Evaluation in Visualization

It is very difficult to get a good solution to a visualization problem on the first try. As we mentioned in the document about how to develop visualizations, several iterations of cycles of conceptualization, prototyping and evaluation are usually necessary to ensure that the final solution is suitable to help target users accomplish their tasks, in the context in which they should do so. There are assessment methods of different nature, complexity and with different application conditions, and it is a good strategy, in most cases, to use several, since different methods can provide different and complementary information. The selection of evaluation methods to use, as well as the way to apply them, will depend on several aspects such as the development phase, the complexity and criticality of the application, as well as the specificity of the application and target users. For example, visualizations that will support decision-making with serious consequences or in critical conditions (such as in monitoring dangerous processes, or in intensive medical care) should be evaluated in a more exhaustive way, using more methods, or methods in their more precisely (e.g., testing involving more participants, target users, or domain experts).

In Visualization we can use evaluation methods from other areas such as Interactive Computing Systems, Image Processing or Computer Graphics, focusing on different aspects, such as usability, image quality, or the algorithms used to generate visual representations; however, to evaluate visual data exploration applications, it is more common to use usability and user experience evaluation methods (User eXperience-UX), which should occur at the second and third design levels in the development cycle. These methods will eventually require some adaptation to the specific nature of these applications which, although interactive, are characterized by having a very significant content of visualizations. According to the ISO (International Standards Organization) definition, evaluating usability involves evaluating the extent to which a product can be used by specific users to achieve specific objectives with effectiveness, efficiency and satisfaction in a specific context of use. In this chapter we cover the main usability evaluation methods that can be used throughout the development process, whether for a simple set of static visualizations, or a sophisticated application for visual data exploration. They are organized, according to a commonly used classification, illustrated in Figure 1, into analytical and empirical methods. The former are carried out by Visualization experts, not involving participants, and the second are carried out with the collaboration of participants, as representative as possible of the target users. The assessment methods we present can be used in the context of formative assessment, or summative assessment. In the first case, they will be used throughout the development process, to improve the solution (identifying problems); in the second case, to evaluate the quality of the final solution (for example, to check whether established requirements are met, or to compare alternatives).

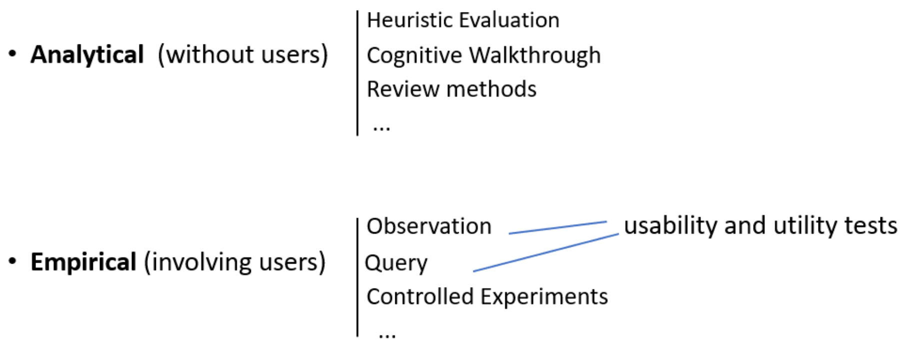

Figure 1. Possible classification of the main usability evaluation methods that have been used in Visualization

It should be noted that a good evaluation strategy in the context of a development process should involve the use of several methods and the results obtained should be used to implement improvements that can solve the identified problems. Occasionally, it may not be possible to solve all problems due to lack of time or other resources, in which case resolution priority must be managed according to the severity of the problem and the cost of its solution; Therefore, serious problems must be resolved, even if this is costly, but less serious problems can be resolved if it is easy, leaving it up to a new version to fix them, otherwise.

Note that, although evaluation is a fundamental activity in the development of a Visualization solution, whether simpler or more complex, focusing only on evaluation may not lead to an overall improvement in usability, since it is only one part of a process iterative. It is always necessary to start from understanding the user's needs, making an adequate characterization of the problems to be solved by users, which corresponds to the first level of design; Only then will it be possible to design potentially useful and usable solutions, evaluating and improving them based on the results obtained, in several cycles of design, implementation and evaluation, until the established objectives are achieved.

A visualization is created to answer a set of questions taking into account a particular set of data and, being something made “to measure”, it is also necessary to evaluate it in a personalized way, and understand if it actually allows not only to answer the questions to be that intends to answer, but adds or allows checking context knowledge and potentially finding existing patterns in the data, which the user may or may not have previously known.

