A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

Evaluation of multimedia applications using inspection methods: The Cognitive Walkthrough case

Julien Huart [a], Christophe Kolski [b], Mouldi Sagar[b]

[a] *Logica S.A., 30 place Salvador Allende, 59650 Villeneuve d'Ascq, France and Laboratoire des Sciences de la Communication, Le Mont Houy, University of Valenciennes, F-59313 Valenciennes cedex, France*

[b] *Laboratoire d'Automatique, de Mécanique et d'Informatique, industrielles et Humaines - UMR CNRS 8530, Le Mont Houy, University of Valenciennes, F-59313 Valenciennes cedex, France*

Abstract

Many evaluation methods are to be found in research literature: they can be formal, automatic, empirical or informal. The informal methods include so-called inspection methods, which provide a good compromise between the cost and implementation time on the one hand, and the results they make it possible to obtain on the other. Amongst these methods, Cognitive Walkthrough enables the detection of a certain number of usability defects and the estimation of the degree of seriousness of the defect. In this article, we concentrate on Cognitive Walkthrough. We are particularly interested in it because, as far as we know, it is the only method based on theory (the theory of learning through exploration, itself inspired by Norman's Action Theory). However, although its usefulness as regards software ergonomics has been recognised, its efficiency in the case of multimedia applications is still far from being proved and very few research projects have been published on the matter. In fact, multimedia documents have characteristics which differ from those of traditional human-machine systems. This article presents a study on the use of Cognitive Walkthrough for the evaluation of several multimedia applications intended for the general public; it reveals the difficulties met by users and the areas in which the method needs to be adapted.

Keywords: Evaluation, Cognitive Walkthrough, Multimedia Application.

1. Introduction

The general quality of current multimedia applications and web sites is often average: they do not really meet user requirements, they do not generally enable easy and efficient communication, they are not very entertaining and a great majority of them have usability problems (Huart et al., 1998; Nielsen, 1997; 1998). It is obvious that the authors of documents are still looking for a "multimedia expression mode" and that the "killer application" has not yet been created. However, the main cause of this phenomenon, which recurs whenever new technology appears (Gibbs, 1994), is to be found in the absence of production methodologies and of tools, and finally in the fact that applications are not evaluated during the design and development stages. It is therefore necessary to make the professionals aware of these problems and provide them with solutions in terms of methodologies and tools.

A particular effort is to be made in the field of evaluation, given that studies in the evaluation of usability show that even when tools exist, the idea of evaluation is not well accepted because it is judged to be expensive, time-consuming, difficult to implement… (Nielsen, 1994a). From this point of view, the user tests often appear to be like scarecrows. This is why inspection methods, which are relatively less cumbersome and costly to implement, could be the way to persuade professionals to adopt evaluation. Heuristic evaluation can be recommended in particular (Nielsen, 1990; 1994a), but at the present time there is a lack of heuristics specific to multimedia, despite the efforts made in this field (Smith, 1996; Bastien et al., 1998; ISO DIS 14915, 1999; Kemp and Buckner, 1999; Ziegler, 1999). This article presents the results of an ergonomic evaluation of multimedia documents for the general public, performed using the Cognitive Walkthrough inspection method

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

(Polson et al., 1992). The aim of the study was to assess the effectiveness of Cognitive Walkthrough in evaluating the usability of multimedia applications (without necessarily going into any one specific point of human-machine interaction; for example, no reference is made to ocular strategies – Mullin et al., 2001).

Section 2 attempts to introduce the general characteristics of multimedia applications. The framework for the evaluation of multimedia documents, the inspection methods and Cognitive Walkthrough (CW) are presented in section 3. Section 4 presents the results of the ergonomic evaluation of four multimedia documents (with different degrees of scenational interactivity). These documents have been evaluated by four experts (two with in-depth experience, two with less experience). The results show inter-individual differences between the different types of experts. Even if the use of CW is a source of relative difficulty for the experts, they recognise its utility and propose several ways for improving the method. Finally, section 5 analyses and comments on the efficiency of Cognitive Walkthrough for the evaluation of multimedia documents.

