Heuristic Evaluation in Visualization: an empirical study

Beatriz Sousa Santos* Samuel Silva Paulo Dias
DETI/IEETA University of Aveiro, Portugal

ABSTRACT

Heuristic evaluation is a usability inspection method that has been adapted to evaluate visualization applications through the development of specific sets of heuristics. This paper presents an empirical study meant to assess the capacity of the method to anticipate the usability issues noticed by users when using a visualization application. The potential usability problems identified by 20 evaluators were compared with the issues found for the same application by 46 users through a usability test, as well as with the fixes recommended by the experimenters observing those users during the test. Results suggest that using some heuristics may have elicited potential problems that none of the users noticed while using the application; on the other hand, users encountered unpredicted usability issues.

Keywords: Visualization evaluation, heuristic evaluation, heuristics sets, empirical study.

Index Terms: H.5.2 [User Interfaces]; Evaluation/methodology; I.6.9 [Visualization]; Information visualization

1 INTRODUCTION

The importance of evaluation in Visualization is well recognized for long in the Visualization research community [32], and several methods have been used to evaluate visualization applications. However, to encourage interest by potential adopters, not only in research, but also in development contexts, the capabilities and limitations of evaluation methods need to be well understood and the methods should be easily applicable by typical developers of visualization applications.

Heuristic evaluation is a usability inspection method used by Human-Computer Interaction practitioners for years [31]. It helps identify potential problems and improve the application before giving it to users and should be, ideally, combined with usability testing. It is applicable throughout the whole development cycle and is considered as a discount method providing useful results even with a low investment. It has been adapted to evaluate visualization applications by developing visualization-specific sets of heuristics. Still, these may be neither easily understandable, nor straightforwardly applicable by practitioners who often are not proficient evaluators, and may foster finding too specific and low-priority usability problems [5]. Additionally, the objective value of heuristic evaluation has yet to be assessed with most of the heuristics proposed in the literature lacking a proper validation of their effectiveness [19].

In this context this paper presents a user study performed in the scope of a more general on-going work meant to better understand heuristic evaluation by gathering further insight on its application and outcomes, identifying potential improvements, and trying to contribute to its dissemination. This study was concerned with the capacity of the method to find issues that are experienced by users, and involved 20 evaluators who had previously analyzed a visualization web application using heuristic evaluation, and a usability test with 46 participants using the same application. The potential usability problems identified by the evaluators were compared with the issues found in the usability test.

The remainder of this paper is organized as follows: section 2 addresses related work and includes a list of research challenges and questions to further understand the relevance of heuristic evaluation in Visualization; section 3 presents the materials and methods used in the study, namely its design, the visualization application used, the potential problems found through heuristic evaluation, and the usability test; section 4 presents and discusses the main results obtained from the usability test and compares them with the potential problems identified by the evaluators. Finally, section 5 presents conclusions and avenues for future work.

2 BACKGROUND AND RELATED WORK

Carpendale [5] addresses the challenges in evaluating information visualizations and provides a thorough overview of quantitative and qualitative methods that can be applied in this context. Many different methods have been used in visualization evaluation: some directly adapted from other fields, others developed taking into consideration the particularities of visualization [8] [1] [2] [27] [23] [24].

2.1 Heuristic evaluation in Visualization

Heuristic evaluation is an analytical usability inspection method generally used in formative evaluation intended to advise designers on how to improve a system and answer the question "can I make it better?" [29]. It is a qualitative approach that uses a set of heuristics as a way of focusing attention on important aspects of the application. Evaluators use the heuristics to guide a structured analysis of the user interface aimed at pinpointing potential usability problems and assigning a corresponding severity grade. The main outcome is a list of categorized potential usability problems intended to support the development team in allocating resources to the most needed fixes. According to Carpendale [5], in Information Visualization it is important to consider precisely what aspects a given set of heuristics will tackle, and while usability heuristics [30] will apply also to visualization applications, they will not address some distinctive visualization aspects. The recognition of these distinctive aspects led several authors to propose sets of heuristics for Visualization, notable examples being the 13 heuristics by Zuk & Carpendale [44] and the 10 heuristics by Forsell & Johansson [10].

References to the relevance of using heuristic evaluation in Information Visualization evaluation (mostly by researchers, but also by practitioners) may be found in the work of several authors. Tory and Möller [41], while addressing the role of human factors in visualization, recommend the usage of methods adapted from user-centered design. Based on their experience and the experience of others, they claim that using heuristic evaluation is "a valuable way to evaluate visualization techniques" [42], e.g. in exploratory phases of research when clear objectives and variables might not yet be well defined and formal laboratory user studies might be inappropriate. Also, according to Isenberg et al. [22] and Freitas et al. [13], formal studies may have an unreasonable cost