1. Analytical methods

Analytical usability evaluation methods do not involve user participation, which, on the one hand, is an advantage, on the other, it is a disadvantage. Effectively, not involving users makes the process more expeditious, as well as less complex and expensive; however, it is simultaneously a limitation, since it is not possible to know the needs and characteristics of users well and they can vary greatly from one another, making it very difficult to anticipate their behavior and identify all the problems that a solution may have. Nevertheless, the use of these methods is very useful. They should be used, as early as possible in the development process, to detect possible problems and improve the solution, before investing time and effort in developing complex prototypes and evaluating with users. There are several analytical methods that can be adapted for evaluation in Visualization, with Heuristic Evaluation being the one that has been most used. Another easily usable analytical method for evaluating data exploration applications is the Cognitive Walkthrough. We briefly describe these two methods below and leave reading suggestions to explore the topic.

1.1 Heuristic evaluation

Heuristic Assessment was initially proposed by Nielsen & Molich in 1990[1]. It is an inspection method in which several experts use sets of rules of thumb (so-called heuristics) to evaluate the usability of user interfaces. These rules are based on characteristics that solutions must have to be usable and provide adequate UX. Analysts must carry out systematic and independent analyzes of all aspects of the user interface to detect potential usability problems, analyzing compliance with sets of known heuristics (for example those proposed by Nielsen-Molich[2]) and prepare a consolidated report that can help development teams to improve the usability of the solution. This method has the advantage of being relatively simple to use, not requiring much time or experience on the part of evaluators to produce useful results; can be used early in the development process and can be applied to a low-fidelity prototype; however, it can also be used to evaluate the usability of final solutions. The main limitation of this method is the fact that it is subjective, i.e., the results will depend on the analysts who carry out the evaluation. To alleviate this, several analysts should be involved. Some studies indicate that a number of evaluators between three and five will be adequate for most cases. Note that, when potential usability problems could cause errors with serious consequences, we should consider increasing this number, as well as involving highly experienced evaluators. Adapting this method to be used in Visualization may be done using sets of specific heuristics, with several sets having been proposed for this purpose.

[1] https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/theory-heuristic-evaluations/

https://www.interaction-design.org/literature/topics/heuristic-evaluation

[2] https://www.nngroup.com/articles/ten-usability-heuristics/

How to perform a Heuristic evaluation

As we mentioned previously, heuristic evaluation is carried out by several evaluators who begin by carrying out an independent analysis, verifying compliance with a set of heuristics. Whenever any of the heuristics appears not to be fulfilled, the evaluator registers a potential usability problem and the heuristic, or unfulfilled heuristics, assigning a degree of severity to the problem according to a chosen scale. After everyone has their analyses, the evaluators communicate and discuss their results and produce a report aggregating all the results and conclusions. This procedure is important to ensure independent assessments and maximize the number of problems found. Assessment results can be recorded in a written report or by having assessors verbalize their comments to an observer who records them.

Although the severity of usability problems depends on several aspects (the frequency with which the problem occurs, the impact if it occurs and its persistence), it is common to combine them into a single classification as a global assessment, in order to facilitate prioritization and decision-making in its resolution. A commonly used scale is the scale, also proposed by Nielsen and Molich, with five levels: 0 – I do not consider it to be a problem; 1- “cosmetic” problem, to be resolved only if possible; 2- minor problem, with low priority for resolution; 3- serious problem, with high resolution priority; 4- usability catastrophe, with an unavoidable resolution. The first level (zero) is reserved for use in the results integration phase, when an evaluator does not consider it to be a problem, some aspect highlighted by another evaluator as a potential usability problem.

A good way to summarize the results of the various evaluators would be to produce a table as in Figure 2 in which the potential problems found by each of the evaluators are presented through an explanatory name of the problem, the severity by each evaluator and the median of the values for each issue. This helps quickly understand which aspects require urgent intervention and which can be left for a new version. The report must also include a detailed description, and illustration of each issue, the unmet heuristics and possible ways to improve.

Potential usability issue	Grade by evaluator				Severity
Description	**Alice**	**Bruno**	**Carol**	**David**	**Median**
Unreliable Internet access	1	3	2	0	1,5
Unclear error messages	3	2	3	3	3
Undistinguishable colour scale levels	2	2	2	3	2
Difficult to see time filter	2	2	3	3	2.5
Confusing map navigation	1	1	2	1	1
Difficult to define Region of Interest	2	2	1	2	2

Figure 2. Example of a table resulting from heuristic evaluation by four evaluators of a visual data exploration application with a short description, the severity [0,4] and the median for each found potential usability issue.

Heuristics list

The first list of usability heuristics was proposed by the method's authors and includes ten rules that help to systematically analyze the interaction aspects of an application. They are quite generic, meaning they also apply to visual data exploration applications. However, these applications have a very significant content of visualizations that are not specifically treated by those heuristics, and specific heuristics have been proposed to be used in Visualization. In this context, an appropriate strategy could be to carry out a heuristic evaluation using more than one set of heuristics, seeking to ensure a more complete analysis of the visualization and interaction aspects. We present below two sets of heuristics that allow an analysis from both points of view: the heuristics of Nielsen and Molich and the heuristics of Zuck et al.

