhere in the analysis space.



idVariant	Chromosome	Position	idSNP	reference Allele	alt Allele	Gen	Phenotype	Action	
68189	13	32906766	rs41293475	C	T	BRCA2	Breast_cancer	Details	Add to Curation
68417	13	32907401	rs56328701	G	C	BRCA2	Breast_cancer	Details	Add to Curation
68553	13	32910630	rs55816687	A	T	BRCA2	Breast_cancer	Details	Add to Curation
69228	13	32912553	rs80358656	C	T	BRCA2	Breast_cancer	Details	Add to Curation

Fig. 1 *Example of a user interface designed with the massiveData-ID patterns*

The contributions of the researchers in HCI are critical in order to be able to understand the consumption needs of Big Data and to improve the way in which the user interacts with the data. The proposed patterns become contributions that capture design solutions to common problems of interaction with the data, which are aimed at improving the user experience in a data analysis environment. In this experiment, we evaluate the improvements in usability of the UIs designed with massiveData-ID patterns versus UIs that do not use these patterns. We describe the experimental design below.

4 Empirical Study Planning

4.1 Goals, Research questions, and Hypotheses

From a general perspective, we want to know the extent to which UIs that are designed with massiveData-ID patterns allow users to achieve specific goals in a massive data analysis environment. Specifically, we aim to evaluate the *usability in use*, which is defined by the standard ISO/IEC 25010 as "the degree to which specified users can achieve specific objectives with effectiveness in use, efficiency in use and satisfaction in use in a specific context of use". Therefore, according to the Goal/Question/Measure (GQM) template [35], the goal of this study is:

Analyse	*the massive-data analysis UIs designed with massiveData-ID* patterns
for the purpose of	*evaluating the usability in use*
with respect to	*effectiveness in use, efficiency in use, and satisfaction in use*
from the point of view of	*the researcher*
in the context of	*analysts searching literature.*

Table 1 GQM template applied to define the experiment goal

The results of this study may be useful for researchers involved in academia and/or industry, primarily those dealing with the challenges of designing UIs to interact with large amounts of information.

From this research goal, we derived several research questions (RQs). Since the standard ISO/IEC 25010 defines three attributes (i.e., effectiveness in use, efficiency in use, and satisfaction in use) to measure the usability in use, we define the RQs and the corresponding null hypotheses (H_0) based on these attributes.

RQ1: Is *effectiveness in use* affected by the use of massiveData-ID patterns in the UIs? The derived hypothesis from the question is **H_{01}**: The effectiveness in use interacting with UIs that incorporate massiveData-ID patterns is similar to the effectiveness in use observed when UIs do not incorporate these massiveData-ID patterns.

RQ2: Is *efficiency in use* affected by the use of massiveData-ID patterns in the UIs? The derived hypothesis from the question is **H_{02}**: The efficiency in use interacting with UIs that incorporate massiveData-ID patterns is similar to the efficiency in use observed when UIs do not incorporate these massiveData-ID patterns.

RQ3: Is user *satisfaction in use* affected by the use of massiveData-ID patterns in the UIs? The derived hypothesis from the question is **H_{03}**: The satisfaction in use interacting with UIs that incorporate massiveData-ID patterns is similar to the user satisfaction in use observed when UIs do not incorporate these massiveData-ID patterns.

4.2 Factors and treatments

Our experiment studies a factor with two levels (also known as *treatments* or *alternatives*). The selection of the factor depends on the purpose of the experiment [36]. In this experiment, the factor is the *Design of UIs for analysing large volumes of data*. The levels or treatments are the factor alternatives that help us answer the questions of the research hypotheses. We work with two treatments:

- **Treatment without patterns** (T.*without-patterns*): This treatment represents UIs for data analysis. This treatment has been operationalized through real existing UIs that do not incorporate the massiveData-ID patterns described in Section 3. T.*without-patterns* is referred to as *control treatment*.
- **Treatment with patterns** (T.*with-patterns*): This treatment represents UIs for data analysis, which are similar to the control treatment but whose design has been modified to incorporate the massiveData-ID patterns described in Section 3. We took the control treatment UIs and redesigned them by incorporating the patterns, and changing the visual characteristics, but still maintaining the functionality of the original UI.

These two treatments become the *independent variables* and are what produce the effects in the *response variables* that we want to measure in the experiment. The response variables are described in the following section.

4.3 Response variables and metrics

The *response variables* (also known as *dependent variables* or *explicative variables*) are the characteristics to be measured in the study, and their value depends on the response of the experimental subjects when they use the treatments. The response variables derive directly from the RQs.

RQ1 is related to the response variable: *effectiveness in use*. According to ISO/IEC-25010, effectiveness in use is "the degree to which specified users can achieve specified goals with accuracy and completeness in a specified context of use". We are specifically interested in the proportion of the completed tasks. We asked users to perform several tasks to solve an experimental problem. The tasks are expressed in the form of open-ended questions. To answer them, the user must click on the interface in certain cases or select an answer from a list in other cases. Depending on the case, we save the answer by registering the place where the user clicked or the selected answer from the list. Then, we evaluate the saved answer by assigning one of the two values: success (result=1) or failure (result=0). Therefore, we calculate the effectiveness by dividing the percentage of tasks successfully completed by the total number of tasks proposed in the experimental problem. The higher the percentage obtained, the greater the effectiveness.

RQ2 is related to the response variable: *efficiency in use*. According to ISO/IEC-25010 efficiency in use is "the degree to which specified users expend appropriate amounts of resources in relation to the effectiveness achieved in a specified context of use". Efficiency in use can be understood to be the amount of time (seconds) that a subject spends to complete a task. During the experiment, the time spent by the user in completing each task (from the beginning of the task until it is completed) is automatically timed in seconds. Since the task completeness is related to the effectiveness, we calculate the efficiency in use for each subject by dividing the effectiveness in use achieved by the subject by the time consumed to complete the task. The lower the number of seconds, the greater the efficiency.

RQ3 is related to the response variable: *satisfaction in use*. According to ISO/IEC-25010, satisfaction in use is "the degree to which users are satisfied in a specified context of use". To measure this variable, we used the IBM CSUQ questionnaire [37], which is widely applied to evaluate the satisfaction of the user in the context of usability studies because of its acceptable reliability. A coefficient alpha exceeding 0,89 has been proved [37]. The questionnaire contains 19 questions, and each question is evaluated qualitatively using

a 7-point Likert scale from "strongly disagree" (1) to "strongly agree" (7). Additionally, the questionnaire contains three open-ended questions that ask for information about the negative aspects, positive aspects, and recommendations for the evaluated UIs. To measure the total satisfaction of each subject, we add the scale points selected by the subject. Thus, if a survey was marked "strongly agree" for all 19 questions, this survey would reach a maximum score of 133 (7 points from the Likert Scale by the 19 questions from the questionnaire). Therefore, the greater the total sum, the greater the satisfaction. The users fill out the CSUQ questionnaire after completing all of the proposed tasks for a problem.

Table 2 summarizes the response variables, the metrics for measuring them, and how they are related to each RQ and the hypotheses.

RQ	Hypotheses	Response variable	Metric	Comments
RQ1	H_{01}	Effectiveness in use	$\frac{\sum_{i=1}^{N} n_i}{N} x100\%$	N = number total of tasks. n_i = the result of task i. t_i = the number of seconds taken to perform the task i.
RQ2	H_{02}	Efficiency in use (task/sec.)	$\frac{\sum_{i=1}^{N} \frac{n_i}{t_i}}{N}$	
RQ3	H_{03}	Satisfaction in use	CSUQ questionnaire	

Table 2 Summary of RQs, hypotheses, response variables, and metrics

4.4 Experimental subjects

The experimental subjects (*experimental units*) are the people participating in the experiment. The experiment is composed of two replications: Replica 1 (R1) was conducted with 24 undergraduate students from the undergraduate degree program of the Information and Documentation at the University of Valencia, and Replica 2 (R2) was conducted with eight researchers (three undergraduate students, four master students, and one PhD from the Genome Research group) from the Polytechnic University of Valencia. The academic profile of the experimental subjects is summarized in Table 3.

Replication	Undergraduate	Master	Ph.D.	Total
R1	24	0	0	24
R2	3	4	1	8

Table 3 Academic profile of the experimental subjects by replica

The subjects of R1 work frequently with tasks that are related to searching, organizing, and evaluating information. The workplaces of these subjects are libraries and information and documentation centres. One of the most frequent tasks for them is to search for scientific literature using the Scopus[4] search website. As academic researchers, the subjects of R2 are focused on studying the human genome (specifically in identifying the genetic variations of DNA that cause diseases). Consequently, one of the most frequent tasks in the researchers' work is to search for clinical evidence using PubMed[5], which is a website for searching for biomedical bibliography. Table 4 shows the frequency of use of the search websites for the subjects in each replication (R1 and R2). In R1, most of the subjects reported that Scopus is the search website that they always (23) or regularly (1) use to search for bibliographic information, while Scopus is occasionally or never used. In R2, all eight subjects reported that PubMed is the search website they always use to search for bibliographic information. This is consistent with the fact that the subjects in R2 focus on studying biological and medical issues; therefore, they use PubMed more than Scopus.

