Chapter 8
Evaluations

Systems that employ visual representations of information are thought of as being used by a particular category of users who have to carry out a specific task in a determined context. It is therefore a good idea to evaluate how these systems affect their users. An evaluation should provide the designer of an application with the data essential for understanding if, and under what conditions, it satisfies the users' needs, if it responds to their expectations and if users can effectively draw some benefit from the activity. A serious and rigorous evaluation is essential in the development process of a system that uses visual representations. Unfortunately, very often assessment is only marginally considered or even omitted. We believe, however, that the assessment of this type of system is of fundamental importance. A correct evaluation with the final users of the system can reveal potential problems and indicate which actions must be carried out to improve the quality. For this reason, we have dedicated an entire chapter to this, too often neglected, activity.

8.1 Human–Computer Interaction

Those who create systems that use visual representations, just like anyone who creates any type of software that presents a user interface, have to answer to the users of their own application sooner or later. At the beginning of the 1980s, when the first personal computers began to circulate outside universities and military research laboratories, the software available required a considerable level of competence to use. The user interfaces were complicated, as very few reflected humans communication methods, but, above all, because of the little care given to the aspect of user interaction. Interfaces of the early software constrained the users to adapt to the system, rather than having the system to adapt to the operating mode of the users.

The problem became evident primarily when the software started to be used by regular people, and not only by the professionals in the data centers. To study the problem from a scientific point of view, a discipline called *Human–Computer Inter-*

R. Mazza, *Introduction to Information Visualization*,
DOI: 10.1007/978-1-84800-219-7_8,

action (HCI) came about in the 1980s. This discipline uses analytic and empirical techniques to evaluate the effects of user interaction with computers.

A typical process of evaluation of a system in HCI usually has the following objectives [16]:

- *to assess the functionality of the system*, which means verifying that the system fulfills all of the functions requested by the user and defined in the phase of user requirements specification;
- *to analyze the effects of the system on the final users*, through a methodology that evaluates the aspects linked to the human factors, such as usability of the graphical interface, simplicity, and level of acceptance by the users;
- *to identify every possible problem that could arise with the final users of the system*, such as preventing an unpredicted result or anything that could be misleading to the users.

The evaluation of a system can be carried out during the design phase of an application or with a functional prototype. In the first case, we speak of *formative evaluation*, directed at identifying potential problems and indicating how to possibly improve the system design. In the second case, we speak of *summative evaluation*, which is often carried out with a sample of final users using a prototype of the system, to identify possible improvements to be applied in the final version of the system.

8.2 Evaluation Criteria

Engineering's best practices teach us that, before realizing any artifact, it's necessary to have performed a minimum of rational design of the system to be produced. This principle is valid in many disciplines: Before building a bridge, the engineer has to plan carefully, performing all the static and dynamic calculations to avoid having the bridge collapse when cars cross or in the presence of wind. Luckily, the design of a visual representation is not as critical as the building of a bridge. It is precisely this that causes many programmers to develop a visual representation without even the minimum of preliminary design. But, to avoid wasting time (and money), rigorous planning is essential for a successful project.

In the case of applications based on visual representations, two phases are very important: the specification of the requirements and the evaluation. All serious projects should start with a rigorous specification of the requirements, collected from potential users of the system through interviews, questionnaires, etc. Before starting a project, it is necessary to know your goals. Even the evaluation has to be carefully planned, and in the project phase we should have also plan the evaluation strategy. The evaluation should at least consider whether the product meets the specific requirements of the users, but that alone is not sufficient: The product must be effective and efficient and serve a purpose before it will be adopted by the final user.

Let's suppose, for example, that we have to evaluate the visual representation of data collected in a discussion forum, a representation that we implemented in Section 2.1 and is shown in Fig. 2.6. For this representation we could define an analysis in which the following criteria are evaluated:

Functionality. Does the visual representation provide all of the functionalities requested by the instructors and identified during the requirements elicitation?

Effectiveness. Does the visual representation provide the instructors with a better knowledge of the number of messages read and written in a discussion forum than the traditional interfaces provided by the tool? In particular, does the use of visual representations allow the instructors to have information on the number of messages sent and read with better accuracy and precision than other tools? Or, is there additional information that is made available exclusively by the visual representations?

Efficiency. Can the visual representation provide the instructors with information more rapidly than the tools provided by the system?

Usability. Is the interaction with the graphical interface simple and intuitive enough for the instructors?

Usefulness. In what way, and in what context, is the information provided by the graphical representation useful to the instructors?

The criteria just listed are the main objects of study in an evaluation process. However, depending on the type of application, and to limit the resources dedicated to the evaluation (which can be very long and expensive), one can limit the evaluation to only a subset of this criteria. For example, in some contexts it can be appropriate to evaluate the functionality, usability, and effectiveness of a representation, assuming that the usefulness is intrinsically derived from the new functionalities provided by the system and that the efficiency is not a critical factor to success. In the following sections, we will see some techniques in which the above-mentioned criteria can be evaluated.