Nielsen and Molich[3] heuristic lists

There are ten general principles to guide the design of interactive systems to obtain a usable solution. They are called "heuristics" because they are general rules and not very specific usability directives.

[3] https://www.nngroup.com/articles/ten-usability-heuristics/

1: Visibility of system status – users must be informed about what is happening; Adequate feedback must be provided within a reasonable period of time.

2: Correspondence between the system and the real world – user terminology must be used, with familiar terms, phrases and concepts and follow real-world conventions.

3: User control and freedom – users often select options by mistake; There must be an obvious and quick way to exit an unwanted state, as well as undo an action or recover it (undo and redo).

4: Consistency and standards - different terms, situations or actions should not be used with the same meaning; Existing conventions must be followed.

5: Error Prevention – An error-preventing design is preferable to clear error messages.

6: Recognize instead of remember – objects, actions and options must be clearly visible; the user should not have to remember information from one part of the dialogue to another and the instructions for use should be visible or easily accessible whenever necessary.

7: Flexibility and efficiency of use – there must be "accelerators" that provide more efficient interaction for experienced users; It should also be possible to customize frequent actions.

8: Aesthetic and minimalist design - dialogue should not contain irrelevant or rarely necessary information; supplementary information competes with relevant information and reduces its relative visibility.

9: Help recognize, diagnose and recover from errors - error messages should use simple, clear language accurately indicating the problem, suggesting a solution.

10: Help and documentation - although it is preferable if the application can be used without documentation, it may be necessary to provide help and documentation; Help must be easy to search, concise and focused on the user's task, indicating concrete steps to be taken.

Zuk et al. heuristics

This list of specific heuristics includes useful general principles to guide the design of visualizations to achieve a solution that is effective. They facilitate a more detailed analysis of specific aspects not addressed by previous heuristics and are based on studies in the areas of Vision and Visualization Sciences. They address aspects of visual perception, a fundamental theme in Visualization, but very complex, as well illustrated by Adelson's illusion, which we can see in Figure 3. In this image, squares A and B appear to have a very different color, although they have exactly the same gray level, demonstrating that what we understand from an image is not just related to the levels of light intensity that reach our retinas and alerting for the need to take perceptual aspects into consideration in the design of a visualization.

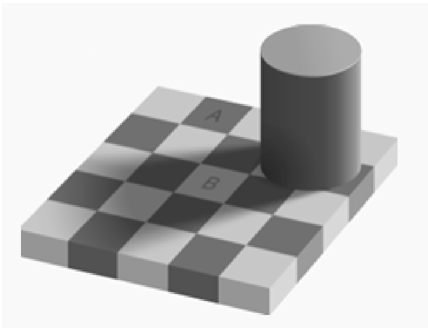

Figure 3. Adelson´s visual illusion: A e B have exactly the same grey level; however, they are perceived as having very different levels[4].

The heuristics of Zuk et al. were obtained by analyzing a set of works carried out by researchers in the field of Visualization and include the following principles:

1: The visual representation must be long enough to be perceived – for example, the smallest bar in a bar graph must be long enough to be easily visible; Figure 4 shows an example (on the left) where this does not always happen.

2: Color in general does not provide order information – for example, the levels of a scale represented by red, green and blue will not be perceived as having an order; if you want to convey an order, progressive levels of gray or the same tone should be used (for

[4] https://michaelbach.de/ot/lum-adelsonCheckShadow/

example, starting with a lighter shade of gray, gradually darkening to black); This will allow you to easily understand that there is an order, although the direction of the scale may be ambiguous (for example, black represents which end of the scale?).

3: The perception of color varies according to the size of the colored item – there is a complex relationship between color, size and shape in perception; the color of very small areas is difficult to perceive, so we should avoid using color in very small visual marks in visualizations; the perceived size also varies with color.

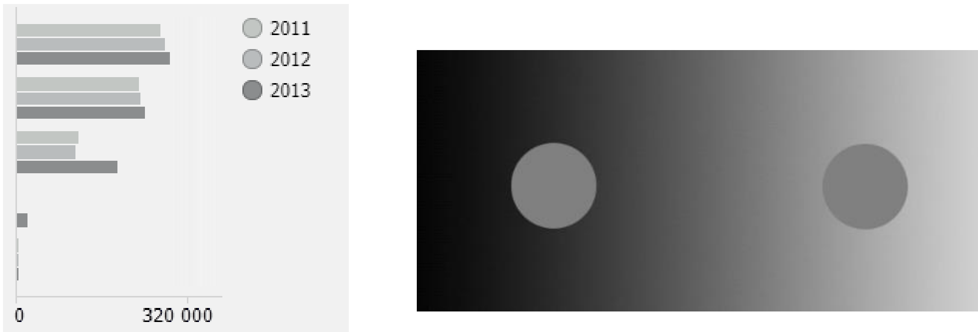

Figure 4. Examples illustrating the Zuk et al. heuristics 1 and 4: on the left, some bars are too small and almost indistinguishable; the left circle seems lighter than the circle on the right, while they have the same colour.

