A Review of Quantitative Empirical Approaches in Human-Computer Interaction

Javier Fernández Serrano | Silvia T. Acuña | José A. Macías

Departamento de Ingeniería Informática, Universidad Autónoma de Madrid
Calle Francisco Tomás y Valiente 11, 28049 Madrid, España
javier.fernandezs01@estudiante.uam.es, {silvia.acunna, j.macias}@uam.es

ABSTRACT

Experimentation plays a major role in the development and validation of new theories in the field of human-computer interaction (HCI). In this paper, we address quantitative empirical studies, an approach that has become the standard methodology for conducting research in HCI. This research provides an in-depth review, gathering a representative and relatively extensive collection of primary studies. We aim to identify relevant research lines in HCI in which quantitative approaches have been employed in order to construct a consistent and comprehensive taxonomy of quantitative approaches and form a broad picture of the procedures, tools and techniques employed. A systematic mapping study methodology has been adopted to produce reliable results in a reproducible and agile manner.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems - *Human factors*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces - *Evaluation/methodology*; A.1 [**General Literature**]: Introductory and Survey

General Terms

Experimentation, Human Factors

Keywords

Human-Computer Interaction, Empirical Process, Experimental Study, Review, Systematic Mapping Study

1. INTRODUCTION

Human-computer interaction is a broad multidisciplinary area of study, with presence within both the research community and industry. As a relatively young discipline, it is strongly connected to more mature fields that actively provide new insights and experiences. The increasingly computerized world in which we live makes HCI a key domain and the mere publication of results is no longer good enough. Methodology and justification matter now more than ever.

HCI deals with a particular part of reality and, as concluded in other areas, empirical evidence is an inescapable step in the discovery process. Although the vast majority of HCI researchers agree on this, there is still debate between qualitative and quantitative approaches. Quantitative approaches are perceived as the only valid paradigm [26], and are hence more widespread in HCI literature than qualitative research.

Despite the importance of quantitative empirical studies in HCI, we have not found any review that specifically addresses this topic from a general and theoretical perspective. A number of empirical HCI reviews have been found[1] [10, 20], but either they have a narrow focus (e.g., mobile usability) or they do not pay enough attention to empirical aspects or use *ad hoc* review protocols. Outside the HCI area, namely in the software engineering (SE) field, there is abundant literature relating to empirical statistical issues [12, 19] and good practices [23].

The main goal of this paper is to unveil the existing link between HCI and empirical approaches. First of all, we aim to illustrate the various branches of HCI studied from an empirical standpoint, either currently or in the past, reporting currently active lines of research, as well as the most relevant findings. The second goal of this review addresses experimental procedures in HCI, including an overview of experimental stages (design, execution and data analysis) and empirical activities. We will also discuss the suitability of selected techniques and procedures, as well as pointing out flagrant errors. One last goal is to exemplify the use, in the context of HCI, of the systematic mapping study (SMS) methodology, which has been successfully employed in SE to produce reliable results in a reproducible and agile manner [6].

This paper is structured as follows. Section 2 describes the SMS methodology and final implementation, emphasizing the main points where our execution (slightly) diverges from the standard steps. Section 3 answers research questions raised in Section 2. Finally, Section 4 synthesizes and discusses the results.

Interaccion '14 Puerto de la Cruz. Tenerife. Spain

http://dx.doi.org/10.1145/2662253.2662309

[1] A parallel review search was conducted by means of the same protocol used to identify primary studies. This will be discussed in the following sections.

2. METHODOLOGY

Systematic literature reviews (SLR) are conceived as a means of acquiring knowledge in an exhaustive, unbiased and reproducible manner. SLR is best suited to disciplines requiring an accurate and audit-proof state of the art such as medicine, where SLR has been widely used to examine existing evidence.

Kitchenham proposes a well-known list of SLR guidelines aimed at SE researchers in [22]. This methodology, though comprehensive and rigorous, is very time consuming. The methodology employed in this review is a lighter version of Kitchenham's approach, known as systematic mapping study (SMS), which is especially well-suited for empirical "gap" identification [4].

2.1 Research Questions

The first step in a SMS is to define the research question(s) that the review aims to answer. The research questions addressed in this paper are as follows:

RQ1 What HCI-related topics does empirical research address?

RQ2 What are the main empirical findings in HCI thus far?

RQ3 What types of quantitative empirical studies are performed in HCI?

RQ4 What are the main characteristics of participants in HCI empirical studies?

RQ5 How are empirical studies designed in HCI?

RQ6 What statistical tools are used in quantitative empirical HCI research?

The above questions cover a broad spectrum of empirical knowledge [4]. Questions RQ1 and RQ2 aim to discover the coverage and relevance of empirical research in HCI. Questions RQ3 to RQ6 look at how the empirical process is instantiated in HCI.

2.2 Search Strings

Research questions need to be coded so as to be processed by search engines for information retrieval.