Replica	Search website	Always	Regularly	Occasionally	Never
R1	PubMed	0	0	2	22
	Scopus	23	1	0	0
R2	PubMed	8	0	0	0
	Scopus	0	0	1	7

[4] https://www.scopus.com/results/handle.uri
[5] https://www.ncbi.nlm.nih.gov/PubMed/

Table 4 Frequency of use of Search websites for subjects belonging to replica 1 and replica 2

4.5 Experimental objects

To observe the effects produced by the two treatments (i.e., T.*with-patterns* and T.*without-patterns*), we defined two problems that we refer to as P.*scopus* and P.*pubmed*. These involve analysing massive bibliographic data in two scenarios: "scientific literature" and "biomedical literature", respectively. We selected these scenarios since the subjects of each replication are experts in one of these scenarios (see Table 4). The subjects of R1 are experts in scientific literature, while the subjects of R2 are experts in biomedical literature. Therefore, we can analyse whether the effects yielded by the treatments are affected by the previous knowledge of the subject in the domain.

The P.*scopus* problem focuses on analysing scientific literature by using Scopus. This is a bibliographic database that to date contains approximately 1.4 billion cited references and covers approximately more than 22,000 indexed titles from more than 5000 publishers. The goal of this problem is *searching and analysing bibliography related to the "Interaction Design" topic and analysing the information displayed on the screen*. To do this, the subject is asked to perform several tasks that have been expressed as questions in order for the subject to better understand them, as shown in Table 5.

Id	Question
1	Search for the topic: "*Interaction Design*". How many results were obtained?
2	What range of years were the results published in?
3	What were the two years containing the greatest number of documents?
4	Who is the author with the largest number of published documents?
5	What country contains the largest number of authors?
6	What proportion of the total number of documents does the set of documents represent between 2008 and 2018?
7	How many of the documents are "articles"?
8	How many citations does the most cited article have?
9	Where are the authors with the most cited articles from?

Table 5 Questions for conducting the P.scopus problem

To answer the questions, we provide two UI prototypes (one for each treatment):

1) For T.*without-patterns*, the prototype consists of UIs that simulate the Scopus UIs through images of the real existing Scopus UIs.

2) For T.*with-patterns*, the prototype consists of images of UIs that are analogous to the existing Scopus UIs but that are designed with massiveData-ID patterns, as shown in number 2 of **Fig. 2**. Note that, the patterns change the visual aspect of the original UIs, but the functionality remains.

The P.*pubmed* problem consists of analysing clinical bibliography using PubMed. PubMed is a search engine that is administered by the National Center for Biotechnology Information (NCBI[6]) that to date (April 2018) contains more than 20 million citations of complete medical articles stored in MEDLINE, the premier bibliographic database of the U.S. National Library of Medicine [38]. The goal of this problem is described as: *searching bibliography about the "Cri-du-chat" disease and analyzing the resulting clinical documents to identify the mutations and genes causing the disease*. To achieve this, the subject is asked to answer the questions shown in Table 6.

Id	Question
1	Search for the *"Cri du chat Syndrome"* disease. How many clinical documents are there?
2	What range of years were the clinical documents published in?
3	Filter the results of the last 10 years. How many results were published?
4	How many genes in the "Gene" database are related to the list of the PubMed clinical articles?
5	How many genetic variations of the "ClinVar" database are related to the set of PubMed clinical articles?

Table 6 Questions for conducting the P.pubmed problem

[6] https://www.ncbi.nlm.nih.gov/

Although the P.*pubmed* problem contains fewer questions than the P.*scopus* problem, the P.*pubmed* problem involves a more complex analysis since it involves dealing with biological and clinical terms as well as identifying relationships between documents and genetic elements (i.e., genes and genetic variations). To answer the questions, we provide two UI prototypes (one for each treatment):

1) The UI prototype related to T.*without-patterns* consists of UIs that simulate the PubMed UIs using images of the real existing PubMed UIs.

2) The UI prototype related to T.*with-patterns* consists of images of UIs that are analogous to the existing PubMed UIs but that are designed with massiveData-ID patterns.

Note that the developed UI prototypes are web pages based on images that are enriched with interactive UI controls (e.g., lists, buttons) placed over the image. The behaviour of the interactive controls was implemented using JavaScript to provide the functionality to interact with the UI and answer the questions. For instance, the UI prototype in **Fig. 2** consists of a background image where some circles in the "Documents (Matching and citations)" chart have been converted to UI controls that respond to click events. Thus, when the user clicks on the largest (area) circle (corresponding to the most cited document) of the chart (**Fig. 2**a), the script corresponding to the click event is executed by automatically selecting the corresponding article from the document list (**Fig. 2**b), presenting the content of the article (**Fig. 2**c), indicating the corresponding author (**Fig. 2**d), and indicating the article's place of origin (**Fig. 2**e).

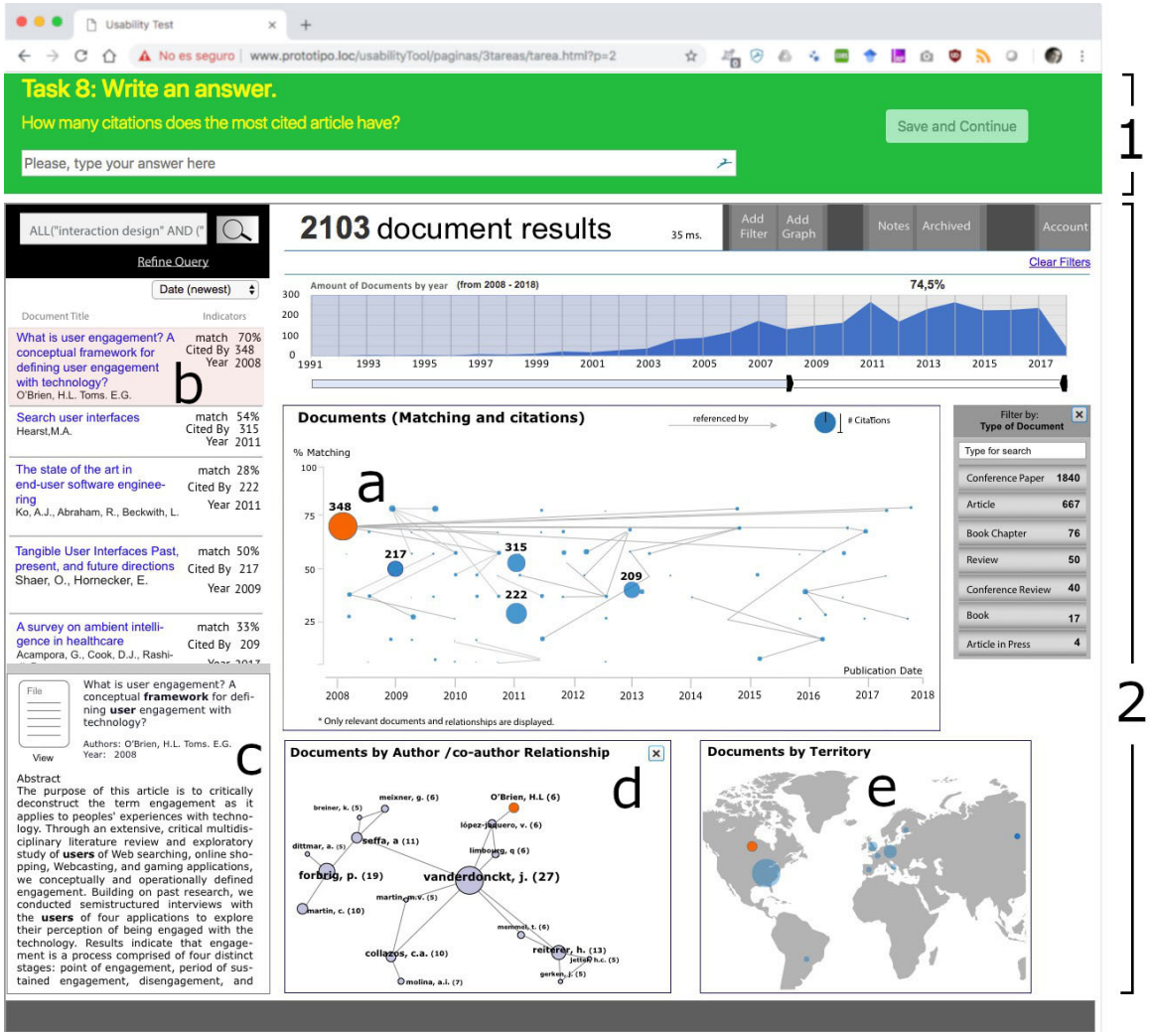


Fig. 2 *Example of the user interface of the evaluation platform containing 1) the task panel and 2) the user interface prototype which is designed with massiveData-ID patterns to perform the tasks for the Scopus problem. The user interface contains interactive controls to show: (a) the resulting documents based on the number of citations and matches with the search term; (b) list of resulting documents; (c) detailed content of*

a selected article; (d) network of authors of the set of documents; (e) geographical location of the place of origin of the article.

Evaluation platform. – We built an on-line evaluation platform that integrates and structures the prototypes that implement the experimental objects. Through this platform, the user is able to perform the experiment from beginning to end while his/her answers to the tasks are automatically captured as well as the time in seconds spent on performing them. In a single interface, the platform shows both the sequence of tasks and the corresponding UI prototype to perform the task, as shown in the web browser screen in **Fig. 2**. While the upper part labelled with the number 1 shows a static panel containing the tasks to perform and the button to save the answer and continue to the next task, the lower part labelled with the number 2 shows the UI prototype to perform the task. At the end of each treatment, the evaluation platform shows the CSUQ questionnaire and collects the results of each subject.

4.6 Experiment design

To assign treatments to experimental subjects, Vegas et. al. [39] describe three experimental design types: independent measures, repeated measures, and matched-pairs. We chose the *repeated measures* design (a.k.a. *within-subjects* design) since both treatments are applied to all subjects in our experiment. As blocked variable, we have used the problem. We are not interested in studying whether the use of a specific problem gets better results than another. That is why we have blocked the problem by applying both problems to both treatments.