4: Local contrast affects the perception of color and gray levels – local contrast affects the perception of color due to its role in the process of color constancy (the ability to perceive the color of an object consistently under various lighting conditions); this implies that the same color is also perceived differently according to the background it is on (including gray levels, as we can see in figure 16.4 (on the right); we must, therefore, take this aspect into account when designing a visualization.

5: Consider people with color perception problems – there is a significant percentage of users with this type of problem (the most common is color blindness, which implies difficulty in distinguishing greens and reds), so we should avoid conveying information only through color, or, alternatively, be careful to use a simulator (for example Coblis[5]) to check how the visualization is perceived by users with these characteristics.

[5] https://www.color-blindness.com/coblis-color-blindness-simulator/

6: Pre-attentive processing increases with the field of vision – Pre-attentive processing corresponds to the process that occurs in the subconscious in which important information is filtered, making it possible to take advantage of this type of rapid processing of visual stimuli in visualization design. Spontaneously perceiving visual characteristics (such as color, shape or movement), i.e., without requiring conscious analysis, becomes easier when the field of vision is larger.

7: Quantitative assessment requires variation in position or size – so that users can perceive a quantitative value, with the maximum possible precision, from a visual representation, it must be represented by the length of a visual mark or the position along of an axis. Figure 5 shows the results of a widely used study on the relative difficulty of evaluating a quantitative value as a function of the coding mechanism.

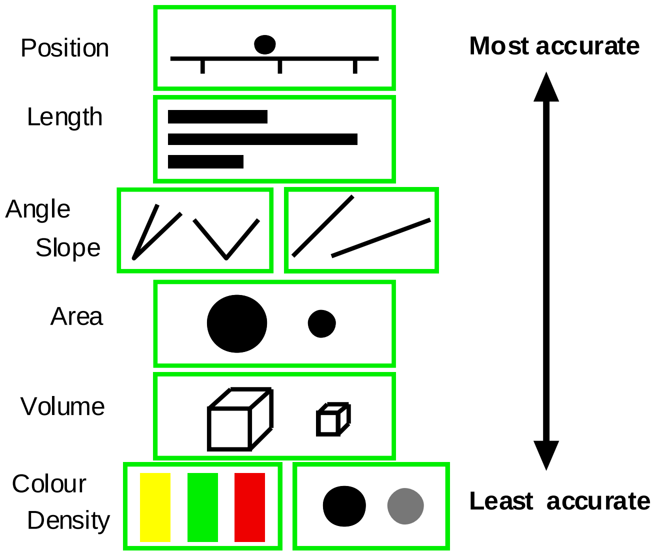

Figure 5. Relative accuracy of quantitative values visual encoding methods according to the study by Cleveland and McGill (adapted from Spence, 2014).

8: Preserve data dimensionality in visual representation – maintaining data dimensionality helps to understand it, allowing you to obtain meaningful insights and make informed decisions; In general, we should only use three-dimensional representations if the data is also three-dimensional (for example, data obtained from a patient's Computed Tomography); in other situations, in general, we should avoid it.

9: Represent the greatest amount of data in less space – condensing high-dimensional data into a smaller space, keeping as much information as possible (i.e., dimensionality reduction), aims to reduce the complexity of the visual representation without losing critical patterns, present in the original dataset.

10: Remove the superfluous – space in a visualization is an important asset and must be used well; irrelevant or rarely necessary information competes with relevant information and reduces its relative visibility; you must carefully choose what is represented in the visualization; allowing faster insights and better understanding of data, facilitating decision making, as the most important information is presented more clearly.

11: Consider Gestalt laws – Gestalt laws[6] are principles that describe human visual perception and how similar elements are grouped, patterns recognized and complex images understood; are useful for organizing visual elements in a visualization so that it is easy to understand (for example, the law of Proximity implies that you must place the name of an attribute close to the axis on which it is represented so that it is easily understandable that it refers to this axis).

12: Provide multiple levels of detail – the ability to analyze data with different levels of detail allows you to support diverse analytical needs and facilitate a deeper understanding of the data, allowing users, for example, to identify patterns and recognize relationships that may not be immediately apparent. apparent at first analysis.