2. Characteristics of multimedia applications

Multimedia applications link images in two or three dimensions, video or sound sequences and animations with the traditional components of human-machine interfaces. Multimedia can be defined as *"a communication technology which tends to bring together all the pluri-sensorial and computer data into supports of the same type"* (Durand, 1997). Web sites and CD-ROMs intended for the general public such as *The Louvre*, *Paintings and Palace*, *Microsoft Art Gallery* and *Microsoft Encarta* are examples of these.

A document is *"an organised structure of information parts of the smallest level"* (Leleu-Merviel, 1997; Durand et al., 1997a) supported by a medium enabling its diffusion. The further characteristic of the multimedia document is the multiplicity of paths and the possibility for the user to intervene directly on the diffusion of the content.

The multimedia document, as an extension to a hypertext (Conklin, 1987), is also called a hyperdocument, that is to say *"an informative content made up of a nebula of fragments whose meaning is constructed using each of the paths determined by the reading"* (Leleu-Merviel, 1997). The notion of the hyperdocument thus assumes the separation between a multimedia document (which we will also call content) and its support (or medium).

The multimedia product therefore has the following characteristics:

- A computer component which supports the multimedia document and is a part of the product's data (the logical data),
- A pluri-sensorial component with media (perceptive data) such as videos, slide-shows, animations, and sounds,
- Finally, the presence of an interface between this logical data and the user: this adds an essential element - interactivity: "*Interactivity is composed of the exchange of data between the two structures of two different devices, one of which may be a human user*" (Balpe, 1997).

In terms of usability, the main aim of interactivity is to provide the users with a coherent cognitive guide, that is to say to enable them to reach their goals using the shortest and least costly path from the cognitive point of view. In the case of multimedia applications, interactivity can provide certain advanced reading possibilities to increase the comfort and entertainment given to the users and which, in addition to the cognitive guiding of the users, introduce a new notion characteristic of multimedia (Leleu-Merviel, 1997): the scenational schema refers precisely to *"a series of pre-planned events/states linked together by different paths for the user to take. In this way, the user can follow a different path with each interactive session"*. The schema is made up of a set of fragments taken from a scenario. The scenario is defined by an architecture of narrative macro-structures; it represents the « story » told in the document; The notion of fragment is taken from a very old document dealing with scenario design in cinema (Eisenstein, 1929).

In the case of multimedia applications, different possibilities for scenational schemas can be distinguished, assimilated to varying degrees of navigational freedom: for example, when the user is obliged to follow a path defined by the document designer using keys, or else when the user has to create his or her own path in the application (this is the case especially in three dimensional applications). By extension of the term interactivity defined above, Durand et al. (1997b) have therefore defined six degrees (also called "levels" by the authors) of scenational interactivity (Table 1). These degrees go from the linear reading of audio-visual type contents to the virtual visit; they are often linked to current multimedia documents, even though a degree of scenational interactivity is always preponderant. The reader will find further explanations concerning the notions of interactivity and degree of interactivity in (Wills, 1994) and (Dholakia et al., 2000).

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

Table 1
Degrees of scenational interactivity

DEGREE OF INTERACTIVITY	EXPLANATION
Audio-vision **(Degree #0)**	This degree corresponds to an absence of scenational interactivity. It is a non-interruptible linear diffusion schema. The users merely follow the presentation performed for them, and their only possible action is to stop (as one would leave a cinema for example). The introductions to certain cultural or game CD-ROMs correspond to this degree of interactivity (for example, *Paris, Virtual Visit, Microsoft Age of Empire*).
Reading **(Degree #1)**	The scenational schema of this degree is suitable for a linear interruptible diffusion. This type of interactivity can be compared to the use of a VCR: one can stop, rewind, watch a sequence again. Even though the structure remains sequential, the user is able to skip certain sections, as with a book. A certain number of CD-ROMs are organised in the form of slide-shows which correspond to this degree of interactivity (for example, the viewer in *CD-Photo Kodak*, or else the normal running of *MS Powerpoint* presentations).
Consultation **(Degree #2)**	The inseparable information units adopt a matrix-type indexed structure, each component of this structure corresponds to a specific item or identifier. Reading is therefore sequential by cell, each cell being identified by one – or several – indexes: this is the classic functioning found in audio CDs. The indexes provided by a great number of multimedia applications correspond to this degree. Web sites are situated between this degree of interactivity and the next.
Navigation **(Degree #3)**	The information units – which are still frozen and inseparable – can be linked together according to numerous and varied paths from which the user has to choose. Nevertheless, all of these paths have been pre-defined by the designer. Cultural CD-ROMS such as *The Louvre¸ Paintings and Palace*, and *Musée d'Orsay, virtual visit* correspond to this degree.
Exploration **(Degree #4)**	The reader is no longer guided by pre-defined paths, but structuring tools, such as link networks, make it possible to generate individualised paths which have not been defined by the designer. This is the case in a certain number of games such as *Civilization*, and in applications such as *Adibou*
Virtual visit **(Degree #5)**	Any modification of diffusion is linked to a reader action; if no action is carried out by the user, the programme remains frozen. The so-called "virtual reality" games such as *Rogue Squadron, Versailles, plotting in the royal court* are examples of this.

