



On the Adoption of Empirical Methods and Systematic Reviews in the Brazilian Symposium on Human Factors in Computing Systems


Mariela I. Cortés  [State University of Ceará | mariela@larces.uece.br]


Sávio Freire  [Federal Institute of Ceará | savio.freire@ifce.edu.br]


Lucas Vieira Alves  [State University of Ceará | lucas.vieira@aluno.uece.br]


Matheus Lima Chagas  [State University of Ceará | matheus.chagas@aluno.uece.br]

Marx Haron Barbosa  [State University of Ceará | marx.barbosa@aluno.uece.br]

Adson Damasceno  [State University of Ceará | adson.damasceno@aluno.uece.br]

Andressa Ferreira  [State University of Ceará | andressa.magda@aluno.uece.br]

Eliakim Gama  [State University of Ceará | eliakim.gama@aluno.uece.br]

José Paulo Rodrigues Moraes  [State University of Ceará | jpaulo.moraes@aluno.uece.br]

Abstract *Context:* Empirical studies (ES) and systematic reviews (SR) play an essential role in the Human-Computer Interaction (HCI) field as its focus is on evaluating the end-user and usability of software solutions and synthesizing the evidence found by the HCI community. Even though the adoption of empirical evaluation techniques and SR has gained popularity in recent years, the consistent use of a methodology is still maturing. *Goal:* This study aims to provide a qualitative and quantitative assessment of the current status of ES and SR presented in the research papers published at the proceedings of the Brazilian Symposium on Human Factors in Computing Systems (IHC Symposium). *Method:* We conduct an empirical study on the papers over the 18 editions in the IHC Symposium to answer four research questions. Our study proposes a protocol to identify and assess ES and SR reported in the papers published at the IHC Symposium. *Results:* From the sample of 259 studies, we find 131 ES and SR (~51%). We have characterized and categorized the ES into case studies, experiments, and surveys. Further, we found evidence that these studies' quantity and quality have been increased over the IHC Symposium editions, and almost half of these studies give detailed information making possible their replication. *Conclusion:* We hope that each study's characterization can support the conduction of new ES and SR by the HCI Brazilian community, producing more reliable results and reducing or eliminating biases.

Keywords: *empirical evaluation, quality assessment, human-computer interaction research.*

1 INTRODUCTION

The Human-Computer Interaction (HCI) field requires that researchers and practitioners understand the psychological, organizational, and social factors of the combined human and computer systems to build competitive software interfaces and evaluate their effects (Valverde, 2011). By performing empirical methods, software practitioners can assess their software interfaces and the techniques used to develop these interfaces. Moreover, empirical methods are able to expand scientific knowledge, which will guide the development of new technologies and contribute with information that will help the decision-making process, both in the industry and in the area of services (Sjoberg et al., 2007). Systematic reviews have been also used by the HCI field for synthesizing the evidence reported in the technical literature to provide background on specific topic and identify some gaps that require further investigation.

In general, the application of empirical methods and systematic reviews requires reliable procedures and practices (Wohlin et al., 2012; Malhotra, 2015). The adoption of a standardized methodology brings consistency to a body of work and facilitates the review and comparison of research from different studies (MacKenzie, 2013).

In this sense, researchers from the HCI field have used

(i) empirical methods based on systematic observations and experiments to capture and understand the user experience when users are performing a task assisted by software (Lazar et al., 2017) and (ii) systematic reviews to summarize the results from the application of these methods. Both, empirical studies and systematic reviews, can provide scientific evidence on the use of tools and techniques and supply the community with data employing which other research can repeat the original study (Basili, 1996).

Some work have identified the quality of the empirical studies and systematic reviews performing empirical studies to assess the protocol used in these work (Zannier et al., 2006; Silveira Neto et al., 2013; Barbosa et al., 2017; Kitchenham et al., 2019). They analyzed the protocols described in the papers published in conferences in the Software Engineering (SE) field. As HCI, SE is also considered a social process in that its methods, tools, and paradigms are affected by the experience, knowledge, and capability of their users (Juzgado and Moreno, 2001), requiring empirical evidence from empirical-based evaluations and validations. However, these studies did not assess the protocols used for researchers in the HCI field. Besides, HCI researchers have used *ad hoc* protocols to perform systematic reviews (Serrano et al., 2014). Then, assessing the HCI field's protocols can reveal improvement points to providing more reliable, replicable, and sound

evidences from empirical studies and systematic reviews.

The Brazilian Symposium on Human Factors in Computing Systems (IHC¹, the acronym in Portuguese) is the main forum in the field of Human-Computer Interaction in Brazil. We used the term IHC Symposium to identify the symposium in the remainder of this paper. IHC Symposium is reached the 18th edition in 2019 and annually gathers researchers and practitioners interested in scientific investigation and practices related to creating, building, and evaluating computing solutions to be used by people, and providing a landscape on the evidences reported by the community. In this context, resembling what befalls with other social sciences, the application of empirical evaluation has to be exploited to gain a better understanding of social, physical, and cognitive environments and their effects as part of the interface design process to develop methodologies to aid appropriate HCI design. Also, performing systematic reviews have supported the HCI community to understand the current position on determinate topic.

This work *provides a qualitative and quantitative assessment of the current status of the empirical evaluations and systematic reviews presented in the research papers published at IHC Symposium proceedings*. Therefore, we performed an empirical study, which was divided into two phases: quality assessment and classification process. In the first one, we examined **259** papers published in the 18 editions of IHC symposium to determine whether it comprises an empirical evaluation or a systematic review. In the second one, only the papers that include empirical studies or systematic reviews, i.e, **131** papers (~51%) out of 259, were analyzed from **our perspective**, in order to classify them into experiments, case studies, surveys, or systematic reviews. As a result, we aim to show the evolution and maturing of the utilization of empirical methods and systematic reviews in HCI research, in quantitative and qualitative aspects.

For supporting the quality assessment and the classification processes, we developed a protocol which includes checklists based on quality assessment criteria from the literature. Thus, we aim at improving the conduction of the studies and consequently aiding to produce more reliable results by reducing or eliminating biases. All the material used in the execution of our study is available to the scientific community². We expect that the proposed research protocol can be fully reproduced by any researcher, as indicated by Munafò et al. (2017).

We found evidence that case studies are the most used empirical methods by the HCI Brazilian community for assessing its tools and methods. We found a clue that there is a certain level of uncertainty on the use of empirical methods and systematic reviews by this community. Moreover, the quantity and quality of empirical studies and systematic reviews appears to have increased in the IHC Symposium lifetime, and almost half of these studies have provided information for their replication.

The remainder of this paper is organized as follows. Section 2 provides an overview of the types of empirical studies, and systematic reviews, and discusses our preliminary

work. Section 3 describes our research method, its composition, and how it works out. Section 4 presents the characterization of each study's type and Section 5 answers the research questions. Finally, Sections 6 and 7 mention threats to validity and conclusions, respectively.

2 BACKGROUND

This section presents some concepts related to empirical methods and systematic reviews. Lastly, it discusses on (the lack of) related work and presents our preliminary study.

2.1 Empirical Methods and Systematic Reviews

A variety of empirical methods are available for HCI researchers and practitioners for assessing interfaces and software (Lazar et al., 2017). These methods provide a framework to validate theories, verify hypotheses, answer research questions by observations or experiments. One of the goals of empirical evaluations is to provide means that can be integrated with practical experience and human values in the decision-making process, regarding the development and maintenance of software (Kitchenham et al., 2004).

Depending on the purpose of the evaluation and the conditions for the empirical investigation, different types of investigations (strategies) may be carried out (Wohlin et al., 2012). The most frequently used by HCI community include observations, field studies, surveys, usability studies, interviews, focus groups, controlled experiments, and case studies (Shneiderman et al., 2016; Lazar et al., 2017). Systematic reviews have been used for HCI community to synthesize the evidences found by the empirical studies (Lazar et al., 2017). In our work, we choose to assess systematic studies and the following empirical studies: experiments, case studies, and surveys as they have a well-defined protocol that must be followed by researchers during an execution in order to summarize or collect empirical evidence. Below, we presented briefly each one of these types of studies considered in our work:

- **Experiment.** An experiment is a method widely used in many areas of science to test the established hypothesis by finding the effect of variables of interest on the outcome variables (Gergle and Tan, 2014; Malhotra, 2015). In HCI, the manipulated variable is typically a property of an interface or interaction technique that is presented to participants in different configurations (MacKenzie, 2013). Experiments are mostly in a laboratory environment. Based on randomization, subjects are assigned to different treatments, while others are keeping constant. The effect is measured and a statistical analysis is applied. Experiments can be oriented by humans or by technology. In the first case, humans apply treatments to objects, whereas in the second one, tools are responsible for performing the experiment, providing greater control. It is recommended to use experiments to confirm theories and widespread acknowledgment or to evaluate models (Wohlin et al., 2012).

¹<http://comissoes.sbc.org.br/ce-ihc/eventos/ihcs/>

²<http://spl.it.to/ydo9X1C>

- **Case study.** It is an empirical enquiry used in various sciences such as sociology, medicine, and psychology. Case studies are used to investigate a single entity or contemporary phenomenon within its real-life context and specific time-space (Wohlin et al., 2012). Usually, the phenomenon can be difficult to distinguish clearly from its environment. In HCI, case studies, in which researchers study a small number of participants (possibly as few as one) in depth, can be useful tools for gathering requirements and evaluating interfaces (Lazar et al., 2017). In the case study processes, researchers should not interfere, directly or indirectly, since it is a purely observational study of real-world scenarios. Qualitative data is often collected in a case study from multiple sources such as interviews, discussions, or observations. The case study approach provides researchers with examples of study designs that could be adapted, with the additional benefit of becoming aware of possible issues prior to deployment (Olson and Kellogg, 2014). An advantage of case studies is that they are easier to plan and more realistic, but their results are difficult to generalize and to interpret.
- **Survey.** A survey is a method of gathering information by asking questions to a subset of people, the results of which can be generalized to the wider target population (Olson and Kellogg, 2014). A survey is conducted to collecting information from a large scale of a population, consistently and systematically, from or about people, to describe, compare, predict, or explain their knowledge, attitudes, and behavior (Fink, 2003). A survey can be useful for determining the characteristics of a population, comparing groups, and making explanatory claims about a population (Wohlin et al., 2012). Surveys are classified into three types: descriptive, exploratory, and explanatory. Other empirical studies may use a survey to collect data, verify hypotheses, and analyze data from a population. In a survey, the primary techniques to collect quantitative or qualitative data are interviews and questionnaires. The questionnaire can be used to detect trends and may provide valuable information and feedback on a particular process, technique, or tool. Surveys also allow you to make statistically accurate estimates for a population, when structured using random sampling (Lazar et al., 2017).
- **Systematic Literature Review.** A systematic literature review provides a comprehensive and valid landscape of the current position of literature in an area, both the identification, analysis, and interpretation of all available evidence related to a specific research question in a way that is unbiased and (to a degree) repeatable. The systematic reviews are secondary studies that are methodologically undertaken with a specific search strategy and well-defined methodology (Wohlin et al., 2012). As reported by Keele (2007), the most common reasons to conduct a systematic review are: (i) to summarize the existing evidence on a specific topic, (ii) to identify any gaps in current research to suggest areas for further investigation, and (iii) to provide a background to position new research activities appropriately. These guidelines also cover three phases of a systematic re-

view: planning, conducting, and reporting the review.