13: Integrate text whenever relevant – text, for example in the form of visualization titles, as well as axis captions, is essential for a visualization to be understandable; we must use carefully selected notes; Annotations are informative elements that provide valuable context, explanation, and highlighting. They are essential for improving the effectiveness, clarity and overall impact of visualizations, allowing for better understanding and interpretation of data and phenomena.

---

[6] https://www.interaction-design.org/literature/topics/gestalt-principles

1.2 Cognitive Walkthrough

While Heuristic Evaluation is probably the method that has been most used to analytically evaluate a visualization or a visual data exploration application, there are other methods that can easily be adapted, such as Cognitive Walkthrough, or the use of indices based on the human visual and cognitive systems characteristics.

Cognitive walkthrough[7] is a task-based usability inspection method that involves a multidisciplinary team of evaluators (including usability experts and programmers, among others) who go through each step of a task flow and answer a set of questions, with the objective of identifying aspects of the user interface that may present difficulties for users. This method is especially suitable for evaluating ease of learning, so it can be used with benefit to evaluate visualizations and applications that should be especially intuitive and easy to learn to use. In its original version, it included four questions that should be answered at each step of the task under analysis; however, a simpler version, the Streamlined Cognitive Walkthrough, includes only two questions:

1. Will the user know what to do in this step?

2. If the user performs the correct action, will they be able to recognize that they are progressing towards their goal?

The team takes note of situations in which the answers are not affirmative to these aspects, which need to be improved. Like heuristic evaluation, cognitive walkthrough allows you to find problems at an early stage of development, without having to organize a usability test, which can be expensive and time-consuming. It corresponds to an analysis more focused on the ease of learning of just a set of tasks to be performed by a user, while heuristic evaluation corresponds to a more general analysis of usability. Note that the use of these analytical methods can be very useful, but does not eliminate the use of empirical methods, as previously mentioned.

---

[7] https://www.nngroup.com/articles/cognitive-walkthroughs/
https://www.interaction-design.org/literature/article/how-to-conduct-a-cognitive-walkthrough

2. Empirical Methods

Empirical usability evaluation methods involve participants who are as representative of the target users as possible. We can divide them into Survey methods (Interviews or Questionnaires), Observation, Utility or Usability Tests and Controlled Experiments. The first two types are often used together in usability testing and controlled experiments, where we ask participants to use a visualization or application, while we observe them and ask them some questions, usually when they have finished using it. However, they can also be used in isolation, in this case, more frequently, within studies to establish usability requirements or evaluate usefulness. In visualization, since solutions are very often developed to support domain experts in very specific situations, it becomes especially interesting to carry out utility tests to find out whether the solution can be used with benefit by these users.

We can consider that usability tests are generally used in an “engineering approach”, to improve and evaluate a solution and controlled experiments, whose planning and implementation is more complex, are mainly used in a research context to test hypotheses (for example, whether visualization technique A is more suitable than B, for specific users to perform a certain task, in a certain context). Empirical methods allow direct and valuable information to be obtained about how users use the solution under development and the problems they encounter; however, this does not mean that analytical methods are not used; As we mentioned previously, we should desirably use both types of methods, as they are complementary.

Below we briefly describe the main empirical methods in their most used variants in Visualization. It is important not to forget that studies involving users involve issues of ethics and data privacy that must be taken care of by consulting an Ethics Committee.

2.1 Inquiry

We can divide survey methods into Interviews and Questionnaires; In either case, they must be carefully prepared so that at the end of the process it is possible to obtain results that answer the questions that motivated the questionnaire or interviews. These methods are widely used in Social Sciences and there is an extensive bibliography on how to prepare and analyze the data collected, such as the book by G. Larossi (2011) [8]. Questionnaires have the advantage of being able to easily reach a large number of participants; however, interviews, although longer, are more flexible, being easily adaptable to the particular circumstances of each participant. In Visualization, these methods are often used to identify requirements, to understand the problem and the target users. They are also used in evaluation, more frequently integrated into user tests and controlled experiments; however, questionnaires can be used to evaluate views with the great advantage of being sent to participants by email and posted on a website. They can also be used on crowdsourcing platforms, with studies suggesting that this type of approach can produce useful assessment results in a short time.

There are questionnaires for usability evaluation, already widely used and with well-known results, generic enough to also be used with benefit in the evaluation of visual data exploration applications. We mention two well-known ones: the System Usability Scale (SUS) and the Questionnaire for User Interface Satisfaction (QUIS)[9]. Both must be answered by participants after using the application and include a set of questions that ask users' opinions regarding various aspects (ease of use and learning, as well as satisfaction). Answers are given on an ascending scale of 1 to 10 and allow a generic indication of the usability of the application to be obtained.