Most search engines support queries constructed by means of logical operators, namely "AND" and "OR" [22]. Although some provide useful operators like "NEAR", we decided to use only logical operators to assure query portability. It should be noted that logical operators are not always available from the basic query entry mode, and, depending on the search engine, the "Advanced search" (SpringerLink), "Command search" (IEEE Xplore) or even "Expert search" (Science Direct) modes typically have to be used. Finally, it is important to remember that search strings can be applied to titles, abstracts, keywords or meta-data (among other fields) and not necessarily to search the full text.

Queries constructed using Boolean syntax may follow a semantic pattern. Although not mandatory [22], it is a common practice among reviewers. For instance, [6] include one or more synonyms in each *maxterm*[2]. Then, *maxterms* are linked by means of "AND" operators. Dieste et al. go beyond this intuitive strategy in [11], making their search strings conform to the PICOC (Population-Intervention-Comparison-Outcome-Context) pattern.

[2]This terminology is borrowed from *Boolean algebra*.

The number of search strings is not *a priori* bounded, because *maxterm* size and quantity is not prescribed. It is commonplace to restrict the *maxterm* size to one atomic string [6, 11], so synonym variability is covered by increasing the total number of search strings. However, this strategy is not feasible if we have three or more *maxterms*, each intended to cover as many synonyms as possible, because the resulting number of search strings would be too large to report individual results. Keeping *maxterm* size equal to one helps to determine which synonyms produce better results, but this goal is outside the scope of this paper.

A single search string was considered, with the following three-maxterm structure: <HCI> AND <experimental-study> AND software. The last maxterm consists of a simple one-word string, and does not therefore have to be enclosed between quotes[3]. It limits searches to software context (neither the term *usability* nor *accessibility* is used exclusively in HCI). The <HCI> and <experimental-study> maxterms convey the HCI topic and experimental features, respectively, to which we expect primary studies to conform. Specifically, their structure is as follows:

- <HCI> = ("human computer interaction" OR "computer human interaction" OR "human machine interaction" OR "man machine interaction" OR *usability* OR *accessibility* OR "human factors" OR "user experience" OR "user centered" OR "user centred" OR "affective interaction" OR "affective computing" OR "affective design")
- <experimental-study> = ("empirical study" OR "empirical research" OR "empirical investigation' OR "empirical evaluation" OR "experimental study" OR "experimental research" OR "experimental investigation" OR "experimental evaluation")

For the <HCI> maxterm several "human-computer[4] interaction" synonyms were included, as well as sub-branches of HCI and American/British English spelling variants. Furthermore, acronyms like HCI or UX (which stands for *user experience*) took no part in the discussion. Finally, the <experimental-study> maxterm simply consists of strings formed by the Cartesian product of {empirical, experimental} and {study, research, investigation, evaluation}.

2.3 Inclusion and Exclusion Criteria

Inclusion criteria are defined to contain general known requirements with which every selected paper should comply, whereas exclusion criteria comprise perhaps unforeseeable exceptions (prior to document search) to the former. Two filters were applied, always taking into consideration both types of criteria, targeting different parts of the paper: the first filter was applied only to document-level information (e.g., type of publication, language, etc.) and titles, whereas the second covered the full text.

Inclusion criteria were as follows:

- The full text of the document is written in English.
- The document was published in a journal or conference proceedings.

[3]For instance, "software" produces the same results as *software*, but "software hardware" and *software hardware* are not equivalent.

[4]Most search engines are not hyphen-sensitive, so hyphens were removed from search strings.

- The title refers to an empirical study, explicitly including one of the following words: "empirical", "experimental", "empirically" or "experimentally".
- The title refers to HCI, to one of its subareas or to relevant concepts in HCI. Publication in a HCI-area journal or conference proceedings will suffice for conditional inclusion.
- The abstract refers to obtained results.
- The paper presents, at least, one quantitative study.
- The paper references hypotheses or conjectures.
- The paper uses inferential statistics.
- The paper discusses design and execution of the empirical study.
- The paper discusses the results of the empirical study.

Exclusion criteria were as follows:

- The document is not publicly accessible.
- The title references a *review* or a *case study*.
- The title references *meta-analysis*.
- The title references usability outside the HCI context.
- The title references a specific experimental study phase (e.g., experimental design).
- The title references HCI in a very technical context (e.g., wireless communications).
- The title references HCI-born disciplines that today have a substantial body of knowledge (e.g., e-learning).
- The paper presents results about meta-experimentation or HCI education.
- The paper procedures are not relevant for HCI.

2.4 Selection Process

A total of five well-known search engines were used in this review. All searches conducted in each search engine were undertaken in a 24-hour span to retain results accuracy, while full result reports were kept for prospective audit in several formats.

The results are summarized in Table 1. The "Retrieved" column contains the number of papers provided by each search engine immediately after search strings were entered. The "Candidates" column contains the number of papers retrieved from each search engine whose title and document-level information alone comply with inclusion and exclusion criteria. The "Selected" column represents the number of candidate papers that comply with inclusion and exclusion criteria after full-text inspection. Papers included in the above columns may be counted more than once, as some of them were indexed by more than one search engine.