8.3 Evaluating Visual Representations

The evaluation of systems that make use of visual representations, just like other systems involving direct interaction with humans, is an extremely complex task. In particular, it is very difficult to create an evaluation model that gives an objective judgment of the effectiveness and usefulness of a certain type of visualization. Two users placed in front of the same visual representation could express completely different and contrasting judgments. Experience, prior knowledge, and perceptive and cognitive ability may differ from person to person, which can bring about discord in judgment.

Chaomei Chen on his inspiring article on top 10 unsolved information visualization problems [9] put as the first three problems issues related to human factors (namely: usability issues, understanding elementary perceptual-cognitive tasks, prior knowledge), highlighting that there is still a lot of work to do on defining evaluation methods that involve real users and perceptual-cognitive tasks.

Diferent from a common user interface, a system that uses visual representations must be evaluated not only in terms of the usability and effectiveness of the interface, but also for the information that it manages to communicate to the users through perceptual and cognitive processes. For example, a crucial aspect could be to comprehend if users manage to decode the graphically codified information, if they can recognize visual patterns, if they manage to identify "interesting" values and elements, etc.

Not having their own evaluation methodology, systems that use visual representations have adopted techniques from human–computer interaction. These techniques, which have been in use for years, can essentially be subdivided into two categories: analytic evaluations and empirical evaluations.

8.3.1 Analytic Methods

Analytic evaluation methods come from psychological models of human–machine interaction and are mainly based on cognitive and behavioral studies. This type of evaluation is carried out by experts who verify whether a certain system is compliant with a series of principles called *heuristics* (from which the name *heuristic evaluation* originates). For example, some heuristics have been defined on the principles of the usability and accessibility of the graphical interface for common use applications. An evaluator use the system and judge its compliance with the heuristics.

Another type of test carried out by experts is called a *cognitive walkthrough*: An evaluator defines a series of possible *scenarios* of use and simulates the behavior of a user who uses the system to perform predetermined tasks. During the use, the evaluator has to identify possible problems that could originate from every task.

These types of evaluations are often used to judge the usability of the interfaces of the software systems, particularly in the initial phases of development, to identify possible problems and indicate modifications to improve the aspect of the interaction with the user. However, because of the difficulty in defining a series of heuristics for visual representations, these techniques are rarely adopted in information visualization.

8.3.2 Empirical Methods

Empirical evaluation methods make use of experiments that make use of functioning prototypes of systems and involve the final users of the application. Experiments can be divided into *quantitative studies* and *qualitative studies* [56], based on the type of data collected. The technique used to collect quantitative data is the controlled experiment, while, for qualitative data, we have a wider range of options at our disposal, including interviews with users, direct observations, and focus groups.

8.3.2.1 Controlled Experiments

Examples of quantitative studies are the *controlled experiments* (also called *experimental studies*), defined in detail in some classic HCI texts such as [30]. These experiments aim to evaluate a certain system property by verifying a series of *hypotheses*, which can be confirmed (or infirmed) through a series of variables that can be measured quantitatively during the experiments performed by the users. The experiments have to be performed in a "controlled" environment, meaning that the person that coordinates the experiment has to systematically manipulate one or more conditions of the experiment (called *independent variables*), to study the effect of this change on other variables (*dependent variables*). The experiments should be conducted with a representative sample of users: the *test users*. During the experiment, the test users are asked to use the prototype to carry out a particular operation; at the same time, a series of "measurements" are carried out directly on the prototype or by observers (for example, notes are taken of the time required to complete a task, or the performance on accomplishing a specific task).

Controlled experiments can be useful for evaluating the functionality, effectiveness, and efficiency of a visual representation. Again, using the example that represents the messages exchanged between students of an online course (Fig. 2.6), to demonstrate that the visual representation is effective, the following hypothesis can be formulated:

> Lecturers of an online course that use the proposed graphical representations have a better knowledge of which students (1) are more active in posting messages on the forum, (2) read all the messages but don't actively participate in the discussions and (3) neither read nor write messages in the forum, compared to lecturers who use the traditional interfaces provided with the e-learning system without the support of visual representations.

To demonstrate this hypothesis, it is necessary to perform a series of experiments in which a number of dependent variables are measured. These variables have to be directly linked to the hypothesis that we wish to verify. We can define the following **dependent variables**:

1. knowledge of the students who are more active in initiating new *threads of discussion*;
2. knowledge of the students who have read the majority of the messages without consistently taking part in posting new messages;
3. knowledge of the students who have contributed to neither the reading nor the writing of messages in the forum.