It is essential to say that without a sound and proven protocol, it can be challenging to carry out efficient and effective systematic reviews or empirical research (Malhotra, 2015; Lazar et al., 2017). Thus, a research methodology must be complete and repeatable, which will enable comparisons to be made across various studies when followed in a replicated, systematic, or empirical study. The empirical and systematic study process phases include the definition and design of the study, research conduct and analysis, and interpretation and reporting of the results. Moreover, sound and proven protocols bring the possibility to assess the obtained results from an empirical or systematic study. Then, biased or misleading results can be identified easily, revealing the quality of the study. In the next section, we present some related work that assesses the quality of empirical and systematic studies.

2.2 Related Work

We looked for studies related to our paper's goal, i.e., those that performed an empirical study for analyzing the case studies, experiments, surveys, or systematic reviews conducted in the HCI field. Although this kind of analysis is common in other areas, such as Software Engineering (Zannier et al., 2006; Silveira Neto et al., 2013; Barbosa et al., 2017; Kitchenham et al., 2019), to the best of our knowledge, we only found our preliminary study with this goal.

In our preliminary work (Damasceno et al., 2019), we performed an empirical study to assess the quality of a sample composed of papers published over the 17 editions until 2018 in the IHC symposium proceedings. For this, we developed a protocol that involves the definition of classification and quality assessment processes. A checklist composed of questions related to empirical studies (case study, experiment, and survey) and systematic reviews was used to identify the study's quality reported in each paper of the sample. We found that (i) the quantity of empirical studies and systematic reviews has increased over the lifetime to the IHC Symposium, and (ii) an increase in the soundness of the empirical validations and systematic reviews before the 14^o IHC Symposium (IHC 2015).

Although those results reveal initial evidence on the quality of empirical studies and systematic reviews in the IHC Symposium, in this paper, we extend these results by increasing their external validity by adding 28 papers from the 18^o IHC Symposium (IHC 2019) to the previous sample. Moreover, we performed new analyses, including:

- A further analysis of the basic characteristics used to the identification of empirical studies and systematic reviews;
- A set of recommendations for each empirical study (experiment, case study, and survey) and systematic review taking into consideration the results obtained from the quality assessment process; and
- A detailed investigation considering if the papers provide enough information for repetition or replication of their results.

3 RESEARCH METHOD

This section presents the protocol used for setting the our study to evaluate the empirical studies and systematic reviews reported in the IHC Symposium proceedings.

3.1 Protocol

To perform the qualitative and quantitative assessment of the empirical studies and systematic reviews reported in the IHC Symposium proceedings, we developed a protocol based on the one proposed by Barbosa et al. (2017).

The protocol is composed by three main phases (*planning*, *execution*, and *reporting*), as shown in Figure 1. The phase of *planning* aims to provide the set of activities required for planning the assessment. Then, research questions and hypothesis are defined in this phase. Moreover, the classification protocol and its review process are described. We used the classification protocol to define whether a study performed an empirical study (experiment, case study, and survey) or a systematic review. Finally, the quality criteria and its review process are defined, resulting in a checklist to drive the study assessment.

In the *execution* phase, the studies are selected following a criterion and they are distributed for researchers who will analyse a subset of selected studies. Then, the classification is performed and reviewed following the *classification process* defined in the previous phase. After that, each study is analysed following the checklist. In the last phase *reporting*, the obtained results are detailed.

As we said, we needed to adapt the protocol defined by Barbosa et al. (2017) as:

- The *classification protocol* previously defined did not avoid the risk of a wrong classification by the author to guarantee a sound quality assessment. For example, the protocol only considered the classification defined by the study's author, and it did not foresee that this classification could be wrong;
- The research questions and some checklists' items were updated because the criteria used to define a systematic review or the kind of empirical evaluations (*experiment*, *case study*, and *survey*) was unclear. We detected it when we tried to classify the papers of IHC Symposium for the first time, and this process was laborious, requiring a set of meetings to analyze and discuss the checklists' items;
- Still in the checklist, we realized that significant questions for defining systematic reviews and empirical studies were not considered. For example, in the checklist proposed by Barbosa et al. (2017), some features related to a specific empirical study should have a standard for all types of empirical evaluations.

3.2 Research Questions

We formulated the following research questions (RQs):

RQ1: Has the quantity of performed empirical evaluations and systematic reviews increased over the IHC Symposium lifetime (since 1998 until 2019)? *Rationale:* Empir-

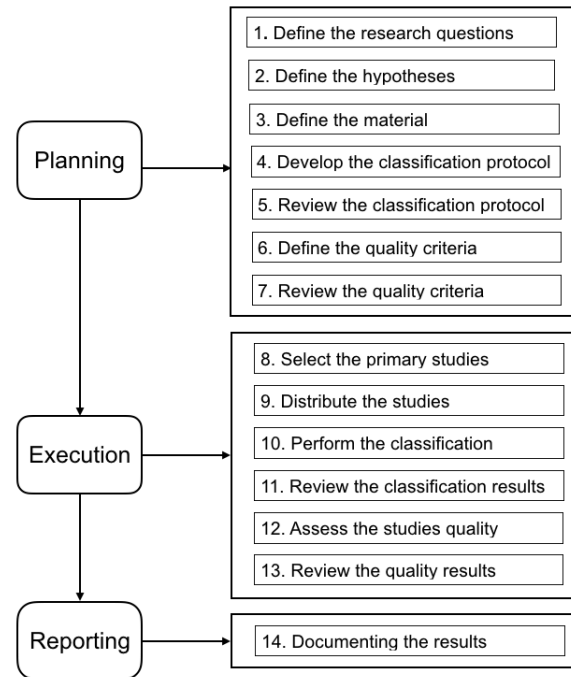


Figure 1. Protocol process.

ical evaluations are necessary for the HCI field as they allow that the theory can be compared to the real world, and systematic reviews are crucial for putting together the evidences collected from empirical studies. In this sense, it was expected that the number of studies using empirical assessments and systematic reviews had been raised in the considered period.

RQ2: Were the protocols used in papers published at IHC Symposium lifetime (since 1998 until 2019) to conduct empirical studies and systematic reviews sound and proven? *Rationale:* This question aims at investigating whether the protocol used to run an empirical study or systematic review reported in the papers under analysis were, in fact, adequately defined by their authors. As explained in Section 2, our study only evaluates the quality of the studies classified as case studies, experiments, surveys, or systematic reviews.

RQ3: What were the most widely used (empirical or systematic) study in papers published at IHC Symposium lifetime (since 1998 until 2019)? *Rationale:* This RQ has the purpose of identifying the most used study types considered in our analysis: experiment, case study, survey, and systematic review. There was an initial expectation that the most widely study was one of those major types.

RQ4: How many studies published at IHC Symposium lifetime (since 1998 until 2019) provide enough information for replication or repetition of their results? *Rationale:* The availability of data for further analysis are considered to reproduce the results of the study (Wohlin et al., 2012). In this sense, it was observed if the data are available for further analysis, and the research protocol defined in the paper under analysis includes the data collection, population definition, and analysis mechanisms.

3.3 Hypotheses

Our study was planned to take into account the following hypotheses, which address the qualitative and quantitative as-

pect of our study:

H_{01} . The quantity of empirical evaluations and systematic reviews has not increased over the IHC Symposium proceedings.

H_{11} . The quantity of empirical evaluations and systematic reviews performed has increased over the IHC Symposium proceedings.

H_{02} . The quality of empirical evaluations and systematic reviews has not increased over the IHC Symposium proceedings.

H_{12} . The quality of empirical evaluations and systematic reviews has increased over the IHC Symposium proceedings.

In order to statistically evaluate our hypotheses, we followed the assessment process used by Barbosa et al. (2017). We decided to use the same process as the strategy is simple and effective to compare two populations with different size.

First, we divided our sample into two populations (P_1 and P_2), where P_1 represents the population of studies from the former editions and P_2 represents the remaining studies from most recent editions. To explore all the scenarios, all possible divisions in two were made in the timeline. In the first comparison, P_1 consisted only of studies from the 1998 edition, while P_2 contained studies from 1999 to 2020. In the second comparison, P_1 considered articles from the 1998 and 1999 editions, while P_2 considered remaining articles starting from the edition of 2000. At each comparison, the articles from the oldest edition from P_2 were moved to P_1 until the last comparison, where P_2 represented only the 2019 edition. In each scenario, we added the quantity and quality rate of empirical studies and systematic reviews that made up each population. We calculated the quality rate by the responses given to each checklist's question (shown in Table 1). We further explain how the quality rate is calculated in Subsection 3.6.

After that, we performed the Mann-Whitney U Test (Malhotra, 2015; Wohlin et al., 2012) for each comparison and calculated the effect size between them using Vargha-Delaney A (Vargha and Delaney, 2000). While the test allows us to accept or reject the hypotheses, the effect size is a real number ranging from 0 to 1 that relatively represents the number of times that a value in one population is greater than the others. We used \hat{A}_{21} to represent the effect size of P_2 in relation to P_1 . Thus, the greater \hat{A}_{21} , the bigger the values in P_2 are in comparison to the ones in P_1 .

Finally, we can reject H_{01} and accept H_{11} if the ratio given by the number of studies over the total number of papers in P_2 is greater than in P_1 . Otherwise, H_{11} is rejected, and H_{01} is accepted. Moreover, aiming to know whether H_{02} can be rejected and, consequently, H_{12} can be accepted, while we performed the test to the quality rates obtained by the papers of both populations.

3.4 Variables

The independent variables of this work are the checklists, the researchers' evaluation, the period under analysis, and the chosen conference. The dependent variables explored in this work are the quality of the studies depending on the study's type (experiment, case study, survey, and systematic review) defined in the paper and assessed by the researchers.

3.5 Material

A set of eight researchers collected the peer-reviewed technical papers published across the all IHC Symposium editions. Originally an annual event, between 2003 and 2010, the IHC Symposium became biennial, alternating its accomplishment with the Latin American Congress of Human-Computer Interaction (CLIHIC), thus totaling 18 editions until 2019. We used the following inclusion and exclusion criteria:

- Inclusion criteria: Research papers published from 1998 until 2019 in the main track of the conference.
- Exclusion criteria: Papers not available, short papers, invited talks, panels, banners, tutorials, or tool sessions papers were not included in our analyses.

Due to these criteria, we obtained a population size of 526 papers. From this population, we randomly drew a sample of these papers because their number was high. Thus, for each year, we got a half for the number of papers. Whether this number was not an integer, we leveled it up. For instance, in 2016, 33 papers were published. By the process of choice, 16.5 would be chosen, but for not being an integer, we rounded it to 17. At finished of this process, we obtained 271 papers. However, 12 papers were not available in online IHC Symposium proceedings, leaving 259 papers in our sample.

A calculation of the sample confidence level was made. For a 95% confidence level accurate, and a confidence interval of 5%, the minimum confidence sample size expected was 223 papers. As our sample size has 259 papers, representing 49.24% of the population, this amount is inside of the confidence level range suggested.

3.6 Methodology

The research design in this study applied a sequential strategy comprising two phases. Firstly, we carefully examined each of the papers in the population (259) to determine whether it contained the main components required in an empirical evaluation or a systematic review. Secondly, we analyzed the papers that contained an empirical study or a systematic review from the previous phase to determine its type, according to the quality assessment process (Subsection 3.6.2). In both phases, each paper was independently reviewed by at least two researchers to mitigate personal bias.

