Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

On the evaluation of usability design guidelines for improving network monitoring tools interfaces[☆]

Sofia A.M. Silveira [a], Luciana A.M. Zaina [a], Leobino N. Sampaio [b], Fábio L. Verdi [a,*]

[a] Computer Science Department, Federal University of São Carlos, Sorocaba, SP, Brazil
[b] Computer Science Department, Federal University of Bahia, Salvador, BA, Brazil

ARTICLE INFO

Article history:
Received 30 March 2021
Received in revised form 16 December 2021
Accepted 11 January 2022
Available online 19 January 2022

Keywords:
Usability
Network-monitoring tools
Network management
Design guidelines
User-centered design
Graphical user interface

ABSTRACT

Network monitoring tools are vital to network administrators, helping them make decisions and accomplish their tasks. In general, those tools are developed with a focus on technical aspects not taking into account important usability principles. On the other hand, the Human–Computer Interaction community presents great potential for the improvement of interfaces in network management tools suggesting that usability guidelines can guide software developers during user interface design. The goal of this work is to evaluate how different usability design guidelines can assist software developers in elaborating network monitoring tools interfaces with improved usability, creating a better experience for network administrators. To do that, we engage in an experimental study, where 52 software developers prototyped user interfaces based on different scenarios and applied 12 guidelines for usability design in network monitoring tools. Through the quantitative and qualitative analysis as well as Fisher's Exact Test, we demonstrate that the level of complexity of the scenarios for the creation of the prototypes had no significant effect on the acceptance of the guidelines. We conclude that the guidelines were used by most participants and are relevant to assist the software developers to create interfaces with a focus on usability in network monitoring tools.

1. Introduction

Mostly designed by and for network specialists, current network management applications' interfaces still tend to lack usability aspects in their design and to present data manipulation resources in a purely technical perspective (Falschlunger et al., 2016). Even the design of advanced tools overlooks the existing diversification of users' profiles interested in network information, encouraging the development of friendly system interfaces to facilitate access, manipulation, and understanding of network data (Nielsen, 2012).

The visualization and usability resources, discussed by the Human–Computer Interaction (HCI) community, hold considerable promise for contributing to more advanced user interfaces of management applications (Nielsen, 2012; Falschlunger et al., 2016; Ward et al., 2010.; Bajpai and Schönwälder, 2015; Guimarães et al., 2016). They can benefit non-specialists and those with in-depth knowledge looking for increasing productivity when carrying out daily monitoring tasks (Guimarães et al., 2016). As a result, recent initiatives have suggested HCI-based usability guidelines (i.e., heuristics) to support the designing of user interfaces for network management tools (Verdi et al., 2020).

User interface design refers to the building interfaces that mediated the user interaction with the software (Sharp et al., 2019). The decisions regarding the visual elements (e.g. buttons, menus) used on the interface as well as the interaction type available (e.g. touch, voice) can impact on the software usability. Usability heuristics and guidelines are commonly used in the user interface design process to provide best practices and rules of thumb to developers produce more friendly and useful software (Nielsen, 1994). Taking into account the best practices, developers construct prototypes of different fidelities which are the deliverables in the process of user interface design. An effort to create new heuristics is evident in the literature since Nielsen (Nielsen, 1994) Heuristics and guidelines do not cover specific characteristics of different types of software and applications, ignoring elements of particular domains (e.g., network, education). Therefore, in the field of network monitoring, it is of paramount importance to define such specific heuristics and guidelines and make evaluations to verify their benefits for usability.

Despite their benefits, it is still fundamental to evaluate to which extent such guidelines can attract software developers on user interface designing when their goal is to help users

[☆] Editor: A. Bertolino.
[*] Corresponding author.
E-mail addresses: sofia.silveira@dcomp.sor.ufscar.br (S.A.M. Silveira), lzaina@ufscar.br (L.A.M. Zaina), leobino@ufba.br (L.N. Sampaio), verdi@ufscar.br (F.L. Verdi).

https://doi.org/10.1016/j.jss.2022.111223
0164-1212/© 2022 Elsevier Inc. All rights reserved.

perform their tasks with greater ease and efficiency. This concern stems from the lack of guidelines' effectiveness on existing user interface designs. Most of the current literature initiatives describe evaluation experiences, which findings are limited for some reasons. Firstly, the studies do not involve real users or rely on few opinions collected in restricted scopes. Secondly, the evaluations do not consider those directly involved with the design (i.e., software developers). Thirdly, the studies do not exploit users' feedback to improve the evaluated guidelines. Finally, there are quite a few initiatives proposed for network management. Most of them are focused on other knowledge fields, e.g., network security or generic heuristics.

Hence, based on the discussion mentioned above, this work aims to evaluate *how different design guidelines can effectively assist software developers in building user interfaces that allow network administrators to perform their tasks with greater ease and efficiency*. To accomplish this goal, we conducted an evaluation study based on our previous work (Verdi et al., 2020). It introduced 12 guidelines for user interface design in network management tools derived from an investigation involving nine network specialists in the area. This follow-up study evaluated them guided by two related Research Questions (RQ) as follows:

- **RQ1: How did the software developers use the guidelines?**
- **RQ2: What was the feedback provided by the software developers?**

To answer them, we conducted an experimental study based on medium-fidelity prototypes (Sharp et al., 2019) that involved software developers (i.e., participants), following a practical approach. Its main goal was to evaluate, control, or improve the management process supported by an application to instruct the participants to use the assessed guidelines. To this end, we created a catalog that extends those proposed in Verdi et al. (2020) by adding new data to support developers in understanding how to apply them. Indeed, we included a corresponding data type affected by each guidelines, issues to be considered about the interaction, additional notes, and the limitations and restrictions to its application.

Besides answering the main research question, this work presents other contributions, as follows: (i) It presents an evaluation study focused on developing network monitoring tools, an area rarely addressed in the literature. So, this experience can guide future initiatives; (ii) The evaluation relied on a study with software developers, supported by experimental methodologies; (iii) Different from related works, we refined the guidelines according to the analysis of the prototypes implemented by the developers. Consequently, we could observe patterns of error in applying the guidelines and modify them as necessary; (iv) We introduced a usability evaluation methodology that used a catalog to support developers during the analysis.

The remaining sections of this paper are organized as follows. Section 2 summarizes the fundamental concepts related to this work. Section 3 describes the study, presenting its steps involving planning, conduction and analysis. Section 4 presents the results of the study, focusing on each Research Question. Section 5 explains which and how guidelines were refined and the motivation to the refinement. Section 6 discusses the study findings regarding each Research Question. Section 7 addresses threats to validity and how we mitigated each of them. Finally, Section 8 presents our final considerations.

2. Related work

The use of graphical visualizations to analyze problems has become essential, especially to analyze and understand a large amount of data such as that collected in the network environment, coming from routers, switches and servers (Keim and Zhang, 2011; Falschlunger et al., 2016). Furthermore, the advent of virtualization, where the number of elements in the network becomes even more significant, promoted an increase in the amount of information (Ogu et al., 2014; Jain and Paul, 2013). Recent reports in the literature point to the need for further investigation into user interaction problems presented by the monitoring tools (Guimarães et al., 2016).

When evaluating user interfaces of monitoring tools, Pretorius, Calitz, and Van Greunen (Pretorius et al., 2005) use the Eye Tracking technique combined with traditional interface evaluation methodologies. The use of Eye Tracking revealed that a specific important region of a data visualization of the monitoring tool was too small, making legibility difficult with impact negatively on usability. It also showed that users always prefer blue to view graphics over texts.

Studies about network monitoring and management tools are being conducted with tools aimed at network administrators and tools that target ordinary citizens to monitor the network in their homes. The study elaborated by Yang and Edwards (2010), focused on User Experience related to network management tools, conducted interviews with 24 home network users, including those with only informal knowledge of networks to expert users. The participants reported some usability problems regarding the difficulty of understanding and using the tools since they require sophisticated technical knowledge in computer networks. Furthermore, the lack of visualization, such as a visual map of the home network, was also a problem pointed out by the participants, especially when faced with issues such as lack of connection or slow network speed. Inconsistent user interfaces of management tools were also identified as a problem. According to the authors, this issue occurs due to the lack of guidelines for developing this type of software, resulting in different interfaces depending on the supplier, the device, and the operating system (Yang and Edwards, 2010).

The literature review conducted by nones and Rusu (2017) showed that many authors do not use a formal methodology to develop new sets of usability heuristics. However, following a formal development process is extremely important to guarantee the efficiency and effectiveness of the heuristics set in the usability evaluation. The authors of this work also emphasize that, when creating a new set of heuristics so that they are effective and efficient, it is necessary to: determine the specific attributes of the application to evaluate these attributes based on the new set of heuristics; identify existing usability heuristic sets to determine how they can help on the definition of new ones, those to be reused, and the elements to use to define heuristics; specify the new set of heuristics following a standard template to obtain well-defined and easy to understand heuristics; validate the new set of heuristics to determine which ones make it possible to find usability problems and which detect specific usability problems related to the application.

The work presented in Vikström (2018) proposes a new set of heuristics for usability in network management systems. The authors start by refining Nielsen's (Nielsen, 1994) generic heuristics and then modifying them to solve some previous network management issues. However, the proposed heuristics are tested in a network management system named "Music", which is not a well-known system. So, they may not be applicable to other network management systems.

Thus, in a previous work (Verdi et al., 2020), we carried out an experimental study with nine network administrators active in the market, followed the fundamentals of the study of usability in the light of HCI techniques. Through video recordings and notes, we collected the difficulties and problems that the administrators faced when performing usual network management tasks in the Nagios management tool, such as congestion analysis, web traffic measurement, and flow analysis in a router, among others. Based on the qualitative data collected, a thorough analysis was carried out to elaborate the "guidelines for usability design in network monitoring tools", a set of guidelines specific for the network monitoring tools domain. However, these guidelines have not yet been evaluated by developers of network management tools. Such evaluation is essential to verify if the guidelines are actually useful for designing these tools and if they need to be refined or changed. Therefore, this article presents this evaluation in detail.

3. Evaluation of the guidelines

This study aimed to evaluate how a catalog of guidelines can help software developers create user interfaces considering aspects of usability in network monitoring. To this end, it took into account the perspectives of *acceptance and application* of the 12 guidelines for usability design available in the catalog. The following sections present the ethical aspects, the steps of planning, conduction, analysis, and the threats to the validity of the study.

3.1. Ethical aspects

Our study considered the Regulation document 510/2016 of the Health National Board in Brazil.[1] It regulates non-invasive studies with human beings. Our planning, execution and analysis steps took into account the recommendations of the document 510/2015. We also applied a *Term of Informed Consent* (see Appendix A) to the participants which covered these recommendations. We outlined below the ethical aspects that are recommended from the regulation document and the procedures we applied in our study to fulfill these recommendations:

- Individuals should be informed about the research goals and the researchers responsible for the study. We prepared an invitation message to the participants informing them who are the researchers in charge of the study as well as the study's aim, i.e. to evaluate the guidelines. We attached to the invitation message, the *Term of Informed Consent* (see Appendix A) from which the participants could see a brief explanation about the study purposes.
- The procedures of data gathering have to be explained to the individuals as well as which data will be collected and at what time. First, the participants had access about the data collection instruments and procedures in the *Term of Informed Consent* (see Appendix A). Before running the study, we conducted an explanation supported by slides informing the participants we will examine the prototypes created by them to see the application of the guidelines. We also informed the collection of their feedback about the guidelines. We make clear that our evaluation focused on the guidelines and not the participants' expertise.
- Make clear to the individuals that the data collected will be strictly used for scientific purposes. The second paragraph of the *Term of Informed Consent* (see Appendix A) clarifies that all the data collected will be anonymized and used exclusively for academic purposes.
- Participants should be guaranteed confidentiality and privacy of the raw data collected that will only be accessed by the researchers. In addition to clarifying about keeping the anonymity of the participants' data, we explained before the study conduction starts that the raw data would be kept in local storage in an institutional computer with access only by the researchers in charge.
- The right to access the results have to be kept to the participant whenever they wish. During the study presentation, we informed all the participants they could have access to the data and the research results.
- The study has to be carried out in an appropriate space. We carried out the study in the same laboratory where the participants used to take the course. The laboratory had computers in adequate numbers for the number of participants. The necessary software was previously installed in all the computers to guarantee that all the participants had the same settings for the study.
- The participant fatigue and stress should be mitigated through short sessions of collected data. First, the profile and demographic data were collected from a questionnaire whose fulfillment lasted 5 min. To mitigate the participants' fatigue, we conducted a short session study lasting up to 2 h and collected the participants' feedback about the guidelines from a concise questionnaire.
- The participant has the right to not participate and to discontinue participation at any time without penalty. We carried out the study making sure that the participants would not feel pressured nor influenced. The study was not related to the subject of the course given by one of the authors. Moreover, the author who was the course professor did not attend the experiment.
- Participation in the study should be voluntary without any financial compensation. The participants were invited to freely and optionally participate in the study, and a few chose not to participate. It is worth mentioning that we rewarded students with an extra point in their grades as a way to thank them for participating in the study. On the other hand, not participating in the study did not affect students' grades.