A set of 38 duplicate-free papers were retrieved. Only 17 of them qualified for scanning, thorough reading and summarizing. The papers were examined in strict alphabetical order in order to avoid selection bias. One of the selected papers was eventually excluded because it reported an experiment whose procedure did not meet the HCI standards intended for this review. A total of 16 primary studies qualified.

Primary studies are presented in Table 2 along with their respective sources. The fact that only one out of sixteen primary studies was indexed by more than one search engine strongly supports our documentation sources selection.

Table 1: Results of the SMS

Search Engine	Retrieved	Candidates	Selected
ACM Digital Library	6743	21	14
IEEE Xplore	186	4	4
ScienceDirect	12278	18	8
SpringerLink	676	6	2
Web of Knowledge	626	31	14

Table 2: Primary study list

Alias	Reference	Source(s)
[AND07]	[1]	ACM Digital Library Web of Knowledge
[BOL09]	[2]	Web of Knowledge
[BRO12]	[3]	ScienceDirect
[CAS06]	[5]	ACM Digital Library
[CON09]	[7]	SpringerLink
[COR05]	[8]	ACM Digital Library
[DEC11]	[9]	IEEE Xplore
[FAN07]	[13]	ScienceDirect
[FAR12]	[14]	ACM Digital Library
[GER09]	[15]	ACM Digital Library
[GRY08]	[16]	Web of Knowledge
[HAR95]	[17]	Web of Knowledge
[JOH05]	[18]	Web of Knowledge
[KIM03]	[21]	ScienceDirect
[KOM11]	[24]	Web of Knowledge
[KUN07]	[25]	ScienceDirect

3. RESULTS AND DISCUSSION

This section provides factual answers to and deeper insights into the questions raised in section 2.1. Sections 3.1 and 3.2 outline research interests and findings appearing in HCI. The remaining sections focus on empirical study categorization, population sampling, experimental design and statistical tools.

3.1 (RQ1) What HCI-related topics does empirical research address?

Table 3 categorizes the topics addressed by primary studies. Three major HCI topics have been identified as empirically relevant: usability, interaction and human factors. The resulting taxonomy matches a tree pattern, with well-documented HCI subareas as interior nodes and primary studies as leaves.

3.1.1 Usability

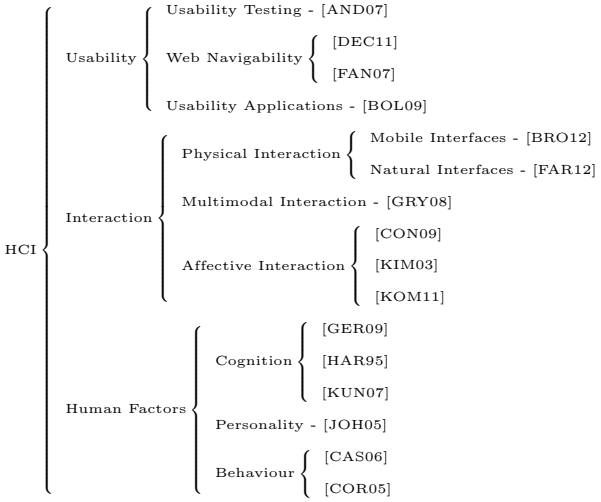
Usability is largely addressed in the context of web applications, focusing on navigability, which may be enhanced through content structure or inclusion of auxiliary tools. This applies to [FAN07], where various semantic approaches to knowledge presentation are compared. On the contrary, [DEC11] aims to prove the beneficial effects of enhanced guidance and orientation.

Two completely different dimensions of usability were targeted by [AND07] and [BOL09]. [AND07] studies which of a set of remote testing techniques performs better when detecting usability problems. [BOL09] offers an innovative perspective of usability as a tool to enhance the perception of corporate values by consumers.

3.1.2 Interaction

Empirical studies on interaction are divided into three major topics, depending on which main type of interactive dimension they address. Physical interaction refers to ways of using computers employing the human body, with or without direct contact. Both [BRO12] and especially [FAR12] explore the naturalness of physical interaction through movement and gestures, respectively. However, [BRO12] focuses on different strategies available for mobile devices and thus

Table 3: Primary study categorization



makes a broader comparison, which is not confined to natural interfaces.

Two of the three papers targeting the affective dimension of HCI employ agents. The agent in [CON06] is designed to act intelligently, helping and motivating students. Contrarily, importance is attached to agent appearance in [KOM11] and no user feedback is addressed. Finally, no agent is used in [KIM03] and emotions are conveyed through web element composition.

Last but not least, multimodal interfaces are addressed in [GRY08], leveraging multiple interaction channels, namely text, speech and images, to reinforce learning.

3.1.3 Human Factors

This branch of empirical HCI covers papers that, to some extent, have to do with psychology and sociology. A first recognizable area of human factors relates to cognition, and more precisely perception [HAR95] [KUN07], (working) memory and cognitive styles [GER09] (image/verbal). Icons play a central role in testing perception in both [HAR95] and [KUN07].