3. Inspection methods used as evaluation methods

Cognitive Walkthrough is one of the inspection methods with which an expert analyses and criticises the interactive system to be evaluated. Before presenting these methods, it is necessary to specify the framework of multimedia document evaluation.

3.1. Framework for multimedia document evaluation

The evaluation of a human-machine interface consists in checking and trying to validate it. In this sense, "any evaluation consists in comparing a model of the object evaluated to a reference model which makes it possible to draw conclusions" (Senach, 1990) (Fig. 1). In the case of the ergonomic evaluation of a human-machine interface, researchers are mainly concerned with its usefulness, which determines whether the interface enables the user to achieve his or her work aims, and especially with usability, which accounts for the quality of human-machine interaction in terms of ease of learning and use, as well as the quality of documentation. Since the beginning of the eighties, the notions of assessment objectives, usefulness and usability have given rise to various definitions which may sometimes appear to be contradictory (although they then gradually converge). These definitions and the surrounding arguments will not be discussed in this article ; the interested reader will find in (Shackel, 1991; Grudin, 1992; Nielsen, 1993; Scapin and bastien, 1997; Bastien and Scapin, 2001) more detailed information on these notions. The basic data allowing a comparison with the reference model is based, in this case, on ergonomic criteria (Bastien and Scapin, 1993; Bastien et al., 1999; Vanderdonckt, 1999).

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

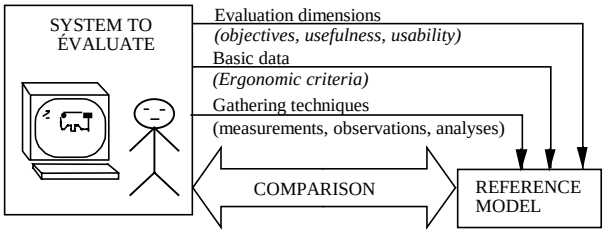

Fig. 1. Overall evaluation principle (according to Senach, 1990)

In addition, any evaluation must be based on the aspects considered to be relevant to it. Consequently, in the case of multimedia applications, the basic data and the dimensions of the evaluation differ from those necessary in a classic evaluation of a human-machine interface. Indeed, the multimedia document (independently from its support) conveys information through a presentation, a speech: as a communication tool, it must be the creator of meanings, the carrier of understandable information, it must provoke interest and action.

The use of static images, animated images and sounds also contributes a greater aesthetic and emotional dimension to the human-machine interaction: multimedia is in fact one of the first media to be able to rely on relational and rational signals. The presence, not only of images, but also of sounds and sometimes sensations acts on the subconscious and on the emotive character of individual users. The addition of the relational component is one of the notable points of multimedia interfaces, especially as regards classic human-machine interfaces which, even though some do attempt to be user-friendly, are above all intended to be functional. By emphasising the relational factor, multimedia interfaces implicate the person using them. They go beyond the simple, merely rational relationship.