3.2. Planning

We opted for an experimental study with the development of medium-fidelity prototypes. They are artifacts that describe the organization of user interface components and how the information is arranged (Sharp et al., 2019). Such prototypes are developed using the interface prototyping technique. Indeed, they do not implement the software's functional requirements and guide the developer to the elements that will conduct the user–software interaction. Therefore, the study's main goal was to instruct the participants to use the guidelines in the catalog for building prototypes to collect data on the acceptance and application of the guidelines.

The study was organized into three steps. Firstly, a *recruitment (i)* step invited students to participate in the study and collected information about the profile of these participants. Then, the *preparation (ii)* step was intended to level out the participants' knowledge about network monitoring, usability, and prototyping tools. Finally, the *conduction (iii)* step designed a set of artifacts to support each step based on the participants' prototypes.

3.2.1. Recruitment

We applied the profile questionnaire[2] in the *recruitment (i)* step to collect the profile of the participants in advance. Initially, it presented the *Term of Informed Consent* (Appendix A) and, in case of acceptance, the process went forward through further questions. In addition to demographic data, the questionnaire

[1] Available at: http://conselho.saude.gov.br/resolucoes/2016/Reso510.pdf (in Portuguese).

[2] http://bit.ly/3cUg1os.

collected data on participants' prior knowledge of network monitoring tools and the use of prototyping tools (Table 2). In total, 52 participants agreed to participate in the study. We grouped them into pairs according to the participants' level of knowledge. For each pair, knowledge in subjects, such as *network management* and *prototyping development*, were considered. If a student had "good" or "very good" knowledge in a subject, then the other student in the pair could not have "good" or "very good" knowledge in that same subject. However, not necessarily all pairs had a student with "good" or "very good" knowledge in a specific subject. It is worth noting that even though the class was aimed at juniors, there were also upper year students and some students who were interns or employees. This is the reason why, in general, students' knowledge on each of the subjects we evaluated is not homogeneous.

We conducted a pilot study with a doctoral student and a professor from the computer networks area through the approach described in the preparation (ii) and conduction (iii) steps. Initially, we had planned to divide participants into two groups, with group 1 starting to prototype based on the scenario of medium complexity and group 2 based on the high complexity scenario, and then swapping the scenarios halfway through the experiment. However, based on the results of the pilot study, we realized that it would not be possible to apply the two scenarios (i.e., medium and high complexity) for the two groups of participants. We observed that the time required for both groups to perform both tasks would make the prototyping activity tiring, which could compromise the data collection. As a result, we separated the participants into two groups in order to observe whether the scenario's complexity interfered with the prototypes. However, each group would elaborate the prototypes based on only one of the scenarios.

3.2.2. Preparation

The *preparation (ii)* step is made of three parts. The first consisted of a 4-hour training prepared with the foundation on monitoring networks based on the NagiosXI tool[3]. This training phase enabled non-expert participants to have contact with the NagiosXI's main features. For the second part, we presented the main study topics (i.e., monitoring tools, usability, and prototyping activity) and the catalog with the 12 guidelines through a set of slides. Finally, we conducted a warm-up for the participants to apply some guidelines and get used to the prototyping tool to prevent prototyping tools from introducing difficulties in the activity. Through this approach, the participants could build a medium-fidelity prototype using Justinmind[4] prototype tool. It is worth mentioning that to support the participants during the warm-up, we provided a simple scenario in the context of our study: *"An equipment's network interface is down, that is, a physical failure has occurred. The network administrator must indicate which interface is down as soon as possible to restore the connection. You should think about which screens the network administrator would have to go through until he reaches a conclusion".*

3.2.3. Conduction

For the *conduction (iii)* step, the participant built prototypes with the help of the catalog containing the 12 guidelines. Afterward, we evaluated their acceptance based on the feedback provided, which helped assess the evaluated guidelines' acceptance and usefulness. To accomplish this goal, we elaborated a questionnaire[5] based on the *Technology Acceptance Model* (TAM). This approach is commonly adopted to analyze the acceptance of given information technology by a group of participants (Venkatesh and Davis, 2000). Besides the questionnaire, we set up two scenarios with different complexity levels, defined according to the difficulty level in using the monitoring tool to guide participants on this step (see Table 1). Consequently, we arranged the participants into two groups. One group started by creating prototypes using the medium complexity scenario, whereas the other by using the high complexity scenario. To assist in further analysis of the prototypes, we developed user interface baselines based on both scenarios by applying the guidelines.

The *preparation (ii)* and *conduction (iii)* steps were conducted based on a catalog containing the 12 guidelines accompanied by a *cards*. Each *card* describes the name, purpose, data set that could be used to view the information, information regarding interactivity, an explanation of when to avoid its use, and an illustrative image with the application of the guideline. Fig. 1 shows an example of a card containing one of the guidelines used. The complete catalog can be found in Appendix B.

3.3. Execution

Since the study focused on evaluating the use of the guidelines, the participants needed to have some knowledge of both Computer Networks and Software Development. Therefore, we decided to invite undergraduate students enrolled in the Distributed Systems course from the Computer Science program of the Federal University of São Carlos (UFSCar), campus of Sorocaba, Brazil, which audience comprises students at an advanced stage (i.e., juniors). To this end, we sent the profile questionnaire to the students, and a total of 52 agreed to participate. In a pre-analysis of the profiles, we noted that, in general, the participants had little knowledge about prototyping tools, usability, and management tools for network monitoring.

After recruiting students, a computer networks professor held a training course presenting to students the main concepts of network monitoring. He also explained the main features of a network monitoring related tools by exploiting the NagiosXI. This activity lasted 4 h and was carried out at Federal University of São Carlos. The goal of this activity was to level the participants' knowledge about network monitoring tools, in order to guarantee that all participants had the basic knowledge of what is a network monitoring tool and what network monitoring tools user interfaces look like. This knowledge was necessary so that the participants could later create prototypes of a monitoring tool.

In the following week, we carried out the warm-up and the experiment on the same day. The warm-up, which lasted approximately 20 min, relied on the Justinmind prototyping tool, the low-level complexity scenario, and the guidelines catalog. A prototype was provided with part of a pre-developed interface, allowing participants to download and change it, as the researcher demonstrated how to use the tool. The purpose of the warm-up was to introduce participants to the tool and demonstrate how to apply the guidelines in the prototype. Right after the warm-up, the experiment was conducted by two researchers: a Professor and a junior researcher in the Software Engineering field. The study took place in the same laboratory where the participants used to take the course, according to the steps described in Section 3.2.

After the warm-up, the participants were arranged into the previously defined pairs, as described in Section 3.2. They started building prototypes by applying the evaluated guidelines but beforehand informed that the main objective was to produce

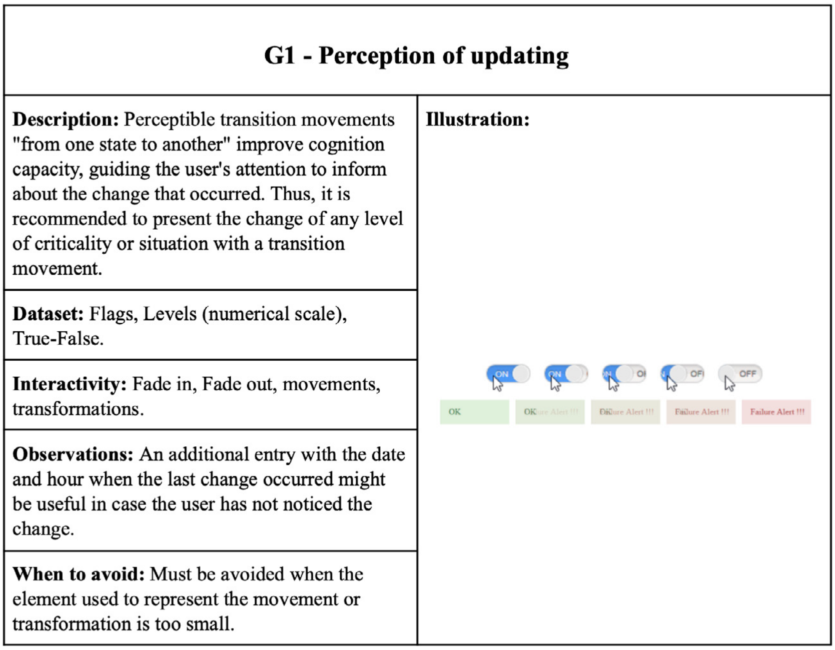
[3] NagiosXI was pointed as one of most used network monitoring tools in our previous study (Verdi et al., 2020). As such, we keep NagiosXI as our choice for this paper.

[4] https://www.justinmind.com/.

[5] http://bit.ly/3ryVidJ.

Table 1
Scenarios used in the study.

Scenario complexity	Description	Instructions on what to prototype
(S1) Medium Complexity Scenario: Router flow Analysis	Josh Baker is a Junior Network Administrator. Josh's boss asked him to show a graph or table representing all router R6 flows in the last 24 h, regardless of the source or destination of these flows, in a network monitoring tool used by the company. The data flow will be used in an audit, identifying security violations and anomalies with the help of an adjacent system. The goal of this task is that Josh is able to extract a sample with all the router flows. To accomplish this, he must find and adjust options that satisfy his need until he reaches his goal.	Considering you are the software developer of Josh's network monitoring tool, you should think about which screens he would have to go through until he finds router R6 flows. Imagine he is already logged into the network monitoring tool. To prototype these user interfaces, you should follow the recommendations in the "Catalog of recommendations for usability design in network monitoring tools". Whenever you use a recommendation from the catalog, you should add a comment with the number of the recommendation you have used at the location in the prototype.
(S2) High Complexity Scenario: Congestion Analysis	Robert Brewer is a Senior Network Administrator. Robert, using the network monitoring tool, realizes there is a network congestion consuming almost the entire available bandwidth and causing high latency on some nodes. Robert must identify the highest traffic generators and consumers and the location of the bottlenecks. Robert must take measurements of traffic and latency in order to identify sources and destinations of flows in the congestion scenario. Then, Robert will be able to make a decision to divert traffic and/or relieve the equipment causing the bottleneck, thus returning services to normal.	Considering you are the software developer of Robert's network monitoring tool, you should think about which screens he would have to go through until he identifies traffic generators and consumers and the location of the bottlenecks. Imagine he is already logged into the network monitoring tool. To prototype these user interfaces, you should follow the recommendations in the "Catalog of recommendations for usability design in network monitoring tools". Whenever you use a recommendation from the catalog, you should add a comment with the number of the recommendation you have used at the location in the prototype.



G1 - Perception of updating

Description: Perceptible transition movements "from one state to another" improve cognition capacity, guiding the user's attention to inform about the change that occurred. Thus, it is recommended to present the change of any level of criticality or situation with a transition movement.

Dataset: Flags, Levels (numerical scale), True-False.

Interactivity: Fade in, Fade out, movements, transformations.

Observations: An additional entry with the date and hour when the last change occurred might be useful in case the user has not noticed the change.

When to avoid: Must be avoided when the element used to represent the movement or transformation is too small.

Illustration:

Fig. 1. Example of a card used on the catalog for G1 guideline.

solutions concerned with usability issues. We followed the approach described in Section 3.2, which consisted of splitting the pair into two homogeneous groups for high and medium complexity scenarios. However, one of the pairs misunderstood which scenario they were supposed to use, so they developed a prototype using the high complexity instead of the medium complexity scenario. As a result, 14 pairs used the high complexity scenario, whereas 12 the medium complexity.