To carry out a controlled experiment, the test users are split into two groups. One group uses the interfaces provided by the e-learning system, while the other uses the same interfaces but are aided by the visual representation that is being evaluated. Both groups of test users are requested to perform the same operations; the users' performance on the dependent variables is analyzed. The performance must be numerically measurable to be able to treat it with a statistical approach (calculating, for example, the average, variance, and standard deviation). Examples of performance can be the grade of accuracy with which each user responds to the questions or completes a particular operation (usually it is encoded with a real number between 0 and 1) or, in cases in which it is necessary to measure efficiency, the performance could be the time (in seconds) needed to answer a question or accomplish a specific task.

Another variant of controlled experiment involves two (or more) alternative solutions of graphical representations being compared empirically. This is useful when the project being worked on produces more than one solutions. In this case, the test users are divided into a number of groups, each group use a specific variant, and all groups carry out the same operations. The final analysis on the values measured in the empirical test indicate which proposal to choose.

8.3.2.2 Qualitative Methods

Qualitative evaluation methods are based on the collection of qualitative data from the test users, obtainable through questionnaires, interviews, and user monitoring. Qualitative methods differ from quantitative ones, described in the previous section, for the ability to analyze the phenomena from the user's point of view, rather than elaborating values measured in experiments. With qualitative methods, it is possible to evaluate the usefulness of a certain representation. Among these methods we can mention the *users' observation*, the *collection of the users' opinions*, and, finally, the *focus group*.

The users' observation consists of asking a certain number of test users to use the application's prototype and observe how users interact with it. Users can be asked to carry out some tasks, or to respond to a certain number of questions. In contrast to the controlled experiments, the aim of the observation consists of identifying possible problems that can rise when using the system; for example, a functionality that is not very clear or a certain visual mapping that could be interpreted incorrectly. This method involves a very expensive and engaging verification, particularly, when a large number of users is participating in the test.

Controlled experiments and the users' observations can be helpful when evaluating the functionality, efficiency, and effectiveness of the visual representations, but they are inappropriate for revealing problems that can manifest during the observations and for eliciting information on preference, impressions, and attitudes. The only way to understand if a certain visual representation can be useful for a certain type of user is to ask the user explicitly. A representation that provides a certain type of information very effectively and efficiently but that is of no use to users doesn't serve any purpose. For this reason, the **collection of the users' opinions** is a very important empirical technique, is relatively convenient, and can be carried out in various ways through interviews and questionnaires. Interviews are one of the most often used evaluation techniques in the social sciences, in market research, and also generally for other reasons in HCI. The key to the success of an evaluation based on interviews lies in the ability of interviewers to capture the most interesting comments from users. For example, to evaluate a certain visual representation, a number of users could be asked to use the prototype of the system, perhaps for a certain period of time, long enough to acquire a certain familiarity with the application. Successively, these users could be asked a certain number of predetermined questions, to obtain comments on general impression of the tool, their opinions on the facility of its use, its usefulness, etc.

A **focus group** is a technique that can help to investigate group attitudes, feelings, and beliefs of users on a proposed visual representation through group interviews [42, 46, 24, 25]. The interviews are carried out by bringing together a sample selection of test users and discussing as a group the functionalities offered by a visual representation. The conversation is led by a moderator, whose role is to facilitate the discussion, stimulate the interaction among participants, and keep the discussion focused on the aspects of the representation to be evaluated, besides collecting all participants' comments. The interesting aspect of this technique is that through the discussions and group interaction, in which each participant brings his or her own competence and personal experience, attention can be drawn to problems and situations that hadn't been foreseen during the system design. Prior to organizing a focus group, it is necessary to plan the meeting in detail: A demo has to be prepared in which the details of the system's functionalities are shown (unless these are already known); a video and audio recording system must be provided to be able to analyze the dialogues and interactions afterwards; the group of test users to involve and the number of focus group sessions to activate must be considered; finally, a series of questions must be prepared. The composition of the group of participants is a vital aspect, in terms of both the number and type of people. The ideal would be to have a group of potential users with a heterogeneous background, so as to cover every possible type of user. The number of participants is also a vital factor: Specialists in focus groups suggest groups of not less than 4 and not more than 12 participants per session. On the other hand, it is not easy to retrieve test users for this type of experiment, for which the composition of the group of participants is a compromise among the number of test users available, the competence and background of these users, and, finally, practical and logistic aspects. As to the questions to ask the group, it is necessary to prepare a number of questions that will elicit a series of critical comments on the representation.

8.4 Conclusion

We have dedicated this final chapter to an often neglected activity that should be part of every visual representation project: the empirical evaluation of the system conducted with potential users. A correct evaluation can reveal potential problems and indicate which actions have to be carried out to improve the quality of the visual representation. Empirical evaluations can be performed in the form of quantitative studies (such as controlled experiments) and qualitative studies (such as interviews), with the aim of providing some feedback on the functionality, effectiveness, efficiency, and usefulness of a visual representation.