The benefits of using this design are mainly based on maximizing the sample size and counterbalancing the effects of learning and fatigue [39]. The size of the sample is maximized since each experimental subject applies all treatments. Therefore, there will be repeated measures for the same experimental subject (one for each treatment). The learning effect, also known as "carryover effect", is controlled since the experimental subjects apply the treatments in different orders. Therefore, the knowledge learned in the first treatment is not applied or transferred to the subsequent treatments. The effect of fatigue is also controlled by applying the treatments in a different order since the problem and the tasks to be performed by applying a treatment are different from those of the subsequent treatments. This prevents the subject from getting tired or bored doing the same tasks. Moreover, this design prevents the subjects from confusing the treatment with the problem since the sequence order in which both the problem and the treatment must be performed is clearly defined. Table 7 shows the configuration of our repeated-measures experiment.

	Session 1		**Session 2**	
Group	**First Problem**	**First Treatment**	**Second Problem**	**Second Treatment**
G1	P.*scopus*	T.*without-patterns*	P.*pubmed*	T.*with-patterns*
G2	P.*pubmed*	T.*with-patterns*	P.*scopus*	T.*without-patterns*
G3	P.*scopus*	T.*with-patterns*	P.*pubmed*	T.*without-patterns*
G4	P.*pubmed*	T.*without-patterns*	P.*scopus*	T.*with-patterns*

Table 7 Design of the experiment

The design is structured in two *sessions* (two periods of time) within the same day; the second starts when the first one is over. Each session applies one treatment to one subject with one problem. We have defined four random-balanced *groups* of experimental subjects (i.e., the same number of experimental subjects in randomly assigned group) to assign them to the four possible problem-treatment sequences. Each sequence is made up of Session 1 (First Problem, First Treatment) and Session 2 (Second Problem, Second Treatment), as shown in Table 7. Consequently, each group has a different sequence order to apply the treatment in a problem. For instance, an experimental subject of G1 will first execute the P.*scopus* using T.*without-patterns* (Session 1), and then he/she will execute the P.*pubmed* using T.*with-patterns* (Session 2). In this way, the experiment is designed in such a way that all of the experimental subjects apply each treatment to one problem within each session.

4.7 Experiment procedure

The flow diagram in Fig. **3** summarizes the procedure that we followed to conduct the experiment. The procedure is strictly based on the experiment design configuration explained in Table 7. The procedure has been clearly labelled with numbers to explain each step.

Before starting the experiment, we explained the goals of the experiment to the experimental subjects as well as the role they played in it. We also randomly created the four groups of subjects (i.e., G1, G2, G3, and G4) and made sure that the number of subjects in the groups was balanced.

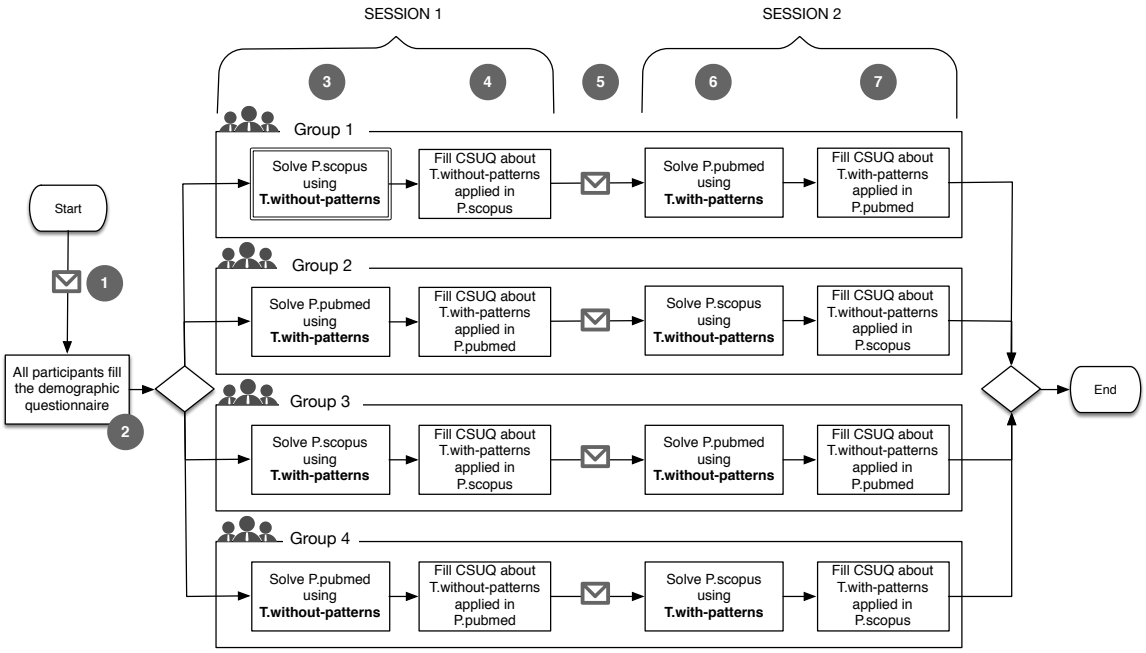


Fig. 3 *Experiment procedure*

Step 1. –The experimenter mailed each subject a URL with an evaluation platform that combines the treatment and problem for Session 1, depending on the group to which the subject belongs. We had 4 URL, one per group. For example, the URL received by a subject from the G1 allowed her/him access to T.*without-patterns* to solve the P.*scopus* problem as indicated in the double-lined "square" of Step3.

Step 2.- Once the experimental subjects accessed the evaluation platform, they answered a *demographic questionnaire*. The questionnaire collects information about the experimental subject including questions about academic profile, knowledge of bibliography search engines, and their expertise using them. The questions were the same for all experimental subjects independently of their group. Since the questionnaire was automated using Google Web Forms, the answers could be automatically saved as an Excel spreadsheet. The demographic questionnaire was filled out once at the beginning of the experiment. After the subjects completed the questionnaire, they continued to the next step.

Step 3.- The subjects started Session 1. The evaluation platform automatically redirects the experimental subject to the corresponding treatment and problem for Session 1 depending on the group to which she/he belongs. First, the platform displays a brief explanation about the problem to be solved, a brief description about the UI prototype, and the instructions to perform the tasks by using the UI prototype. Once the instructions were read, the subject started the sequence of the tasks that compose the experimental problems. Each response and click of the subject and the time (in seconds) spent to answer were automatically stored through the evaluation platform.

Step 4. - Once the subjects finished the experimental problem, the evaluation platform redirected them to the CSUQ questionnaire. Here, the subjects evaluated the satisfaction of using the assigned treatment. This CSUQ questionnaire was built as a web page using Google Web Forms, and the answers could be automatically saved as an Excel spreadsheet.

Session1 ends when the corresponding questionnaire is completed.

Step 5. - Between the two sessions, we had a slot of three minutes to send subjects an email with the URL of the evaluation platform for Session 2. For instance, the experimental subjects from G1 received the URL to access T.*without-patterns* to solve the P.*pubmed* problem.

Step 6.- The subjects started Session 2. This step is similar to Step 3. However, each subject interacts with the problem and treatment corresponding to Session 2.

Step 7. – In the same way as Step 4, the experimental subjects filled out the CSUQ questionnaire regarding the satisfaction of using the treatment applied in Session 2.

Once the experiment was over, one of the experimenters analysed the data to extract conclusions.

4.8 Threats to validity of the experiment

The analysis of threats to validity protects the results of the experiment by avoiding the inadequate selection of statistical tests, sample sizes, and other topics that influence the veracity of the results. To assess validity, there are four validity categories [40]: Conclusion, Internal, Construct, and External. Below, we describe the threats identified in our experiment and the considerations taken to try to mitigate them.

Conclusion Validity. - The threats that affect the ability to draw the correct conclusions about the relationship between the treatments and the outcome of the experiment. These threats are:

Low statistical power.- This threat is related to the power of the statistical test to reject the null hypothesis when it is false [41]. A low power means the statistical test increases the risk of drawing erroneous conclusions, that is, the risk of concluding there is no effect when, in fact, there is a high one. We used the *Linear Mixed Model* (LMM) as the statistical test, but it was not possible to calculate the statistical power for this test. Nevertheless, to face this threat, using G*Power[7], we have estimated how many subjects and how much effect size we needed to get a power of 95%. We found that a sample of 16 subjects and a *medium* effect size (d=0.50) result in a power of 95%. Therefore, our experimental sample size (R1 + R2 = 24 +8 = 32) is more than enough to guarantee significant effects and produce a power of 95%, decreasing the likelihood of drawing erroneous conclusions.

Violated assumptions of statistical tests. – This threat concerns the preconditions to be satisfied before applying a statistical test. We applied the LMM statistical test, and the assumption to be satisfied was the *normality of the residue*. Before applying the LMM, we applied the Shapiro-Wilk' W test to verify the residue normality. All residues had a normal distribution.

Reliability of measures. – This threat is associated with the measure's trustworthiness. Objective measures are more reliable than subjective measures since there is no intervention of human judgement. In our experiment, we had three response variables. Two of them (*effectiveness in use* and *efficiency in use*) are measured objectively since they are calculated and registered automatically by the evaluation platform. The *satisfaction in use* is the only response variable that is subjectively measured; it depends on the qualitative answer of the subjects. However, we minimized this threat by operationalizing the subjects' answers using a 7-point Likert scale.