Other properties of the image give it a certain importance in multimedia communication: the image is polysemic and can be interpreted differently according to the individual and the context; it has a power of persuasion and is often taken as being a proof. All of these properties form the underlying principle in interactions between the various media which constitute multimedia documents: this is especially the case of icons: in the new multimedia interfaces, even though some of the older icons are still used, three new data elements are added to the problem : firstly, thanks to technical progress and because of their very nature, icons using the 16x16 pixel or 32x32 pixel size are no longer sufficient for multimedia documents; secondly, an increasingly great proportion of the users of multimedia documents have no knowledge of micro-computing and therefore are not familiar with the symbols (such as the explorer, the desktop, etc.), obviously causing problems of usability. Finally, the creator of multimedia documents is no longer a specialist in the field of human-machine interface design and does not necessarily know (or wish to use) the interface design graphic standards; we should also add that norms and standards specific to multimedia remain rare and lacking in detail, and they are not easily accessible for the multimedia creator. Consequently, a rupture is often noticed between the multimedia applications and the ergonomic standards in current use, without necessarily altering the usability of the multimedia documents. The evaluation methods for such documents, which are generally based on existing standards, must therefore take this state of affairs into account.

The dimensions of evaluation must therefore be based especially on the information, its presentation, the logic of interaction, the structure with which the user is confronted, and also on usability. For example, Garzotto et al. (1995) suggest as evaluation dimensions the *content* (the pieces of information included in the application), the *structure* (organisation of the data offered to users), the *presentation* (how the application content and functions are shown to readers), the dynamics (how users interact with individual pieces of information) and the types of interaction (which concerns navigation through the document).

The criteria and basic data used in human-machine interface engineering are not longer sufficient as regards these dimensions. Some authors therefore suggest criteria relating to the association of media, their aesthetic aspect and the presentation modes (Kouroupetroglou et al., 1994; Mendes et al., 1998; Vanderdonckt, 1998; Leulier et al., 1998). Nevertheless, there are currently few classifications of evaluation criteria covering all of these dimensions, apart from the framework established recently by the ISO DIS 14915 (1999).

Finally, collection techniques and more specifically the evaluation methods for human-machine interfaces have limits as regards multimedia: they are intended to detect usability defects and can no longer be applied totally, or are so with reduced efficiency because they do not take the specificity of the multimedia field into account; moreover, they can no longer be based on heuristics and adapted recommendations. This is especially the case of inspection methods which are not particularly straightforward to implement as such. However, according to the type of multimedia application, in some cases the experience of the evaluators can no doubt compensate or extend the boundaries of certain evaluation methods.

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

3.2. Classification of evaluation methods

Several classifications of evaluations methods exist in research literature, for instance (Senach, 1990; Dix et al., 1993; Sweeney et al., 1993; Mack and Nielsen, 1994; Kolski, 1997). According to Mack and Nielsen (1994), there are, in all, four types of method to evaluate usability: the formal, automatic, empirical and informal methods:

- Formal methods use formal and metric models to take usability measurements; they are difficult to use and their usefulness has yet to be proved. For example Mc Call et al. (1997) and Forse (1989) have suggested qualimetry models for computing systems; in the case of multimedia documents, no such tool exists at the present time.
- The automatic methods use systems to take measurements, which are still limited for the moment. One of the first approaches is that of Tullis (1988) who suggested the "Display Analysis Program"; the automatic screen evaluation tool measures the six following parameters: the global density of information, the local density, the number of distinct information groups, the average size of the groups, the number of items, the display complexity. Other approaches are based on rules used by an inference motor to evaluate the presentation of information: SYNOP (Kolski and Millot, 1991), KRI/AG (Löwgren and Nordqvist, 1992) or ErgoVal (Farenc et al., 1996). In the case of web sites, a certain number of tools have been proposed recently in order to measure accessibility, readability and coherence of pages by analysing the HTML code (Layaïda and Keramane, 1995; Bowers, 1996; Scholtz and Laskowski, 1999; Cooper, 1999).
- The empirical methods include interviews, questionnaires and user testing, for which a panel of users try the product. These methods are potentially the best, ((Nielsen and Mack, 1994) indicates that an evaluation carried out with 5 users makes it possible to detect 80% of the usability problems), but it is difficult to select a representative sample and to recreate real conditions of use in the laboratory; moreover, problems connected with the time and cost of implementation are considerable. In the case of user tests, the users can be asked for example to think aloud, to use the interface in pairs - Paired-User Testing (Wildman, 1996), or to work directly with the designer: co-operative evaluation (Wright and Monk, 1991). Within the framework of multimedia application evaluation, several propositions have been made (Kouroupetroglou et al., 1994; Trigano, 1997; Scholtz and Downey, 1998, Hû and Trigano, 1999), often based on the classification of usability criteria, such as (Bastien and Scapin, 1993). For example, WAMMI (Website Analysis and MeasureMent Inventory) (Kirakowski and Cierlik, 1998), is an on-line questionnaire enabling web site users to give an overall mark for the usefulness and usability of the site. It is made up of an open question, along with twenty statements which the user marks on a 5-point scale, going from « I don’t agree at all » to «I totally agree ». The results established from the analysis of WAMMI questionnaires are interesting as they go beyond the measurement of usability appreciation. However, any improvement to a site still requires further systematic evaluations.
- The informal methods correspond to inspections based on heuristics (Nielsen, 1993), on recommendations (Bastien and Scapin, 1993; INA, 1994; Vanderdonckt, 1994) and on rules (Duffy et al., 1993; Ruokamo and Pohjolainen, 1998) and are performed solely by experts: although they are less expensive in terms of time and cost, they are also less efficient than empirical methods (Nielsen and Mack, 1994). They are generally used for prototypes, especially when the intervention of real users is not necessary or impossible (Grislin and Kolski, 1996). We concentrate on these methods below.