The pairs were instructed to add comments to their prototypes in the Justinmind tool, pointing out where the guidelines were applied. Hence, it was possible to see if participants indeed understood and used the guidelines. We should point out that these comments are not part of the guidelines, instead they were used as a tool to help us later analyze the prototypes created by the participants. During the experiment, the participants produced a total of 26 prototypes, as presented by the example depicted in Fig. 2. Each pair generated at least two and at most four user interfaces in each prototype. The number of user interfaces per prototype was not limited, so each pair defined the number they considered most appropriate to obtain a complete solution. In the end, the participants answered the feedback questionnaire individually.

3.4. Analysis

The analysis relied on data sources generated by the participants' prototypes, the responses to the feedback questionnaire, and the baseline prototype. In respect to the participants' profiles, Table 2 details them and show how the participants were organized according to Medium and High complexity scenarios, respectively. The "Scenario" column identifies the scenario used by participants when developing the prototypes. As the tables

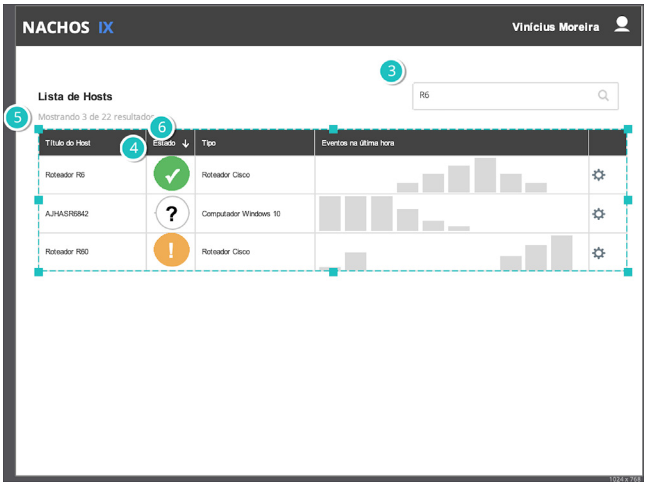
Table 2
Participants profile - S: Scenario; Pair: Pair number; Id: Participants Id; (i) Network Management; (ii) Network Foundations; (iii) Software Prototyping; (iv) UI design; W: Work; F: Field of work; A: Age.

S	Pair	Id	(i)	(ii)	(iii)	(iv)	W	F	A
Medium Complexity (S1)	P1	1	△	□	◯	△	No	–	25
		2	◊	★	◊	◊	No	–	24
	P2	3	◯	◯	△	△	No	–	25
		4	△	△	◊	□	Intern	BPM Dev.	23
	P3	5	△	□	△	◯	No	–	24
		6	□	□	◊	□	No	–	22
	P4	7	□	□	△	□	No	–	21
		8	□	◊	◊	□	No	–	21
	P5	9	◯	△	□	△	No	–	23
		10	△	□	△	△	No	–	21
	P6	11	□	□	□	△	No	–	21
		12	△	△	△	△	No	–	22
	P7	13	△	△	□	□	No	–	23
		14	□	◊	△	□	No	–	21
	P8	15	◯	□	□	◊	No	–	22
		16	△	□	◯	△	No	–	21
	P9	17	△	□	◯	△	No	–	21
		18	□	□	◊	□	No	–	21
	P10	19	□	◊	◊	□	No	–	20
		20	△	□	△	△	Intern	IT	23
	P11	21	△	★	□	□	No	–	20
		22	△	□	□	□	No	–	20
	P12	23	△	△	□	□	No	–	21
		24	◊	◊	□	◊	No	–	21
High Complexity (S2)	P13	25	△	◊	◊	□	No	–	24
		26	△	□	△	□	No	–	21
	P14	27	△	△	◊	△	Intern	IT Consulting	25
		28	△	△	△	△	Employee	Android Dev.	22
	P15	29	△	□	◊	△	No	–	27
		30	◯	□	△	△	No	–	20
	P16	31	◯	□	◯	◯	No	–	24
		32	△	□	◊	◊	No	–	22
	P17	33	△	□	□	△	Intern	AI/Data Sci.	22
		34	△	□	△	△	No	–	21
	P18	35	△	□	□	□	No	–	21
		36	◯	◊	◯	△	No	–	22
	P19	37	△	□	□	□	No	–	23
		38	△	□	△	□	No	–	20
	P20	39	□	□	□	◊	Intern	Marketing	23
		40	△	□	◯	△	No	–	20
	P21	41	△	◊	△	△	No	–	25
		42	□	◊	◊	◊	No	–	23
	P22	43	□	◊	□	◊	No	–	20
		44	△	□	△	◯	No	–	25
	P23	45	△	◊	◯	◯	No	–	20
		46	◯	□	□	◊	No	–	22
	P24	47	△	△	◊	◊	Intern	Software Dev.	21
		48	□	□	△	□	Intern	Java Dev.	21
	P25	49	□	◊	□	△	No	–	21
		50	◊	◊	□	◊	No	–	22
	P26	51	□	□	□	□	No	–	24
		52	□	□	◊	△	No	–	25

Knowledge degree: No knowledge (◯), Basic (△), Intermediate (□), Good (◊), Very good (★).

show, Participants 1 to 24 belong to S1, and participants from 25 to 52 belong to S2. The "Pair" column describes how the pairs were formed. As it shows, every two participants are part of a different pair. Among the 52 participants who took part in the study, 53.8% and 55.8%, had basic knowledge and intermediate knowledge in Network Management, respectively. As for Software Prototyping, 32.7% of the participants had intermediate knowledge, whereas 38.5% had basic knowledge in UI design. The table also shows that most participants are only undergraduate students, while a few also work, whether as interns or employees. Participants' ages range from 20 to 27.

For the analysis, we firstly evaluated the solutions developed by each pair considering the guidelines applied. To this aim, we inspected the comments added by the pairs, identifying the application of a guideline and comparing to the application of the same guideline in the baseline solution. The researchers classified the use of the guidelines into one of the following categories:



NACHOS IX

Vinícius Moreira

Lista de Hosts

Mostrando 3 de 22 resultados

R6

Título do Host	Estado ↓	Tipo	Eventos na última hora	
Roteador R6	✓	Roteador Cisco		⚙
AJHASR6842	?	Computador Windows 10		⚙
Roteador R60	!	Roteador Cisco		⚙

Fig. 2. Example of a prototype produced by the participants.

- (i) applied the guideline correctly, which means the pair pointed out the use of the guideline and it was correctly applied;
- (ii) applied the guideline incorrectly, when students point out a specific part of their prototype saying they have applied a certain guideline but they do not follow the guideline specification;
- (iii) applied the guideline correctly without mentioning, which refers to the correct application of the guideline but without a comment pointed out by the pair;
- (iv) applied the guideline partially correctly, when students understood the overall meaning of the guideline but missed an important detail that the guideline specifies to make the application completely correct;
- (v) did not apply the guideline, which means participants chose not to apply that guideline.

Secondly, we verified the acceptance of the guidelines, using the TAM questionnaire. Considering the different questions composing TAM, it was possible to analyze the guidelines' use and relevance in the developers' view.

4. Results

This section discuss the results according to each research question stated in Section 1.

4.1. (RQ1) how did the software developers use the guidelines?

The answer to this question comes from the analysis of which guidelines applied to develop the prototypes. So, they were classified into one of the five categories, as discussed in Section 3.4. In general, the pairs created prototypes that covered the relevant interfaces related to the scenario given to them.

In some of the prototypes, we observed that the same pair did not maintain the standard of interfaces, which is an important guideline related to usability (Nielsen, 1994). For instance, a pair did not keep the standard when using the yellow color both in conjunction with the success symbol and the failure symbol, in addition to the fact that the yellow color is generally associated with "alert" (Verdi et al., 2020).

Table 3 presents the result of the analysis on the application of each guideline. It shows the number of pairs that applied that guideline (NP), the total number of times that the guideline was applied (NA), the results from the perspective of the categories of application (i.e., (i) to (iv)), and the findings we observed from the analysis. Besides that, Fig. 3 illustrates the frequency of the categories per guideline. The results revealed that G12 was the least applied and G2 the most applied, the latter with the greatest number of proper applications, i.e., 27 correct applications and three correct applications without mentioning (see (i) and (iii) in Table 3). On the other hand, G4 had a high incorrect application rate (see (ii) in Table 3).

These results show that the participants applied some of the guidelines more times than others in each solution. Fig. 4 presents the use of the guidelines per prototype. G2 - "Perception of colors" was applied by 23 out of the 26 prototypes, most of the time being applied more than once. Comparing Fig. 4 with the data presented in Table 3, we concluded that guidelines G2 and G3 were not only the most used guidelines but also had the highest correct application rate. guidelines G1 and G10 had few correct applications. guidelines G9 to G12 were applied only by a few of the pairs.

We developed an approach to compare the application of the guidelines in each prototype to the baseline elaborated by the researchers (see Eq. (1)).

$$\frac{\mathrm{Corr} \times 5 + \mathrm{NoMention} \times 4 + \mathrm{Partial} \times 3 + \mathrm{Incorr} \times 2 + \mathrm{LackNot} \times 1}{(\mathrm{Corr} + \mathrm{NoMention} + \mathrm{Partial} + \mathrm{Incorr} + \mathrm{LackNot} + \mathrm{LackEss}) \times 5} \tag{1}$$

Eq. (1) calculates the average considering the categories related to the application or not application of the guidelines and assigning a weight to each category as follow: the correct applications (*Corr*) – 5, the correct applications without mentioning (*NoMention*)– 4, the partially correct applications (*Partial*) – 3, the incorrect applications (*Incorr*) – 2, the lack of application of "not essential" guidelines (*LackNot*) – 1 and the lack of application of "essential" guidelines (*LackEss*) – 0.

Correct applications award 5 points because the guideline was correctly applied according to the its specification, rewarding the participants the maximum score in that application. Hence, a correct implementation means students understood the guideline and could apply it correctly. A correct application without mentioning is also correct. However, since participants did not add a comment pointing out in the prototype where the guideline was used, we cannot guarantee that participants actually understood its meaning and forgot to add a comment or if they accidentally added a certain feature in the prototype without noticing that the feature was the application of that guideline. That is why a correct application awards 5 points, while a correct application without a comment awards only 4 points. Partially correct applications award 3 points because participants understood the overall meaning of the guideline but did not include an important detail stated in the guideline specification. An incorrect implementation awards 2 points because even though it was not correct, participants tried to implement it, whereas lack of implementation means students chose not to implement that guideline. Regarding the lack of application, we analyzed if each guideline that was not applied was "essential" or "not essential". The "essential" guidelines are G1 to G9, which were applied by the researchers in the baseline solution. The "not essential" guidelines are G10 and G11, since they were not applied in the baseline. These guidelines were not applied in the baseline because they did not fit the proposed solution. However, not using them had no negative impact on the baseline, hence they were considered not essential. Therefore, no implementation of "not essential" guidelines awards 1 point, since the baseline solution also did not implement them and no implementation of "essential" guidelines awards 0 points, since these recommendations should have been implemented, according to the baseline prototype.

Table 3
Application of the guidelines - G: guideline; NP: number of pairs which applied G; NA: number of times that G was applied; Categories of application: (i), (ii), (iii), (iv).