Empirical studies on personality try to validate a complex psychological model, (see [JOH05]), where personality traits, along with computer experience, are believed to explain the construct of computer self-efficacy. Contrarily, in this context, hypotheses about behaviour rely on observation rather than models. In fact, [CAS06] merely observes undo mechanism interpretation. Finally, [COR05] is an interesting study about whether experts trust software tools to do much of their work and under what circumstances.

3.2 (RQ2) What are the main empirical findings in HCI thus far?

The main findings are summarized in Table 4. With the exception of [GRY08] and [KOM11], all primary studies report successful results regarding targeted hypotheses. However, not all of them are equally valuable since some authors *fish* for results [23]. For instance, [DEC11] emulates a between-group design by means of an execution and replication, which is not fair because the replication is conducted

Table 4: Main results

Primary Study	Empirical Results
[AND07]	Remote asynchronous testing results are not affected by competency level of testers.
[BOL09]	High(low) usability enhances(damages) brand value perception.
[BRO12]	Accelerometer interface is the one preferred by users of a scroll-shooting mobile game.
[CAS06]	Not all undo mechanisms spring to mind with equal frequencies.
[CON09]	Emotions predictor performs well even without goal-predicting information.
[COR05]	Experts check automatic-solver step-by-step solution more frequently than novices.
[DEC11]	Guidance mechanisms enhance web usability.
[FAN07]	Usage-oriented navigation structures are more usable than subject-oriented ones.
[FAR12]	Gestural interfaces cause more fatigue.
[GER09]	Cognitively tailored content presentation enhances task performance.
[GRY08]	Multimedia interfaces do not enhance high functioning autistic users' learning.
[HAR95]	All icon, background and transparency level factor combinations affect foreground legibility.
[JOH05]	Computer self-efficacy model that includes personality traits matches empirical data.
[KIM03]	Successful linear regression of emotion levels on aesthetic web design factors.
[KOM11]	Dog-like robot's appearence is misleading for attitude-conveying sound interpretation.
[KUN07]	Icon type affects both abstract and concrete learner's performance.

Table 5: Empirical study categorization

Type of Study	Primary Studies	Total
True experiment	[BOL09], [BRO12], [DEC11], [FAN07], [FAR12], [GER09], [HAR95], [KOM11]	8
Quasi-experiment	[AND07], [COR05], [GRY08], [KUN07]	4
Correlational study	[JOH05], [KIM03]	2
Observational study	[CAS06]	1
Classifier validation	[CON09]	1

employing a revised and enhanced procedure. Also, [CAS06] adds value to results by using a post-hoc analysis.[5]

Of all reviewed studies, only [BRO12] would fall into the evaluation, as opposed to the research category. It is based on a concrete commercial device (iPod Touch), rather than *ad hoc* newly developed technology (e.g., the HTML extension developed by [GER09]), whose features largely drive the course of the study. Following a similar reasoning [BRO12] could be called a performance study [26].

3.3 (RQ3) What types of quantitative empirical studies are performed in HCI?

A categorization of primary studies according to the main empirical approach employed is given in Table 5.

A first level of classification splits primary studies into experiments or non-experiments depending on whether or not they aim to introduce intervention (i.e., have independent variables), respectively. Additionally, experiments are traditionally divided into true experiments or quasi-experiments depending on whether or not the assignment of subjects or participants to factor conditions is randomized. This categorization is well-documented in the literature, with different names (randomized controlled trial/quasi-randomized trial [22] and quantitative experiment/quasi-experiment [26]). Three factors that cannot be randomized have been identified: expertise [AND07] [COR05], impairment [GRY08] and learning style [KUN07].

The simplest non-experimental study measures a quantitative dimension from a population sample without pre-

[5]Notwithstanding, [CAS06] proposes an experimental design to replace the post-hoc analysis for future work.

vious intervention. Such a study does not necessarily lack hypothesis formulation or statistical analysis, as intervention (treatments assignment) is meant to isolate cause over effects, but is not required to statistically infer knowledge about a population. In this paper, this type of study is referred to as observational, as in [CAS06]. However, authors' comments on their own papers have not always been taken into consideration. For instance, [KIM03] claims:

> (p. 901) "In the second study, we conducted a *controlled experiment* to identify key design factors [...]"

whereas, in actual fact, not only was not a controlled experiment the second study, but it also failed to qualify even as a quantitative study, as it is generally regarded in the literature [BOL09] [DEC11].

Correlational studies search for linear dependencies between random variables without introducing intervention. Although [KIM03] and [JOH05] both belong to this category, their goals are different. [KIM03] performs regression analysis on one or more output variables, but does not look for any relation among the regressed variables. Contrarily, [JOH05] addresses a joint regression in which some of the considered variables count as inputs and outputs at the same time.