3.3. Inspection methods

The general aims of the inspection methods are to intervene early in the interface design procedure, to identify, qualify and quantify usability problems and finally to be incorporated into a usability life cycle (by making suggestions for repair, and for correction priorities, and by estimating the cost) (Virzi, 1997). Amongst these methods, the following can be noted:

- Heuristic evaluation, which is the most informal method: during the evaluation, usability specialists compare each element of dialogue with principles and heuristics (Nielsen, 1994b). Several authors have suggested heuristic tables with this in mind (Karat et al., 1992; Prümper, 1993).
- The Guidelines Review which enables usability experts to test the interface using a check-list. This method requires a high degree of expertise (Wixon et al., 1994).
- Pluralistic Usability Walkthroughs (Bias, 1994), which correspond to a meeting at which users, designers and usability experts go through the task scenario step by step.
- Consistency inspection (Karat, 1994): “neutral” designers compare an interface with their own design standards. This makes it possible to judge consistency between families of products.

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

- During the use of Standard inspections, an expert in an interface standard (Windows, Openlook, Macintosh…) checks that this standard is respected by the interface. This method makes it possible to improve the homogeneity of the interface in relation to interfaces on the market which use the same standard (Simpson, 1999).
- Formal Usability Inspections (Kahn and Prail, 1994) are the review of an interface performed by the designer of the product and a team of peers. It is organised in several stages (preparation, review, presentation of results, …) and with four profiles: a regulator who directs the meetings, a designer responsible for the design and modification of the interface, interface inspectors and a note-taker. The inspectors are instructed to present a list of defects detected which is debated under the leadership of the regulator, and then validated.
- Cognitive Walkthrough, based on Norman's action theory (Norman, 1986) provides a predictive approach for a human-machine interface expert (cf. below).

Most of the inspection methods are intended for use by experts: in this article, we shall concentrate on Cognitive Walkthrough, which is described in the following section. We have chosen this method because it is one of the few with an exploration-linked theoretical basis. It would therefore appear to have great potential for the evaluation of multimedia applications (in which the user often explores the document in question progressively).

4. Cognitive Walkthrough (CW)

Cognitive Walkthrough, described in detail by Polson et al. (1992), is one of the evaluation methods said to be based on a theory (Lewis, 1997). It is in fact linked to a Software Engineering method of the same name which consists in simulating code series in order to check whether they correspond to the implementation of the specified functions. The method is based on a theory of learning through exploration developed by Polson and Lewis, which was itself inspired by Norman's action theory (Norman, 1986).

The aim of the method is to evaluate the usability of a system and to enable the designer to find the causes of usability problems very early on in the design process (Abowd, 1995) without any user intervention (Lewis and Rieman, 1994). The analysis concentrates on 2 points : firstly, the ease with which a user can perform a task with a minimum of knowledge of the system; secondly, the ease of learning through exploration of the interface.