Guidel. (G)	NP	NA	(i)	(ii)	(iii)	(iv)	Findings from the application of the guideline
G1 - Perception of updating	7	8	3	4	1	0	Even though the pairs applied the guideline, it was not applied in the way the guideline proposes, which is to demonstrate the change in critically levels or in the state of the machines. The incorrect applications were movements in general, such as graphs that change and a scroll bar on the monitoring tool page.
G2 - Perception of colors	25	27	21	3	3	0	Solutions that applied this guideline incorrectly did not demonstrate the degree of status (success, alert and problem).
G3 - Finding specific information in a large set of data	23	24	16	1	6	1	The most recurrent application was with search bars, but some pairs applied with a filter of results (with selectors). One pair applied incorrectly, pointing to an entire table as the guideline, while another pair applied partially correctly, explaining that the data presented on the screen were the result of a filter, but did not prototype the filter, which would effectively be the application of the guideline.
G4 - Obtaining more detailed information	18	18	8	7	0	3	Incorrect and partially correct applications were those in which the pairs pointed the entire graph as the guideline, sometimes with no value or showing all the values of the graph points, but without explaining or demonstrating that the values would only appear when the mouse pointer hovers over a determined point in the graph. This was essential, considering that the guideline explicitly states that specific values of points in the graph should appear as the mouse pointer hovers over them.
G5 - Sorting information	19	19	15	2	1	1	One pair presented data ordered in a table and pointed out the use of the guideline, but the application was only partially correct, since they did not use symbols to demonstrate how the ordering is done. Another pair did not understand the guideline, pointing to the use of a data filter, instead of a way to order them.
G6 - Spying before going deeper	19	20	15	3	2	0	The incorrect applications were those in which the prototype element pointed out by the pair did not present a summary view of the data as a starting point. One of the applications pointed to the title of the "Tables" tab, another pointed to a table on the page of a specific router (complete information, not a preview before delving into it).
G7 - Starting point	19	19	13	2	4	0	An incorrect application pointed to the title of a specific page with information from only one router. The solution did not present the information in a dashboard format as instructed by the guideline. In another incorrect application, the solution presented specific data and graphics without showing an overview of all equipment. In addition, it was not the start page of navigation and cannot be considered a starting point, as described in the guideline.
G8 - Use of metaphors to inform about status and incidents	18	19	15	2	1	1	The partially correct application used symbols only for some of the equipment and incorrectly associated the symbol "!" to "problem", instead of "alert", but they understood that the guideline was about the use of symbols to demonstrate the state of the network equipment. We also found the use of symbols with unconventional meanings. For example, a pair pointed out that they used this guideline when they used the pencil symbol, which represented editing, but this symbol does not represent levels.
G9 - Notifications	4	4	2	1	1	0	The incorrect application pointed to the use of the guideline in a large part of the screen showing the latest updates. However, a notification should be just a warning about something important to the network administrator. In addition, the notification should not disturb the screen, that is, it should not be intrusive, but should be shown in the corner of the screen.

(*continued on next page*)

We applied this formula to each of the solutions developed by the pairs. The guideline "*G12 - Suitably arranged data*" was not considered in any of the solutions for this analysis, since the prototypes were developed for desktop format and G12 is related to devices of different screen sizes (i.e., mobile devices). The score obtained by the baseline prototype after the calculation was approximately 0.85. The baseline did not achieve a score of 1 since the authors did not apply all the guidelines. Fig. 5 shows the comparison between the baseline and the pairs' solutions (P). Prototypes P8 and P11 achieved scores higher than the baseline score. In both solutions, we found many occurrences of correct applications of the guidelines. Also, these prototypes correctly applied one or both of the "not essential" guidelines (G9 and G10), which were not applied in the baseline. On the other hand,

Table 3 (*continued*).

Guidel. (G)	NP	NA	(i)	(ii)	(iii)	(iv)	Findings from the application of the guideline
G10 - Help filling in fields	9	10	3	3	4	0	A correct application added a placeholder in the search field, defining that the entry should be the name of a router. An incorrect application pointed to a selector, but since it is a selector there is no explanation on how to fill in the field. Some pairs pointed out field labels, which cannot be considered a help to fill in the field.
G11 - Gradual display of information	4	4	2	1	0	1	It was expected that the network topology, a list or dictionary would be presented to assist the network administrator. The partially correct application pointed to a table listing the subnets, but this information is not necessary to understand the other data on the screen, since the dashboard presents only a summary of the system. The incorrect application presented the network topology on a separate page, but did not allow it to be viewed on pages with other information.
G12 - Suitably arranged data	2	2	0	2	0	0	This guideline was used by only few of the pairs because the solution instructions were to design for a desktop device and not for different devices. Two pairs pointed out the use of responsiveness and mentioned that the system itself was already responsive, but did not prototype mobile screens or screens of different sizes, only screens desktop, and it was not possible to see the responsiveness of the prototyped elements.

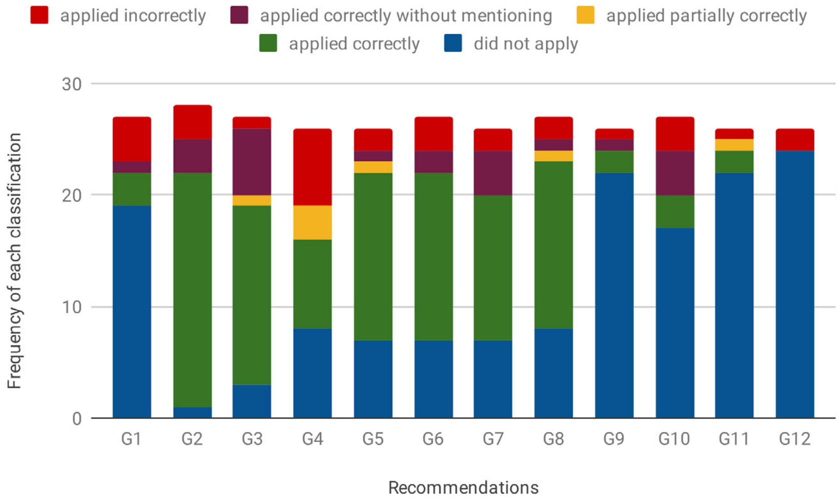


Fig. 3. Frequency of categories per guideline.

prototype P4 applied very few guidelines, and many were applied incorrectly. In general, we see that only three solutions scored less than or equal to half mark (0.6).

Fig. 6 presents the comparison between the level of complexity of scenarios 1 and 2 and the application of the guidelines. For this, the classifications "*applied correctly*", "*partially applied correctly*" and "*applied correctly without mentioning*" were merged, polarizing the classification of the guidelines used between "*applied correctly*" and "*applied incorrectly*".

4.2. (RQ2) what was the feedback provided by the software developers?

This section presents the results regarding the developers' perception when using the 12 guidelines. We carried out an analysis using the Technology Acceptance Model (TAM) responses collected from the participants individually. TAM questions are divided into two dimensions (see Table 4): the perceived ease of use and the perceived usefulness. The ease-of-use is related to a person's perception of adopting a technology with no effort. In contrast, the usefulness dimension represents how much a person considers using a specific technology to improve their performance. We added another question to the TAM questionnaire to check how easily the participants considered remembering the guidelines (F7). Furthermore, we adopted a six-point Likert scale going from *Strongly Disagree* to *Strongly Agree* (Likert, 1932). We chose to use a 6-point Likert scale (i.e. strongly disagree; largely disagree; partially disagree; partially agree; largely agree; and strongly agree) without using a central point option (neutral, neither agree nor disagree, I do not know) because according to Johns (2005), a central point option can motivate the respondents to not pinpoint their opinion on the issue. In that work, the author argues that the central point, or neutral, is commonly used by the respondent to avoid a possible conflict of opinion with the researcher.

For each TAM question, the participant selected his/her degree of agreement. Fig. 7 presents, for each TAM question, the total of participants who selected each option in the Likert scale.

The acceptance responses to the guidelines (i.e., TAM answers) were also pooled to analyze each TAM dimension's degree of agreement. For this, the Relative Strength Index (Wilder, 1978) was used, which provides the agreement factor considering the influence of disagreements. We also calculated the individual degree of agreement for each TAM question (IDA) and the factor of agreement of each TAM dimension (FDA). IDA results from the agglomeration formula (see Eq. (2)) (Wilder, 1978). To obtain the *Ag* value, Eq. (3) considers the frequency of all agreement

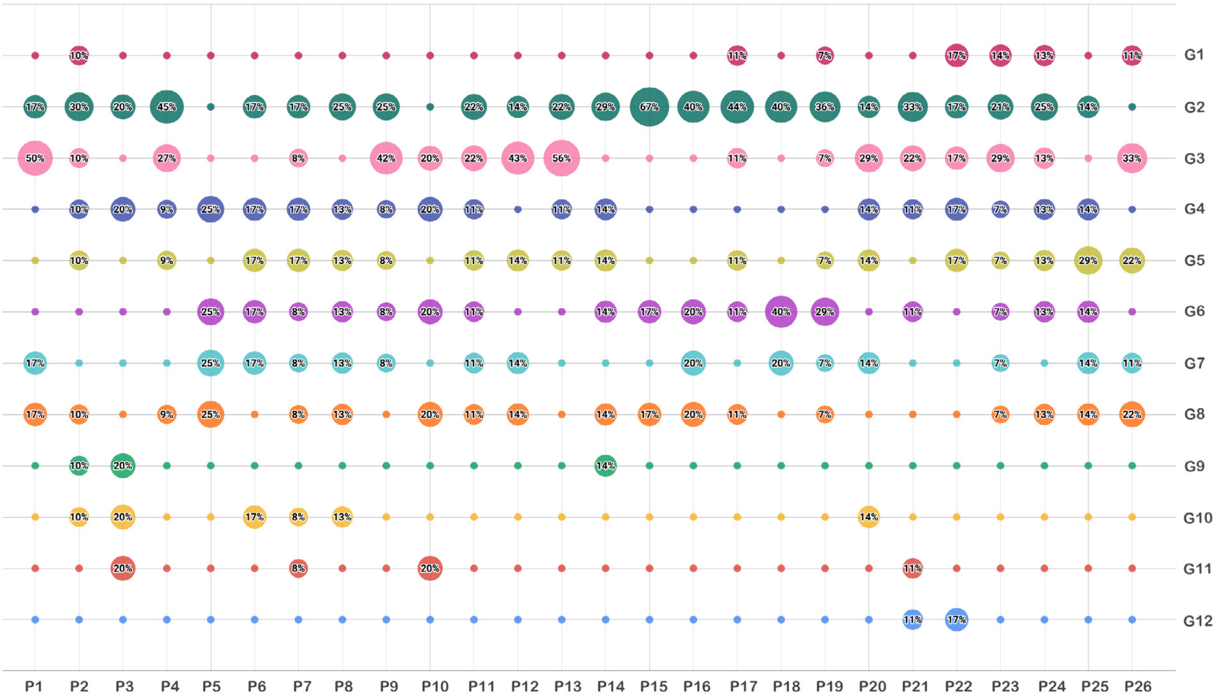


Fig. 4. Frequency of application of each guideline per prototype.

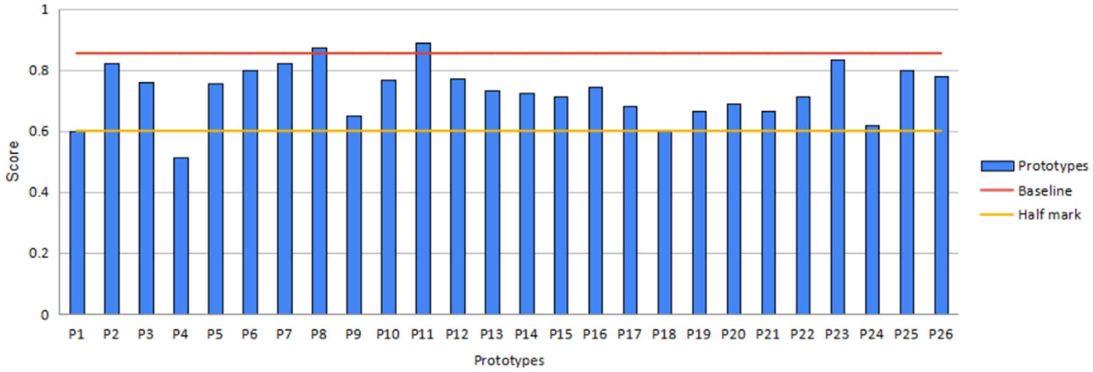


Fig. 5. Comparison between the solutions developed by the pairs and the baseline.

Table 4
TAM questionnaire.