Random heterogeneity of subjects. - This threat refers to the risk of having a group of subjects that is "very" heterogeneous or "very" homogeneous. In very heterogeneous groups, the individual differences may be greater than the differences produced by the treatments, and in very homogeneous groups, the group of subjects may not be representative of the study population. We avoided this threat by making two replicas of the same experiment where the group of each replica was homogeneous (with similar knowledge and background), but when the groups of each replica are compared, they are heterogenous (each group belonged to different domains).

Internal Validity. - The threats that, without the researcher's knowledge, can affect the conclusions about a possible causal relationship between the treatment and outcome. The threats are:

Maturation. – This concerns the causes that make the subject behave differently as the time passes. A negative cause can be that the subject gets tired or bored. To minimize this threat, the experiment was designed with two different problems in a short time, preventing the subject from getting bored. However, since the test is taken on-line, the user can get distracted or tempted to do other things different from the test. To avoid these effects, we asked the subjects to assign a specific time and the suitable conditions to perform the test exclusively. Moreover, the experimenters were in the same room to ensure that the subjects participated in the experiment the whole time.

Testing. - This threat appears when the experiment is repeated with the same subjects. Therefore, the subjects have previous knowledge about how it is conducted. This did not affect our experiment since each replication had different subjects.

Instrumentation. - This means that experimental objects (forms, questionnaires, UI prototypes) that are not properly designed can affect the experiment negatively. To avoid this threat, the evaluation platform together with the integrated questionnaires, UI prototypes, and the tasks were pre-evaluated by a researcher who was independent from the group of experimenters. As a result of the pre-evaluation, we solved ambiguity issues and changed the order of the tasks to improve coherence before conducting the experiment.

Selection. - This threat deals with the effects derived from the way in which the experimental subjects are selected. Our experiment was not affected by this threat since the subjects were selected taking into account their knowledge and experience in managing bibliography search engines.

[7] http://www.gpower.hhu.de/

Construct Validity. – The threats concerning the generalization of the experimental results to the concept or theory behind the experiment:

Mono-operation bias. - This threat means that when using a single independent variable, subject, or treatment, the experiment may under-represent or underestimate the construct and thus not give the full picture of the theory. We avoided this threat by using two treatments (T.*with-patterns* and T.*without-patterns*) in two different and relevant search contexts (i.e., Scopus and PubMed). Thus, the results obtained can be generalized for subjects who have similar profiles to the subjects we recruited in our replications.

Interaction of different treatments. – This threat means that there is no way of concluding whether the effect is due to the treatments or due to a combination of them. Our experiment was not affected by this threat since only one treatment was applied per group for each session. A combination of treatments was not possible in the same group per session. Therefore, the effects are produced directly by the treatments.

Hypothesis guessing. - This threat occurs when the experimental subjects figure out the purpose and intended result of the experiment. To avoid this threat, we did not explain to the experimental subjects the specific goals of the experiment or the response variables we aimed to measure.

Evaluation apprehension. – This threat occurs when the experimental subjects are afraid of being evaluated and try to be more attentive and focused on the evaluation, which is a behavior that affects the outcome of the experiment. To avoid this threat, before starting the experiment, we explained to the experimental subjects that all of the answers were important regardless of whether they were positive or negative. In the instructions of the tasks, we had the following text: *"Important: This is not a skill test (there are no correct or incorrect answers)."*

Experimenter expectancies. – This threat occurs when the experimenter consciously or unconsciously biases the results based on what she/he expects from the experiment. Unfortunately, this experiment is affected by this threat because the experimenter is the same one who created the evaluation platform, randomly assigned treatments to the experimental subjects, and performed the analysis of the data.

External Validity. – The threats concerning the limitations of generalizing the experimental results to industrial practice:

Interaction of selection and treatment. – This threat refers to the effect of not having a representative sample of the target population that we want to generalize. To reduce this threat, we ensured that the experimental subjects belonged to an environment in which "literature search" is a predominant activity, as is the case for "scientific research" and "library management".

Interaction of setting and treatment. – This threat refers to the effect of not having the representative material (e.g., tools) to carry out the process. Our experiment is affected by this threat since we used UI prototypes instead of real UIs. However, to minimize this threat, we designed and used high fidelity prototypes that, in the case of the control treatment, the prototype has UIs that look and work in a way similar to real UIs. The UIs of the prototype incorporate the same visual design, content, and interactions that the real UIs incorporate. To do this, each UI was designed upon an image of the original UI on which we added the interaction necessary to perform the experiment-specific tasks.

5 Data analysis

To contrast the hypotheses and make decisions about whether or not to reject the null hypotheses, in this study we define the *significance level* at 0.05, which is interpreted as a 5% probability of rejecting the null hypothesis (H_0), given that it were true [41] (also known as type I error or "false positives"). In this study, each response variable is analysed by following two steps: 1) descriptive analysis and 2) analytical analysis.

First Step. - We apply descriptive statistics to observe how the response variables behave between the treatments in the subjects' sample. To do this, we will visually inspect and compare the data distribution of each treatment by using box-and-whiskers plots and explain the differences or similarities in data observed in terms of descriptive measures such as median, mean, standard deviation (SD), and quartiles. This allows us to have an overview of the response variable behaviour regarding the treatments.

Second Step. - We apply inferential statistics to determine whether the sample-based observations in the first step reflect population-level parameters. To do this, we apply the *Linear Mixed Model* (LMM) statistical method, also known as *multilevel model*, *linear hierarchy model*, or *random coefficient model* [42]. We have chosen this statistical method for two reasons: 1) This method deals with correlated data produced by the repeated measures; and 2) The method is suitable when analysing more than one repeated measure [43] as is the case of our experiment where we have two repeated measures: the *treatment* and the *problem*.

To apply the LMM statistical test, the residue's normality assumption must be satisfied. Therefore, we applied the Shapiro-Wilk's (W) test to the LMM residues corresponding to each response variable and replica, as shown in Table 8. The W test is widely recommended for assessing whether the data distribution

follows a normal distribution [44]. Most of the residues fulfilled the normality assumption, but for those that did not (specifically the residues corresponding to the satisfaction variable in R1), we transformed the residues by applying the *Two-Step approach* [45] to achieve the residue's normality. Consequently, the W test applied to the transformed residues confirmed the normality of these residues.

	R1		**R2**	
Treatment	**T.*with-patterns***	**T.*without-patterns***	**T.*with-patterns***	**T.*without-patterns***
Effectiveness	0,997	0,521	0,932	0,932
Efficiency	0,422	0,552	0,111	0,194
Satisfaction	0,999(*)	1,000(*)	0,939	0,214
(*) p-value from a residue previously transformed using Two-Step approach [45].				

Table 8 Resulting P-values from the Shapiro-Wilk normality test applied to the residue value from the Lineal Mixed Model (LMM) for Replica 1 (R1) and Replica 2 (R2)

The resulting *p-values* of applying the LMM test help us analyse whether or not the treatments yield different effects in the subjects. If the p-value is less than or equal to the significance level (p-value<=0,05), we conclude that the treatments yield "different effects" (i.e., there are significant differences between the treatments). Otherwise, the treatments yield the "same effects".

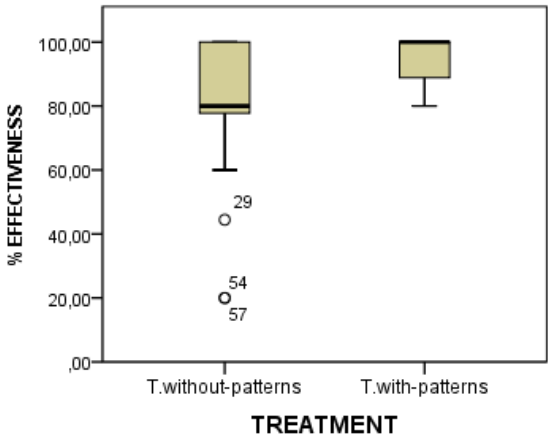
If the effects are different, we aim to measure how big or small the difference is since this measure will allow us to know the magnitude of the effect yielded by the UIs based on patterns. To do this, we estimate the *effect size*, which is defined as the quantitative magnitude of the effect in the studied sample by using the Cohen *d* index because this index is well recommended when assessing the differences between treatments [46]. The resulting *d* value allows us to define whether the effect size is *small (d=0.20)*, *medium (d=0.50)*, or *large (d=0.80)*, according to the conventions established by Cohen [47].

We used the SPSS statistical program to carry out both Step 1 and Step 2. After analysing the effects produced in each response variable, we draw conclusions considering the effects in: a) each replica separately; and b) the total sample, by combining the individual participants' data from both replicas. This is an increasingly popular approach for synthesizing data in disciplines such as medicine, where it has been shown to have potential advantages [48].

We combined the replicas by adding a *moderator variable* that moderates the relationship between the treatment and the response variable. Therefore, we analyse the impact the moderator variable causes on the relationship between the treatments and the response variable under study.

5.1 Effectiveness

In this section, we deal with the research question RQ1 and H_{01}: *The effectiveness in use interacting with UIs that incorporate massiveData-ID patterns is similar to the effectiveness in use observed when UIs do not incorporate such massiveData-ID patterns*. The analysis of effectiveness for both replicas (R1 and R2) is described below.