2.2 Observation

Observation is a method widely used in usability testing and controlled experiments, also in Visualization. As the name implies, it involves observing participants, usually

---

[8] The Power of Survey Design: A User's Guide for Managing Surveys ... - Giuseppe Larossi

[9] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

https://digital.ahrq.gov/health-it-tools-and-resources/evaluation-resources/workflow-assessment-health-it-toolkit/all-workflow-tools/questionnaire while they perform tasks with the help of visualizations or the application under evaluation. There are several types of observation, more or less complex and expensive to implement, with advantages and disadvantages that must be weighed before using observation in specific situations. A simple possibility consists of direct observation, in which the observer observes, as discreetly as possible, what the participant is doing and takes note of relevant aspects for later analysis. This way of observing is simple, but the observer must be careful not to interfere, except when strictly necessary (for example, when the participant requests help or is having great difficulty in carrying out the task, in order not to waste time unnecessarily). Indirect observation through a glass mirrored on only one side, as found in some usability laboratories, has the advantage of allowing the observation of the more natural behavior of participants who, not seeing the observers, more easily forget that they are watching. be observed. It is also possible to record participants' behavior on video for later analysis. Another form of observation that can help obtain more relevant information is to ask the participant to “think out loud”, that is, to justify what they are doing. However, it has the disadvantage of changing the participant's behavior and, sometimes, what they are transmitting is what they are doing and not why they are doing it. It is also possible to use a combination of several of the ways described.

In any case, it will always be necessary to inform and request permission from participants to be observed. The so-called “informed consent” is a fundamental step in any study with participants and should be one of the first steps in this type of study.

2.3 Utility tests

As we mentioned earlier, in visualization, solutions are very often developed to support domain experts (e.g. doctors, engineers or scientists) in very specific situations. As these solutions are tailor-made for these situations, evaluating the extent to which they actually allow you to provide relevant information is extremely important. It is, therefore, a good strategy to conduct utility tests to find out whether the solution can be used with benefit by its intended users. These tests, being focused on evaluating the practical value of an application or visualization technique to support users' goals, generally include the use of observation methods and surveys of domain experts who are asked to explore the solution and provide feedback. They may also involve trying to understand whether the solution provides complex, qualitative, unexpected and relevant insights from the visual analysis of the data. The tests are carried out in a case study format, individually, in which the application is presented to the user, who is instructed to use the "think aloud" protocol. In this way, when interacting with the visualization, in a freer or more structured way, domain experts verbalize insights, frustrations, eventual discovery of new patterns or verification of context knowledge. As the results of utility tests are predominantly qualitative, their analysis must consolidate and harmonize the different case studies, reflecting positive points, needs and gaps in the solution.

These tests allow you to better understand the practical benefits and limitations of the solution, ultimately guiding improvements in design and implementation that lead to better results for target users. They additionally complement other forms of evaluation, as they provide a measure of how useful the visualization or application actually is, validating its relevance to the context in which it was created.

2.4 Usability tests

So-called Usability Tests are fundamental in the evaluation of interactive systems in general, as well as applications involving visualization. They allow you to detect problems that might otherwise not be identified. They aim to test the ease of use of a final design or solution with a group of participants representative of the target users. They usually involve observing participants as they perform specific tasks[10] and can be carried out based on prototypes with increasing degrees of fidelity throughout the development process. They provide insights into users' goals, preferences, as well as difficulties; They help reduce development costs and adapt the solution to users, increasing satisfaction and accessibility. They are crucial for determining usability and

[10] https://www.interaction-design.org/literature/topics/usability-testing

ensuring that it ends up meeting users' expectations. They can be used in formative assessment, as they allow us to understand which improvements are necessary, or in summative assessment, to determine the degree of usability, for example to compare alternatives or to ensure that usability requirements are verified.

The main objectives of a usability test are to determine whether users can complete tasks successfully and independently; evaluate their performance and level of satisfaction while trying to complete tasks, identify problems and their severity, as well as find solutions to these problems.

How to perform a usability test

Carrying out a test to evaluate the usability of a visualization application involves a set of steps; Firstly, it will be necessary to carefully plan the test, defining exactly what you want to evaluate, identifying the target users and defining the usability measures to be used and how to collect the corresponding data; At this stage we must also decide whether to take place in person or remotely, what tasks the participants will need to carry out, identify the necessary resources, the number of participants and how to recruit them, as well as the documentation to prepare. It is very important to carry out a pilot test to rehearse the entire test (for example, check that the tasks are understandable and the documentation is complete and contains no errors); The results of the pilot test must be used to correct any problems detected, before carrying out the usability test with participants. After carrying out the test, it will be necessary to analyze the recorded data to report the results that should allow the application development team to implement improvements that can resolve identified problems. This process may have to be repeated more than once to validate the improvements made. In what follows we present relevant aspects to consider in the phases of a usability test [11].