Correlational studies provide a predictive model and so [CON09] does, albeit taking a different approach. In regression, only ratio data are used and importance is attached to the statistical significance of relations between variables, whereas, in classification, nominal data are also used and quality of the model is empirically assessed in terms of prediction *accuracy* (see p. 298 of [CON09], last paragraph).

3.4 (RQ4) What are the main characteristics of participants in HCI empirical studies?

Table 6 summarizes sampling data collected from participants in reviewed studies. Question marks (?) in cells denote unknown information. Age fields considered in columns are comprehensive, meaning no single study addresses an age descriptive statistic not shown in this table. Some studies present additional facts which are not taken into account because they are too specific to a topic or area. For instance, both physical interaction studies, [BRO12] and [FAR12], report the number of right- and left-handed participants.

Looking at Table 6 we find that there is a notably large amount of unknown cells. A possible reason for generalized loose reporting is the apparent absence of empirical guidelines in the area of HCI. In fact, of all reviewed papers, only [DEC11] follows a series of guidelines for this purpose and, more surprisingly, these were conceived for SE, not HCI.

Of course, there are several cases where not all data need to be explicitly presented or, at least, missing data are not so harmful. For instance, participants in [CON09] are sixth or seventh grade students so ages are both lower and upper bounded. There are also cases where it is not recommended to extract certain information from the sample on ethical grounds. This is the case of [GRY08], where IQ measurements were not taken from non-autistic participants.

Apart from the aforementioned information gaps, some papers fail to accurately report information. For instance, [COR05] claims:

> (p. 664) "The sample was balanced with respect to [...] gender, education, age and profession",

but it is uncertain what balanced actually means in this context.

Another worrying aspect of Table 6 refers to sample background, as most papers rely on young participants in higher education who are not sufficiently representative of the heterogeneous population interacting with computers. This could be because the overwhelming majority of authors are academics (working at universities and research institutes), who take advantage of their access to students for research purposes. In fact, only [AND07] and [HAR95] report a combined effort by academia and industry.

3.5 (RQ5) How are empirical studies designed in HCI?

Table 7 summarizes the most relevant design dimension data collected from experimental primary studies.

A general lean toward optimal design has been observed throughout the reviewed papers. This tendency is perceived in variable complexity, especially in the rather large number of factors considered. An increase in the number of factors generally introduces threats to validity, so special care must be taken. Interestingly, the experiments considering more factors (up to three factors [HAR95] [FAR12]) employ a within-subjects design. This type of treatment assignment uses the same participants for all factor levels, which introduces the so-called "learning effect" [23]. However, not all primary studies take correct measures to prevent this effect. While [DEC11] cancels out learning effects, [FAR12] does not mention counterbalancing at all, and, even though "training effects" are addressed in [GER09] (p. 604), an explicit reference to treatment order randomization is missing. Interestingly, [KUN07] recognizes that "test scores revealed the presence of order/practice effect" and consequently addresses this problem "by counterbalancing the order of presenting the three sections in the lesson and counterbalancing and randomizing the test items", and finally concludes that

> (p. 1460) "In both field and final test, the student's academic and work schedules dictated the random assignment",

but this "schedule randomization" is not valid unless item and lesson order was indeed randomized, which is not stated in the text. Because participants schedules are likely to be correlated (they are reported to study in the same university and even at the same school), the experiment may contain undesired correlations.

Another point worth mentioning is the existence of just one primary study using block design [BOL09]. It should be noted that [BOL09] is also the only study with no parameters, which strongly suggests that other studies could benefit from blocking the effects of some of the variables now treated as parameters.

3.6 (RQ6) What statistical tools are used in HCI quantitative empirical research?

Table 8 shows the statistical tools employed by reviewed studies. Mean difference tests are mainly used by experimental studies, where treatment and control group means are contrasted to assess causality. Asterisks (*) denote non-parametric mean difference tests, i.e., which make no assumption regarding the underlying population. Parametric

Table 6: Sampling data

Primary Study	No. Subjects		Age				Background	Inclusion & Exclusion Criteria
	Female	Total	Min.	Max.	μ	σ		
[AND07]	10	24	19	30	25.13	3	University students	For AE condition, usability evaluation training was required
[BOL09]	?	120	?	?	?	?	Potential users of a certain website	No previous use of web sites used in the study
[BRO12]	13	36	19	43	26	4.36	University students Graduates	?
[CAS06]	?	29	?	?	?	?	Mostly, undergraduates not taking computer courses	Ideally, undergraduates taking computer courses
[CON09]	?	66	?	?	?	?	6^{th} grade students 7^{th} grade students	Disturbing and disturbed students' data were discarded
[COR05]	?	46	23	58	33.3	?	Balanced education Balanced profession	Half experts, half non-experts in solver implementation
[DEC11] (1 replication)	?	84/16	?	?	?	?	M.Sc. students / Ph.D. students	"sufficiently competent to perform the level of experimental tasks required"
[FAN07] (1 replication)	?	134/ 99	?	?	?	?	Undergraduate business students with homogeneous POM training	12-week instruction on a POM course
[FAR12]	10	20	11	40	29	?	?	?
[GER09]	?	89	18	21	19	?	University students	?
[GRY08] (clinical/control)	?/2	10/10	?	?	12.83/ 9.58	?	Normal IQ for age / Non-autistic children took no IQ test	High functioning autistic teenagers / typically developing children
[HAR95]	?	14	?	?	?	?	University students	Not colour-blind Not acquainted with icons
[JOH05]	156	313	?	?	19.4	2.1	University students 47% employed Computer-acquainted	Database skills required
[KIM03]	?	515	?	?	?	?	"were in their twenties" "gender was balanced" Undergraduates	?
[KOM11]	3	20	19	24	?	?	University students Not familiar with robots	No hearing problems
[KUN07]	39	53	?	?	?	?	Graduate students	?