	TREATMENT	
	T.without-patterns	T.with-patterns
Mean	78.93	96.18
Median	80.00	100.00
SD	22.87	6.62
Minimum	20.00	80.00
Maximum	100.00	100.00
Q1	72.23	88.89
Q3	100.00	100.00

	p-value (LMM)
Treatment	0.000

Effect Size	1.02

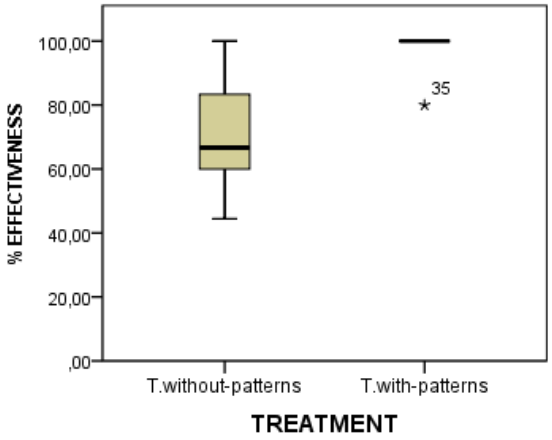
Fig. 4 *Box and whisker plots along with their descriptive statistics, p-value resulting from Lineal Mixed Model (LMM) method, and effect size index for effectiveness in Replica 1 (R1)*

The box-plot in **Fig. 4** compares the distribution of the effectiveness percentages between the two treatments for R1. The difference between the medians indicates that T.*with-patterns* has better values for effectiveness than T.*without-patterns*. The mean value of effectiveness for T.*with-patterns* is 96.18% with SD=6.62 and 78.93% with SD=22.87 for T.*without-patterns*. This means that the treatments produce different effects in effectiveness, with the T.*with-patterns* getting better average results for effectiveness, exceeding T.*without-patterns* by 17.25% (96.18% - 78.93%).

The low effectiveness obtained with T.*without-patterns* may be caused by the subjects of the outliers that negatively affect the total effectiveness (points 29, 54, and 57 in the plot). Two of them reached 20% effectiveness (points 54 and 57) and the other reached 44% effectiveness (point 29). With regard to the median value and the maximum value of the effectiveness percentage, we can say that by using T.*without-patterns,* 50% of the subjects got effectiveness values ranging from 80% to 100% (range between Median and Maximum) whereas by using T.*with-patterns*, the same percentage of subjects reached exactly 100% effectiveness. Although the subjects using T.*without-patterns* are able to obtain 100% effectiveness, only a few can actually obtain it.

In both treatments, the distribution of effectiveness is asymmetrical since the length of the range Q2–Q1 is different to the length of the range Q3–Q2. However, this asymmetry evinces better effectiveness results when using T.*with-patterns*. In T.*with-patterns*, the most frequent effectiveness value is 100% and is reached by around 50% of the subjects. In T.*without-patterns*, the most frequent values of effectiveness range between 72.23% and 80% and are reached by around 25% of the subjects (range Q1-Median).

We applied the LMM statistical method to look for significant differences. The p-value=0.000<0.05 and the effect size of 1.02 for R1 indicate that there is a significant difference between the treatments and this difference is *large* in magnitude, with T.*with-patterns* getting better results. These results agree with the descriptive analysis.



TREATMENT		
	T.without-patterns	T.with-patterns
Mean	71.11	97.14
Median	66.67	100.00
SD	19.03	7.56
Minimum	44.44	80.00
Maximum	100.00	100.00
Q1	60.00	100.00
Q3	88.89	100.00

	p-value (LMM)
Treatment	0.025

Effect Size	1.79

Fig. 5 *Box-and-whisker plots along with their descriptive statistics, p-value resulting from Lineal Mixed Model (LMM) method, and effect size index for effectiveness in Replica 2 (R2)*

The box-plot in **Fig. 5** compares the effectiveness between the two treatments for R2. The difference between the medians indicates that the experimental subjects achieved better effectiveness percentages when they used T.*with-patterns*. The comparison of means indicates that the subjects using T.*with-patterns* reached a higher effectiveness average (97.14%, SD=7.56) than the subjects using T.*without-patterns* (71.11%, SD=19.03). In summary, the UIs designed with patterns achieve 26.03% more effectiveness than conventional UIs designed without patterns.

The T.*without-patterns* data distribution indicates that 50% of the experimental subjects using the treatment got effectiveness between 60% and 88.89% (Q1 and Q3, respectively). From this range, most of the experimental subjects got effectiveness percentages between 60% and 66.67% (Q1 and Median, respectively), whereas fewer subjects got effectiveness values between 66.67% and 88.9% (Median and Q3, respectively).

The T.*with-patterns* data distribution indicates that all of the experimental subjects reached 100% effectiveness, except one subject (point 35 in the plot) who reached 80% effectiveness.

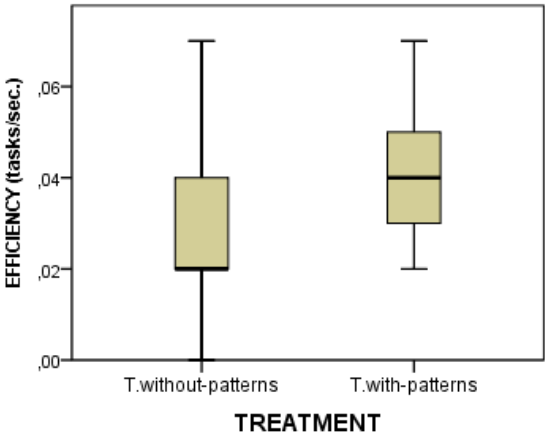
The p-value=0.025<0.05 of LMM and the effect size (d=1.79) for the effectiveness variable in R2 indicate that there is a significant difference between the treatments and that this difference is *large* in magnitude, with T.*with-patterns* getting better results (an effectiveness average of 97.14%).

Based on the analytical and descriptive results, we conclude that **H_{01}** is rejected for both R1 and R2. This means that, in both replicas, the effectiveness in use when interacting with UIs that incorporate massiveData-ID patterns is *different* from the effectiveness in use observed when UIs do not incorporate such massiveData-ID patterns, with the UIs based on patterns achieving the highest effectiveness values. The difference in effectiveness between the treatments is *large* in magnitude in both replicas.

5.2 Efficiency

In this section, we answer the research question RQ2 by analysing H_{02}: *The efficiency of the user when using the UIs designed with the massiveData-IDP is similar to the efficiency obtained when using conventional UIs that do not incorporate the patterns*.

Using the formula for efficiency in use indicated in Table 2, we calculated the average number of tasks per second performed by each experimental subject using the treatments.



TREATMENT		
	T.without-patterns	T.with-patterns
Mean	0.03	0.04
Median	0.02	0.04
SD	0.02	0.01
Minimum	0.00	0.02
Maximum	0.07	0.07
Q1	0.02	0.03
Q3	0.04	0.05

	p-value (LMM)
Treatment	0.002

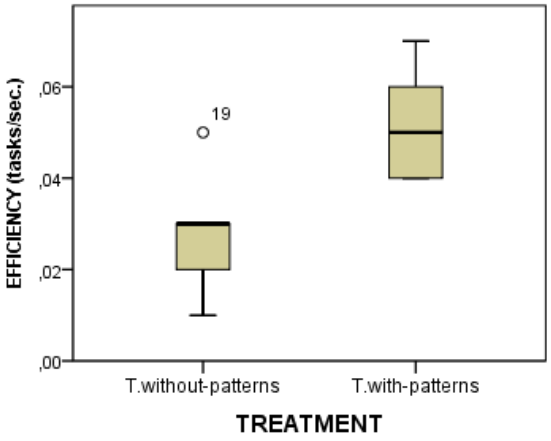
Effect Size	0.92

Fig. 6 *Box-and-whisker plots along with their descriptive statistics, p-value resulting from Lineal Mixed Model (LMM), and effect size index for efficiency in Replica 1 (R1)*

The box-and-whisker plot in **Fig. 6** compares the efficiency obtained (measured in tasks per second) for each treatment. The median comparison shows that T.*with-patterns* gets a better score (0.04 tasks/sec.) than T.*without-patterns* (0.02 tasks/sec.). This observation is fully supported by the means value where the subjects using T.*with-patterns* are able to perform approximately two-and-a-half tasks per minute (Mean=0.04 tasks/sec.=2.4 tasks/min, SD=0.01), whereas the subjects using T.*without-patterns* are able to perform approximately two tasks (Mean=0.03 tasks/sec.=1.8 tasks/min., SD=0.02).

T.*with-patterns* achieves more concentrated and better efficiency scores than those obtained with T.*without-patterns*. By analysing the range Maximum-Minimum, we can say the T.*without-patterns* distribution is fairly disperse with efficiency values ranging from 0 tasks/min. to 4.2 tasks/min. (Minimum=0 tasks/sec. and Maximum=0.07 tasks/sec., respectively), whereas the T.*with-patterns* results distribution is more concentrated with values ranging from 1.2 tasks/min. to 4.2 tasks/min. (Minimum=0.02 tasks/sec. and Maximum=0.07 tasks/sec., respectively). Although both treatments reach a maximum value of 4,2 tasks/min. (Maximum= 0.07 tasks/sec.), the difference between the treatments is in the minimum value. This difference indicates that, in the worst case, the subjects who used T.*with-patterns* achieved at least 1 task in one minute (Minimum= 0.02 tasks/sec. = 1.2 tasks/min.), while no task was achieved by the subjects who used T.*without-patterns* in the same time period of one minute (Minimum= 0.0 tasks/sec. = 0 tasks/min.).

From the analytical analysis perspective, the p-value=0.002<0.05 calculated with the LMM for the efficiency variable in R1 means that, there is a statistically significant difference between the treatments and that this difference is *large* (*d*=0.92) in magnitude.