at the design stage of the development cycle, and heuristic evaluation (as well as other discount evaluation methods) is more adequate. Munzner [27] [28] refers to heuristic evaluation as an adequate evaluation method at the visual encoding/interaction design level of the model the author proposes for the visualization design and validation process. At this level, "the threat is that the chosen design is not effective at communicating the desired abstraction to the person using the system" [27] and using heuristic evaluation is a way to systematically confirm that a set known guidelines is not disregarded by the design. Lam et al. [24] propose seven concrete evaluation scenarios most commonly encountered (exemplified with existing cases) to help visualization researchers and practitioners selecting appropriate evaluation approaches. They propose methods adequate to each scenario, and heuristic evaluation is mentioned in one of the scenarios (CDA: Collaborative Data Analysis); however, we believe it can be useful in other scenarios, as the new scenario identified by Isenberg et al. [23] (QRI: Qualitative Result Inspection) in their review of the most common evaluation practices reported in papers published at the IEEE Visualization conference along ten years. Besides using heuristic evaluation to fix usability problems during design phases, as recommended by these authors, Sedlmair et al. [35] also conducted various heuristic studies during the development process of all the information visualization tools they developed in the context of large companies and concluded that these studies did not replace usability studies with target users, but undeniably helped to spare "valuable experts' time".

2.2 Establishing the value of heuristic evaluation

While all previous authors refer to heuristic evaluation on a positive note, the real value of the method to successfully predict actual usability issues found by users in visualization applications is still not well known, since most heuristics lack a systematic validation [11]. Such an assessment is not only important for objectively securing the value of heuristics for the design and development, but it is also paramount to establish a common ground for the comparison among works [11] [19] and support building a more solid theoretical background for Visualization [6]. Yet, the validation of heuristics may become a daunting task.

The complex problem of assessing usability evaluation methods (not specifically in visualization) was addressed by Hartson et al. [16] who reported several measures used to assess usability evaluation methods: thoroughness (the proportion of real problems found on a user interface), validity, reliability, effectiveness, cost effectiveness and downstream utility. Thoroughness was used in most studies comparing usability evaluation methods reviewed by the authors, and though in theory it is not possible to know exactly all the usability problems of an application, it is reasonable to assume that if many users use it, a fair approximation will be achievable. In this vein, Hearst et al. [18] discuss the subject of evaluating the relation between the outcomes of heuristic evaluations and the problems felt by users in the context of Information Visualization and compare the outcomes of a heuristic evaluation with those gathered through a questionnaire.

2.3 Challenges and research agenda

The reviewed literature encourages a continued effort of the community to further understand the relevance and benefits of considering heuristic evaluation in Visualization. This, however, entails addressing a number of intertwined questions, some of them being:

1.How can we measure the advantages of using heuristic evaluation? Proof needs to be provided showing how the consideration of heuristic evaluation positively impacts the design and development of concrete interactive systems.

2. How can we perform a systematic validation of heuristics? Systematic approaches need to be devised to support the assessment of any set of heuristics to establish their reliability and usefulness. Many heuristic sets are proposed, but seldom validated [11].

3. How can we analyze the outcomes of heuristic evaluation in close relation with the real usability issues that will later affect users? How can we go beyond problem counting and actually pinpoint the usability issue felt by a user to the heuristic violation previously reported? This will enable, for instance, the identification of potential flaws of heuristic sets and foster their improvement and/or expansion [19].

4. Can we find a common ground for comparing results across peers and more easily profit from a joint effort? The repeated use of the same set of heuristics, for instance, along with a critical analysis of its use, would enable comparing evaluations among visualization techniques, for the same visualization and iteratively build and/or improve knowledge [11].

5. How can we decide which heuristics are adequate, in each case? In the presence of different sets of heuristics, how do we choose the most appropriate for our case? Can we mix heuristics from different sets?

6. What is the impact of adopting domain specific heuristics vs general heuristics? The specificities of visualization advise specific heuristics. But exactly how much do we gain? Are domain specific heuristics useful without domain specific experts? In many cases, domain specific heuristics are proposed, but serve no more than a few evaluations.

The analysis of these questions led us to the conclusion that efforts are primarily needed in advancing the methods and tools that enable answering questions 2 and 3, i.e., validation of heuristics and identification of their connection with problems felt by users. To this effect, in this work, we follow the rationale of Hearst et al. [18] combined with the suggestions of Hornbaek [20] to extend the comparison of usability evaluation methods beyond problem counting, identifying aspects of design that can be improved and enabling evaluators to suggest solutions and comment on the evaluation methods. As a consequence, we collected usability issues and recommended fixes through a usability test of a specific visualization application, instead of using a questionnaire, as we consider the usability test involving users and observers as a more effective method to collect real usability issues affecting a user interface. Besides, we had asked evaluators to rate the understandability of the set of heuristics they used in their evaluation of the application, as described in [39].

3 The Study: Materials and Methods

While the overarching goal of this on-going research, including this study as well as previous work [39], is to better understand the value of heuristic evaluation as a method to evaluate the usability of visualization applications, the study reported in this paper had the specific objective to assess if the main potential usability problems previously found by 20 evaluators analyzing a specific application through heuristic evaluation are in fact experienced by users. Whereas in the previous study [39] the potential usability problems found using heuristic evaluation were compared with the results of a questionnaire concerning the application, the study presented in this paper compares those potential problems with the results of a usability test organized to understand what issues users come across while using the

application. This was done as a triangulation of methods and since the usability test was expected to provide a better proxy to the "ground truth" than the results of the questionnaire. This section presents the hypotheses initially considered, the experimental design of the study, the main potential usability issues previously found using heuristic evaluation, the usability test and its participants, as well as the collected data.