In this context, we have compiled a checklist (see Table 1) based on the literature (Kitchenham et al., 2009a; Runeson et al., 2012; Wohlin et al., 2012; Linåker et al., 2015) to capture the general properties common to any empirical evaluations and systematic reviews and to identify specific characteristics related to these types of studies reported in the papers. The checklist was verified on the basis of literature about empirical research methods and systematic reviews (Lazar et al., 2017; MacKenzie, 2013). All researchers participated in the elaboration of this checklist. Each checklist's question is evaluated as "Yes" (indicating that data for the specific question is clearly available), "Partially" (data is vaguely available), or "No" (indicating that data is unavailable) with a corresponding score of 1, 0.5, or 0, respectively. We defined this scale because a simple Yes/No answer may

be misleading. This decision is also in line with the quality checklist recommendations presented by (Kitchenham et al., 2009b). Some of the questions were not applicable (N/A) to some studies, then, these studies were not evaluated for those questions. To assess a paper, we added the scores for each question, and found the percentage over the number of applicable questions for that paper. We called the resulting number the **quality rate**.

A pilot project was conducted to validate the checklist and synchronize the understanding of each question by each researcher. Renowned research publications from each kind of empirical studies (case study (Karlström and Runeson, 2006), experiment (Wohlin and Wesslen, 1998), and survey (Linåker et al., 2015)), and systematic reviews (Kitchenham et al., 2007) were used as material in this process. Our goal was to identify whether the checklist was able to identify high-quality papers for each type of empirical evaluation, and systematic reviews. Moreover, 8 papers (2 for each kind of study) of IHC proceedings were analysed after the improvements applied to the checklist. The papers used in the pilot were not part of the sample.

The last seven authors are students from the Academic Master's in Computer Science. All the authors have some experience level in software engineering and development, and have recently completed the course *Empirical Studies in Software Engineering* as a requirement of the program. Moreover, a PhD student and a PhD in Computer Science integrate the team of researchers.

In both phases, each researcher from each pair, individually, answered the questions about the identification and the categorization of the study performed in the papers. Whether the title, abstract, keywords, introduction, and conclusion were not sufficient for this identification, the entire paper was read. Next, the answers were compared, and the pair of researchers tried to solve the detected inconsistencies. When a pair of researchers could not fix a divergence, the first author supports in solving it.

3.6.1 Categorization Process

In this process, as illustrated in Figure 2, papers were randomly assigned to pairs of researchers. In order to answer RQ1, each study was analyzed to determine if any *empirical evaluation* was used to validate the contribution or reported a systematic review. In this sense, general questions were formulated (see Table 1) to identify evidence on these study's types (empirical or systematic review studies) in the validation process. These questions were aiming to characterize the utilization of these study's types in the verification process, regardless of a specific type. Each pair of researchers analyzed individually their group of papers.

3.6.2 Quality Assessment Process

In order to answer RQ2 and RQ3, a quality assessment process was applied to the papers that included empirical studies or systematic reviews, according to the previous phase. The studies were classified as experiments, case studies, surveys or systematic reviews in the previous phase. The eight researchers were randomly divided into four teams of two

members, and each team was randomly assigned to analyze the two editions of the IHC Symposium. The papers of the 2018 and 2019 editions were shared later randomly between the teams.

Each team of researchers answered the questions about systematic reviews, and the four types of empirical study. This classification was done individually by each researcher. Classifying a particular research paper in one of the types aforementioned might be not trivial, because the paper may fit into two types simultaneously. In this scenario, studies were classified according to the type used to validate their contribution.

This process was documented collaboratively in an online spreadsheet in Google Docs. When the researchers entered the data that had been extracted from the papers into the scheme, they provided a short rationale to justify why each paper was supposed to be in a particular category. From the final spreadsheet, the frequencies of papers in each category are calculated.

4 CHARACTERIZATION OF THE STUDIES

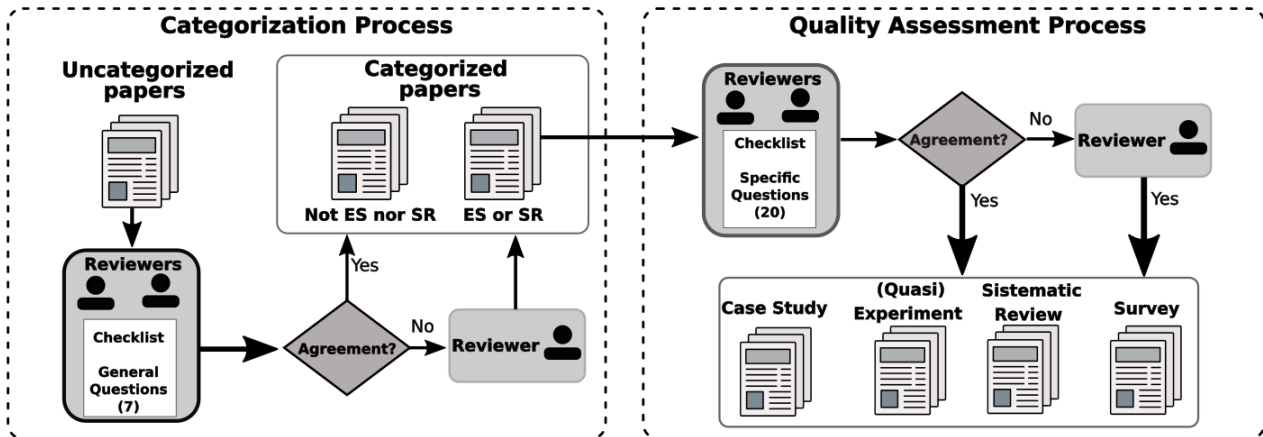
This section presents the characterization of each type of study based on the quality evaluations carried out by us. From this characterization, we draw conclusions and indicate suggestions on how empirical studies and systematic reviews in HCI can be improved.

4.1 Distribution of Studies per Type

From the sample composed of 259 studies, 131 (~51%) of them reported empirical studies or a systematic review in the researcher's perspective. Table 2 presents the percentage distribution of the empirical studies or a systematic review taking into consideration the authors' and the researchers' perspectives. We can notice that the studies labeled with "Others" reached 40.15% (104) of the total of studies, where these involve ones that do not fit into any of the four types evaluated from the authors' perspective. It is essential to say that other types of empirical study can be reported in these studies, but they are not considered in the scope of our work.

The papers that did not perform an empirical study or a systematic review, or had low quality indicated by the researchers' perspective represent almost 49.42% (128) of the total of studies. The "low quality" characteristic can make a replication of studies hard or unviable, compromising the validity of the study's findings. We further investigate the possibility of replicating HCI studies published at the IHC Symposium in Section 5.4.

Regarding the types of studies, we identified evidence of discrepancies in the classification of empirical studies and systematic reviews, which suggests a departure from the author's perspective on each study's requirements and adequacy. The most significant discrepancy was found in the *Case Study* type, where the author's perspective represents 18.15% (47) of the total of studies, while the researcher's perspective was 34.36% (89). This finding may indicate a concern about the extent to which authors are able to apply



Legend: ES: Empirical Studies SR: Systematic Reviews

Figure 2. Classification and quality assessment process.

Table 1. Checklists used to evaluate the studies.

General Questions
<p>Q1. Are the research objectives and the research questions defined in advance?</p> <p>Q2. Is the research protocol explicit? (data collection, population definition, analysis mechanisms)</p> <p>Q3. Are threats to validity conducted in a systematic way, and are countermeasures taken to reduce them?</p> <p>Q4. Is a pilot study presented in the research?</p> <p>Q5. Is the data presented or made available for further analysis?</p> <p>Q6. Does the research make evident the method of analysis applied in your data?</p> <p>Q7. Do the findings of the study answer the research questions and / or hypotheses that have been raised?</p>
Experiment
<p>Q1. Are the hypothesis (null hypothesis and alternative hypothesis) presented?</p> <p>Q2. Does any method was used to prove/reject the hypothesis?</p> <p>Q3. Are the independent, dependent variables and their metrics presented?</p> <p>Q4. Were quantitative methods applied to interpret the results?</p> <p>Q5. Are the treatments presented?</p> <p>Q6. Is the sample randomized?</p>
Case Study
<p>Q1. Is the case and its analysis units explicitly defined and presented? (size, domain, process and subjects)</p> <p>Q2. Does the researcher avoid any interference in the process, technique and methodology used in the case study?</p> <p>Q3. Is triangulation applied? (multiple methods of collection and analysis, multiple authors and various theories)</p> <p>Q4. Are ethical issues handled appropriately (personal intentions, integrity, confidentiality, consent and approval of the review board)?</p> <p>Q5. Does the study report provide implications for practice?</p>
Survey
<p>Q1. Does the study specify and thoroughly describe its sampling method (e.g. probabilistic or non probabilistic sampling methods)?</p> <p>Q2. Does the study describe how the questionnaire was designed (e.g. the number of questions, type and wording of the questions, translations, etc.)?</p> <p>Q3. Is the questionnaire of the study available (e.g. attached to the report or included as an appendix, etc.)?</p> <p>Q4. Does the study provide information on its response rate?</p> <p>Q5. Does the study formally assess its trustworthiness (e.g. through calculating measurement error, sample frame error, error of selection, etc.)?</p>
Systematic Review
<p>Q1. Is it possible to identify the population, the intervention and the outcome in main research question?</p> <p>Q2. Are the reviews inclusion and exclusion criteria described?</p> <p>Q3. Is the search strategy defined in the study?</p> <p>Q4. Did the reviewers assess the quality/validity of the included studies?</p>

Table 2. Distribution of the Studies (%) - Authors' vs. Researchers' perspectives

Perspective	Case Study	Experiment	Systematic Review	Survey	Others	Subtotal	No Empirical Evaluation nor Systematic Review
Author	18.15% (47)	10.42% (27)	5.79% (15)	3.09% (8)	40.15% (104)	77.6% (201)	22.39% (58)
Researcher	34.36% (89)	5.02% (13)	8.49% (22)	2.70% (7)	0.0% (0)	50.58% (131)	49.42% (128)

empirical studies or systematic reviews in practice, respecting their characteristics.

to/ydo9X1C.

The classification of the studies according to authors' and the researchers' perspectives is available at <http://split>.

4.2 Considerations about Studies in general

Figure 3 shows the percentage of scores for each general question. These scores were defined by the researchers following the checklist (Table 1) while analyzing each study. This set of questions reveals important elements in any empirical study or systematic review. Regarding the research questions and objectives, captured in Q1, their definition is reported in 77.22% (101) of studies, indicating the authors' concern about approaching their research problem and the study type performed.

The development of the research protocol involves a series of steps to reduce the possibility of research bias. These steps define a basic process and procedures that will be followed during the study conduction. In approximately half of the studies (50.97%, 67), an explicit and complete research protocol (Q2) is detailed involving the data collection, population definition, and analysis mechanisms. Moreover, a partial description is presented in 30.12% (39) of studies.

The validity of the results is an essential concern for any empirical study and systematic review since it will provide complete information about the research results' limitations and applicability. Few studies in the sample have detailed threats to the validity of the findings (Q3), in the vast majority of studies (75.68%, 99) did not indicate this aspect. Researchers have a responsibility to discuss any limitations of their study (Kitchenham et al., 2002) to relate which any sources of bias may have compromised the design and analysis elements of the study.

Pilot studies are intended to identify any problems in the protocol and testing the instruments used to support the process, such as questionnaires. Pilot studies may also contribute to reliability assessment. However, data about pilot study (Q4) was presented in only 12.74% (17) of studies in the sample.

It may be indicative of a worrisome scenario because research that cannot be replicated is useless. (MacKenzie, 2013). The findings become more reliable (and one has greater confidence in the theories they support) if studies are replicated (i.e., are repeated or conducted in different settings). Similar findings in replications increase the confidence in the results. In the sample, the availability of data for further analysis (Q5) is full or partially presented in 72.10% (94) of them. Moreover, in 62.93% (82) of studies, the method of analysis applied in the data (Q6) was explicitly evident, and the obtained findings answer the research questions or hypotheses (Q7) that have been raised in 73.36% (96) of studies.