Dimension	Question	
Usefulness	U1	Using the guidelines enables me to create the solution more quickly.
	U2	Using the guidelines improves my ability to build the solution.
	U3	Using the guidelines increases my productivity when developing the solution.
	U4	Using the guidelines enhances my effectiveness on the development of the solution.
	U5	Using the guidelines improved my perception about the best practices to build the solution.
	U6	I consider the guidelines useful to create the solution.
Ease-of-use	F1	Learning how to use the guidelines was easy for me.
	F2	I found it easy to use the guidelines the way I wanted to.
	F3	The orientation regarding the use of the guidelines are easy to understand.
	F4	I understood what happened during my interaction with the guidelines.
	F5	It was easy for me to become skillful at using the guidelines.
	F6	The guidelines give me flexibility to create the prototypes.
	F7	I consider the guidelines easy to remember.

responses to the question (i.e., from strongly agree to partially agree). On the other hand, to obtain the *Dis* value, Eq. (4) considers the frequency of disagreement responses obtained in the same question (i.e., from strongly disagree to partially disagree). If the value of *Dis* is equal to zero, the final IDA value is considered 100. Table 5 presents the results obtained through this approach.

$$IDA = 100 - (\frac{100}{\frac{Ag}{Dis} + 1}) \tag{2}$$

$$Ag = SA + LA + PA \tag{3}$$

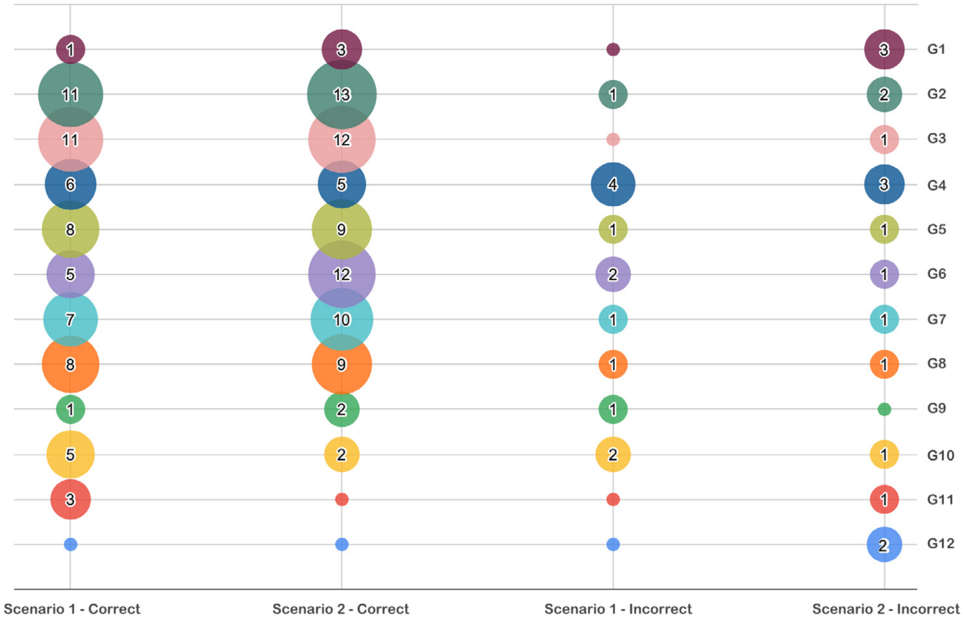


Fig. 6. Correct and incorrect applications in each task.

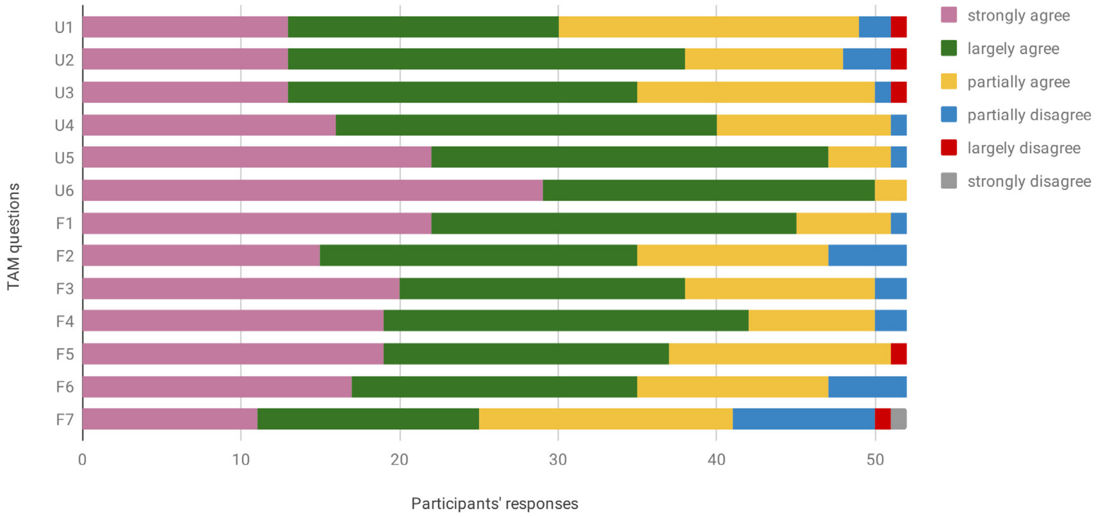


Fig. 7. TAM questionnaire responses.

Table 5
IDA and FDA Results — Strongly agree (SA), Largely agree (LA), Partially agree (PA), Partially disagree (PD), Largely disagree (LD), and Strongly disagree (SD).

Question	SD	LD	PD	PA	LA	SA	Ag	Dis	IDA	FDA
U1	0	1	2	19	17	13	49	3	94.23	**96,48**
U2	0	1	3	10	25	13	48	4	92.31	
U3	0	1	1	15	22	13	50	2	96.15	
U4	0	0	1	11	24	16	51	1	98.08	
U5	0	0	1	4	25	22	51	1	98.08	
U6	0	0	0	2	21	29	52	0	100	
F1	0	0	1	6	23	22	51	1	98.08	**92,58**
F2	0	0	5	12	20	15	47	5	98.38	
F3	0	0	2	12	18	20	50	2	96.15	
F4	0	0	2	8	23	19	50	2	96.15	
F5	0	1	0	14	18	19	51	1	98.08	
F6	0	0	5	12	18	17	47	5	98.38	
F7	1	1	9	16	14	11	41	11	78.85	

$$Dis = SD + LD + PD \tag{4}$$

Regarding the FDA, its calculation involves the arithmetic mean of the values obtained from calculating the degree of agreement of the questions (IDA) for the calculated factor. Therefore, we calculated the usefulness and ease-of-use FDA's factors based on the IDA arithmetic mean of questions from U1 to U6 and from F1 to F7, respectively. Hence, the values obtained from the IDA and FDA calculation were interpreted according to Table 6 (Wilder, 1978). Considering the results in Table 5, we found that the vast majority obtained very strong degrees of agreement (i.e., above 90). The exception is in question F7. The results of the usefulness dimension revealed that the participants found the guidelines useful. In particular, we observed that question U6 obtained a level of agreement of 100, demonstrating that no participants selected the disagreement options. On the other hand, in the easy-of-use dimension, question F7 showed a moderate agreement (i.e., 78.85) due to the large amount of partial agreement and disagreement responses. This result indicates that many of the participants do not find the guidelines easy to remember.

Table 6
Interpretation of Degree of Agreement values.

Degree of Agreement value	Interpretation
≥ 90	Very strong agreement
80 - 89.99	Substantial agreement
70 - 79.99	Moderate agreement
60 - 69.99	Fair agreement
50 - 59.99	Slight agreement
40 - 49.99	Slight disagreement
30 - 39.99	Fair disagreement
20 - 29.99	Moderate disagreement
10 - 19.99	Substantial disagreement
≤ 9.99	Very strong disagreement

In addition to the guidelines' acceptance, we also asked participants, when answering the Feedback questionnaire, to indicate how frequent the guidelines when building the solution and how useful they considered each of the guidelines. In these questions, the participants indicated the use and perceived relevance of each of the 12 guidelines. Table 7 presents the results of the participants' responses. One of the highlights is the guideline "*G4 - Obtaining more detailed information*", which was used and considered relevant by the vast majority of participants, 51 out of 52. The guideline with the greatest difference between use and relevance was "*G11 - Gradual display of information*", with a difference of 31% (16). The only guideline that was not considered relevant by most participants was "*G12 - Suitably arranged data*". This perception is due to the fact that the participants were not asked to create solutions that were adaptable to different devices.

Finally, we verified whether the complexity of the scenario provided to the participants (Scenario 1 and 2) influenced the acceptance of the guidelines. To do this, we applied Fisher's Exact Test to each of the TAM questions. Fisher's Exact Test (Fisher, 1922) seeks to test statistical significance between two elements, observing the significance of the deviation from the null hypothesis. The goal is that the results obtained in Fisher's Exact Test, called *p-value*, reject the null hypothesis and accept the alternative hypothesis. The test can be used on samples of any size, yet it is used more commonly on samples smaller than 30.

To calculate Fisher's Exact Test and show the results, we created Table 8. It contains all the responses for each TAM question (columns from 1 to 6) mapped into two groups, one for each scenario (in the rows medium and high complexity). Following that, we counted the number of participants who selected each of the six options in the Likert scale (from strongly disagree - 1 to strongly agree - 6) for each question. Then, they were separated into two rows: one referring to Group 1 participants (medium complexity scenario) and another referring to Group 2 participants (high complexity). Thus, columns 1 to 6 sustained our testing. Since the sample was small, we adopted a 95% confidence interval to mitigate errors in the analysis of the results. The Fisher's Exact Test was performed using the online tool (Vasavada, 2016), since it allows the calculation with tables larger than 2X2. It is worth noting that we defined the following null and alternative hypotheses to support the test results analysis.

- **H0: There is no influence of the scenario's complexity on the acceptance of the guidelines**
- **HA1: There is an influence of the scenario's complexity on the acceptance of the guidelines**

An important remark is that both hypotheses were tested by replacing "TAM questions" with each TAM question (i.e., U1, …,U6; F1, …,F7). Table 8 presents the *p-values* obtained in each TAM question. Since all the *p-value* results were above the confidence interval (i.e., 0.05), it is not possible to reject the null hypothesis for each of the TAM questions. This means that, from Fisher's Exact Test, there was no statistical significance found to determine that the scenarios' complexity influenced the acceptance of the guidelines.

5. Refining the guidelines

After the evaluation, we considered that some guidelines required refinement, mostly due to a potential misunderstanding on how to use them. Consequently, we elaborated a new version of *G1, G2, G4, G6, G7, G10,* and *G11* guidelines Appendix C. Such a refinement was done by rewriting the title so that the guideline's purpose becomes more clear for software developers. In what follows, we explain what triggered our motivation for refining each of the mentioned guidelines.

In general, these guidelines were refined because the rate of incorrect and partially correct applications was high compared to the total number of applications. At the same time, the incorrect and partially correct applications followed a pattern, i.e., the prototypes which guidelines were applied did similar mistakes.

The guideline's title was one aspect subject to refinement. In some cases, they were not consistent with the remaining information (e.g., description, illustration) on how the guidelines were supposed to be applied. Therefore, the experiments showed the correct and incorrect use of the same guideline by a pair at different prototype's locations. Based on these findings, we decided to change the titles to prevent developers from getting confused with the real guideline's purpose, so ensuring that they are consistent with its description. Table 9 summarizes the refinements made.

6. Discussion

Considering our results and the RQs stated in Section 1, we pinpoint discussions in the following subsections.

6.1. RQ1 - how did the software developers use the guidelines?

Based on the data analyzed, we observed that G2 and G3 were the most used guidelines. Yet, G4, G5, G6, G7, and G8 were widely used. G2 was correctly applied more often, followed by guidelines G3, G5, G6, G7, and G8. In particular, G2 and G3 were applied more than once in most of the prototypes. G4 had a lower rate of correct applications. G1, G9, G10, G11 e G12 were applied by a few participants and, in most cases, were only used once in each prototype. By comparing the prototypes built by the participants to the baseline (Fig. 5), we observed that the developers could use and understand the guidelines since most of them obtained scores close to the baseline score.

We also analyzed whether the scenarios' complexity influenced the correct or incorrect use of the guidelines. From a qualitative analysis of the prototypes, we concluded that it did not significantly interfere, except for guideline G6, which had many correct applications for prototypes related to the S2 than the S1. Similarly, the complexity of the scenario also does not influence the non-application of guidelines, except for G10, which was much less applied in the prototypes linked to S2 compared to those built to the S1.