Participants

The correct selection of participants is a fundamental aspect for us to obtain significant results; We should look for participants who are representative of the target users,

[11] https://www.interaction-design.org/literature/article/the-basics-of-recruiting-users-for-usability-testing

considering factors such as age, gender, education and experience in the application domain and with visualization. Recruiting appropriate participants can be complex and we must use criteria to include or exclude potential participants based on relevant characteristics (e.g., to identify potential participants who have vision problems, lack knowledge of the application domain, or be involved in the design or application development) to avoid biased results. Sample size is also important; In general, it is considered that between 3 and 5 will be an adequate number of participants, in most cases, to discover the main usability problems[12].. If statistical significance is desired, a much higher number of participants will be needed, possibly several dozen, depending on what is intended to be measured); Note, however, that usability tests generally do not have this objective. The compensation to be offered to participants is also an important aspect; It can be just symbolic (like a small gift or a snack), or monetary, based on the time spent and salary level of the participants.

Tasks to give to participants

When selecting tasks for usability testing of a visualization application, we must create realistic scenarios based on the target users' workflow, specifying the context and including all necessary details; however, we should not give too specific instructions such as “select option x”, or follow “link y”, to avoid users being directed to the answer. Since it is not possible to include tasks corresponding to all aspects of the application in a test, we must choose them carefully, and it is a good strategy to include tasks that test the most important functionality and aspects that have been considered potentially problematic, for example in a prior analytical assessment. Usability testing should begin with a period of familiarization and simple tasks and we should provide one task at a time so that participants are able to complete the tasks and do not become disoriented.

[12] https://www.usability.gov/how-to-and-tools/methods/recruiting-usability-test-participants.html

Usability measures

We can use several common measures to evaluate different aspects of viewing and interaction[13]; In general, measures of effectiveness, efficiency and satisfaction are used. Some of the most used effectiveness measures are completion rate (percentage of users who completed tasks); the success rate (percentage of participants who successfully completed the task); the error rate (number of errors participants made while carrying out the task); types of error (provide information on aspects that are not very usable).

Some of the most commonly used efficiency measures are task time (time spent by users to complete tasks) and number of selections or links followed (measures user interactions while completing tasks).

We can use several common measures to evaluate different aspects of viewing and interaction; In general, measures of effectiveness, efficiency and satisfaction are used. Some of the most used effectiveness measures are completion rate (percentage of users who completed tasks); the success rate (percentage of participants who successfully completed the task); the error rate (number of errors participants made while carrying out the task); types of error (provide information on aspects that are not very usable).

Some of the most commonly used efficiency measures are task time (time spent by users to complete tasks) and number of selections or links followed (measures user interactions while performing tasks).

Satisfaction measures correspond to opinions, and it is also possible to record users' feelings and emotional responses in relation to the application. As we mentioned previously, there are well-known questionnaires that allow users to obtain a general assessment of the usability of the application.

Protocol and documentation

When moderating a usability test, it's crucial that we follow best practices to ensure the session runs smoothly and provides valuable insights. We must warmly welcome participants and create a pleasant environment during the test, to ensure that they feel

[13] https://www.nngroup.com/articles/usability-metrics/

comfortable during the session; We must clearly communicate the purpose of the usability test, the tasks involved and what is expected of the participant and ask for their informed consent. While conducting the test, the moderator must maintain neutrality, providing clear instructions to avoid influencing participants and maintaining objectivity throughout the session. It may be a good option to encourage them to express their thoughts and actions aloud while carrying out tasks to better understand their decision-making process; however, as we mentioned previously, in addition to being able to change natural behavior, sometimes what participants say may not be very useful. It is good practice to debrief participants after each task to get immediate feedback or resolve any issues. It is also important to ensure that sessions do not exceed the agreed duration, respecting the participants' time.

The documentation to be prepared may involve, in addition to informed consent and a list of tasks for participants, a guide for moderators. This last document must include precise instructions on what to communicate to participants at the different stages of the test (for example at reception), record during observation (for example types of errors made), questions to ask during the interview, or questionnaires that must be handed out to the participants. All of this documentation must be carefully prepared, reviewed and tested before being used in a test with participants.

Final report

The final report must include a summary of the context and objectives, methodology, results and recommendations. There are several report templates that can be easily adapted. Context describes what was tested (e.g. a website or application), where and when the test was performed, the equipment used, what was done during the test (including documentation), the test team and a brief description of the problems encountered, as well as what worked well. The methodology used must be described in such a way that it is possible to replicate the test. We must describe the protocol used in the testing sessions, the tasks that users performed and the usability measurements collected. The report must also include the relevant characteristics of the participants and tables summarizing the responses to the questionnaires, never identifying the participants, as well as an analysis of what the experimenters recorded through observation or interviews. Tasks that had the highest and lowest completion rates should be identified, including a summary of effectiveness and efficiency measures such as successful completion rates and times per task. We must also include summaries of satisfaction results; Illustrative comments from participants may also be included. The data collected must be made available in some form and visualizations used to present the most important aspects. The report should include as specific a description as possible of the usability problems found, as well as recommendations for their solution. Although most usability test reports focus on problems, it is also useful to report positive results. What works well must be maintained or improved.