During a preliminary stage of preparation, the evaluator chooses the human tasks to be analysed; each task must be described and associated to a sequence of actions. The targeted population group must be identified by associating it to basic characteristics which could influence the evaluation validity considerably. The initial goals of the user are described. The evaluation can then begin.

For the evaluator, the implementation of the CW method then consists in simulating the cognitive behaviour of the user when confronted with each of the tasks chosen during the preparation stage. At each step of the task, the evaluator fills in a specific questionnaire (Fig. 2). This questionnaire includes questions concerning the current goals (hereafter called type 1.i questions), the choice and execution of actions (type 2.i), the system information feedback (type 3.i). For each problem encountered, the evaluator fills in a "problem description" form (Fig. 3). The analysis of the results makes it possible in principle to highlight the problems encountered by the user when performing the tasks (problems linked to goals and actions).

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

COGNITIVE WALKTHROUGH FOR A STEP

Task: **Action**:

1. Goal structure for this step.

1.1 What are the appropriate goals for this point in the interaction ?

1.2 Will the user have this goal ?

2. Choosing and executing the action.

2.1 Is it obvious that the correct action is a possible choice here ?

2.2 Are there other actions that might seem appropriate to the current goal ?

2.3 If there is a label or description associated with the correct action, is it obviously connected to one of the current goals for this step ?

3. Modification of goal structure.

3.1 Assume the correct action has been taken. What is the system's response ?

3.2 Will users see that they have made progress towards the current goal ? What will indicate this to them ?

3.3 Are there any current goals that have not been accomplished, but might appear to have been based on the system response ? What might indicate this ?

3.4 Does the system response contain a prompt or cue that suggests any new goal or goals ? If so, describe the goals.

Fig. 2. Typical evaluation form taken from (Polson et al., 1992)

Since it was proposed by Polson et al., the CW method has been the subject of many uses and evaluations within the framework of interactive systems, and not specifically multimedia ones: Wharton et al. (1994) illustrate the functioning of the method using an evaluation of the implementation of a “call transfer” service proposed by a telephone operator; John and Packer (1995) relate the experiment of a novice user of the method who evaluated the interface of a training system; finally, Kelley and Allender (1995) compare the results of CW used in the evaluation of human-machine interfaces with the results of other methods. These research projects, and especially the comparison of evaluation methods with practical cases carried out by Karat (1994), have shown that the use of CW obtains good results when the evaluator has an in-depth knowledge of the system, which moreover is also recognised by Wharton et al. (1994). On the other hand, the method takes a long time to apply and presents the particularity of missing general and recurrent problems when compared to heuristic evaluation.

CW has rarely been the subject of studies in the case of multimedia document evaluation. Although the hyperdocuments inherit some properties from human-machine interfaces, they also have their own characteristics which the evaluation of multimedia applications presented in section 3.1 has shown up. Questions must therefore be asked concerning the results obtained by the method during the evaluation of multimedia documents: can CW be applied in its current form? Does it make it possible to go beyond the detection of usability defects? Is it adapted to the various degrees of scenational interactivity offered by hyperdocuments? The study presented in the following section attempts to provide an answer to these questions.