Table 7: Experimental design

Primary Study	Experimental Units	Factors (levels)	Parameters (values)	Response Variables	Design (comments)
[AND07]	Software system	Remote testing method (LAB, RS, AE, AU)	System type (email client Mozilla Thunderbird 1.5)	Task completion Task completion time Number of usability problems identified	Between-group
[BOL09]	Information-intensive and brand-intensive web sites	Usability (original, modified)	×	Brand value perception	Between-group (brand domain as blocking variable)
[BRO12]	Mobile device video game	User interface type (accelerometer, simulated button, touch gesture)	Level difficulty Video game genre (scroll-shooter) Device (iPod Touch)	Game performance Interface preference	Within-subjects
[COR05]	Problem-solving tasks	Expertise level (expert, non-expert) Task difficulty (easy, difficult)	Problem area (planning and scheduling)	Solving strategy chosen Access to explanation frequency	Between-group (expertise) Within-subjects (task difficulty)
[DEC11]	Web Information System (WIS)	Navigation model (control/proposal)	Type of WIS (conference manager)	Perceived ease of use Effectiveness Efficiency	Within-subjects
[FAN07]	Web site for knowledge acquisition	Navigation structure (pure usage, subject-oriented, mixed) Task complexity (simple, complex)	Field for knowledge acquisition (production and operations management)	Task accuracy Navigation time Usability experienced	Between-group Factorial
[FAR12]	User interface	Task difficulty (simple, complex) Input device (mouse, gesture) Output device (big screen, small screen)	Interface metaphor (desktop)	Completion time Naturalness Fatigue	Within-subjects Factorial
[GER09]	Commercial web site	Content presentation (original, tailored)	Hypertext content (sony laptops)	Task accuracy Task completion time User satisfaction	Within-subjects
[GRY08]	Gamified learning software	Autism (yes, no) Type of interface (simple, multimedia)	Autism disorder targeted (pragmatics)	Completed scenarios count Game score Scenario completion time	Between-group (autism) Within-subjects (type of interface)
[HAR95]	Desktop scenario (foreground-background combination)	Icon type (solid, line, text) Transparency level (0%, 50%, 75%, 90%) Background type (text page, wire frame, solid image)	Number of icons in the foreground (12)	Legibility error rate Response time	Within-subjects Factorial
[KOM11]	Sound-equipped agent	Agent's appearance (PC, dog-like, machine-like) Auditory attitude (positive, negative)	Nature of auditory attitude (bipolar) Nature of sound (synthetic) Sound intensity (60 dB)	Correct attitude interpretation rate	Within-subjects
[KUN07]	Instructional program lesson	Icon type (pictorial, abstract, drawing) Learning style (abstract, concrete)	Lesson topic (digital video camcorder)	Quiz score	Within-subjects Factorial

Table 8: Main statistical tools

Use	Tool	Primary Studies
Mean difference test	Mixed ANOVA	[COR05]
	Multivariate ANOVA	[FAN07]
	One-way ANOVA	[AND07], [COR05]
	Repeated measures ANOVA	[COR05], [FAR12], [GRY08], [HAR95], [KUN07]
	Paired t-test	[BOL09], [CON09], [FAR12], [GER09], [KUN07]
	One-sample t-test	[CON09]
	Cochran's Q*	[KOM11]
	Friedman test*	[KOM11]
	Mann-Whitney-Wilcoxon*	[DEC11], [GRY08]
	Wilcoxon signed rank test*	[DEC11], [GER09] [GRY08]
	Z-test	[BRO12]
Correlation effect size estimation	Pearson's r	[KIM03], [KUN07]
	Spearman's ρ	[GRY08]
	Kendall's τ	[KUN07]
Sphericity test	Greenhouse-Geisser test Mauchly's test Huyn-Feldt	[KUN07]
Post-hoc analysis	Student-Newman-Keuls test	[HAR95]
	Tukey's range test	[AND07]
Normality test	Kolmogorov-Smirnov test	[DEC11], [JOH05]
	Shapiro-Wilk test	[GRY08]
Internal consistency estimation	Cronbach's α	[DEC11], [FAN07], [KIM03], [KUN07]
Regression analysis	Partial Least Squares	[JOH05]
	Stepwise regression	[KIM03]
Uniformity test	χ^2 test	[CAS06], [GRY08]
Independence test	Fisher's exact test	[AND07], [CON09]
Classifier validation	N-fold cross-validation	[CON09]
Clustering	Ward's clustering	[KIM03]
Mean difference effect size estimation	Cohen's d	[CON09]

tests, like ANOVA and t-tests, rely on further hypothesis for their results to be safely credited, but are generally preferred because of their greater statistical power [12].