From this analysis, we can bring the following **recommendations**:

- Identify the threats to validity that can affect the empirical study or the systematic review. After that, plan and perform mitigation actions for each one of the threats. A list of threats can be found in Wohlin et al. (2012). This process should be described in the paper preferentially in a dedicated section; and
- Before performing the empirical study or systematic review, it is important to perform a pilot study to verify the study's protocol's same inconsistencies. This study

should be detailed in the paper to bring more quality to the process performed in that study.

We found evidence that the analyzed studies have reported the research questions, the objective, the protocol, the collected data, the analysis method, and the findings to answer the authors' research questions. However, these studies have not detailed the pilot study and the threats to validity. It reveals that the researchers have described aspects that allow a good understanding of the performed study, but this understanding is not complete as the studies' limitations have not been deeply detailed.

4.3 Considerations about Experiments

In the sample, 27 (10.42%) of the studies were categorized as experiments by the authors' perspective, while the researchers' perspective identified only 13 (5.02%) experiments. Through the evaluation of the answers to questions related to the experiment type, we found the result presented in Figure 4.

Considering the related questions about the hypothesis presentation (Q1) and the method to reject or validate the hypothesis (Q2), the researchers' evaluation finds that 69.23% (9) of the experiments have presented null and alternative hypotheses in formal suitability. In the same proportion, the methods used to validate or reject them are also described. It demonstrates that the authors understand that the hypothesis definition is part of the planning level to the experiment goal execution and is substantial to the goal achievement's statistical analysis.

The third more expressive result was pointed out in question Q3 that deals with the empirical study variables and their metrics. In 84.62% (11) of the studies, the authors elected independent and dependent variables and the metrics to evaluate them in the analysis. It indicates that the authors recognize that the treatments' effect depends on measuring the variables and increasing the quality of the study analysis.

The quantitative methods (Q4) were identified in 92.31% (12), raising the hypothesis that the authors were concerned with basing their results in a quantitative statistical model. It demonstrates that the authors know that statistical analysis methods are essential for draw a meaningful conclusion from an experiment.

The question Q5 refers to the presentation of experiment treatment. The researchers found that all evaluated papers described this presentation. It can indicate that researchers are aware of the treatment is fundamental as a basis in the experiment.

Almost 50% (06) of all experiments did not randomize the data sample (Q6), which did not avoid the possibility of bias from the authors' perspective, or could indicate the possibility of quasi-experiments.

From this analysis, we **recommend** that when to assign participants in an experiment to groups following a random strategy, this is a quasi-experiment. It also provides valuable findings (Wohlin et al., 2012).

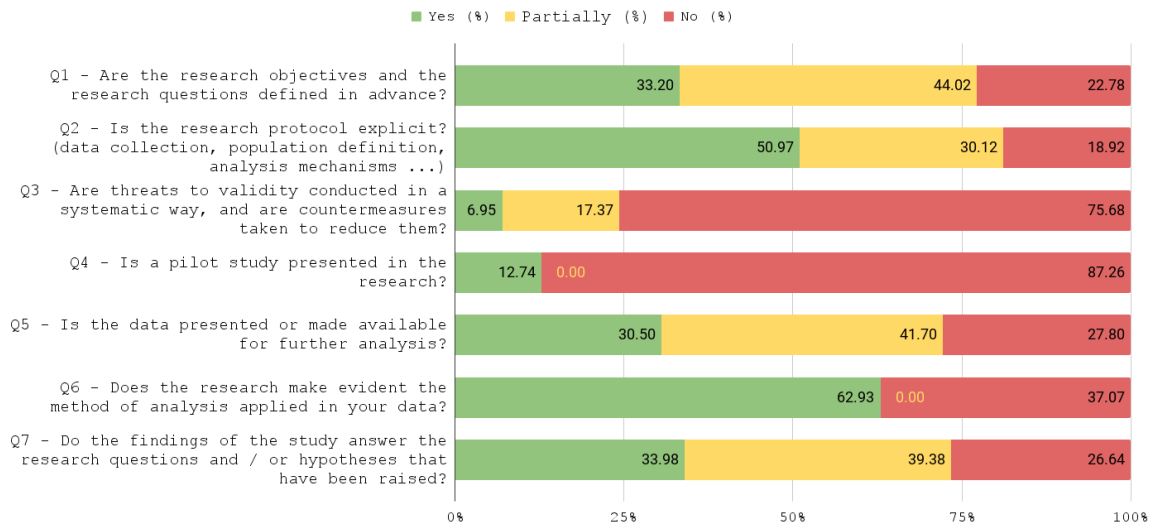


Figure 3. Answers to general questions

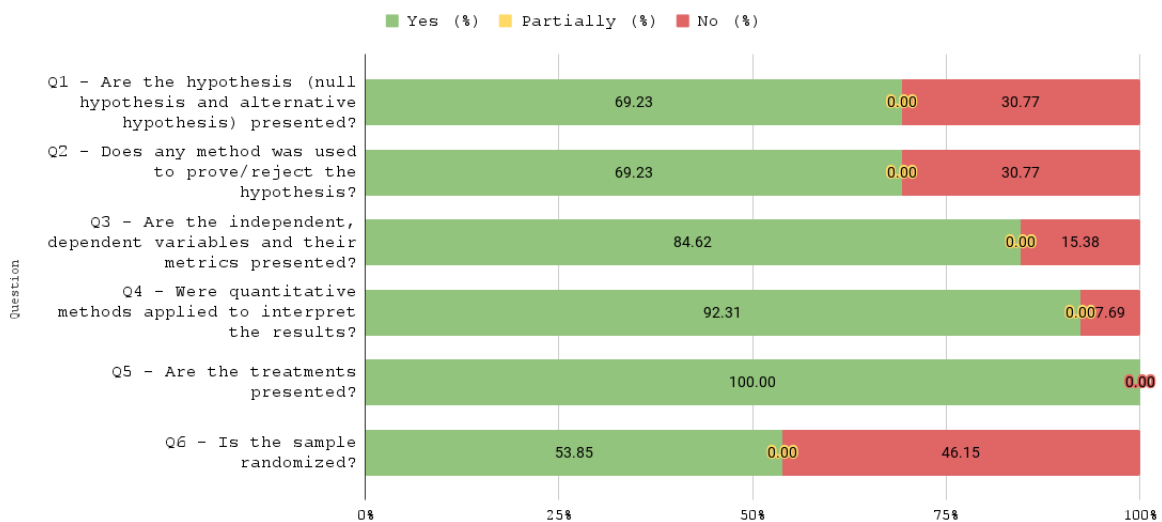


Figure 4. Answers to specific experiments questions

We found a clue that the analyzed studies reporting experiments have described the hypotheses, the method to prove or reject them, the independent and dependent variables, the quantitative methods, and the treatments. However, neither all studies did not randomize the used sample. It can reveal that the Brazilian HCI researchers thoroughly report experiments, but half of them are quasi-experiments.

4.4 Considerations about Case Studies

Regarding case studies, there were 28 papers with case studies included in which the authors’ perspective was confirmed by the researchers’ perspective. This confirmation was given through scores extracted from the specific checklist for this type of empirical study. In a total, 89 papers with case studies were found and analyzed by researchers, corresponding to approximately 67.9% of the 131 selected studies.

Figure 5 shows the frequency of the scores for each question in the case study checklist for all case studies identified

by the researchers. The first question (Q1) assessed whether the authors provided precise details about the cases of their analysis. 87% (77) of the authors provided some detail, while only about 12% (11) did not provide any detail. This data shows that the authors have been more attentive, exposing in a more precise way the object analyzed in their work.

In question Q2, the authors who avoided any interference in the data collection of their study were scored positively. The authors who interfere in the process, but justify such interference, were partially scored. Finally, negative scores were attributed to authors who interfere in the process used and do not attest to the justification for this. As can be seen in Figure 5, a higher percentage of authors avoids any interference, avoiding possible bias in their studies. This point contributes mainly to the reliability of the analyzed studies.

The question Q3 investigates the use of data triangulation. A positive score was provided if more than one data entry was used. When only one entry was used, the evaluation process scored partially. Finally, a negative score was provided

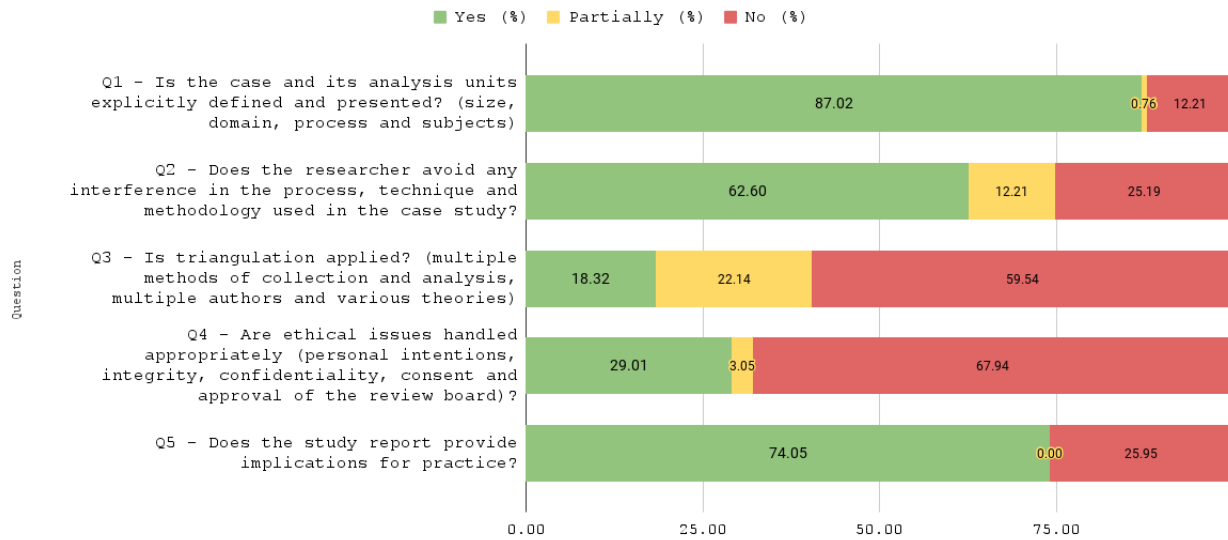


Figure 5. Answers to specific case study questions

when there was no data entry. This can be seen in Figure 5, most of the authors did not do well at this point, which can lead to a decrease in the safety of the results of the studies.

In question Q4, the studies that deal with ethical issues in their text were scored positively. However, the partially scored studies do not address these issues, but they justify the reason or do not yet need it. A negative score was attributed to studies that did not address ethical issues in their study, although it was necessary. An example is a study involving human participants that does not guarantee the anonymity of the participants. As shown in Figure 5, most researchers ignore this point in their studies.

Question Q5 concerns the usefulness of the study. The case studies, ideally, provide a lot of results for practice. A positive score was provided for those studies that provide a useful contribution and a negative score for those that do not. As can be seen in Figure 5, most studies did well at this point, which leads us to believe that researchers are directing their studies towards beneficial results for the practice.

From this analysis, we can bring the following **recommendations**:

- The use of other collection and analysis methods can increase the value of results from case studies. How to add these methods is discussed by Runeson et al. (2012); and
- Ethical issues are required for demonstrating that the researcher protected the participants from harm. Moreover, the methods used for mitigating these issues should be reported in the paper in detail.

We found a clue that the analyzed studies reporting case studies have described the size, domain, process, and subjects presented in the case, evidence for not interference of the researcher, and implications of the practice results. However, we found evidence that these studies have not conducted in detail triangulation neither approached ethical issues. It can reveal that the case study context is detailed, allowing the complete understanding of its conduct, but the results can be limited to only on data source.