3.1 Hypotheses

At the onset and based on previous exploratory work we formed the following hypotheses to be examined in this study:

H1- All main potential usability problems found previously by the evaluators should be experienced by the users of a usability test in a more or less extent.

H2- The usability test should reveal usability issues not previously found by the evaluators.

3.2 Experimental design

Figure 1 shows the organization of the study: on one hand, 20 evaluators evaluated an application using heuristic evaluation; as a result a list of the twelve most often reported potential usability issues was produced; on the other hand, a usability test concerning the application was performed by 46 participants. The issues found during the test were compared with the potential usability problems reported by the evaluators.

The application used in this study allows the visual exploration of data concerning the football World Championships since 1930 (http://spotfire.tibco.com/demos/spotfire-soccer-2014); it was selected as this dataset was expected to be easy to understand by most participants in the study (evaluators and users).

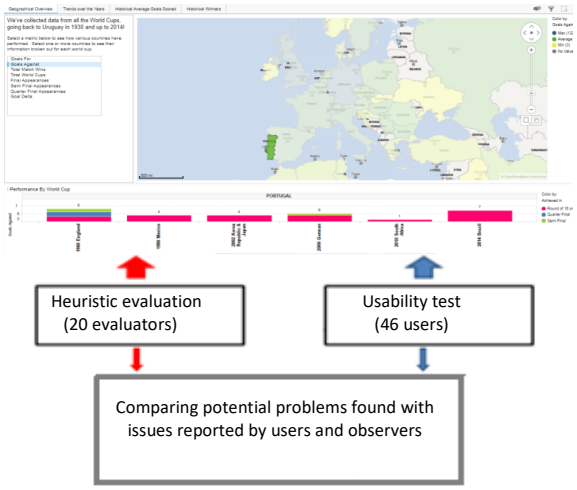

Figure 1: Organization of the study: the potential usability problems of an application found through heuristic evaluation were compared with the issues reported in a usability test

3.3 Heuristic Evaluation

The number of evaluators needed to find a significant percentage of the potential usability problems of an application may vary with several factors, such as the nature or complexity of the application, and the previous experience of the evaluators [17]. Taking into consideration these findings, we involved 20 evaluators in the study. These were students of an MSc program in Informatics Engineering attending an Information Visualization course who had previous experience in using heuristic evaluation to evaluate visualization applications obtained in a mandatory evaluation assignment. This assignment involved using Nielsen's heuristics and the Visualization specific sets by Zuk & Carpendale and Forsell & Johanson, and included presenting as well as discussing their conclusions in class. In the scope of this study the students evaluated the application as an exercise during a 2h laboratory session at the end of the semester. They were informed that their evaluations would contribute to a study about heuristic evaluation in Visualization and were asked to identify at least six potential problems using the heuristics sets they had previously used. More detail about this session can be found in [39]. All evaluators used the Nielsen's heuristics, 10 evaluators also used the Zuk & Carpendale´s and 5 others used the Forsell & Johanson's heuristics sets (see the Appendix for a summary of these heuristics sets). A total of 140 non-consolidated potential problems were found. The descriptions were analyzed to classify the problems into different categories and recognize repetitions. Most of these potential problems were identified and classified using the Nielsen's heuristics; 34 were classified using Zuk & Carpendale's heuristics, and 21 were classified using Forsell & Johanson's heuristics. Of the second set of heuristics, the most used heuristics were Z5 (Consider people with color blindness) and Z1 (Ensure visual variable has sufficient length) to classify 5 problems respectively. Z6 (Pre-attentive benefits increase with field of view) and Z7 (Quantitative assessment requires position or size variation) were not used. In a preliminary study Z6 was considered by the evaluators the least understandable heuristic of this set and thus this result was not a surprise. From the latter set, the most used heuristics were F4 (Orientation and help) and F9 (Remove the extraneous) to classify 4 potential problems each. F10 (Data set reduction) was not used by any of the evaluators, possibly because it was not properly understood as several evaluators mentioned filter-related potential problems and classified them using other heuristics.

This analysis allowed identifying the following potential problems more often found by the evaluators (number of occurrences inside parentheses):

- The application may be too slow (12)
- Colors may be difficult to discriminate (12)
- Meaning of colors may be difficult to understand (12)
- Zoom and Scroll may be difficult to use (9)
- Some application aspects may have too much information (9)
- Help may be insufficient (5)
- The lack of Undo may have a negative impact (6)
- Filters may have be difficult to use (5)
- Multiple country selection may be difficult to find (4)
- Documentation may be insufficient (4)
- Small items may be difficult to select (3)
- Charts under the map may have low visibility (2)

After concluding their evaluation, the students were asked to give their opinion about the overall usability of the application in a 5-level Likert-like scale (1- very poor to 5 - very good). The median value was 3 (reasonably usable).