Since parametric tests behave robustly to non-"extreme assumption violations", researchers are encouraged to use them [12]. However, very few authors report having checked whether violations really were moderate [GRY08] [KUN07]. Interestingly, [DEC11], [GER09], [JOH05] failed to meet normality assumptions, whereas [KUN07] only provided graphical justification (see pp. 1468-1469). Assumptions other than normality, like the repeated measures ANOVA sphericity assumption [KUN07], are rarely addressed.

All primary studies achieve high statistical significance, generally reporting *p-values* under 0.05. However, this is not sufficient to assess whether findings are of *practical* significance, which should be reported through *effect size* estimations explicitly discussed in relation to the specific area of study [19]. All studies conducting any form of correlation analysis report effect sizes by means of Pearson's r, Spearman's ρ or Kendall's τ, the latter two being rank-based and, hence, non-parametric. With the exception of correlation analysis and [CON09], which uses Cohen's d and Cohen's effect size criteria, no study addresses practical significance.

4. CONCLUSIONS

Despite HCI being well-positioned in today's information society and the empirical process increasingly taking root in neighbouring domains, the HCI community has few surveys about this research methodology, and none presents an overall state of the art.

This review addresses general questions about quantitative empirical HCI research. A SMS was successfully conducted and a sufficient number of quality primary studies were retrieved. We first examined HCI branches in the context of empirical research, providing a taxonomy of topics that included usability, interaction and human factors as top-level categories. Then we explored the main resulting findings in empirical HCI and identified a number of different quantitative empirical approaches, including experimental and correlational studies. Next, we analysed empirical procedures and pointed out some common flaws encountered: poor data and procedures recording, unrealistic sampling, absence of blocking variables and disregard for test assumption verification.

Future work may take a closer look at some specific topics or empirical flaws encountered. Additionally, it might be interesting to estimate the saturation point where results would not significantly differ if new primary studies were considered.

Some traits of this review could question its correctness and generalizability:

- **Internal validity**: Our own leanings and likes are non-negligible, though efforts were made to avoid them. Another source of bias is inclusion and exclusion criteria and search string arbitrariness. Again, experts assessed both steps, so arbitrariness is, at least, justified.
- **External validity**: Generalizability for purposes other than providing an overview of the field is limited because of the small number of primary studies reviewed ($N = 16$) and the limited scope of the investigation (journal or conference papers written in English).

Empirical approaches are and are perceived to be capital for extending the body of knowledge in HCI. The empirical process covers almost all of the most recognizable HCI branches, *accessibility* being a notable absence. Furthermore, researchers are acquainted with a variety of quantitative tools and techniques, which denotes that they are highly concerned with their own empirical instruction. However, they are not generally fully aware of some of the technical details involved. This is understandable given that the empirical process in HCI is work in progress, without published guidelines.

Overall, HCI offers good empirical opportunities. Even though from the empirical standpoint it is still a relatively unexplored and immature area, the groundwork has definitely been laid.

5. ACKNOWLEDGEMENTS

The work reported in this paper has been supported by the Spanish Ministry of Science and Innovation projects TIN2011-24139 and TIN2011-23216 and also by *Universidad Autónoma de Madrid* through the postgraduate fellowship *Beca de inicio de estudios de posgrado*.

6. REFERENCES

[1] M. S. Andreasen, H. V. Nielsen, S. O. Schrø der, and J. Stage. What happened to remote usability testing? An Empirical Study of Three Methods. In *Proceedings of the SIGCHI conference on Human factors in*

computing systems - CHI '07, page 1405, New York, New York, USA, Apr. 2007. ACM Press.

[2] D. Bolchini, F. Garzotto, and F. Sorce. Does Branding Need Web Usability? A Value-Oriented Empirical Study. In *Proceedings of the 12th IFIP TC13 Conference on Human-Computer Interaction*, pages 652–665. Springer-Verlag Berlin, 2009.

[3] K. Browne and C. Anand. An empirical evaluation of user interfaces for a mobile video game. *Entertainment Computing*, 3(1):1–10, Jan. 2012.

[4] D. Budgen, M. Turner, P. Brereton, and B. Kitchenham. Using Mapping Studies in Software Engineering. In *Proceedings of the 20th Annual Workshop of the Psychology of Programming Interest Group (PPIG'08)*, volume 2, pages 195–204, Lancaster University, UK, 2008.