4.5 Considerations about Systematic Reviews

According to Table 2, only 15 (~41%) papers were classified as a systematic review in authors’ perspective, but 22 papers reported systematic reviews from researches’ perspective. This finding was obtained after the checklist application showed in Table 1 considering the 131 papers classified by the researchers as empirical studies or systematic reviews. More specifically, when we analyzed the answers given for each question related to systematic review configuration, we can perceive that most of these studies have strong characteristics associated with this configuration. Figure 6 presents this detail.

Question Q1 was related to the primary research question formulation. All assessed papers present at least one for the following elements: population, intervention, and outcome. These elements limited the scope of a systematic review, avoiding unmanageable primary studies (Wohlin et al., 2012). About inclusion and exclusion criteria (Question Q2), only 13.64% (3) of papers did not present these criteria that support the researchers to include or exclude primary studies in their review. Moreover, inclusion and exclusion criteria should be based on the main research question.

The majority of papers classified as systematic review (90.91%, 20) described their used search process (Question Q3), increasing the possibility to find relevant primary studies. By this process, other researches can understand how the primary papers were identified and can update the review adding new primary papers published posteriorly. In Question Q4, the quality assessment of primary studies was considered. We identified that 54.55% (12) of systematic reviews did this evaluation driving the interpretation of findings and the recommendations for further research (Kitchenham et al., 2015).

From this analysis, we **recommend** that the studies’ quality assessment considered in a systematic review is necessary for indicating how these studies were conducted. Besides reporting the summary of these studies, the paper should have a section where the quality assessment is

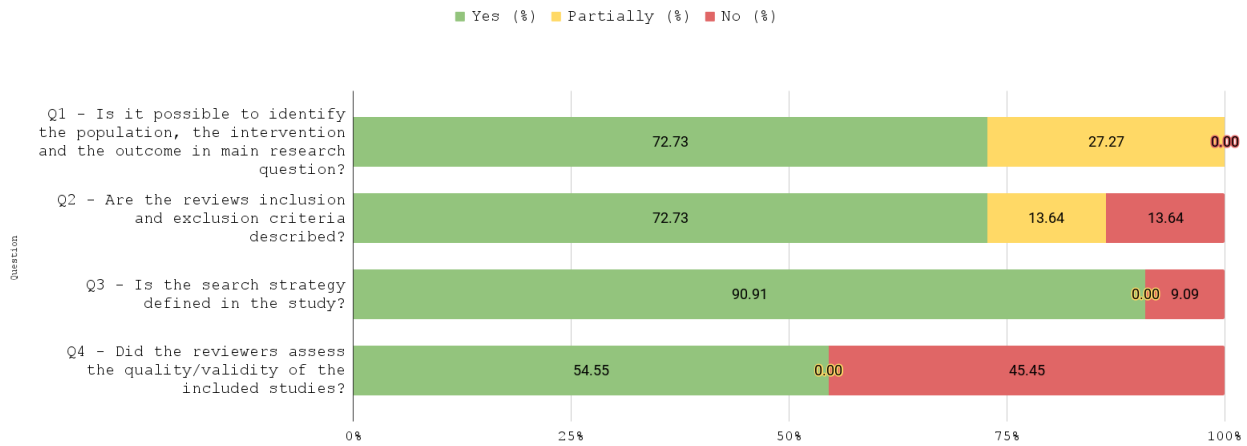


Figure 6. Answers to specific systematic reviews questions

presented and discussed. This assessment can be used to define what primary studies will be included in the review. Studies’ quality assessment is optional for systematic mapping study, as indicated by Kitchenham et al. (2015).

We found evidence that the analyzed studies reporting systematic reviews have described the population, intervention, and outcome in the main research question, the inclusion and exclusion criteria, and the search strategy. However, half of these studies did not perform the quality assessment of the primary studies. It can reveal that the Brazilian HCI researchers provide a complete view of the topic approached in systematic reviews, but these studies’ quality is not always assessed. This assessment is not mandatory for systematic mapping studies.

4.6 Considerations about Surveys

From the sample analysis, through scores extracted from the checklist, eight studies were found in the authors’ perspective as surveys, while seven were confirmed in the researcher’s perspective, as shown in Table 2.

Figure 7 shows the results for the specific questions that characterize a survey study. We can notice that questions Q2 and Q4 reached 100% (7). It demonstrates the author’s concern to describe how the questionnaires are prepared, reporting the number of questions, type, text, sequence, and groups. Moreover, the authors also provided a response rate.

Regarding questions Q1, Q3, and Q5, the answer “Yes” scored 75% (5), while the answer “No” scored 25% (2). Q1 addresses the sampling method’s specification and details, while Q3 checks whether the questionnaire is available, either in the article or in an external link. Both questions can help in replicating the questionnaire in other contexts. Q5 assesses the reliability of the study.

Among the four studies analyzed, only one study answered 100% of the questions, in contrast, three studies did not answer at least one of the questions. Among these three studies, it was identified that the authors could pay attention to detail who are the participants of their studies, make the questionnaire used in the study available as an attachment or appendix, and evaluate its reliability.

From this analysis, we **recommend** that the HCI Brazilian community continues to follow the right practices for conducting surveys.

We observed that most of the analyzed studies’ authors considered all questions for survey study defined in the checklist (Table 1). This evidence can indicate that the authors know how to perform this study type.

5 ANSWERING THE RESEARCH QUESTIONS

In this section, we provide results to answer the research questions posed in this work.

5.1 The quantity of performed empirical evaluations and systematic reviews (RQ1)

Figure 8 shows the distribution of empirical studies and systematic reviews by year. one can notice that the peak of studies happened in 2019, with 18 empirical studies and systematic reviews performed, followed by 2017 (16 studies) and 2015 (15). On the other side, from 2004 to 2015, one can observe an increase in the number of empirical studies and systematic reviews performed by the HCI Brazilian community.

In the figure, we used the least-squares polynomial method (Levenberg, 1944) to generate a trend line (green line in Figure 8) of this distribution, providing a visual representation of the trend variation and its resistance for quality rate. To show how the line is fitted, we used the Normalized Mean Squared Error (NRMSE) (Sammut and Webb, 2011) to measure the fitting error. The actual trend fits the equation of a line with coefficients of approximately 1.031 and -1.485 ($y = 1.031 * x - 1.485$), generating an NRMSE of 0.0985.

For understanding numerically how the trend line is characterized, such as its direction and steepness, we used the slope factor. In a linear function ($y = mx + b$), the slope is represented by the coefficient m . A positive slope ($m > 0$) indicates that the trend line is increasing (line going up from left to right). On the other side, a negative slope ($m < 0$)

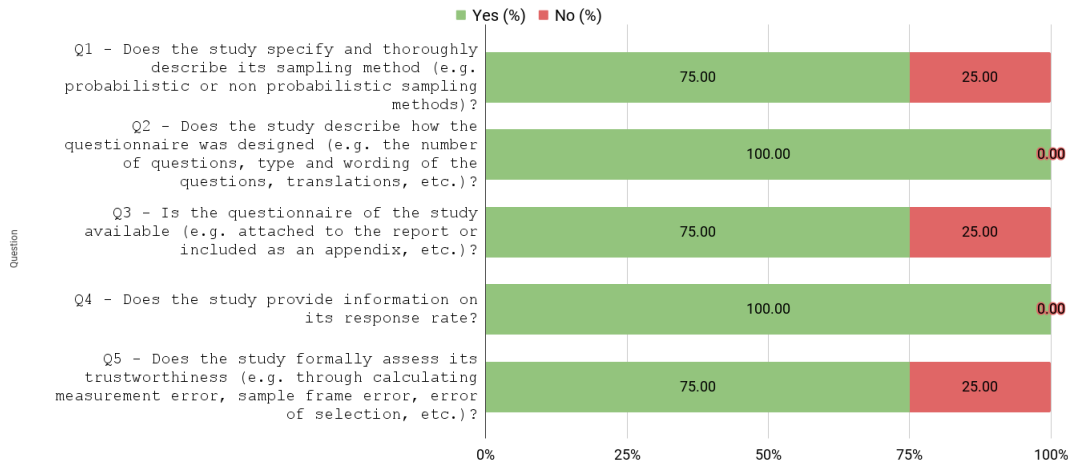


Figure 7. Answers to specific survey questions

indicates that the trend line is decreasing (line going down from left to right). Lastly, a null slope ($m = 0$) indicates that the trend line is constant (horizontal line). Based on that, we can identify an increasing trend in the number of empirical studies and systematic reviews with an approximate growth of 1.031 units of studies per year. It suggests that the number of studies has been growing over the years.

Expanding the analysis of the distribution presented in Figure 8, we performed a statistical test considering the number of studies presented in Figure 9. To answer the RQ1, we considered the following hypotheses previously defined in Subsection 3.3:

- H_{01} . The quantity of empirical evaluations and systematic reviews has not increased over the IHC Symposium proceedings.
- H_{11} . The quantity of empirical evaluations and systematic reviews performed has increased over the IHC Symposium proceedings.

In order to reject the null hypothesis H_{01} , it was necessary to divide the period of the IHC Symposium proceedings in two groups, P_1 and P_2 , and compare them to check if there was an increase in P_2 with respect to P_1 over the years. Thus, 17 configurations were developed and analyzed, represented by the expression $C_i = \{P_{1i}, P_{2i}\}$ where i ranges from 1 to 17 (3), $P_{1i} = \bigcup_{x=1}^i A_x$ e $P_{2i} = \bigcup_{x=i+1}^{17} A_x$. The term A_x belongs to the set S , the set of all years of the conference presented in the sample. Table 2 details the number of papers of the sample (259) distributed (P_{1i} and P_{2i}) along the configurations, indicating the rate of papers where the authors have used any empirical method to validate their contribution, in the researcher’s perspective.

To perform this comparison, we performed the Mann-Whitney U test (McKnight and Najab, 2010), considering a confidence level of 95% ($\alpha = 0.05$), and Vargha-Delaney A measure for effect size (Vargha and Delaney, 2000).

Splitting the conference lifetime by half, P_1 represents the period from 1998 to 2008 and P_2 refers to the period from 2010 to 2019 (see line 8 of Table 3). In this configuration,

the average presence of empirical studies and systematic reviews inside the sample is 23.19% in P_1 and 60.53% in P_2 . Comparing the most recent years group (P_2) to the previous years group (P_1), there is an increase of 37.34% in the average of the studies number. Alongside that finding, since the obtained p-value is less than 5%, this increase represents a significant statistical difference. Moreover, the effect size shows that the magnitude of the difference is medium. Thus, we could reject the null hypothesis, proving that there would be an increase in the number of studies if we had considered only that specific comparison between those two periods. However, though the chances of obtaining an error or a false positive are 5%, many other comparisons were executed and most of them do not present a similar output. Therefore, it would be risky to reject the null hypothesis (H_{01}) and affirm that there was an increase in the number of studies, considering solely that comparison.

Still analyzing Table 3, the comparisons between P_1 and P_2 , showed in the last two configurations, presented a p-value greater than α (5%), indicating that we can not reject the null hypothesis H_{01} with 95% confidence. This finding could be an indication that in recent years, the proportion of the present empirical studies and systematic reviews stayed stable. Even if the obtained p-value were less than 0.05 and the effect sizes were the same, they would show that the difference would be small or negligible.

We found evidence of increasing trend in the in the number of empirical studies and systematic reviews over the years, specially from 2004 to 2015. Moreover, when we divided the all empirical studies and systematic reviews performed by the HCI Brazilian community into two groups, we realized that the quantity of studies performed from 2010 to 2019 is greater than the ones performed from 1998 to 2008. Although it is statistically significant, we only considered a sample of papers in this analysis; then, this result can indicate that the quantity of empirical studies and systematic reviews increases in the IHC Symposium proceedings lifetime.