	TREATMENT	
	T.without-patterns	T.with-patterns
Mean	0,03	0,05
Median	0,03	0,05
SD	0,01	0,01
Minimum	0,01	0,04
Maximum	0,05	0,07
Q1	0,02	0,04
Q3	0,03	0,06

	p-value (LMM)
Treatment	0,001

Effect Size	1,97

***Fig.** 7 Box-and-whisker plots along with their descriptive statistics, p-value resulting from Lineal Mixed Model (LMM) method, and effect size index for efficiency in Replica 2 (R2)*

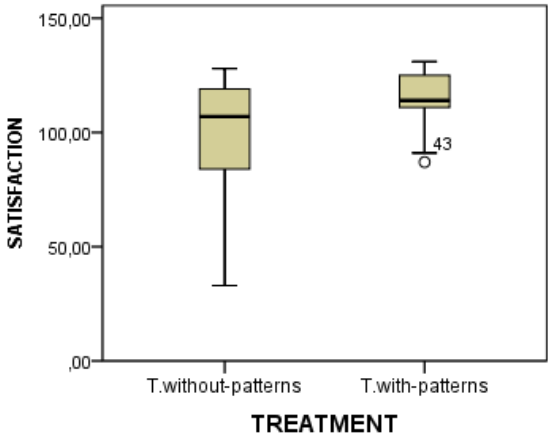
The box-and-whisker plot in **Fig. 7** compares the efficiency obtained by each treatment for R2. The plot shows a clear difference between the medians from the two treatments, with T.*with-patterns* getting better results. While the subjects using T.*without-patterns* performed an average of 2 tasks per minute (Mean=0.03 tasks/sec.=1.8 tasks/min., SD= 0.01) approximately, the subjects using T.*with-patterns* performed 3 tasks per minute (Mean=0.05 tasks/sec.=3 tasks/min., SD=0.01). In the worst case (i.e., the lowest efficiency values reported), the subjects using T.*without-patterns* achieved at least 1 task per minute approximately (Minimum=0.01 tasks/sec.=0.6 tasks/min.), while the subjects using T.*with-patters* achieved at least 2 tasks and a half per minute approximately (Minimum= 0.04 tasks/sec.=2.4 tasks/min.). Considering the highest efficiency values, the subjects using T.*with-patterns* achieved four tasks per minute approximately (Maximum=0.70 tasks/sec. = 4.2 tasks/min.) compared to the three tasks per minute (Maximum=0.05 tasks/sec.= 3 tasks/min.) achieved by a subject using T.*without-patterns*, which exceptionally correspond to outlier 19 in the plot.

The resulting p-value=0.001<0.05 from the LMM stated that, the treatments produced different effects and that this difference is *large* (*d*=1.97) in magnitude, with T.*with-patterns* achieving better results.

We conclude that $\mathbf{H_{02}}$ for R1 and R2 is rejected. This means that, in both replicas, the efficiency in use when interacting with UIs that incorporate massiveData-ID patterns is *different* to the efficiency in use observed when UIs do not incorporate these massiveData-ID patterns. The UIs based on patterns achieved more tasks per minute than the UIs without patterns (an average of approximately four tasks per minute compared to three tasks per minute, respectively). The difference in efficiency between the treatments is *large* in magnitude in both replicas.

5.3 Satisfaction

In this section, we answer the research question RQ3 by analysing H_{03}: *The satisfaction in use interacting with UIs that incorporate massiveData-ID patterns is similar to the user satisfaction in use observed when UIs do not incorporate such massiveData-ID patterns*. The satisfaction in use was measured through the CSUQ questionnaire, as indicated in Table 2.



TREATMENT		
	T.without-patterns	T.with-patterns
Mean	98.44	114.56
Median	107.00	114.00
SD	28.14	12.26
Minimum	33.00	87.00
Maximum	128.00	131.00
Q1	82.00	109.50
Q3	120.50	125.00

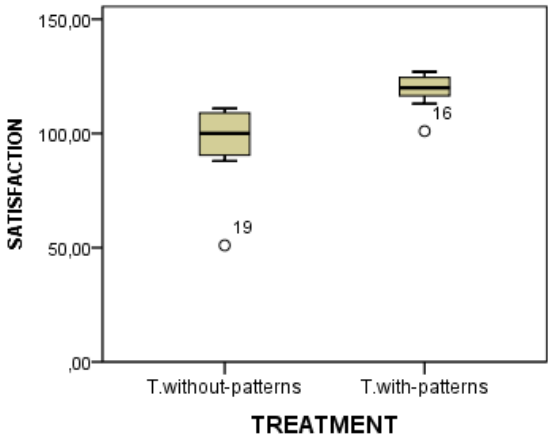
	p-value (LMM)
Treatment	0.005

Effect Size	0.74

Fig. 8 *Box-and-whisker plots along with their descriptive statistics, p-value resulting from Lineal Mixed Model (LMM), and effect size index for satisfaction in Replica 1 (R1)*

Fig. 8 compares the data distribution of satisfaction perceived by the subjects between the treatments in R1. The plot shows a slight difference between the medians. However, the satisfaction scores obtained with T.*without-patterns* are more disperse than those obtained with T.*with-patterns*, as can be seen by comparing the interquartile range (Q3-Q1) between the two treatments. The analysis of the means indicates that the subjects who used T.*with-patterns* perceived a higher average score of satisfaction (Mean=114.56, SD=12.26) than the subjects who used T.*without-patterns* (Mean=98.44, SD=28.14). Although none of the treatments reaches the maximum satisfaction score (133), T.*with-patterns* got the highest satisfaction score of 131 compared to the score of 128 obtained with T.*without-patterns*. Considering the 25% of subjects with the lowest scores of satisfaction (Q1-Minimum) in both treatments, we can say that any satisfaction score obtained with T.*with-patterns* is superior to any score obtained with T.*without-patterns*.

The p-value=0.005<0.05 from the LMM method indicates that there is a significant difference between the treatments, ratifying the difference observed in the visual inspection. This difference is *medium* (*d*=0.74) in magnitude, with the T.*with-patterns* achieving better results.



TREATMENT		
	T.without-patterns	T.with-patterns
Mean	94.42	118.57
Median	100.00	120.00
SD	21.06	8.99
Minimum	51.00	101.00
Maximum	111.00	127.00
Q1	88.00	113.00
Q3	110.00	125.00

	p-value (LMM)
Treatment	0.043

Effect Size	1.49

Fig. 9 *Box-and-whisker plots along with their descriptive statistics, p-value resulting for Lineal Mixed Model (LMM) method, and effect size for satisfaction in use in Replica 2 (R2)*

Fig. 9 compares the data distribution of the perceived satisfaction between treatments for R2. Although none of the treatments reaches the maximum satisfaction score (133), the visual difference between the medians and the mean values shows that T.*with-patterns* got higher satisfaction scores (Mean=118.57, SD=8.99) than T.*without-patterns* (Mean=94.42, SD=21.06). The satisfaction scores in T.*with-patterns* (Maximum-Minimum = 127-101 = 26) are less variable than those in T.*without-patterns* (Maximum-Minimum = 111-51 = 60). Therefore, the satisfaction scores obtained with T.*with-patterns* are higher and less variable than those obtained with T.*without-patterns*.

The resulting p-value=0.043<0.05 obtained from the LMM method confirms that there is a significant difference between treatments and the magnitude of this difference is *large* (*d*=1.49), with the T.*with-patterns* yielding the best satisfaction scores.

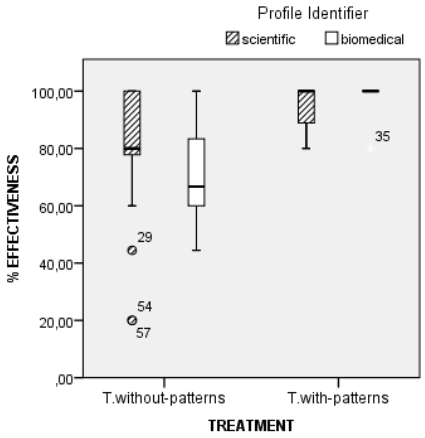
We conclude that **H_{03}** is rejected for both R1 and R2. Note that the satisfaction in use when interacting with UIs that incorporate massiveData-ID patterns is *different* from the user satisfaction in use when UIs do not incorporate these massiveData-ID patterns, with the UIs based on patterns yielding the highest satisfaction scores. The difference in satisfaction between treatments is more pronounced in R2 than R1 since the magnitude of the effect in R2 is *large*, whereas in R1 the magnitude is *medium*.

5.4 Replica combination

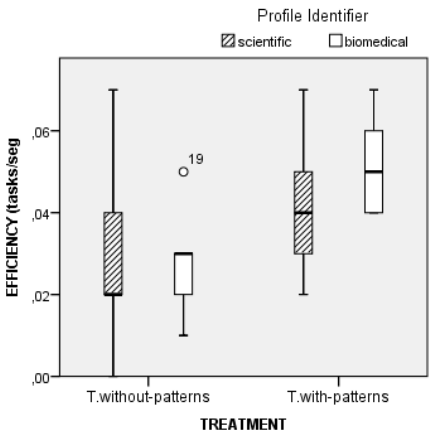
In addition to studying the effect produced by the treatments in each replica separately, we analysed the effects produced in the whole sample. To obtain a single set of observations, we combined the individual observations for each replica by adding the *profile* as a moderator variable [49] that contains two possible values: *scientific* and *biomedical*. This variable represents the differences between the experimental subjects in accordance with the use preferences of the search websites shown in Table 4. The scientific profile represents the R1 subjects and the biomedical profile represents the R2 subjects.

Fig. 10 shows the box-and-whisker diagrams comparing the response variables (i.e., effectiveness, efficiency, and satisfaction) between T.*without-patterns* and T.*with-patterns*, differentiating the two profiles. The resulting p-values from LMM and the effect size values are also presented in this Figure.