6.2. RQ2 - what was the feedback provided by the software developers?

The second research question concerns the developers view regarding the guidelines.

Table 7
Use and perception of relevance of guidelines — absolute value presented in parenthesis.

	Guidelines	Use	Relevance
G1	Perception of updating	42% (22)	58% (30)
G2	Perception of colors	98% (51)	98% (51)
G3	Finding specific information in a large set of data	75% (39)	88% (46)
G4	Obtaining more detailed information	87% (45)	92% (48)
G5	Sorting information	69% (36)	87% (45)
G6	Spying before going deeper	62% (32)	83% (43)
G7	Starting point	73% (38)	87% (45)
G8	Using metaphPors to inform about status and incidents	69% (36)	79% (41)
G9	Notifications	27% (14)	56% (29)
G10	Help filling in fields	35% (18)	52% (27)
G11	Gradual display of information	27% (14)	58% (30)
G12	Suitably arranged data	13% (7)	42% (22)

Table 8
Count of participants' responses to TAM questions - 1: Strongly disagree; 2: Largely disagree; 3: Partially disagree; 4: Partially agree; 5: Largely agree; 6: Strongly agree.

Question	Scenario	1	2	3	4	5	6	p-value
U1	Medium	0	0	2	7	9	6	0.45
	High	0	1	0	12	8	7	
U2	Medium	0	0	1	7	10	6	0.46
	High	0	1	2	3	15	7	
U3	Medium	0	0	1	8	9	6	0.78
	High	0	1	0	7	13	7	
U4	Medium	0	0	0	7	10	7	0.54
	High	0	0	1	4	14	9	
U5	Medium	0	0	1	2	8	13	0.16
	High	0	0	0	2	17	9	
U6	Medium	0	0	0	2	9	13	0.40
	High	0	0	0	0	12	16	
F1	Medium	0	0	0	3	10	11	0.96
	High	0	0	1	3	13	11	
F2	Medium	0	0	2	4	13	5	0.20
	High	0	0	3	8	7	10	
F3	Medium	0	0	0	4	9	11	0.47
	High	0	0	2	8	9	9	
F4	Medium	0	0	0	3	13	8	0.48
	High	0	0	2	5	10	11	
F5	Medium	0	0	0	7	9	8	0.94
	High	0	1	0	7	9	11	
F6	Medium	0	0	3	7	9	5	0.39
	High	0	0	2	5	9	12	
F7	Medium	0	1	6	9	4	4	0.24
	High	1	0	3	7	10	7	

First, we concluded that the level of complexity of the scenarios does not influence the TAM questions (i.e. in the acceptance of the guidelines). Moreover, the strong agreement on TAM questions shows that, in general, developers have accepted the guidelines well. A more thorough analysis about the developers view regarding the use and relevance of the guidelines was carried out and showed that 7 of the 12 guidelines were marked as used by more than 50% of the developers, whereas 11 of them were marked as relevant by more than 50% of the developers. Therefore, from the developers' point of view, the guidelines were well accepted, being considered useful and relevant in general.

7. Threats to validity

Threats to validity are situations that can occur throughout the work's development, which compromise its validity. They can be internal, external, instrumentation, or conclusion threat.

Internal threat refers to the tiredness or lack of motivation of the participants in the experimental study. It would consequently prevent them from carrying out the study seriously, which could negatively affect the results. Regarding motivation, it was possible to mitigate this threat by giving extra credit points in their grades in the Distributed Systems discipline if the student (i.e., software developer) participated in the study, which was an incentive. Regarding fatigue threat, we mitigated it by performing a short interval between training and the study's conduction.

The external threat refers to the set of participants not representing the population of interest, who are the developers of network management tools, that is, students' participation in the study and not of market developers. This threat could be considered, as the students' lack of knowledge could prevent us from generalizing the results to other environments. However, Salman, Misirli, and Juristo (Salman et al., 2015) demonstrate that students perform similarly to experienced professionals in new activities. Despite having more experience than students, market professionals do not know the use of design guidelines presented in this study and the students. Hence, students can participate in this study without their participation, posing a threat to its validity.

The instrumentation threat is related to the instruments' preparation, that is, to the use of the Justinmind prototyping tool and the participants' knowledge both in network management and in the NagiosXI network monitoring tool. However, we mitigated this threat with an expository class on network management and the NagiosXI network monitoring tool. Subsequently, we trained with the participants to become familiar with the prototyping tool Justinmind and how the study would be performed later.

Finally, the conclusion threat is related to the method used for the data analysis. This threat was mitigated by carrying out a quantitative and descriptive statistical analysis using the TAM questionnaire and the utility perception questionnaire.

8. Conclusions and future work

This work presented a study to evaluate 12 guidelines for usability design in network monitoring tools aimed at software developers. The analysis showed that they are indeed relevant and help the software developers create interfaces focusing on usability in network monitoring tools.

The study described in this work brought relevant contributions. It explored the use of design guidelines for usability in the specific context of network management, showing how they help software developers. Through a qualitative and quantitative analysis, the experiments revealed that the proposed guidelines, specifically the guidelines G2, G3, G4, G5, and G7, are relevant to the development of network monitoring tools. Most participants used them. With the feedback questionnaire analysis, we

Table 9
Guidelines refinements — Original Title (O) → Refined Title (RF).

Id	Title	Motivation
G1	Perception of updating (O) → Movements representing a situation change (RF)	Most of the pairs applied this guideline incorrectly, all of them showing general movements in elements of the prototype, such as changing graphics or the scroll bar of the prototyped page, without representing movements of transition of states and levels of criticality, as proposed by the guideline.
G2	Perception of colors (O) → Colors representing the state of the elements (RF)	All incorrect applications used colors in the prototyped elements, but did not represent the state of the elements (success, alert and problem), since they did not use specific colors (green, yellow and red), which was the objective of the guideline.
G4	Obtaining more detailed information (O) → Obtaining detailed information with mouse pointer (RF)	Most of the applications of guideline G4 were incorrect and partially correct. All of them pointed out an entire graph as the use of the guideline, with no value or showing all the values of the graph points. However, they did not explain or demonstrate that such values should appear only when the mouse pointer hovers over them, which is what the guideline proposes.
G6	Spying before going deeper (O) → Present summary before going deeper (RF)	All incorrect applications of guideline G6 maintained a pattern: they all pointed out elements of the prototypes that did not represent a summary view. However, the guideline proposes the developer presents a summary view to correctly guide users, before they investigate all the data.
G7	Starting point (O) → Dashboard as a starting point on the home screen (RF)	As for guideline G7, the incorrect applications were similar to each other, pointing out elements that were not an overview of everything that was happening on the network and these elements were not on the prototyped system's homepage (right after the user logs in to the tool). Therefore, they did not represent a dashboard with information about all that is happening on the network, as proposed by the guideline.
G10	Help filling in fields (O) → Explanatory text for filling in fields (RF)	guideline G10 had incorrect applications were those in which pairs pointed out elements that were not information on how to fill in a field, often pointing to a title or label of a field instead of a placeholder or explanation on how to fill in a field below it, for example. However, the guideline proposes that, in addition to the label, an explanation of how the user should fill in the field should be presented.
G11	Gradual display of information (O) → Display window with additional information (RF)	Finally, regarding guideline G11, incorrect and partially correct applications presented information that was not necessary for the understanding of other data on the screen or presented information on a separate screen. However, the guideline proposes that the information presented should help the user to understand the other information presented on the screen and be presented in a separate window that can be moved around the screen as needed by the administrator.

observed that the developers accepted the guidelines well, considering them relevant. The research work, both through the graphs and Fisher's Exact Test, demonstrated that, in general, the level of complexity of the scenarios for the creation of prototypes had no significant effect on the acceptance of the guidelines.

In Verdi et al. (2020), the authors sought to understand the difficulties that network administrators encountered when using network monitoring tools, culminating in elaborating 12 guidelines to improve the usability of these tools. This work analyzed these guidelines and their efficiency, observing whether they were well accepted and correctly applied by the developers and if it would be necessary to make adjustments to the guidelines to refine them. This activity was possible by analyzing the prototypes developed by the software developers. Future work involves validating the prototypes generated in this work through a study with network administrators, verifying whether the use of guidelines in elaborating prototypes of network monitoring tools can effectively generate better usability. Therefore, it will be possible to analyze the end user's perspective, that is, the network administrator, ending the last step of validating the guidelines.

CRediT authorship contribution statement

Sofia A.M. Silveira: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing. **Luciana A.M. Zaina:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Supervision. **Leobino N. Sampaio:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing. **Fábio L. Verdi:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partly supported by the National Council for Scientific and Technological Development (CNPq), Brazil under grants 313312/2019-2, 144477/2018-1 and 432064/2018-4; and FAPESB, Brazil under grant TIC0004/2015.

Appendix A. Term of informed consent

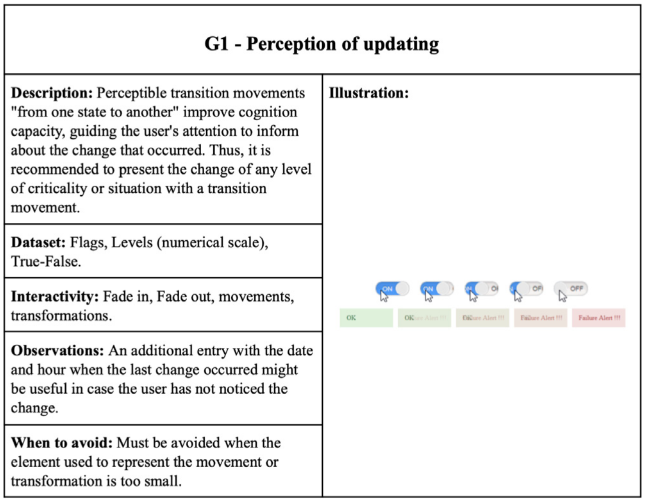
TERM OF INFORMED CONSENT

You are invited to take part in the research "Analysis of guidelines for (Re)Design of Visualization in Network Monitoring Tools". This research is part of an undergraduate research project at Federal University of São Carlos - UFSCar, conducted by student Sofia A. M. Silveira and supervised by professors Fábio Luciano Verdi and Luciana Zaina. The purpose of the research in which you will participate is the use of a set of design guidelines for network monitoring tools. To participate, you should answer a profile questionnaire that will be available in the next step. During the study, you will build low-fidelity prototypes. At the end, we will collect your feedback about the guidelines from a questionnaire.

The data collected from the two questionnaires (i.e. profile and feedback) and during the study will only be analyzed altogether for the purpose of scientific research. We guarantee the anonymity of the participants and access to the results of the research after it is completed. Your participation will not involve any physical risk neither financial expense or gain. You have the right to withdraw from participating in the study at any time without prejudice.

The questionnaires will be accessed and answered by those who agree to take part in this study. In case of accepting to participate, the participant should indicate the acceptance to the consent term and answer to participate in the study by selecting the option "Yes, I accept to take part in the research". On contrary, if you do not accept to participate, you should select the option "No, I do not accept to take part in the research".

Appendix B. Original guidelines



G1 - Perception of updating

Description: Perceptible transition movements "from one state to another" improve cognition capacity, guiding the user's attention to inform about the change that occurred. Thus, it is recommended to present the change of any level of criticality or situation with a transition movement.

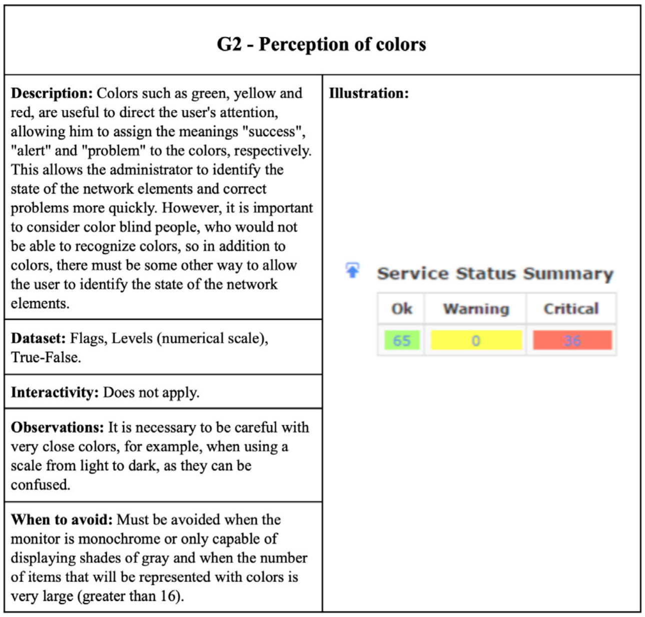
Dataset: Flags, Levels (numerical scale), True-False.