3.4 Usability test

As the main goal of this study was to assess if the potential usability problems identified though the heuristic evaluation translated into actual issues experienced by users while using the application, we designed a set of tasks involving the use of potentially problematic features of the application. By accounting for the potential problems when designing the tasks, we aimed at: (1) ensuring that users were faced with the potential problems during the test; and (2) easing the *a posteriori* match between the

outcomes of both evaluation methods, a potentially challenging task [21].

The test included six tasks implying performing data analysis with the application.

The tasks considered for the usability test consisted in finding:

- how many goals has Portugal scored in all championships
- how many goals has Portugal scored in semi-finals
- in how many championships Portugal score as much as Spain
- where and when did that happen
- how many games did Germany win in the 1954 championship
- how many games did Germany win in the "Round of 16", at that same championship.

Immediately after each task the users were asked about the difficulty they felt in performing it (in a 5 level Likert-like scale) and could add any comments if desired. The task time was also logged.

After performing all the tasks the users answered a set of questions concerning personal data and platform characteristics (e.g., gender, age, screen size, input device), as well as a final set of questions concerning their opinion on the application usability, including a SUS (System Usability Scale- an industry standard usability satisfaction questionnaire [4][25]), an open-ended question about the main shortcomings of the application, and the same a question about the overall usability of the application that had previously been asked to the students who assessed the usability of the application through heuristic evaluation (see section 3.1).

Finally, each observer made a report in collaboration with the user they had been observing. This report had a normalized format and included identifying the main user difficulties observed, any comments considered relevant and the main fixes recommend by the user and/or the observer, all in free-form text.

The test was performed as an exercise during 2h laboratory sessions of a Human-Computer Interaction course to Computer and Informatics Engineering students. In each class half of the students acted as users and the other half observed them. The observers were asked to explore the application and received instructions on how to behave as well as all the documentation they needed during the test (e.g., an informed consent for the users to sign, a script with all the steps to be performed during the test, and a form to register observations). All students were asked to be thorough in their analysis of the application.

The test was performed during six laboratory sessions organized by the authors involving 92 students: 46 (8 female) acting as users aged 19 to 24, and 46 acting as observers; these roles were assigned randomly to avoid bias. The first author was present in all the sessions to clarify any doubts and make sure the experimental conditions were maintained among sessions.

4 Results and Discussion

The usability test most significant results were obtained from the users' and observers' free-text comments (which allowed identifying the main issues encountered by the users), and the application fixes recommended in the final report. In this paper we focus on the analysis of these data.

4.1 Issues and fixes obtained from the usability test

All users made comments. Analyzing these comments and trying to match the issues mentioned with the list of potential problems previously identified it was possible to find the issues mentioned in column "Com." of Table 1.

No specific comments concerning color were made; only a very general comment mentioning that several aspects of the user interface needed to be improved was made by one user. Moreover, no mentions to the lack of undo, neither to poor tooltips were found in these comments suggesting that none of the users felt these issues while using the application, or, at least, did not consider them as having a significant negative impact on the application usability.

The following issues (more high level) not included in the list of main potential issues identified by evaluators were found in the participants' comments:

- Difficult navigation in the map (4)
- Poor organization (3)
- Difficult to find information (2)
- Difficult to find some options (2)
- Not intuitive (2)
- Poor UI design (1)
- Knowledge of countries geographical location needed (1)

The median value of the answers to the SUS questionnaire was 41(<68), clearly considered below average. This is consistent with the median value 2 obtained for the opinion on the overall usability of the application.

Analyzing the 57 application fixes recommended by the users and observers, in their final report, it was possible to find recommendations to improve several aspects of the application. These (e.g., "improve responsiveness") were assigned to the corresponding issue, and the number of occurrences for each is shown in Table 1.

4.2 Matching issues and fixes with potential problems

Table 1 allows comparing the fixes recommended with the issues mentioned in the users' comments and the potential usability problems found by the evaluators. We notice that the more often recommended fix matches the main potential usability problem found by the evaluators and the most mentioned issue in the users' comments. This suggests that it is a serious usability problem. The next most recommended fixes are related with filters and map usability. Filter issues were mentioned by users in their comments and some evaluators also pointed them as potential usability problems. Map usability recommendations encompass improving interaction, navigation, scrolling and zooming, as well as selection of small countries, aspects that have been considered as potential problems by evaluators, while in different problem types, and were also mentioned in the users' comments. The most unexpected result was the complete lack of comments and low number of recommended fixes related to specific color issues, as "colors difficult to discriminate" and "unclear meaning of the colors" were, along with low responsiveness of the site, the most prominent potential problems identified by the evaluators (12 each). This may be explained by the fact that Zuk & Carpendale's set has 4 out of the 13 heuristics devoted to color issues (Z2- Don't expect reading order from color, Z3- Color perception varies with size of colored items, Z4- Local contrast affects color and gray perception, Z5- Consider people with color blindness), which may elicit a detailed analysis of the color usage in the application, resulting in a high proportion of potential usability problems identified by the evaluators that have used this set of heuristics. In fact, analyzing the description of these problems, we noticed that only 4 were classified not using Zuk & Carpendale's heuristics. This result is an indication of why color is, often, such a prominent source for issues in visualization [37], hinting on the users' lack of awareness regarding the problems created and/or potentiated by the wrong use of color and, naturally, emphasizing that color related issues may eventually only be detected, during usability tests, if the proposed tasks strongly rely in color analysis. In this regard, the consideration of heuristics, particularly those regarding color, can help identify color related issues at an early