Carefully planning and executing all aspects of a test is the best way to obtain relevant, representative results that improve the usability of the application and the user experience of future target users.

2.5 Controlled experiments

Controlled Experiments are a fundamental instrument of experimental science (for example, in Physics, Chemistry, or Biology); are used extensively to test and validate hypotheses formulated based on prior knowledge. They are also used in research in the area of Visualization; in this case, involving participants. While usability testing is generally used in a formative evaluation context, controlled experiments are primarily used for summative usability evaluation. Although both are empirical evaluation methods, with many aspects in common, there are significant differences between controlled experiments and usability tests in terms of complexity, nature, objectives and participants, as well as data collection and analysis. Usability tests focus on evaluating how well participants can perform tasks with an application, with the aim of improving it, involving a relatively small number of participants. On the other hand, controlled experiments are more complex; are conducted under carefully controlled conditions by manipulating independent variables, while keeping others constant, defining various experimental conditions. Typically, they involve a much higher number of participants (several dozen), the measurement of dependent variables and analysis of the data to obtain statistical significance of the results. When planning a controlled experiment we can use different experimental designs, according to how participants use the different experimental conditions: within groups, when all participants use all experimental conditions, and between groups, when each participant uses only one experimental condition. We can also use some combination of these designs. Despite these differences, the aspects mentioned above remain relevant for usability tests, and in this case, they must also be carefully planned based on the objectives, which will generally be to test a set of hypotheses. We must meticulously define the independent variables (those that we can control) and dependent variables (those that we measure), as well as the secondary variables (that we cannot control, but can influence the result), the tasks to be asked of the users, the experimental protocol, including the documentation to support experimenters and participants during the experiment, as well as the final report. Carrying out a pilot test becomes even more vital than in the case of usability tests, as many more participants will be involved in a more complex process. The analysis of the data obtained also becomes much more complex, requiring specific knowledge of data analysis and statistics. It is especially important to have a specialist in data analysis and statistics on the team who, in addition to analyzing the data, also monitors the planning of the experiment to ensure that it will be possible to collect data that can produce relevant results. There is an extensive bibliography on planning experiments that can be consulted to delve deeper into the topic; the book by Douglas Montgomery (2019) is a good reference.

Summary

We have briefly addressed the importance of evaluation in Visualization and described the main usability evaluation methods that have been used, both in the development of solutions and in a research context, in Visualization. These methods can involve participants or not, and it is good practice to use methods of both types, as they can provide different and complementary information. We describe in more detail two methods whose use we strongly recommend in the process of developing a Visualization solution: Heuristic Evaluation and Usability Testing. We also leave reading suggestions to delve deeper into the topic, which is complex.

Main bibliography

Carpendale, S. (2008). Evaluating Information Visualizations. In A. et al. Karren (Ed.), *Information Visualization, Human-centered issues and perspectives* (pp. 19–45). Springer.

Fisher, D. and Meyer, M. (2018). *Making data Visible*, O'Reilly.

Gonçalves, D., Fonseca, M. J., Campos, P. (2017). *Introdução ao Design de Interfaces*, 3ª ed., FCA.

Kirk, A. (2019). *Data Visualisation A Handbook for Data Driven Design*, 2ª ed., Sage.

Kirk, A. (2012). *Data Visualization: A successful design process*. Pack Publishing.

Larossi, G. (2011). *O Poder da Concepção em Inquéritos por Questionário*, Fundação Calouste Gulbenkian.

Montgomery, D. (2019). *Design and Analysis of Experiments*, 10ª ed., Wiley.

Molich, R., and Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM*, 33, 3 (March), pp. 338-348.

Munzner, T. (2014). *Visualization Analysis and Design*. A K Peters/CRC Press.

Telea, A. (2015). *Data Visualization, Principles and Practice* (2nd ed.). Routledge.

Ward, M., G. Grinstein, D. Keim (2010). *Interactive Data Visualization*. A K Peters/CRC Press.

Ware, C. (2013). *Information Visualization, Perception to Design*, 3nd ed., Morgan Kaufmann.

Errey, N., Liang, J., Leong, T. W., & Zowghi, D. (2023). Evaluating narrative visualization: a survey of practitioners. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-023-00394-9

Zuk, T., Schlesier, L., Neumann, P., Hancock, M. S., & Carpendale, S. (2006). Heuristics for Information Visualization Evaluation. First Workshop on Beyond Time and Errors Novel Evaluation Methods for Visualization BELIV'06, 1–6.