Interactivity: Fade in, Fade out, movements, transformations.

Observations: An additional entry with the date and hour when the last change occurred might be useful in case the user has not noticed the change.

When to avoid: Must be avoided when the element used to represent the movement or transformation is too small.

Illustration:



G2 - Perception of colors

Description: Colors such as green, yellow and red, are useful to direct the user's attention, allowing him to assign the meanings "success", "alert" and "problem" to the colors, respectively. This allows the administrator to identify the state of the network elements and correct problems more quickly. However, it is important to consider color blind people, who would not be able to recognize colors, so in addition to colors, there must be some other way to allow the user to identify the state of the network elements.

Dataset: Flags, Levels (numerical scale), True-False.

Interactivity: Does not apply.

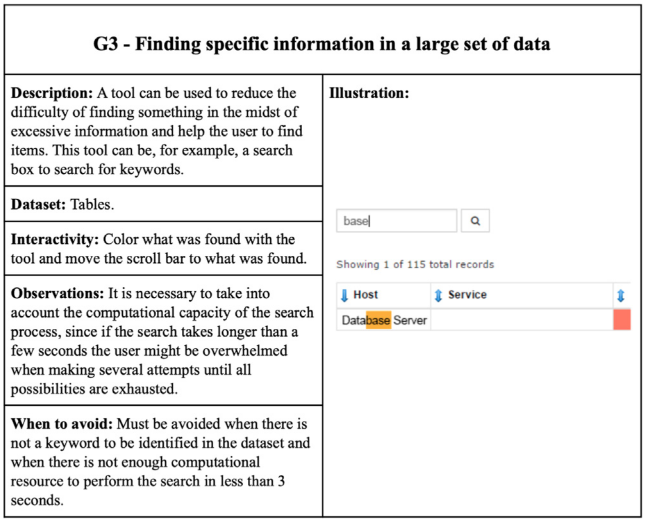
Observations: It is necessary to be careful with very close colors, for example, when using a scale from light to dark, as they can be confused.

When to avoid: Must be avoided when the monitor is monochrome or only capable of displaying shades of gray and when the number of items that will be represented with colors is very large (greater than 16).

Illustration:

Service Status Summary

Ok	Warning	Critical
65	0	36



G3 - Finding specific information in a large set of data

Description: A tool can be used to reduce the difficulty of finding something in the midst of excessive information and help the user to find items. This tool can be, for example, a search box to search for keywords.

Dataset: Tables.

Interactivity: Color what was found with the tool and move the scroll bar to what was found.

Observations: It is necessary to take into account the computational capacity of the search process, since if the search takes longer than a few seconds the user might be overwhelmed when making several attempts until all possibilities are exhausted.

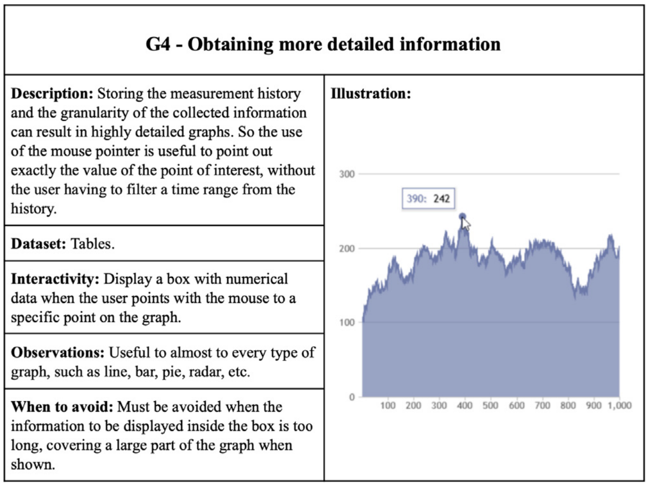
When to avoid: Must be avoided when there is not a keyword to be identified in the dataset and when there is not enough computational resource to perform the search in less than 3 seconds.

Illustration:

base

Showing 1 of 115 total records

Host	Service
Database Server	



G4 - Obtaining more detailed information

Description: Storing the measurement history and the granularity of the collected information can result in highly detailed graphs. So the use of the mouse pointer is useful to point out exactly the value of the point of interest, without the user having to filter a time range from the history.

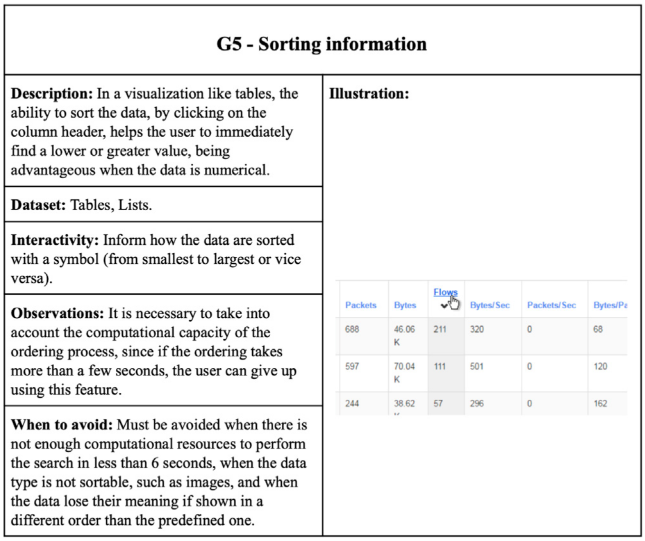
Dataset: Tables.

Interactivity: Display a box with numerical data when the user points with the mouse to a specific point on the graph.

Observations: Useful to almost to every type of graph, such as line, bar, pie, radar, etc.

When to avoid: Must be avoided when the information to be displayed inside the box is too long, covering a large part of the graph when shown.

Illustration:



G5 - Sorting information

Description: In a visualization like tables, the ability to sort the data, by clicking on the column header, helps the user to immediately find a lower or greater value, being advantageous when the data is numerical.

Dataset: Tables, Lists.

Interactivity: Inform how the data are sorted with a symbol (from smallest to largest or vice versa).

Observations: It is necessary to take into account the computational capacity of the ordering process, since if the ordering takes more than a few seconds, the user can give up using this feature.

When to avoid: Must be avoided when there is not enough computational resources to perform the search in less than 6 seconds, when the data type is not sortable, such as images, and when the data lose their meaning if shown in a different order than the predefined one.

Illustration:

Packets	Bytes	Flows	Bytes/Sec	Packets/Sec	Bytes/Pa
688	46.06 K	211	320	0	68
597	70.04 K	111	501	0	120
244	38.62 K	57	296	0	162

G6 - Spying before going deeper

Description: Using a summary view that shows the latest samples (or summarized samples) is important to direct the user to the correct path, avoiding unnecessary steps and, therefore, reducing the number of steps needed to perform a task.

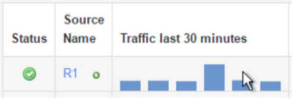
Dataset: Tables, Lists, Flags, Levels (numerical scale).

Interactivity: Does not apply.

Observations: The summarized information to be shown needs to be validated with a key user or with an expert, so that the information shown to the tasks that need to be accomplished.

When to avoid: Must be avoided when it is not possible to show updated enough information (how much updated depends on each case) so as not to cause any false impression and prevent the user from making a wrong decision.

Illustration:



Status	Source Name	Traffic last 30 minutes
✓	R1 ●	

G7 - Starting point

Description: Graphs and information organized as soon as the user enters the monitoring tool provide an overview of everything that is happening on the network. Therefore, dashboards are a good starting point to begin analyzing a situation. Elements such as incident counters or problem counters alert the user to take an initiative.

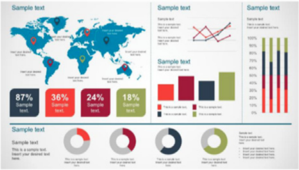
Dataset: Tables, Lists, Flags, Levels (numerical scale).

Interactivity: Does not apply.

Observations: Dashboards must be assembled by the user in order to show what is relevant to him. Dashboard suggestions can be provided by the tool itself, so that the user can make only minor adjustments.

When to avoid: Does not apply.

Illustration:



G8 - Use of metaphors to inform about status and incidents

Description: Symbols that represent the severity level of a problem are well understood, being recognized as "normal", "alert" and "problem" and helping the user to identify problems more quickly, when compared to using only text.

Dataset: Flags, Levels (numerical scale)

Interactivity: Does not apply.

Observations: The style of the symbols may vary according to the theme of the tool, however it is important to keep the colors green, yellow and red respectively with the symbols "✓", "!" and "✗".

When to avoid: Does not apply.

Illustration:

G9 - Notifications

Description: The use of pop-ups is not effective. Therefore, a non-intrusive notification method is recommended, which should appear at the corner of the screen, without disturbing the user. Thus, notifications at the corner of the screen are a succinct way to warn the user about an important event that may impact the functioning of the network.

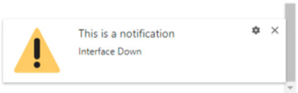
Dataset: Messages, Flags, True-False.

Interactivity: Display the notification with smooth movements when appearing and disappearing. Show detailed information when the user clicks on notification and allow the user to dismiss notification.

Observations: Do not use pop-ups that interrupt the user. Since they are widely used on the web for advertisements, they are often labeled as something that is unimportant and, therefore, users tend not to pay attention to any information presented by pop-ups.

When to avoid: Must be avoided when the monitoring tool is designed to be shown on a panel for a large audience to observe and when there is no user who will interact with the notification.

Illustration:



This is a notification

Interface Down

G10 - Help filling in fields

Description: With the help of only a simple label in the fields it is not possible to understand how they should be filled out. Thus, fields that need to be filled out with some specific rule or syntax should contain help to avoid misunderstandings.

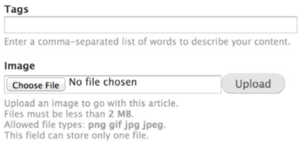
Dataset: Field and value.

Interactivity: Color the field borders red if it is not filled in correctly, for example, in a different format than specified, and specify what is incorrect.

Observations: In addition to the description of how to fill the field, an example of filling can be provided.

When to avoid: Does not apply.

Illustration:



Tags

Enter a comma-separated list of words to describe your content.

Image

Choose File No file chosen Upload

Upload an image to go with this article.
Files must be less than 2 MB.
Allowed file types: png gif jpg jpeg.
This field can store only one file.

G11 - Gradual display of information

Description: In complex cases, providing the network topology, a list or a dictionary can help the network administrator to understand the information seen on the screen. Content needed to support an analysis, when absent, impairs understanding and decision making. Given the large number of networks and hosts, it is difficult for the administrator to remember all the information, so such content must be visible when necessary.

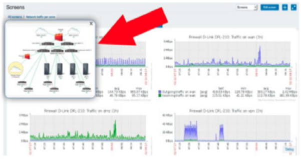
Dataset: Tables, Lists, Images.

Interactivity: The supporting content must be moved around freely, so the administrator can place that content in the best place for him.

Observations: Supporting content can be anything the user finds useful: a topology, a list of routers or IPs. The user must be able to choose what type of information he wants to appear on the floating window.

When to avoid: Must be avoided when the monitor is too small and the support content covers the entire primary screen of the tool.

Illustration:



G12 - Suitably arranged data

Description: The different sizes of graphs can make some information unreadable if the space available for rendering the graph is not big enough. Therefore, the content needs to adapt to the format and size of the user's screen, so that the available space is used, keeping all information readable.

Dataset: Tables, Lists, Flags, Levels (numerical scale), Images.

Interactivity: Does not apply.

Observations: It is necessary to be careful with dynamic data, which makes it unclear how many items will be shown on the screen. The available space needs to be taken into account so that in a graph, for example, it does not end up being incomprehensible because it is too small.

When to avoid: Must be avoided when the content is designed to be shown on a fixed-size panel (TV or screen), very common in network operations centers, making automatic adaptation to the screen format unnecessary.