stage, as a hidden issue. Additionally, while designing a usability test, we argue that one can profit from a set of validated heuristics to tailor tasks around specific potential issues, thus probably improving the overall value of usability tests. This further highlights the importance of the visualization guidelines, from which many heuristics have stemmed [6], bringing them into usability testing design and, also, to reporting, as they can provide a more structured categorization of the issues found [11].

Table 1: Number of occurrences of issues related to different aspects of the application mentioned by evaluators (HE), in users' comments (Com.), and recommended fixes (Fixes) (more often recommended fixes first) (*- mentioned together in HE)

Aspects of the application	HE	Com.	Fixes
high response time	12	14	10
filters difficult to use	5	9	9
difficult scrolling in the map	9*	11	7
lack of documentation	4	5	5
difficult zooming in the map	9*	6	4
difficult selection of small items	3	3	3
lack of help	5	1	2
too much information in some…	9	5	2
unclear meaning of the colors	12	-	1
unclear selection of multiple items	4	-	-
charts with low visibility	2	1	-
lack of undo	6	-	-
colors difficult to discriminate	12	-	-

Another result worth noting is the absence of comments and recommended fixes concerning the lack of undo. Apparently, during the usability test, no one considered that the application should have this functionality. In this case, the heuristics forced the identification of an issue that was not relevant for the evaluated domain, at least in the extent explored by the considered tasks, although we consider it is due to the nature of the application.

Overall, considering the hypotheses explored in this study, these results suggest that H1 is not confirmed (H1- All main potential usability problems found previously by the evaluators should be experienced by the users of a usability test to a greater or lesser extent) as was the case for the color-related potential problems reported by many evaluators, but not mentioned by the users. On the other hand, H2 is confirmed (H2- The usability test should reveal usability issues not previously found by the evaluators), as several higher-level issues not mentioned as potential usability problems, appeared in the comments and in the recommended fixes.

In view of these results it seems that using heuristic evaluation with very specific heuristics leads the evaluators to make a detailed analysis of the application concerning the aspects addressed by those heuristics, perhaps distracting them from other characteristics of the application that are also relevant in its usability (or lack of); this effect may be related to what Cockton and Woolrych [7] observed when they compared the potential problems found by heuristic evaluation with actual problems revealed by user testing of an office application, and in line with the idea that heuristic evaluation may foster finding too specific and low-priority usability problems [5]. This is also in line with the observations of Hermawati and colleagues [19], in their survey, noting that some usability issues are missed by domain heuristics.

Nevertheless, concerning the potential problems pointed out in a heuristic evaluation and not in a usability test, it is difficult to demonstrate that they are not issues at all, as they might be encountered if the test had involved more users or different, more complex tasks. It is probably better to characterize them as less frequent issues.

4.3 Limitations of the study

This study had several limitations concerning the methods and the participants. It involved a relatively low number of participants with a narrow profile. Nevertheless, while the number of participants is a limitation, and it would be advantageous to have a greater number, the profile of the users and observers who participated in the usability test may have had a positive impact in the results given their experience in assessing usability, which probably helped them to criticize and recommend fixes to the application.

Other possible limitations of the study with impact on the results are a consequence of having used only one application and of the way we collected and analyzed the comments and the recommended fixes in the usability test. The free-text format used in order to avoid any bias and collect, as much as possible, the spontaneous opinion of the users and observers made it very complex to match the issues mentioned with the potential usability problems found through heuristic evaluation; nevertheless, while we might have missed some possible matches, we deem that due to the careful content analysis of all the potential problem descriptions, comments, and recommended fixes, we have been able to capture the most relevant existing associations (as the case of low responsiveness and difficult scrolling in the map), as well as the unexpected nonappearances (as the case of color related and undo issues).

Yet another possible limitation of this study is the relatively low experience of the evaluators. Whereas they all had previously attended an explanation about how to use the method, and succeeded in a graded assignment involving using more than one set of heuristics to assess the usability of a visualization application, some of them might have an incomplete understanding of the meaning of some heuristics. In fact, some of the heuristics were considered hard to understand by these students (e.g., heuristic Z6- Pre-attentive benefits increase with field of view), which may be a real problem hampering the usefulness of heuristic evaluation as a method to improve applications during development. However, we argue that many developers of visualization applications, at least in their first years of profession, most likely will not have much more experience than our students, and thus the potential problems we collected may be fairly representative of a real situation.