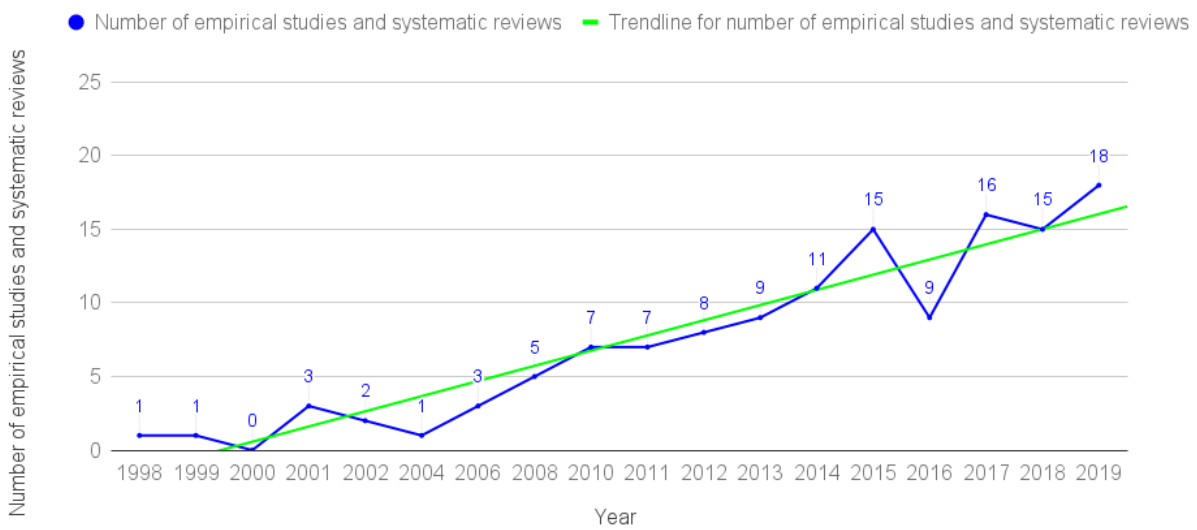


Figure 8. Quantity of empirical studies and systematic reviews performed by year

Table 3. Distribution of studies with empirical evaluation or a systematic review.

Conf. C_i	Sample Size		Emp. Studies or Syst. Reviews		Means		p-value	\hat{A}_{21}
	P_1	P_2	P_1	P_2	P_1	P_2		
1	8	251	1	130	0.1250000	0.5179283	0.0291619	0.6964641 (medium)
2	14	245	2	129	0.1428571	0.5265306	0.0053504	0.6918367 (medium)
3	22	237	2	129	0.0909091	0.5443038	0.0000492	0.7266974 (medium)
4	33	226	5	126	0.1515152	0.5575221	0.0000138	0.7030035 (medium)
5	39	220	7	124	0.1794872	0.5636364	0.0000102	0.6920746 (medium)
6	46	213	8	123	0.1739130	0.5774648	0.0000007	0.7017759 (medium)
7	56	203	11	120	0.1964286	0.5911330	0.0000002	0.6973522 (medium)
8	69	190	16	115	0.2318841	0.6052632	0.0000001	0.6866895 (medium)
9	79	180	23	108	0.2911392	0.6000000	0.0000049	0.6544304 (small)
10	95	164	30	101	0.3157895	0.6158537	0.0000034	0.6500321 (small)
11	110	149	38	93	0.3454545	0.6241611	0.0000096	0.6393533 (small)
12	123	136	47	84	0.3821138	0.6176471	0.0001583	0.6177666 (small)
13	140	119	58	73	0.4142857	0.6134454	0.0014341	0.5995798 (small)
14	162	97	73	58	0.4506173	0.5979381	0.0220355	0.5736604 (small)
15	179	80	82	49	0.4581006	0.6125000	0.0219741	0.5771997 (small)
16	203	56	98	33	0.4827586	0.5892857	0.1592176	0.5532635 (negligible)
17	231	28	113	18	0.4891775	0.6428571	0.1256315	0.5768398 (small)

5.2 The quality of the protocols used to perform empirical studies and systematic reviews (RQ2)

Figure 9 shows distribution and trend line of the quality rate of all empirical and systematic review studies together per year. The values shown in the figure is an average of the sum of the quality rate of all types of empirical evaluations and systematic reviews divided by the number of all studies performed in the same year. One can observe that the peak occurred in 1999 with a quality rate equals to 1.548. Further, an increase in the quality happened from 2004 to 2019, but from 2006 to 2019, the quality rate is relatively uniform.

Similarly to Section 5.1, we used the least-squares polynomial method to generate a trend line and NRMSE to measure the fitting error. Then, we found the coefficients 0.034 and 0.897 the trend line introducing a NRMSE of 0.190. It indi-

cates that the quality of all types of empirical evaluations and systematic reviews increases 0.034 units over the years. We went further and generated the distributions and trend lines for each empirical study and systematic review, as shown in Figure 10. Each slope factor and NRMSE are detailed in Table 4.

Table 4. Slope factor and NRMSE from trend analysis for empirical study and for systematic review

	Quality Rate Trend Analysis	
	Slope	NRMSE
Case Study	0.053	0.215
Experiment	0.053	0.397
Survey	0.035	0.360
Systematic Review	0.114	0.242

Expanding the analysis of the distribution presented in Figure 9, we performed a statistical test considering the quality rate presented in the figure. To answer RQ2, we considered the following hypotheses previously defined in Subsection 3.3:

- H_{02} . The quality of empirical evaluations and systematic reviews has not increased over the IHC Symposium proceedings.
- H_{12} . The quality of empirical evaluations and systematic reviews has increased over the IHC Symposium proceedings.

Similarly to the evaluation about the quantity of performed empirical evaluations and systematic reviews, several comparisons were performed considering partitioning the studies with empirical evaluation and systematic reviews (131) from the sample in two groups, P1 and P2. For each configuration, it was performed an Mann-Whitney U Test and the effect size of the difference was calculated by Vargha-Delaney A measure.

Initially, we considered the partition that was analyzed in Section 5.1 by splitting the conference lifetime by half involving the papers from the early 11 editions to P1 and the rest to P2. Table 5 shows the sizes of the groups P1 and P2, the quality rate average of each group, the p-value, effect size between the P1 (1998 to 2008) and P2 (2010 to 2018, in 2009 did not have the conference) for each study type and for all studies together. In this comparison, for each study type, the obtained p-value is greater than an $\alpha = 5\%$. Thus, we can not reject the null hypothesis H_{02} with 95% confidence for any study type. For the overall comparison, the obtained p-value is also greater than $\alpha = 5\%$. Therefore, we can not reject the null hypothesis H_{02} with 95% confidence considering the overall as well.

After the evaluation of all data partitions it was not possible to reject the null hypothesis, and therefore it was not possible to demonstrate an increase in the quality of the articles, except for the configuration where P_1 represents the period from 1998 to 2014 and P_2 represents the period from 2015 to 2018. Table 6 presents the quality rates and the results of the Mann-Whitney U Test and Vargha-Delaney A measure in this case. Analyzing each study type individually, we can not reject the null hypothesis H_{02} with 95% confidence, since the obtained p-values are greater than $\alpha = 5\%$. However, considering all the studies in the comparison, the obtained p-value is less than 5%. This indicates that the difference in terms of quality rate between these two periods is statistically significant. Moreover, the quality average in P_2 is greater than P_1 , and although the effect size shows that the magnitude of such a difference is small, it is not negligible. Thus, using the same rationale applied previously, we could reject H_{02} and affirm the alternative hypothesis H_{12} , which would mean that the quality of the empirical studies and systematic reviews had increased over the IHC Symposium proceedings, considering the partition of data in this configuration. However, many other comparisons were also executed, and most of them do not have a similar outcome. Considering that risk, we do not reject or confirm any hypothesis related to the quality of the empirical studies and systematic reviews.

In summary, although we found, with 95% confidence, an outcome with a p-value less than 5%, it would not be possible to reject the null hypothesis and affirm the alternative hypothesis, considering all the empirical studies and systematic reviews, since it represents only the comparison between those periods, 1998-2014 and 2015-2019. However, we can point out as evidence that there was an improvement in the quality of the studies published in the symposium from the year 2015 edition.

We found evidence of an increase in trend of the quality of empirical studies and systematic reviews over the years, specially from 2004 to 2019. Moreover, when we divided the all studies performed by the HCI Brazilian community into two groups, we realized that the quality of studies performed from 2015 to 2019 is greater than the ones performed from 1998 to 2014. Although it is statistically significant, we only considered a sample of papers in this analysis; then, this result can indicate that the quality of empirical studies and systematic reviews increases in the IHC Symposium proceedings lifetime.

5.3 The most used study type (RQ3)

Figure 11 illustrates the amount of each study type by the author's and the researcher's perspectives. The author's perspective is represented by numbers in red, while the researcher's perspective in blue. The label "Others" represents the studies types not addressed in our assessment, and the label "Not Empirical Evaluation nor Systematic Review" represents studies that did not present any evaluation that could be regarded as empirical or did not describe a systematic review. This label also includes the studies that did not reach a satisfactory quality level for the researchers' perspective.

Considering the author's perspective, *case study* was the most used empirical evaluation, performed in 47 studies (18.14% of the sample). The second most used was *experiment*, performed in 27 studies (10.42% of the sample). At third and forth, was *systematic review* and *survey*, with 15 and 8 studies in the sample, which represents 5.79% and 3.08% of the total amount, respectively.

Regarding the researcher's perspective, 128 studies were classified as not empirical nor systematic review, involving studies without any empirical evaluation, or a systematic reviews, or with low quality evaluations. Thus, the quality assessment was done with the remaining 131 empirical studies and systematic reviews. In this perspective, the most used study type was *case study*, with 89 studies, representing 34.36% from the complete sample (259 studies). *systematic review* was the second most used study type, with 22 studies, which corresponds to 8.49% of the sample size. *Experiment* was positioned in third, with 13 studies (5.62%) while *survey* was conducted in 7 studies (2.70%).