Illustration:

Appendix C. Refined guidelines

G1 - Movements representing a situation change

Description: Perceptible transition movements "from one state to another" improve cognition capacity, guiding the user's attention to inform about the change that occurred. Thus, it is recommended to present the change of any level of criticality or situation with a transition movement.

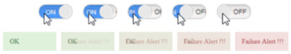
Dataset: Flags, Levels (numerical scale), True-False.

Interactivity: Fade in, Fade out, movements, transformations.

Observations: An additional entry with the date and hour when the last change occurred might be useful in case the user hasn't noticed the change.

When to avoid: Must be avoided when the element used to represent the movement or transformation is too small.

Illustration:



G2 - Colors representing the state of elements

Description: Colors such as green, yellow and red, are useful to direct the user's attention, allowing him to assign the meanings "success", "alert" and "problem" to the colors, respectively. This allows the administrator to identify the state of the network elements and correct problems more quickly. However, it is important to consider color blind people, who would not be able to recognize colors, so in addition to colors, there must be some other way to allow the user to identify the state of the network elements.

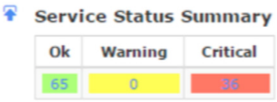
Dataset: Flags, Levels (numerical scale), True-False.

Interactivity: Does not apply.

Observations: It is necessary to be careful with very close colors, for example, when using a scale from light to dark, as they can be confused.

When to avoid: Must be avoided when the monitor is monochrome or only capable of displaying shades of gray and when the number of items that will be represented with colors is very large (greater than 16).

Illustration:



Service Status Summary

Ok	Warning	Critical
65	0	36

G4 - Obtaining detailed information with mouse pointer

Description: Storing the measurement history and the granularity of the collected information can result in highly detailed graphs. So the use of the mouse pointer is useful to point out exactly the value of the point of interest, without the user having to filter a time range from the history.

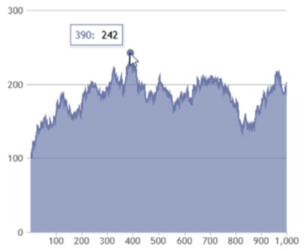
Dataset: Tables.

Interactivity: Display a box with numerical data when the user points with the mouse to a specific point on the graph.

Observations: Useful to almost to every type of graph, such as line, bar, pie, radar, etc.

When to avoid: Must be avoided when the information to be displayed inside the box is too long, covering a large part of the graph when shown.

Illustration:



G6 - Present summary before going deeper

Description: Using a summary view that shows the latest samples (or summarized samples) is important to direct the user to the correct path, avoiding unnecessary steps and, therefore, reducing the number of steps needed to perform a task.

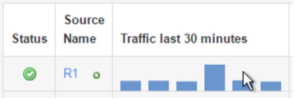
Dataset: Tables, Lists, Flags, Levels (numerical scale).

Interactivity: Does not apply.

Observations: The summarized information to be shown needs to be validated with a key user or with an expert, so that the information shown to the tasks that need to be accomplished.

When to avoid: Must be avoided when it is not possible to show updated enough information (how much updated depends on each case) so as not to cause any false impression and prevent the user from making a wrong decision.

Illustration:



Status	Source Name	Traffic last 30 minutes
✓	R1 ●	

G7 - Dashboard as a starting point on the home screen

Description: Graphs and information organized as soon as the user enters the monitoring tool provide an overview of everything that is happening on the network. Therefore, dashboards are a good starting point to begin analyzing a situation. Elements such as incident counters or problem counters alert the user to take an initiative.

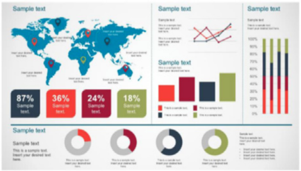
Dataset: Tables, Lists, Flags, Levels (numerical scale).

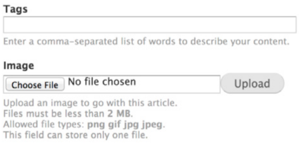
Interactivity: Does not apply.

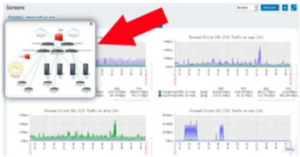
Observations: Dashboards must be assembled by the user in order to show what is relevant to him. Dashboard suggestions can be provided by the tool itself, so that the user can make only minor adjustments.

When to avoid: Does not apply.

Illustration:



G10 - Explanatory text for filling in fields	
Description: With the help of only a simple label in the fields it is not possible to understand how they should be filled out. Thus, fields that need to be filled out with some specific rule or syntax should contain help to avoid misunderstandings.	**Illustration:**  Enter a comma-separated list of words to describe your content. Upload an image to go with this article. Files must be less than 2 MB. Allowed file types: png gif jpg jpeg. This field can store only one file.
Dataset: Field and value.	
Interactivity: Color the field borders red if it is not filled in correctly, for example, in a different format than specified, and specify what is incorrect.	
Observations: In addition to the description of how to fill the field, an example of filling can be provided.	
When to avoid: Does not apply.	

G11 - Display window with additional information	
Description: In complex cases, providing the network topology, a list or a dictionary can help the network administrator to understand the information seen on the screen. Content needed to support an analysis, when absent, impairs understanding and decision making. Given the large number of networks and hosts, it is difficult for the administrator to remember all the information, so such content must be visible when necessary.	**Illustration:** 
Dataset: Tables, Lists, Images.	
Interactivity: The supporting content must be moved around freely, so the administrator can place that content in the best place for him.	
Observations: Supporting content can be anything the user finds useful: a topology, a list of routers or IPs. The user must be able to choose what type of information he wants to appear on the floating window.	
When to avoid: Must be avoided when the monitor is too small and the support content covers the entire primary screen of the tool.	

References

Bajpai, V., Schönwälder, J., 2015. A survey on internet performance measurement platforms and related standardization efforts. IEEE Commun. Surv. Tutor. 17 (3), 1313–1341. http://dx.doi.org/10.1109/COMST.2015.2418435.

Falschlunger, L., Lehner, O., Treiblmaier, H., 2016. Infovis: The impact of information overload on decision making outcome in high complexity settings. In: SIGHCI 2016 Proceedings: Proceedings of The Fifteenth Annual Pre-ICIS Workshop On HCI Research in MIS. Association for Information Systems, Dublin, Leinster, Ireland, pp. 1–5, URL https://aisel.aisnet.org/sighci2016/3.

Fisher, R.A., 1922. On the interpretation of X2 from contingency tables, and the calculation of p. J. R. Stat. Soc. 85 (1), 87–94. http://dx.doi.org/10.2307/2340521.

Guimarães, V.T., Freitas, C.M.D.S., Sadre, R., Tarouco, L.M.R., Granville, L.Z., 2016. A survey on information visualization for network and service management. IEEE Commun. Surv. Tutor. 18 (1), 285–323. http://dx.doi.org/10.1109/COMST.2015.2450538.

Jain, R., Paul, S., 2013. Network virtualization and software defined networking for cloud computing: A survey. IEEE Commun. Mag. 51 (11), 24–31. http://dx.doi.org/10.1109/MCOM.2013.6658648.

Johns, R., 2005. One size doesn't fit all: Selecting response scales for attitude items. J. Elections, Public Opin. Parties, Routledge 15 (2), 237–264. http://dx.doi.org/10.1080/13689880500178849.

Keim, D., Zhang, L., 2011. Solving problems with visual analytics: Challenges and applications. In: Lindstaedt, S., Granitzer, M. (Eds.), I-KNOW '11: Proceedings of The 11th International Conference On Knowledge Management And Knowledge Technologies. Association for Computing Machinery, New York, NY, United States, pp. 1–4. http://dx.doi.org/10.1145/2024288.2024290.

Likert, R., 1932. A technique for the measurement of attitudes. Arch. Psychol. 22, 5–55, URL https://legacy.voteview.com/pdf/Likert_1932.pdf.

Nielsen, J., 1994. 10 Usability Heuristics for User Interface Design. Nielsen Norman Group.

Nielsen, J., 2012. Usability 101: Introduction to usability. J. Usability Stud..

nones, D.Q., Rusu, C., 2017. How to develop usability heuristics: A systematic literature review. Comput. Stand. Interfaces 53, 89–122. http://dx.doi.org/10.1016/j.csi.2017.03.009.

Ogu, E.C., Ayokunle, O., Yaw, M., Achimba, O., 2014. Virtualization and cloud computing: The pathway to business performance enhancement, sustainability and productivity. Int. J. Bus. Econ. Res. 3 (5), 170–177. http://dx.doi.org/10.11648/j.ijber.20140305.12.

Pretorius, M.C., Calitz, A.P., van Greunen, D., 2005. The added value of eye tracking in the usability evaluation of a network management tool. In: Bishop, J. (Ed.), SAICSIT '05: Proceedings of The 2005 Annual Research Conference Of The South African Institute of Computer Scientists And Information Technologists On IT Research in Developing Countries. South African Intitute for Computer Scientists and Information Technologists, P.O. Box 392, South Africa, pp. 1–10.. http://dx.doi.org/10.5555/1145675.1145676.

Salman, I., Misirli, A.T., Juristo, N., 2015. Are students representatives of professionals in software engineering experiments? In: Bertolino, A., Canfora, G., Elbaum, S. (Eds.), ICSE '15: Proceedings of The 37th International Conference On Software Engineering. IEEE Press, Piscataway, NJ, Unites States, pp. 666–667.. http://dx.doi.org/10.1145/2024288.2024290.

Sharp, H., Rogers, Y., Preece, J., 2019. Interaction Design: Beyond Human-Computer Interaction, 5th John Wiley & Sons, Hoboken, NJ, United States.

Vasavada, N., 2016. Fisher's test for exact count data. https://astatsa.com/FisherTest/.

Venkatesh, V., Davis, F.D., 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. Manage. Sci. 46 (2), 186–204. http://dx.doi.org/10.1287/mnsc.46.2.186.11926.

Verdi, F.L., Oliveira, H.T., Zaina, L., Sampaio, L., 2020. Usability matters: A human-computer interaction study on network management tools. IEEE Trans. Netw. Serv. Manag. 17 (3), 1865–1878. http://dx.doi.org/10.1109/TNSM.2020.2987036.

Vikström, E.J., 2018. Heuristic Evaluation of Network Management Systems Using Axis Communications' Network Management System Music in Creation of Usability Heuristics (Bachelor Thesis at Malmö University). Bachelor Thesis at Malmö University, Malmö, Sweden.

Ward, M., Grinstein, G., Keim, D., 2010.. Interactive Data Visualization: Foundations, Techniques, and Applications. Peters, Ltd., United States.

Wilder, J.W., 1978. New Concepts in Technical Trading Systems. Trend Research.

Yang, J., Edwards, W.K., 2010. A study on network management tools of householders. In: Gkantsidis, C., Papagiannaki, K., Salonidis, T. (Eds.), HomeNets '10: Proceedings of The 2010 ACM SIGCOMM Workshop On Home Networks. Association for Computing Machinery, New York, NY, US, pp. 1–6. http://dx.doi.org/10.1145/1851307.1851309.

Sofia Silveira holds a Bachelor's degree in Computer Science from Federal University of São Carlos, Brazil. During her undergraduate program, she conducted research on Software Engineering and User Experience (UX). Currently, Sofia is pursuing her Master's degree at the University of Ottawa, Canada.

Luciana Zaina is an Associate Professor in the Computing Department at the Federal University of São Carlos in Brazil. She has experience in conducting research projects in collaboration with software industries. Her research interests have been mostly in requirement engineering, user experience (UX), empirical software engineering, and UX in startups.

Leobino Sampaio is an Associate Professor of Computer Science with the Federal University of Bahia (UFBA). In 2020, he was a visiting researcher with the Computer Science Department of the University of California, Los Angeles (UCLA) in the United States. His research interests include future Internet architectures and network performance evaluation.

Fábio Luciano Verdi is an Associate Professor in the Computer Science Department at Federal University of São Carlos (UFSCar). He has been working with data centers, cloud computing, SDN, and dataplane programmability. He is a member of the LERIS Research Group and has been leading projects in the area of computer network monitoring, virtual resources and cloud infrastructures. He is currently a research visitor at KTH, Sweden.