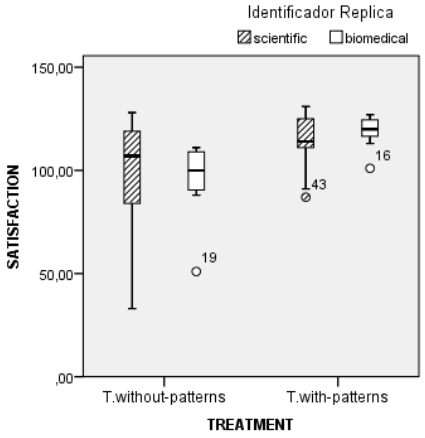
a)



b)



c)



d)

P-values (LMM) and Effect Size values			
	Effectiveness	**Efficiency**	**Satisfaction**
Treatment	0.000	0.000	0.001
Problem	0.098	0.393	0.078
Profile	0.406	0.246	0.960
Treatment * Profile	0.206	0.162	0.591
Effect Size	1.77	1.06	0.87

Fig. 10 *Box-and-whisker plots comparing the treatments considering the "profile" moderator variable for the combined replicas for a) effectiveness, b) efficiency, and c) satisfaction. The p-values resulting from the LMM and Effect Size value for each response variable are shown in d)*

5.4.1 Effectiveness

With respect to effectiveness (**Fig. 10**a), the difference in medians between profiles is more evident for T.*without-patterns* than for T.*with-patterns*.

For T.*without patterns*, the comparison of the medians between profiles indicates that the scientific profile got (80%) higher effectiveness value than the biomedical profile (66.67%). The mean value is 78.93% with SD=22.87 for the scientific profile and 71.11% with SD=19.03 for the biomedical value. The visual comparison of the size of the Q3-Maximum range indicates that the scientific profile reached exactly 100% effectiveness, while the biomedical profile ranged from 88.89% to 100% effectiveness.

For T.*with-patterns*, the median of both profiles is 100% and the mean is 96.18% with SD=6.62 for the scientific profile (**Fig. 4**) and 97.14% with SD=7.56 for the biomedical profile (**Fig. 5**). These mean values are close to 100% effectiveness, and the difference between them is slight. However, the variability of effectiveness values is higher for the scientific profile than for the biomedical profile. In fact, in contrast to the biomedical profile, the scientific profile's effectiveness values are more spread out from the mean value, as can be seen visually when comparing the size of the Q3-Q1 range between the two profiles.

The p-value=0.000<0.05 from LMM applied to the effectiveness variable (**Fig. 10**d) shows that there is a significant difference between treatments and the magnitude of this difference is considered to be *large* (*d*=1.77), where the T.*with-patterns* shows better effectiveness values. The p-value=0.206>0.05 corresponding to the interaction between the treatment and the profile is not significant; therefore, the profile does not affect the effects produced by the treatments. For the aggregated replicas, $\mathbf{H_{01}}$ is rejected.

5.4.2 Efficiency

With respect to efficiency (**Fig. 10**b), the median of the biomedical profile is higher than the media of the scientific profile for both treatments.

In T.*without-patterns*, the comparison of the medians between profiles indicates that the biomedical profile got (0.03 tasks/sec. = 1.8 tasks/min., **Fig. 7**) higher efficiency value than the scientific profile (0.02 tasks/sec. = 1.2 tasks/min., **Fig. 6**). The mean value is the same for both profiles (1.8 tasks/min.=0.03 tasks/sec.), but there is a slight difference in the SD value (0.02 for R1 and 0.01 for R2). Comparing the range between Q3 and the Maximum, we can say that the biomedical profile can obtain a maximum of 1.8 tasks/min. (0.03 tasks/sec., **Fig. 7**), while the scientific profile can exceed this value, being able to obtain a maximum efficiency value of 4.2 tasks/min. (0.07 tasks/sec., **Fig. 8**).

In T.*with-patterns*, the biomedical profile has better values of efficiency that are more concentrated around the mean value than those presented by the scientific profile. The comparison of medians indicates that values of efficiency for the biomedical profile are above 3 tasks/minute (Median=0.05 tasks/sec., **Fig. 7**), while the values of the scientific profile are above 2.4 tasks/minute (Median=0.04 tasks/sec., **Fig. 5**). The difference in the size range Q3-Q1 between profiles shows that the distribution of the biomedical profile data is more concentrated than for the scientific profile.

The p-value=0.000<0.05 from LMM applied to the efficiency variable (**Fig. 10**d) shows that there is a significant difference between treatments, with the T.*with-patterns* achieving better results. The magnitude of this difference is *large* (*d*=1.06) in magnitude. The p-value=0.162>0.05 corresponding to the interaction between the treatment and the profile is not significant; therefore, the profile does not affect the effects produced by the treatments. For the aggregated replicas, $\mathbf{H_{02}}$ is rejected.

5.4.3 Satisfaction

With respect to efficiency (**Fig. 10**c), in T.*without-patterns*, the comparison of medians indicates that scientific profile got better satisfaction scores than the biomedical profile. However, in T.*with-patterns*, the scientific profile achieved lower satisfaction scores than the biomedical profile.

The p-value=0.001<0.05 of LMM applied to the satisfaction variable (**Fig. 10**d) means that there is a significant difference between the treatments and that this difference is *large* (*d*=0.87), with T.*with-patterns* getting better satisfaction scores. The p-value=0.551>0.05 corresponding to the interaction between the treatment and the profile is not significant; therefore, the profile does not affect the effects produced by the treatments. For the aggregation of replicas, $\mathbf{H_{03}}$ is rejected.

6 Discussion

This section discusses the major findings and their meaning based on the contributions made by other authors and also highlights the limitations of our findings. The major finding of this empirical study is that users using UIs that are implemented with massive-ID patterns achieved better scores of effectiveness,

efficiency, and satisfaction than those using UIs implemented without the patterns. In general terms, the proposed patterns provide improvements in usability, allowing the user to achieve the proposed tasks in less time and in an easy, natural, and simple way. The meaning of these findings is detailed for each research question:

RQ1: Is *effectiveness in use* affected by the use of massiveData-ID patterns in the UIs? The data analysis results from both replicas (R1 and R2) indicated that effectiveness in use is higher when UIs are designed with massiveData-ID patterns. From the perspective of the subjects' profile, it is important to highlight that the subjects with the scientific profile obtained the best effectiveness rates. This can be explained by the fact that the subjects of this profile are more familiar with the search and retrieval of information because of their professional training in documentation and library management.

RQ2: Is *efficiency in use* affected by the use of massiveData-ID patterns in the UIs? The data analysis results from both replicas indicate that the use of massiveData-ID affects efficiency in use, allowing the users to perform more tasks per minute. A plausible justification for this result is that the proposed patterns accelerate and enrich the way in which users visualize and navigate through the data, replacing the traditional tabular, flat, and disconnected representation of data with a more dynamic and connected one. The PT1, PT2, and PT4 patterns especially enrich the visualization of data by connecting contents, suggesting relevant information, and facilitating the filtering of results.

RQ3: Is user *satisfaction in use* affected by the use of massiveData-ID patterns in the UIs? The data analysis results indicated that users who used UIs designed with massiveData-ID patterns got higher satisfaction scores. Some positive comments of the subjects about the UIs designed with patterns indicated that the idea of having all of the information in a single UI with connected content and recommendations helped them to perform the tasks. However, a small number of subjects indicated that information overload made it difficult to identify the correct answers. This situation has led us to think that it may be necessary to differentiate between novice and expert users and design UI prototypes for each type of user.

The results obtained contribute to the existing knowledge on patterns mentioned in the section of related works, presenting a set of patterns dedicated to dealing with complex data that have been empirically tested according to the impact produced in massive data analysis environments.

7 Conclusions

In this paper, we evaluate the applicability of massiveData-ID patterns (patterns for analysing massive amount of data) in terms of the impact they cause in two different knowledge domains (biomedical and scientific). To do that, we empirically evaluate the usability in use of the UIs designed with massiveData-ID patterns by measuring the effectiveness, efficiency, and satisfaction of users using such UIs. We consider the resulting impact as the result caused by the set of patterns rather than the isolated result caused by individual patterns. The empirical study consisted of conducting a repeated measure designed experiment considering two replicas, one with 24 experimental expert subjects in scientific literature search and one with eight experimental expert subjects in biomedical literature search. The results of the experiments allowed us to test three hypotheses, one for each usability measure, which was analysed using the Lineal Mixed Model statistical method. For each hypothesis, we evaluate whether using UIs designed with or without massiveData-ID patterns produce similar or different effects in the usability measure.

Our findings suggest that UIs designed with massive-ID patterns help users to achieve higher scores in effectiveness, efficiency, and satisfaction. The relevant findings of this empirical study can be outlined as follows:

a) The massiveData-ID patterns can be used to improve the usability of UIs that handle massive amounts of data.
b) Considering the study of each replica and the aggregated replicas, we conclude that the massiveData-ID patterns implemented in the UIs produce good usability results regardless of the knowledge, background, or expertise of the subject.
c) The impact on effectiveness, efficiency, and satisfaction produced by the set of patterns can be translated to more satisfied users who are able to complete more tasks in less time.
d) The study indicated that the background and knowledge of the subjects do not affect the results produced by the patterns. In other words, the profile of the subjects does not affect the results obtained for effectiveness, efficiency, and satisfaction.

These findings contribute to the existing literature by providing a proven set of patterns for designing UIs that deal with large amounts of data. Thus, the set of patterns becomes a tool that can be used by designers and developers as knowledge to guide the design of UIs that support massive data.

It is important to highlight that our results must be contextualized according to the characteristics of the experiment (i.e., number of participants, profile of the participants, proposed problems, number of replications). Therefore, more replications with similar numbers of participants are desirable in order to be able to confirm the promising results of the patterns.