As one can noticed in Figure 11, there were several studies that their classification differs from the author's and researchers' perspective. Figure 12 shows this difference. Each bar represents a study type classified by the author's perspective, while other classification from the researcher's perspective is represented as part of the bar following the caption of the figure. The authors indicated that 47 studies have a *case*

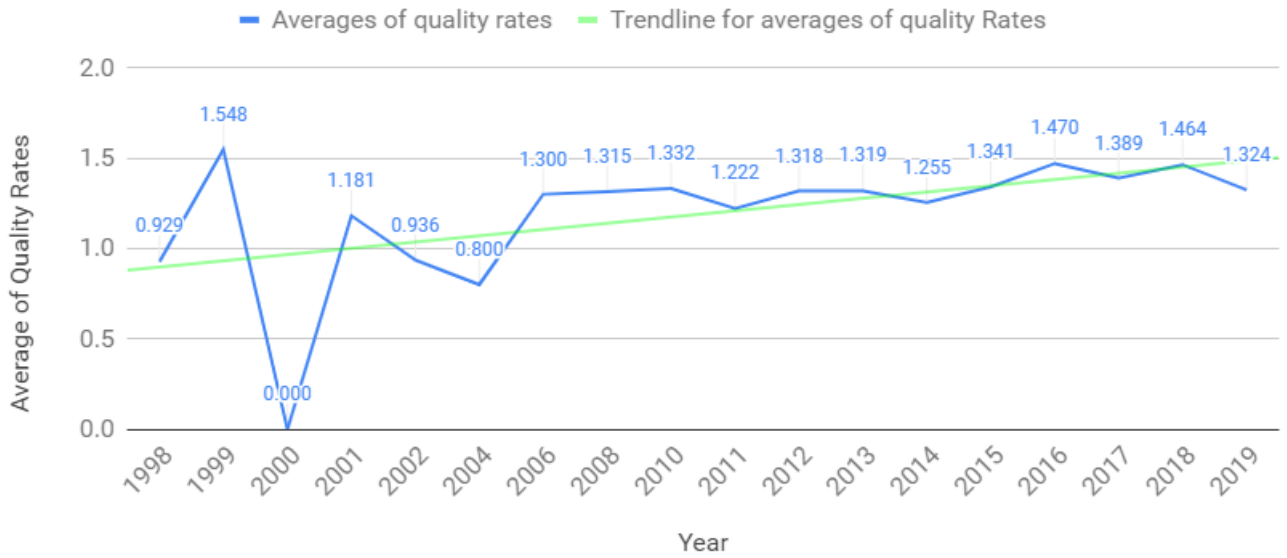


Figure 9. Quality of all empirical studies and systematic reviews performed by year

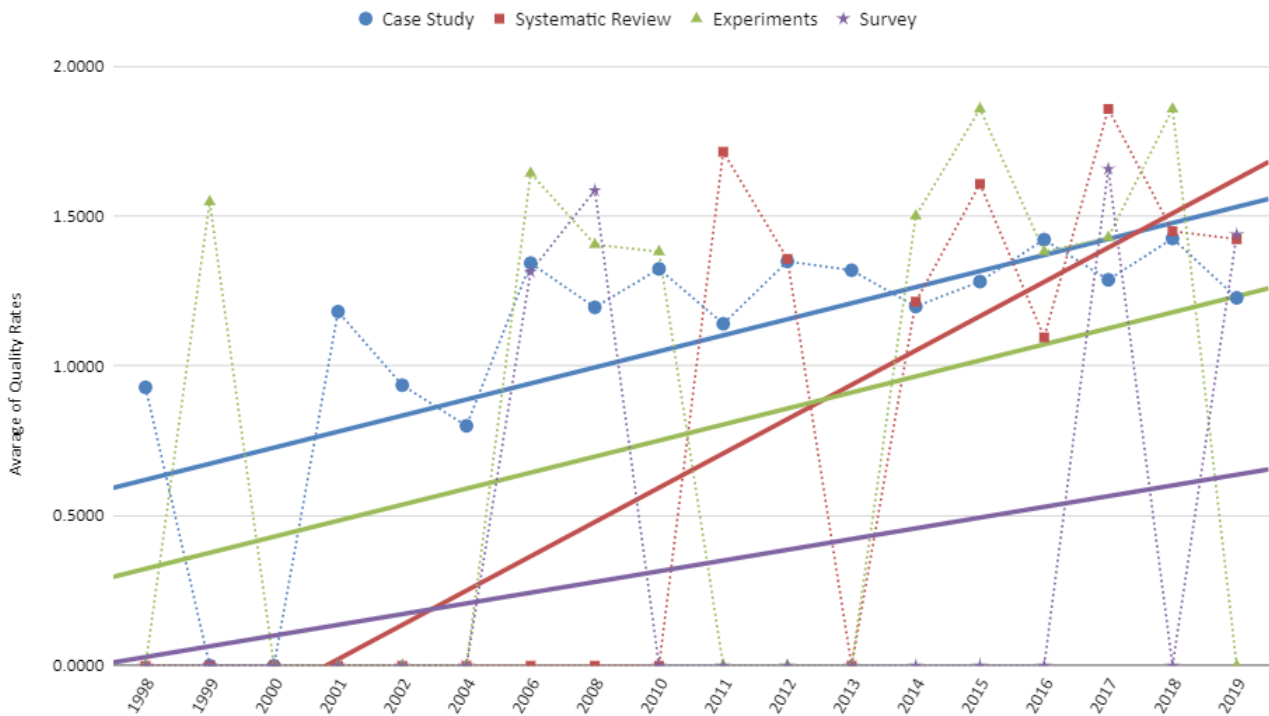


Figure 10. Quality of each empirical studies and systematic reviews performed by year

Table 5. Quality rate averages, p-values and effect sizes between P_1 (1998 to 2008) and P_2 (2010 to 2019).

Study Type	Sample Size		Quality Rate Average		p-value	\hat{A}_{21}
	P_1	P_2	P_1	P_2		
All	16	115	0.6291625	0.6941304	0.2054284	0.5975543 (small)
Case Study	11	78	0.5454545	0.6538462	0.0689576	0.6684149 (medium)
Experiment	3	10	0.8888667	0.7500000	0.3818598	0.3166667 (medium)
Systematic Review	-	22	-	0.7784091	-	-
Survey	2	5	0.7000000	0.8400000	0.4123552	0.75 (large)

study, but the researchers found that 25 were a case study, one were classified as a systematic review, and 21 studies did not even reach the minimum quality of the empirical study to be selected. Similarly, from the 27 studies that alleged to have an *experiment*, only seven were experiments. The other was

10 case studies, one survey, and 9 studies did not overcome the quality level.

From 15 *systematic reviews*, only one divergence was found, which was a study that were not selected as a systematic review nor an empirical study. Finally, eight *sur-*

Table 6. Quality rate averages, p-values and effect sizes between P_1 (1998 to 2014) and P_2 (2015 to 2019).

Study Type	Sample Size		Quality Rate Average		p-value	\hat{A}_{21}
	P_1	P_2	P_1	P_2		
All	58	73	0.6454017	0.7186068	0.0434646	0.6023855 (small)
Case Study	5	44	0.6155556	0.6659091	0.2264002	0.5737374 (small)
Experiment	6	7	0.8055500	0.7619000	1.0000000	0.5 (negligible)
Systematic Review	5	17	0.7000000	0.8014706	0.6555839	0.5705882 (negligible)
Survey	2	5	0.7000000	0.8400000	0.4123552	0.75 (large)

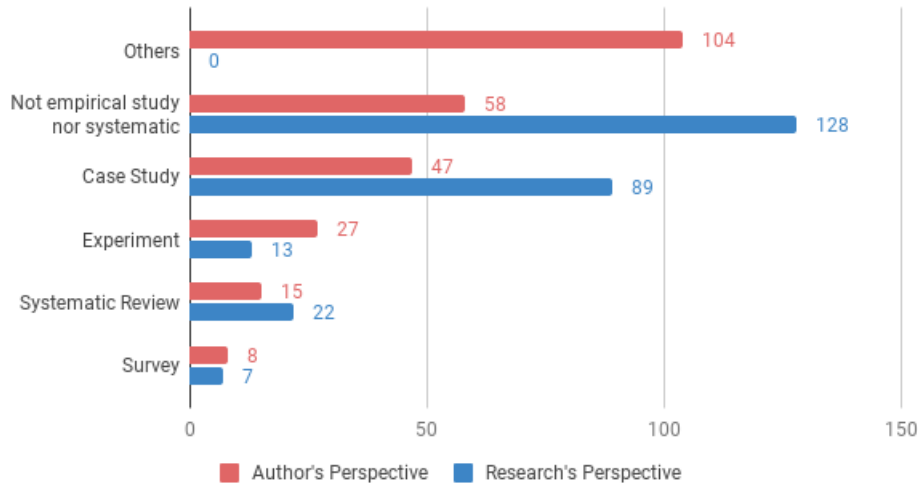


Figure 11. Types of studies included in the sample.

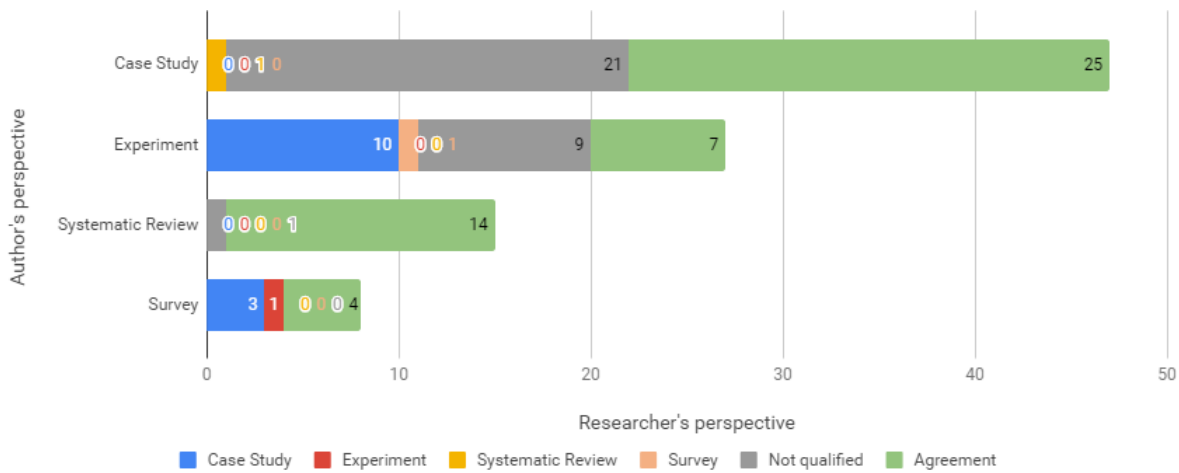


Figure 12. Comparison between author's and researcher's study classification.

veys were reported by the authors, but, according to the researcher's perspective, three were actually classified as case studies and one as an experiment.

This result indicates a lack of understanding of the study types' concepts and techniques, and the level of uncertainty on the use of these types by the Brazilian HCI researchers.

In the sample, case study is the most used empirical method by the HCI Brazilian community for assessing interfaces and software. Concerning the empirical methods and systematic reviews in general, we found evidence of a certain level of uncertainty on the use of these methods by the community, requiring that a review of the protocols used for conducting case studies, experiments, systematic reviews, and surveys.

5.4 The viability of replication (RQ4)

A key component of experimentation is its replication (Juzgado and Gómez, 2010). To consolidate a body of knowledge built upon empirical results, they have to be extensively verified, aiming to check if they can be reproducible. If the same results are reproduced in different replications, it is possible to affirm that such results are regularities existing in the piece of reality under study.

Also, achieving an expected standard of reproducibility or repeatability is crucial (MacKenzie, 2013). This is one reason for advancing a standardized methodology: it enforces a process for conducting and writing about the research that ensures sufficient detail is included to allow the results to be replicated. To answer RQ4, we used the general questions Q2 related to the explicit and sufficiently detailed protocol and

Q5 associated with data availability. These questions were defined in the protocol (Section 3) and address criteria that make possible repetition or replication of results, such as protocol definition and data availability for further analysis.

About the protocol (Q2), 103 (78.63%) of the studies reporting empirical studies or systematic reviews provided a detailed protocol, while the other 28 (21.37%) studies described the protocol partially. It means that researchers should not have problems in replicating the same protocol followed by the majority of the analyzed studies. If skilled researchers care to test the claims, they will find sufficient guidance in the methodology to reproduce, or replicate, the original research. This is an essential characteristic of research (MacKenzie, 2013). On the other side, 62 (47.33%) studies made available all data collected (Q5) from the performed empirical study or a systematic review, but 69 (52.67%) studies only partially presented them. It indicates that replications that intend to increase the studies' external validity may have difficulty accessing the collected data.

We also found 45 (34.35%) studies with detailed protocol and without making available the data, not allowing the study's increase of external validity in replication. Conversely, we realized that only four (3.04%) studies did not present the protocol in detail but make the collected data available. Unfortunately, these studies did not allow the complete replication, only the verification of the analysis performed by them.