5 Conclusion and Future Work

While heuristic evaluation cannot be used *per se* to evaluate a visualization application (as it is generally also the case with any other evaluation method), it has several advantages: it can be used in different development stages, is generally considered to provide useful results with a modest investment, and has been adapted to

be used in Information Visualization evaluation, namely by using visualization specific sets of heuristics. However, it has several potential shortcomings: it has been noted that it may elicit too specific and low-priority potential problems and the heurists sets are not easy to select and apply to a specific case, which may affect the usefulness of the method. The capacity of heuristic evaluation to help predict the issues that will affect users is not well established, and assessing it appears to be a challenging task. Still, we argue that, empirically, it will be possible to increase our understanding of the method, namely the above-mentioned capacity, and have started to work in this direction. The results presented in this paper are a first indication that it is possible to obtain insight using the proposed comparison method, and suggest that from the two hypotheses (H1 and H2) defined at the beginning of the study, one is confirmed (H2-the usability test should reveal usability issues not previously found by the evaluators), whereas the other is not (H1-all main potential usability problems identified previously by the evaluators should be experienced by the users of a usability test in a more or less extent). The confirmation of H2 seems to corroborate the need to use empirical methods in the evaluation, and the non-confirmation of H1is in line with the recognition that heuristic evaluation may elicit too specific potential problems. Moreover, a significant part of the issues found by users and observers were of a higher-level nature. These results also suggest that such a comparison method could be used in a systematic approach to establish the reliability and usefulness of heuristics sets, thus contributing to our questions Q2-"How can we perform a systematic validation of heuristics?", and Q3-"How can we analyze the outcomes of heuristic evaluation in close relation with the real usability issues that will later affect users".

After performing this study, we still believe heuristic evaluation should be part of the "evaluation toolkit" of any InfoVis developer; yet, the currently available visualization-specific heuristics sets may not be easy to select, understand and apply, which may preclude a more widespread usage of the method. Moreover, some of the heuristics may elicit too specific potential issues that are not encountered by many users, which may decrease the efficacy of the method. Thus, comparing the "thoroughness" of heuristic evaluation with different sets of heuristics is a relevant work that may provide interesting results and contribute to answering our above mentioned questions, Q2 and Q3.

On the other hand, we noticed that a positive effect of using the method is making people more aware of existing principles and guidelines. A relevant example can be drawn from how color-related heuristics can elicit awareness to the specificities of using color, an issue that can be overlooked in usability tests, by task design or due to user characteristics and knowledge. Thus, heuristic evaluation is a valuable method to help prepare developers to create InfoVis applications; we have been using it in our Information Visualization courses to foster students' critical thinking and design skills as it is described in [38].

Further work is needed to better comprehend how understandable and applicable are visualization-specific heuristics. A first step towards improving comprehensibility would possibly be to rephrase some of the heuristics making them more specific and explanatory, perhaps expanding them into guidelines and illustrating their application through examples. Another possible direction to facilitate the dissemination and applicability might be to extend the patterns proposed by Elmqvist and Yi [9] as to explicitly include heuristic evaluation. We are already working on studying the applicability of the heuristics sets and the performance of domain specific heuristics as compared to general ones, as well as how to guide evaluators through their selection in a specific case. We agree with Tarrell et al. [40] that a community effort is necessary to develop more understandable visualization-specific heuristics sets, as well as to ascertain their applicability and validity. We also agree that "a more-accepted and more-useful set of visualization-specific heuristics and guidelines … could function essentially as a 'checklist' for designers and evaluators alike. It would also provide impetus for renewed programmatic support in visualization design activities". In this context, considering the potential specificity of the heuristics, we argue that an important route to explore may be the proposal of heuristic sets that are customized based on specific aims of the envisaged application, mixing individual heuristics from different sets, taking out those that might not apply. This work can start, for instance, from the work on taxonomies and decomposition of visualization tasks (e.g., [3] [27]), the characteristics of the data/information involved [36], or existing proposals for visual taxonomies (e.g., [26]), moving from there to suggest heuristics that apply to the specificities of each case. A notable example following this path is the recent work by Alonso-Rios et al. [34].

Finally, we believe that how effective are heuristic evaluations in Visualization is a matter that needs more consideration by researchers and further work is needed, as, for instance, the validation of domain-specific heuristics sets [19]. The two questions we addressed in this study (Q2 and Q3) are concerned with the creation of the methods and tools that should constitute the basis of our research; yet, we are strongly focused on evolving our proposals into a framework that more clearly encompasses the different aspects raised by all the questions we defined and, as long-term goals, to establish more clearly the value of heuristic evaluation in Visualization, and produce guidelines that may assist designers and developers in using the method and selecting the heuristics.

APPENDIX

The following three sets of heuristics were used by the evaluators to identify potential usability problems in the application:

Nielsen's heuristics- Ten Usability Heuristics for User Interface Design [30]:

N1- Visibility of system status;
N2- Match between system and the real world;
N3- User control and freedom;
N4- Consistency and standards;
N5- Error prevention;
N6- Recognition rather than recall;
N7- Flexibility and efficiency of use;
N8- Aesthetic and minimalist design;
N9- Help users recognize, diagnose, and recover from errors;
N10- Help and documentation.