As future steps in this research, we plan to perform more replications of the experiment in order to confirm the obtained results. Furthermore, we plan: i) to individually evaluate the patterns in order to identify the strengths and weaknesses of each pattern and define the suitability of its use, and ii) to evaluate the written content of each pattern to identify whether or not the narrative can help designers and developers understand and implement the pattern.

Acknowledgments

The authors thank the members of the PROS Center Genome group for productive discussions. In addition, it is also important to highlight that the Secretaría Nacional de Educación, Ciencia y Tecnología (SENESCYT) and the Escuela Politécnica Nacional from Ecuador have supported this work. This project has also been developed with the financial support of the Spanish State Research Agency and the Generalitat Valenciana, under the projects TIN2016-80811-P and PROMETEO/2018/176, and co-financed with ERDF.

Bibliography

1. C. Alexander, S. Ishikawa, and M. Silverstein, "A Pattern Language," *Ch. Alexander*. p. 1171, 1977.
2. A. Toxboe, "User interface design pattern library," *UI Patterns*, 2013. Available at. http://ui-patterns.com. Accessed on Feb-05-2018.
3. J. O. Borchers, "Interaction Design Patterns : Twelve Theses," *Position Pap. CHI Work. "'Pattern Lang. INteractoin Des. Build. Momentum,'"* 2000.
4. Å. Granlund, D. Lafrenière, and D. A. Carr, "A Pattern-Supported Approach to the User Interface Design Process," 2001.
5. Yahoo, "Yahoo Design Pattern Library," 2006. Available at. https://developer.yahoo.com/ypatterns/everything.html. Accessed on Apr-03-2017.
6. J. Tidwell, "Common ground: A pattern language for human-computer interface design." O'Reilly Media, 1999.
7. C. E. Iñiguez-Jarrín, J. I. Panach, and Ó. Pastor, "Defining Interaction Design Patterns to Extract Knowledge from Big Data," in *Advanced Information Systems Engineering*, 2018, vol. 10816, pp. 539–553, DOI:10.1007/978-3-319-91563-0.
8. Z. Lu, "PubMed and beyond: A survey of web tools for searching biomedical literature," *Database*, vol. 2011, p. baq036, 2011, DOI: 10.1093/database/baq036.
9. N. Fiorini *et al.*, "PubMed Labs: an experimental system for improving biomedical literature search," *Database*, vol. 2018, Jan. 2018, DOI: 10.1093/database/bay094.
10. J. L. Marill, N. Miller, and P. Kitendaugh, "The MedlinePlus public user interface: studies of design challenges and opportunities.," *J. Med. Libr. Assoc.*, vol. 94, no. 1, pp. 30–40, Jan. 2006.
11. Genomenon, "Mastermind - Comprehensive Genomic Search Engine." Available at. https://mastermind.genomenon.com/. Accessed on Apr-22-2018.
12. C. Wu, X. Jin, G. Tsueng, C. Afrasiabi, and A. I. Su, "BioGPS: building your own mash-up of gene annotations and expression profiles," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D313–D316, 2016, DOI: 10.1093/nar/gkv1104.
13. B. M. Good, E. L. Clarke, S. Loguercio, and A. I. Su, "Linking genes to diseases with a SNPedia-Gene Wiki mashup," *J. Biomed. Semantics*, vol. 3, no. 1, p. S6, 2012, DOI: 10.1186/2041-1480-3-S1-S6.
14. DigitalScience, "Dimensions." Available at. https://app.dimensions.ai/discover/publication. Accessed on Mar-03-2018.
15. VOSviewer, "Visualizing scientific landscapes," *Centre for Science and Technology Studies, Leiden University*, 2015. Available at. http://www.vosviewer.com/.
16. S. M. Douglas, G. T. Montelione, and M. Gerstein, "PubNet: a flexible system for visualizing literature derived networks.," *Genome Biol.*, vol. 6, no. 9, p. R80, 2005, DOI: 10.1186/gb-2005-6-9-r80.
17. J. Borchers, J. By-Buschmann, and Frank, *A pattern approach to interaction design*. Wiley, 2001.
18. M. Van Welie, "Patterns in Interaction Design," 2008. Available at. http://www.welie.com/patterns/. Accessed on Mar-01-2018.

19. A. Seffah and M. Taleb, “Tracing the evolution of HCI patterns as an interaction design tool,” *Innov. Syst. Softw. Eng.*, vol. 8, no. 2, pp. 93–109, Jun. 2012, DOI: 10.1007/s11334-011-0178-8.
20. J. Tidwell, *Designing interfaces: Patterns for effective interaction design*. “ O’Reilly Media, Inc.,” 2010.
21. E. G. Nilsson, “Design patterns for user interface for mobile applications,” *Adv. Eng. Softw.*, vol. 40, no. 12, pp. 1318–1328, 2009.
22. B. Scott and T. Neil, *Designing web interfaces: Principles and patterns for rich interactions*. “ O’Reilly Media, Inc.,” 2009.
23. I. Graham and Ian, *A pattern language for Web usability*. Addison-Wesley, 2003.
24. D. K. Van Duyne, J. A. Landay, and J. I. Hong, *The design of sites : patterns, principles, and processes for crafting a customer-centered Web experience*. Addison-Wesley, 2003.
25. P. Martín-Rodilla and J. I. Panach, “Applications in the Context of Cultural Heritage Data,” 2014.
26. E. Folmer, “Usability patterns in games,” *Futur. Play*, vol. 6, 2006.
27. J. Borchers, “The Aachen Media Space: Design Patterns for Augmented Work Environments,” in *Designing User Friendly Augmented Work Environments*, Springer, 2009, pp. 261–312.
28. M. Schmettow, “User interaction design patterns for information retrieval,” *Eur. 2006*, pp. 489–512, 2006.
29. P. Cremonesi, M. Elahi, and F. Garzotto, “Interaction Design Patterns in Recommender Systems,” in *Proceedings of the Biannual Conference on Italian SIGCHI Chapter*, 2015, pp. 66–73, DOI:10.1145/2808435.2808442.
30. N. Seidel, “Empirical Evaluation Methods for Pattern Languages: Sketches, Classification, and Network Analysis,” in *Proceedings of the 22Nd European Conference on Pattern Languages of Programs*, 2017, p. 13:1--13:24, DOI:10.1145/3147704.3147719.
31. E. Guerra and C. Fernandes, “An Evaluation Process for Pattern Languages,” in *Proceedings of the 8th Latin American Conference on Pattern Languages of Programs*, 2010, p. 18:1--18:11, DOI:10.1145/2581507.2581525.
32. P. Cremonesi, M. Elahi, and F. Garzotto, “User interface patterns in recommendation-empowered content intensive multimedia applications,” *Multimed. Tools Appl.*, vol. 76, no. 4, pp. 5275–5309, Feb. 2017, DOI: 10.1007/s11042-016-3946-5.
33. The Hillside Group, “How to Hold a Writer’s Workshop,” 1994. Available at. https://hillside.net/conferences/plop/235-how-to-hold-a-writers-workshop. Accessed on Dec-18-2018.
34. T. Thimthong, T. Chintakovid, and S. Krootjohn, “An empirical study of search box and autocomplete design patterns in online bookstore,” *SHUSER 2012 - 2012 IEEE Symp. Humanit. Sci. Eng. Res.*, pp. 1165–1170, 2012, DOI: 10.1109/SHUSER.2012.6268796.
35. R. Van Solingen, V. Basili, G. Caldiera, and H. D. Rombach, “Goal question metric (gqm) approach,” *Encycl. Softw. Eng.*, 2002.
36. R. O. Kuehl, *Diseño de experimentos: principios estadísticos de diseño y análisis de investigación*, 2 ed. México, 2001.
37. J. R. Lewis, “IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use,” *Int. J. Hum. Comput. Interact.*, vol. 7, no. 1, pp. 57–78, 1995.
38. U.S. National Library of Medicine, “MEDLINE®: Description of the Database.” Available at. https://www.nlm.nih.gov/bsd/medline.html. Accessed on Jan-18-2019.
39. S. Vegas, C. Apa, and N. Juristo, “Crossover Designs in Software Engineering Experiments: Benefits and Perils,” *IEEE Trans. Softw. Eng.*, vol. 42, no. 2, pp. 120–135, 2016, DOI: 10.1109/TSE.2015.2467378.
40. C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*, vol. 9783642290. United States: Springer, 2012.
41. K. A. Pituch and J. P. Stevens, *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM’s SPSS*. Routledge, 2015.
42. A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 4th ed. Sage Publications Ltd., 2013.
43. H. J. Seltman, “Experimental design and analysis,” *Online at: http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf*, 2012.
44. A. C. Elliott and W. A. Woodward, *Statistical Analysis Quick Reference Guidebook: With SPSS Examples*. Sage Publications Pvt. Ltd., 2006.
45. G. F. Templeton, “A two-step approach for transforming continuous variables to normal: implications and recommendations for IS research,” *Commun. Assoc. Inf.*, vol. 28, 2011.
46. P. D. Ellis, *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the*

Interpretation of Research Results. Cambridge University Press, 2010.
47. J. Cohen, "Statistical power analysis for the behavioral sciences 2nd edn." Erlbaum Associates, Hillsdale, 1988.
48. R. D. Riley, P. C. Lambert, and G. Abo-Zaid, "Meta-analysis of individual participant data: rationale, conduct, and reporting," *BMJ*, vol. 340, p. c221, Feb. 2010, DOI: 10.1136/BMJ.C221.
49. M. S. Fritz and A. M. Arthur, "Moderator Variables." Oxford University Press, 2017.