As we said, we considered that a study could be fully replicated if it satisfied Q2 and Q5 simultaneously, i.e., it received "yes" in both questions. We observed that 58 (44.27%) studies fit in this condition. However, 24 (18.32%) of the studies reported the protocol and made available the data partially. It demonstrates that almost half of the studies reporting empirical studies or systematic reviews provide enough information for their replication.

To further investigate the distribution of these 58 studies over time concerning their replication, we performed a trend analysis over average scores obtained in the quality assessment process (presented in Section 3.6.2) grouped by year. As a result, we obtained the trend line and slope coefficients. The former indicates the best fit of scores average using a single line (shown in Figure 13), and the latter measures the increase or decrease.

Figure 13 shows the distribution of these studies by year and a trend line for Q2 (in blue) and for Q5 (in red). The translucent line represents the distribution of the average scores, and the linear line represents the trend of average scores over the years. The trend lines of Q2 and Q5 are characterized by coefficients 0.018 and 0.687 with NRMSE of 0.284 and -0.008 0.614 with NRMSE 0.270, respectively.

In addition, we can notice that the line behavior of Q2 and Q5 is similar from 1998 to 2008. However, while the quality of Q2 improves from 2008 to 2013, the contrary occurred for Q5 in the same period. From 2014 to 2019, Q2 remained almost constant, while Q5 had a peak in 2016. Considering both lines, one can observe that studies published from 2006 to 2008 and the ones from 2012 to 2015 are the most likely to be replicated.

Regarding the slope value for Q2 and Q5, we found that there is an increasing trend in both analyzed questions with

an increase of about 0.032 units of average score per year to Q2 and approximately 0.008 units of average score per year Q5. It indicates that the studies provide enough information and an increasing trend over the years for repetition or replication of results.

We found that almost 44% of the studies from the sample provide the complete description of the protocol and make the collected data available, allowing replicating these studies. Also, we found a clue that the increase of the studies providing this information is not constant over the IHC Symposium lifetime.

6 THREATS TO VALIDITY

In this section, the threats of our study were evaluated according to the taxonomies defined by Wohlin et al. (2012). We tried to eliminate these threats. When it was not possible, we mitigated their effects.

Construct validity: A threat arises from the questionnaire's construction due to its checklist used to assess the empirical studies and systematic reviews published at the IHC Symposium. To mitigate this threat, we identified the characteristics of *sound* empirical studies and systematic reviews in classical references published in the technical literature. Moreover, we performed a pilot study to validate the questionnaire. The pilot study consisted of assessing renowned publications using the questionnaire. An experienced researcher (the first author) indicated these studies. Another threat that affects the construct validity is the kind of studies considered in our analysis. Although the HCI community has also used other study types, we choose case studies, experiments, systematic reviews, and surveys to have a well-defined and formal protocol, reducing this threat.

Conclusion validity: We identified a threat affecting the conclusion validity associated with our expectation of high-quality papers from renowned authors or the last edition of the conference. To reduce this threat, we randomized sample choice and paper distribution among peer researchers. Another threat is related to the partitioning of the articles we made to confirm or reject the hypotheses. Although we performed a statistical test to analyze each partitioning, the test can lead to false positives. We mitigated this threat by describing our results as indications rather than as factual conclusions.

Internal validity: We recognized a threat presented in the assessment process as it is subjective and could give biased results. To mitigate it, we randomly distributed all papers among all researchers to avoid biased results. Another threat arises from the classification process as it involved subjective decisions by the researchers. A pair of researchers reviewed individual and separated the papers considering the proposed protocol to reduce this threat. Following, they compared the results and tried to solve the divergences among them. When the divergence remains, the first author assisted this process. Another threat affecting the internal validity is related to the authors' level of experience in the HCI field, as these authors



Figure 13. Score distribution and trend over years of Q2 and Q5 when their answers was yes

have a background in empirical studies and systematic reviews in SE. To mitigate this threat, we also analyze case studies, surveys, experiments, and systematic reviews published at the IHC Symposium. These types of studies have a well-defined protocol, and their elements are typical for IHC and SE.

External validity: A threat pointed out from the possibility of generalization of the results. Although we extracted a sample of papers from the IHC Symposium with confidence level range (95%) and confidence level (5%), and representing 49.15% of the population, we do not indicate the generalization of the results.

7 FINAL REMARKS

This paper presents an empirical study that provides a qualitative and quantitative assessment of a sample of the empirical evaluations and systematic reviews presented in the research papers published at IHC Symposium proceedings. In this sense, we formulate four research questions and define a protocol composed of checklists to assess the papers' quality. A sample representing $\sim 49\%$ was obtained from the papers' population along 18 editions spread over 21 years of the symposium.

The protocol allows the studies' classification from the sample into case studies, experiments, systematic reviews, and surveys. Moreover, it also supports us in assessing the quality of these studies. From the results obtained in the assessment, we compare the studies' classification considering the authors' and the researchers' perspectives and analyze each study's characteristics. This analysis reveals the essential findings of the conduction of these studies. Also, we provide some recommendations that can support the new empirical and systematic reviews studies' performances.

A quantitative analysis is performed to investigate if there is some evidence of the quantity and the quality of the empirical studies and systematic reviews have increased in the

IHC lifeline. Further, in the sample, we recognize the most empirical study conducted by the HCI Brazilian community and if the studies can be replicated. It is essential to say that our results can be influenced by the number of pages of the articles under analysis since they were published at a conference where there is a space restriction. Therefore, the empirical studies and systematic reviews reported in these articles may not have been described in detail.

Although our results were found from a sample, they can assist new studies performed in the HCI field by using the protocol as a checklist according to the study type that will be performed. Further, each study type's recommendations also help the community not repeat the same mistakes found in the analyzed studies.

As future work, we intend to continually increase the study's external validity, including all papers published at the IHC Symposium instead of a sample, and analyze the other empirical studies performed in the HCI field.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) Finance Code 001.

References

- Barbosa, D. M., Gadelha, R., Alencar, T., Neves, B., Yeltsin, I., Gomes, T., and Cortés, M. I. (2017). An analysis of the empirical software engineering over the last 10 editions of brazilian software engineering symposium. In *Proceedings of the 31st Brazilian Symposium on Software Engineering, SBES 2017, Fortaleza, CE, Brazil, September 20-22, 2017*, pages 44–53.
- Basili, V. R. (1996). The role of experimentation in software engineering: Past, current, and future. In *Proceedings of*

- the 18th International Conference on Software Engineering (ICSE), pages 442–449. IEEE Computer Society.
- Damasceno, A., Ferreira, A., Gama, E., Moraes, J. P. R., Alves, L. V., Barbosa, M. H., Chagas, M. L., Freire, E. S. S., and Cortés, M. I. (2019). A landscape of the adoption of empirical evaluations in the brazilian symposium on human factors in computing systems. In *Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems, IHC '19*, New York, NY, USA. Association for Computing Machinery.
- Fink, A. (2003). *The Survey Handbook*. SAGE Publications.
- Gergle, D. and Tan, D. (2014). *Experimental Research in HCI*. Olson J., Kellogg W. (eds) Ways of Knowing in HCI. Springer, New York, NY.
- Juzgado, N. J. and Gómez, O. S. (2010). Replication of software engineering experiments. In Meyer, B. and Nordio, M., editors, *Empirical Software Engineering and Verification - International Summer Schools, LASER 2008-2010, Elba Island, Italy, Revised Tutorial Lectures*, volume 7007 of *Lecture Notes in Computer Science*, pages 60–88. Springer.
- Juzgado, N. J. and Moreno, A. M. (2001). *Basics of software engineering experimentation*. Kluwer.
- Karlström, D. and Runeson, P. (2006). Integrating agile software development into stage-gate managed product development. *Empirical Software Engineering*, 11(2):203–225.
- Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3 EBSE Technical Report*. EBSE. sn.
- Kitchenham, B., Madeyski, L., and Brereton, P. (2019). Problems with statistical practice in human-centric software engineering experiments. In *Proceedings of the Evaluation and Assessment on Software Engineering, EASE '19*, page 134–143, New York, NY, USA. Association for Computing Machinery.
- Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2009a). Systematic literature reviews in software engineering - a systematic literature review. *Information and Software Technology*, 51(1):7–15.
- Kitchenham, B. A., Brereton, O. P., Budgen, D., and Li, Z. (2009b). An evaluation of quality checklist proposals: A participant-observer case study. In *Proceedings of the 13th International Conference on Evaluation and Assessment in Software Engineering, EASE'09*, pages 55–64, Swindon, UK. BCS Learning & Development Ltd.
- Kitchenham, B. A., Budgen, D., and Brereton, P. (2015). *Evidence-Based Software Engineering and Systematic Reviews*. Chapman & Hall/CRC.
- Kitchenham, B. A., Dyba, T., and Jorgensen, M. (2004). Evidence-based software engineering. In *Proceedings of the 26th international conference on software engineering*, pages 273–281. IEEE Computer Society.
- Kitchenham, B. A., Mendes, E., and Travassos, G. H. (2007). Cross versus within-company cost estimation studies: A systematic review. *IEEE Transactions on Software Engineering*, 33(5):316–329.
- Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., and Rosenberg, J. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, 28(8):721–734.
- Lazar, J., Feng, J. H., and Hochheiser, H. (2017). *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, Cambridge, MA, 2 edition.
- Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168.
- Linåker, J., Sulaman, S. M., Maiani de Mello, R., and Höst, M. (2015). Guidelines for conducting surveys in software engineering. <http://portal.research.lu.se/portal/files/6062997/5463412.pdf>. Accessed 10 December 2016.
- MacKenzie, I. S. (2013). *Human-Computer Interaction: An Empirical Research Perspective*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition.
- Malhotra, R. (2015). *Empirical Research in Software Engineering: Concepts, Analysis, and Applications*. Chapman & Hall/CRC.
- McKnight, P. E. and Najab, J. (2010). Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.
- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1:0021.
- Olson, J. S. and Kellogg, W. A. (2014). *Ways of Knowing in HCI*. Springer Publishing Company, Incorporated.
- Runeson, P., Host, M., Rainer, A., and Regnell, B. (2012). *Case Study Research in Software Engineering: Guidelines and Examples*. Wiley Publishing.
- Sammur, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.
- Serrano, J. F., Acuña, S. T., and Macías, J. A. (2014). A review of quantitative empirical approaches in human-computer interaction. In *Proceedings of the XV International Conference on Human Computer Interaction*, pages 1–8.
- Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., and Elmquist, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, Boston, 6 edition.
- Silveira Neto, P. A. D. M., Gomes, J. S., De Almeida, E. S., Leite, J. C., Batista, T. V., and Leite, L. (2013). 25 years of software engineering in brazil: Beyond an insider's view. *J. Syst. Softw.*, 86(4):872–889.
- Sjoberg, D. I. K., Dyba, T., and Jorgensen, M. (2007). The future of empirical methods in software engineering research. In *Proceedings of the Future of Software Engineering (FOSE)*, pages 358–378. IEEE Computer Society.
- Valverde, R. (2011). *Principles of Human Computer Interaction Design*.
- Vargha, A. and Delaney, H. D. (2000). A critique and improvement of the cl common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell,

- B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.
- Wohlin, C. and Wesslen, A. (1998). Understanding software defect detection in the personal software process. In *Proceedings Ninth International Symposium on Software Reliability Engineering (Cat. No.98TB100257)*, pages 49–58.
- Zannier, C., Melnik, G., and Maurer, F. (2006). On the success of empirical studies in the international conference on software engineering. In *Proceedings of the 28th International Conference on Software Engineering, ICSE '06*, pages 341–350, New York, NY, USA. ACM.