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

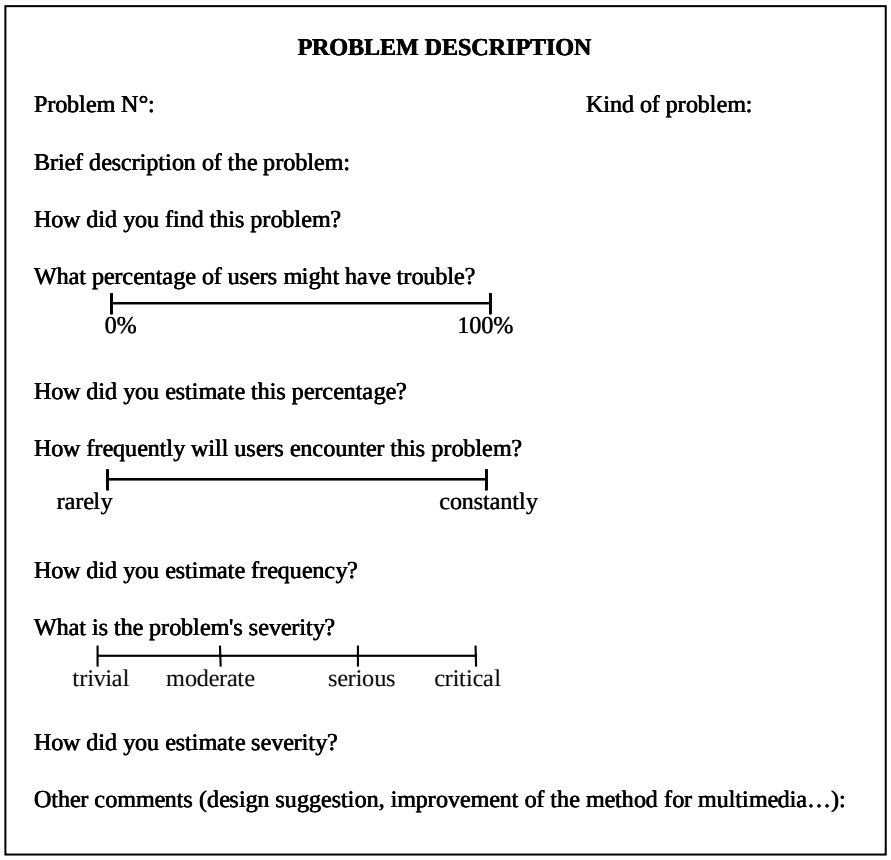

Fig. 3. Problem description template taken from (Polson et al., 1992)

5. Study based on the use of CW for the evaluation of multimedia products

CW is a method which makes it possible to detect usability defects. Despite this, the method has certain characteristics which lead one to envisage its use in and/or development towards the evaluation of multimedia documents: it uses the intervention of an expert, the evaluation is carried out using a review of the tasks to be performed, a form facilitates the task of the evaluator. Its use for the evaluation of multimedia documents has therefore been envisaged. With this in mind, the aim of the study was to assess the effectiveness of CW for evaluating such applications. In other words, it was a question of finding out whether this usability evaluation method is beneficial in its current form, or whether it can be used once it has undergone modifications specific to the multimedia field.

5.1. Multimedia applications evaluated

In order to verify the efficiency of CW over a wide panel of commercialised multimedia applications, the hypermedia chosen for this study were selected according to their degree of scenational interactivity, as explained in §2 (Table 1). Table 2 gives the list of these applications.

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

Table 2
Multimedia applications chosen for the evaluation

MULTIMEDIA APPLICATION	SCENATIONAL INTERACTIVITY	SUBJECT
Comment ça marche (version: 1998) (Company : Nathan, Paris) (called Document 1 in the paper)	Consultation (Degree #2)	The first hyperdocument is a children's encyclopaedia aimed at explaining and illustrating the functioning of natural and artificial systems. This product is intended for the 8 – 14 year age group. More information about *Comment ça marche* can be found at: www.france-edition-opi.asso.fr/docs/cata.67.htm
Le Louvre, Collections & Palais (version: 1998) (Company : Montparnasse Multimedia, Paris) (called Document 2 in the paper)	Navigation (Degree #3)	This hypermedia proposes the tour of a famous museum (Le Louvre) and its collections. Explanations are given concerning a certain number of paintings as well as on the groups of artistic schools. This CD-ROM is intended for the general public (from around 7-8 years of age). More information about *Le Louvre, Collections & Palais* can be found at: www.france-edition-opi.asso.fr/docs/cata.1d.htm
Adibou (version: 1998) (Company : Cocktel Vision, Paris) (called Document 3 in the paper)	Exploration (Degree #4)	This document is an educational software programme used as a backup to school work. The children are provided with exercises and a reminder of theoretic notions. They can also use a recreation area. It is intended for children aged 4 to 7. More information about *Adibou* can be found at: www.cocktel.fr
Versailles, complot à la cour du roi (version: 1998) (Company : Cryo Interactive, Paris) (called Document 4 in the paper)	Virtual visit (Degree #5)	This software is a virtual reality game. Its aim is to provide entertainment in the form of enigmas, at the same time giving a precise view of the historical context. It is therefore also intended to be educational. The users must find objects one at a time: the entire navigation through the document is intuitive. It is aimed at the general public (from around 7-8 years of age). More information about *Versailles, complot à la cour du roi* can be found at: www.cryo-interactive.com