Forsell & Johanson's heuristics- A Heuristic Set for Evaluation in Information Visualization [10]:

F1- Information coding;
F2- Minimal actions;
F3- Flexibility;
F4- Orientation and help;
F5- Spatial organization;
F6- Consistency;
F7- Recognition rather than recall;
F8- Prompting;
F9- Remove the extraneous;
F10- Data set reduction.

Zuk & Carpendale's heuristics- Heuristics for Information Visualization Evaluation [44]:

Z1- Ensure visual variable has sufficient length;
Z2- Don't expect a reading order from color;
Z3- Color perception varies with size of colored item;
Z4- Local contrast affects color & gray perception;
Z5- Consider people with color blindness;
Z6- Preattentive benefits increase with field of view;
Z7- Quantitative assessment requires position or size variation;
Z8- Preserve data to graphic dimensionality;
Z9- Put the most data in the least space;
Z10- Remove the extraneous (ink);
Z11- Consider Gestalt Laws;
Z12- Provide multiple levels of detail;
Z13- Integrate text wherever relevant.

ACKNOWLEDGMENTS

Authors would like to appreciate the anonymous reviewers for their constructive suggestions, and are grateful to the students who contributed to this study as evaluators, users or observers. This work was partially funded by FCT – Foundation for Science and Technology, in the context of the project UID/CEC/ 00127/2013. Samuel Silva's work is funded by Portugal 2020 under the Competitiveness and Internationalization Operational Program, and by the European Regional Development Fund through project SOCA -- Smart Open Campus (CENTRO-01-0145-FEDER-000010).

REFERENCES

[1] K. Andrews, "Evaluating Information Visualisations," in *Proceedings of the Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization BELIV'06*, 2006.

[2] K. Andrews, "Evaluation Comes in Many Guises," in *Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization BELIV'08*, 2008.

[3] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Trans. Vis. Comput. Graph.,* vol. 19, no.12, pp. 2376-85, 2013.

[4] J. Brooke, "SUS: a retrospective," *Journal of Usability Studies, vol.* 8, n.2, pp. 29-40, 2013.

[5] S. Carpendale, "Evaluating Information Visualizations," in *Information Visualization , Human-centered issues and perspectives*, A. et al. Karren, Ed. Springer, 2008, pp. 19–45.

[6] M. Chen, G. Grinstein, C. Johnson, J. Kennedy, and M. Tory, "Pathways for theoretical advances in visualization", *IEEE Comput. Graph. Appl.*, vol. 37, no. 4, pp. 103-112, 2017.

[7] G. Cockton and A. Woolrych, "Understanding Inspection Methods: Lessons from an Ass essment of Heuristic Evaluation," in *People and Computers XV*, A. Blandford and J. Vanderdonckt, Eds. Springer, 2001, pp. 171–192.

[8] G. Ellis and A. Dix, "An Explorative Analysis of User Evaluation Studies in Information Visualisation," in *Proc. of the Workshop on BEyond time and errors: novel evaluation methods for information visualization BELIV'06*, 2006.

[9] N. Elmqvist and J. S. Yi, "Patterns for visualization evaluation," in *Proceedings of the Workshop on Beyond Time and Errors - Novel Evaluation Methods for Visualization BELIV '12*, 2012.

[10] C. Forsell and J. Johanson, "An heuristic set for evaluation in information visualization," in *Proc. International Conference on Advanced Visual Interfaces AVI2010*, 2010, pp. 199–206.

[11] C. Forsell, "Evaluation in Information Visualization: Heuristic Evaluation," in *Proc. 16th Int. Conf., Inf. Vis. (IV),* 2012, pp. 136–142.

[12] C. Freitas, P. R. G. Luzzardi, R. A. Cava, M. A. A. Winckler, M. S. Pimenta, and L. P. Nedel, "Evaluating Usability of Information Visualization Techniques," in *Proc. Symposium on Human Factors in Computer Systems, IHC 2002*, 2002.

[13] C. Freitas, M. S. Pimenta, and D. Scapin, "User-Centered Evaluation of Information Visualization Techniques : Issues and Perspectives," in *Anais do Colóquio em Informática: Brasil / INRIA, Cooperações, Avanços e Desafios*, 2009, pp. 2603–2606.

[14] C. Freitas, M. S. Pimenta, D. L. Scapin, C. Maria, D. Sasso, and M. S. Pimenta, "User-Centered Evaluation of Information Visualization Techniques: Making the HCI- InfoVis Connection Explicit," in *Handbook of Human Centric Visualization*, W. Huang, Ed. Spinger, 2014, pp. 315–336.

[15] L. Hasan, A. Morris, and S. Probets, "A comparison of usability evaluation methods for evaluating e-commerce websites," *Behav. Inf. Technol.*, vol. 31, no. 7, pp. 707–737, 2012.

[16] H. R. Hartson, T. S. T. Andre, and R. R. C. Williges, "Criteria for evaluating usability evaluation methods," *Int. J. Hum. Comput. Interact.*, vol. 13, no. 4, pp. 373–410, 2001.

[17] W. Hawang and G. Salvendy, "Number of People Required for Usability Evaluation: The 10±2 Rule.," *Commun. ACM*, vol. 53, no. 5, pp. 130–133, 2010.