[5] A. G. Cass, C. S. T. Fernandes, and A. Polidore. An empirical evaluation of undo mechanisms. In *Proceedings of the 4th Nordic conference on Human-computer interaction changing roles - NordiCHI '06*, pages 19–27, New York, New York, USA, Oct. 2006. ACM Press.

[6] J. W. Castro and S. T. Acuña. Differences between Traditional and Open Source Development Activities. In *Proceedings of the 13th International Conference on Product-Focused Software Process Improvement*, pages 131–144, 2012.

[7] C. Conati and H. Maclaren. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19(3):267–303, Jan. 2009.

[8] G. Cortellessa, V. Giuliani, M. Scopelliti, and A. Cesta. Key Issues in Interactive Problem Solving: An Empirical Investigation on Users Attitude. In M. F. Costabile and F. Paternò, editors, *Proceedings of the 10th IFIP TC13 International Conference on Human-Computer Interaction*, volume 3585 of *Lecture Notes in Computer Science*, pages 657–670, Berlin, Heidelberg, Sept. 2005. Springer Berlin Heidelberg.

[9] V. de Castro, M. Genero, E. Marcos, and M. Piattni. Empirical study to assess whether the use of routes facilitates the navigability of web information systems. *IET Software*, 5(6):528–542, 2011.

[10] D. M. Dehn and S. Van Mulken. The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies*, 52(1):1–22, Jan. 2000.

[11] O. Dieste, N. Juristo, and M. D. Martínez. Software industry experiments: A systematic literature review. In *Proceedings of the 1st International Workshop on Conducting Empirical Studies in Industry (CESI 2013)*, pages 2–8, 2013.

[12] T. Dybå, V. B. Kampenes, and D. I. Sjøberg. A systematic review of statistical power in software engineering experiments. *Information and Software Technology*, 48(8):745–755, Aug. 2006.

[13] X. Fang and C. W. Holsapple. An empirical study of web site navigation structures' impacts on web site usability. *Decision Support Systems*, 43(2):476–491, Mar. 2007.

[14] F. Farhadi-Niaki, R. GhasemAghaei, and A. Arya. Empirical study of a vision-based depth-sensitive human-computer interaction system. In *Proceedings of the 10th asia pacific conference on Computer human interaction - APCHI '12*, pages 101–108, New York, New York, USA, Aug. 2012. ACM Press.

[15] P. Germanakos, N. Tsianos, Z. Lekkas, C. Mourlas, M. Belk, and G. Samaras. Proposing Web Design Enhancements Based on Specific Cognitive Factors: An Empirical Evaluation. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 602–605. IEEE, Sept. 2009.

[16] O. Grynszpan, J.-C. Martin, and J. Nadel. Multimedia interfaces for users with high functioning autism: An empirical investigation. *International Journal of Human-Computer Studies*, 66(8):628–639, Aug. 2008.

[17] B. L. Harrison, G. Kurtenbach, and K. J. Vicente. An experimental evaluation of transparent user interface tools and information content. In *Proceedings of the 8th Annual Symposium on User Interface Software and Technology*, pages 81–90, 1995.

[18] R. D. Johnson. An empirical investigation of sources of application-specific computer-self-efficacy and mediators of the efficacy-performance relationship. *International Journal of Human-Computer Studies*, 62(6):737–758, June 2005.

[19] V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. Sjøberg. A systematic review of effect size in software engineering experiments. *Information and Software Technology*, 49(11-12):1073–1086, Nov. 2007.

[20] D. J. Kim and C. K. Coursaris. A Meta-Analytical Review of Empirical Mobile Usability Studies. *Journal of Usability Studies*, 6(3):117–171, 2011.

[21] J. Kim, J. Lee, and D. Choi. Designing emotionally evocative homepages: an empirical study of the quantitative relations between design factors and emotional dimensions. *International Journal of Human-Computer Studies*, 59(6):899–940, Dec. 2003.

[22] B. Kitchenham. Procedures for Performing Systematic Reviews. Technical report, Keele University Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1, 2004.

[23] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. E. Emam, and J. Rosenberg. Preliminary Guidelines for Empirical Research in Software Engineering. *IEEE Transactions on Software Engineering*, 28(8):721–734, 2002.

[24] T. Komatsu and S. Yamada. How Does the Agents' Appearance Affect Users' Interpretation of the Agents' Attitudes: Experimental Investigation on Expressing the Same Artificial Sounds From Agents With Different Appearances. *International Journal of Human-Computer Interaction*, 27(3):260–279, Feb. 2011.

[25] M. L. A. Kunnath, R. A. Cornell, M. K. Kysilka, and L. Witta. An experimental research study on the effect of pictorial icons on a user-learner's performance. *Computers in Human Behavior*, 23(3):1454–1480, May 2007.

[26] V. Tscherter, P. Ravasio, and S. Guttormsen-schär. The Qualitative Experiment in HCI: Definition, Occurrences, Value and Use. *ACM Transactions on Computer-Human Interaction*, V:1–24, 1986.