[18] M. A. Hearst, P. Laskowski, and L. Silva, "Evaluating Information Visualization via the Interplay of Heuristic Evaluation and Question-Based Scoring," in *Proc. 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 5028–5033.

[19] S. Hermawati and G. Lawson, "Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus?," *Appl. Ergon.*, vol. 56, pp. 34–51, 2016.

[20] K. Hornbæk, "Dogmas in the assessment of usability evaluation methods," *Behav. Inf. Technol*., vol. 29, no. 1, pp. 97–111, 2010.

[21] K. Hornbæk and E. Frøkjær, "Comparison of techniques for matching of usability problem descriptions," *Interact. Comput.*, vol. 20, no. 6, pp. 505–514, 2008.

[22] P. Isenberg, T. Zuk, C. Collins, and S. Carpendale, "Grounded Evaluation of Information Visualizations," in *Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization BELIV'08,* 2008.

[23] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Moller, "A systematic review on the Practice of Evaluating Visualization," *IEEE Trans. Vis. Comp. Graph.*, vol. 19, no. 12, pp. 2818–2827, 2013.

[24] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical Studies in Information Visualization: Seven Scenarios.," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 9, pp. 1520–1536, 2012.

[25] J. R. Lewis, "The System Usability Scale: Past, Present, and Future," *Int. J. Hum. Comput. Interact*, vol. 34, no. 7, pp. 1–14, 2018.

[26] E. Morse, M. Lewis, and K. Olsen, "Evaluating visualizations: using a taxonomic guide," *International Journal of Human-Computer Studies,* vol. 53, no. 5, pp. 637-662, 2000.

[27] T. Munzner, "A Nested Process Model for Visualization Design and Validation," *IEEE Trans. Vis. Comp. Graph.*, vol. 15, no. 6, pp. 921–928, 2009.

[28] T. Munzner, *Visualization Analysis and Design*. AK Peters, 2014.

[29] J. Nielsen and R. Molich, "Heuristic Evaluation of user interfaces," in *CHI '90 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1990, pp. 249–256.

[30] J. Nielsen*, Usability Engineering*, Morgan Kaufmann, 1993.

[31] J. Nielsen, "Discount usability: 20 years". Retrieved June 2018, from: http://www.nngroup. com/articles/discount-usability-20-years/.

[32] C. Plaisant, "The Challenge of Information Visualization Evaluation," in *Proc.working conference on Advanced Visual Interfaces AVI2004*, 2004, pp. 109–116.

[33] D. Quiñones and C. Rusu, "How to develop usability heuristics: A systematic literature review," *Comput. Stand. Interfaces*, vol. 53, March, pp. 89–122, 2017.

[34] D. Alonso-Ríos, E. Mosqueira-Rey, and V. Moret-Bonillo, "A Systematic and Generalizable Approach to the Heuristic Evaluation of User Interfaces," *Int. Journal Human. Comput. Interact.,* online 2018.

[35] M. Sedlmair, P. Isenberg, D. Baur, and A. Butz, "Evaluating Information Visualization in Large Companies: Challenges, Experiences and Recommendations," in *Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization BELIV'2010*.

[36] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of the 1996 IEEE Symposium on Visual Languages*, 1996, pp. 336-343.

[37] S. Silva, B. Sousa Santos, and J. Madeira, "Using color in visualization: A survey," *Computers & Graphics*, vol. 35, no.2 pp. 320-333, 2011.

[38] B. Sousa Santos, B. Q. Ferreira, and P. Dias, "Using Heuristic Evaluation to Foster Visualization Analysis and Design Skills," *IEEE Computer Graphics and Applications*, vol. 36, no. 1, pp. 86–90, 2016.

[39] B. Sousa Santos, S. Silva, B. Quintino Ferreira, and P. Dias, "An exploratory study on the predictive capacity of heuristic evaluation in visualization applications," in *Lecture Notes in Computer Science,* vol 10271, 2017, pp. 369–383.

[40] A. Tarrell, A. Fruhling, and R. Borgo, "Toward Visualization-Specific Heuristic Evaluation," in *Proceedings of the Fifth Workshop on Beyond Time and Errors novel evaLuation methods for Visualization BELIV'14*, 2014, pp. 110–125.

[41] M. Tory and T. Möller, "Human Factors In Visualization Research," *IEEE Trans. Vis. Comput. Graph.*, vol. 10, no. 1, pp. 1–13, 2004.

[42] M. Tory and T. Möller, "Evaluating visualizations: do expert reviews work?," *IEEE Computer Graphics and Applications*, vol. 25, no. 5, pp. 8–11, 2005.

[43] T. Zuk and S. Carpendale, "Theoretical analysis of uncertainty visualizations," in *Visualization and Data Analysis, SPIE*, vol. 6060, 2006, pp. 606007–14.

[44] T. Zuk, L. Schlesier, P. Neumann, M. S. Hancock, and S. Carpendale, "Heuristics for Information Visualization Evaluation," in *First Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization BELIV'06*, 2006, pp. 1–6.