5.2. Questions studied

In order to centre the experiment on specific elements, it was important to go beyond the general questions posed in §4; the following more detailed questions were therefore put forward:

- **Q1**: is it true that evaluators who use CW on multimedia applications have more difficulty in discerning the cognitive problems such as disorientation and the formulation of sub-goals to achieve a task than with self-evidence and predictability problems, and more generally classical usability problems (such as signifiance of codes, legibility, compatibility, and so on, as defined by Bastien and Scapin, 1993; Scapin and Bastien, 1997)?
    - *Disorientation has been defined as the tendency to lose one's sense of location and direction in a non-linear document by Conklin, 1987; Head et al., 2000.*
    - *Predictability expresses how well users anticipate an operation's outcome (according with Garzotto et al., 1995).*
    - *Self-evidence is defined by Garzotto et al., 1995 as "how well users guess the meaning and the purpose of whatever (content or navigational element) is being presented".*
- **Q2**: is it true that the method appears to be efficient (that is, makes it easier for the evaluators to discover errors) when the presumed user of the interface has an explicit task?

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

- **Q3**: is it true that the method does not seem to be adapted for the evaluation of software based on exploration (and which have a high degree of scenational interactivity)?
- **Q4**: is it true that when software has few ergonomic defects (in the sense of software ergonomics, cf. Bastien and Scapin, 1993), the detection of the defects is far more accurate? On the other hand, is it true that when software has a great number of defects, the major (serious or critical) defects hide the minor (trivial or moderate) defects which are then not detected?
- **Q5**: is it true that the number of defects detected is greater in the first phase of a task breakdown and decreases as the evaluation advances, regardless of the defects truly present in the application?
- **Q6**: is it true that the more the degree of scenational interactivity increases, the more considerable is the phenomenon described in Q5 (i.e. the detection of defects lessens as the evaluation progresses)?

In fact, Q1 to Q6 can be considered as being questions we asked ourselves, not only during the study, but also during discussions with colleagues, clients and students. Even though they seem to be heterogeneous, or in certain cases fairly close to each other, they represent the “Frequently Asked Questions” we felt it was important to attempt to answer.

5.3. Study design

The study involved four evaluators who successively reviewed four multimedia applications. The first two evaluators had in-depth expert knowledge, whereas the second two had a considerably lower degree of expert knowledge:

- The first evaluator is an expert in the field of industrial ergonomics and human-machine system design. He is a Doctor of Ergonomics, a university lecturer and researcher, aged 46 (he will be referred to as Expert 1). He has more than 15 years of experience in the field of ergonomic design and he has taken part in many projects and studies concerning the design and evaluation of interactive systems. He is familiar with inspection methods and has in-depth knowledge of CW.
- The second evaluator is an expert in the field of design and evaluation of interactive systems. He is a Doctor of Computer Science and Human-Machine Interaction, aged 38 (he will be referred to as Expert 2). He has more than 15 years of experience in this field and has also taken part in many university and industrial projects and studies regarding the design and evaluation of interactive systems. Like Expert 1, he has used inspection methods in many projects and has in-depth knowledge of CW.
- The two other evaluators have considerably less experience than the first two. They are both ergonomists, male, aged 23 and female, 24 respectively (they will be referred to as experts 3 and 4). Over the last two years, they have participated in several projects dealing with the analysis of working conditions for human operators, from the physical and the cognitive point of view (situations with or without the use of human-machine interfaces). Some projects concerned the use of inspection methods, including CW, for the evaluation of interactive systems.

The study was structured around three major phases (Carretier et al., 1999; Huart, 2000) (Fig. 4):

- A phase of learning the CW method: firstly, the aims of the research project were presented and discussed with the evaluators. In order to enable the evaluators to learn the method, it was first presented to them. Then they went on to increase their knowledge of CW using basic articles presenting the method, especially (Polson et al., 1992). A French translation of the questionnaires and forms to be filled in during the evaluation was also given to them, inspired by (Polson et al., 1992). To complete and validate their learning, a preliminary phase was performed: the four evaluators were asked to evaluate a commercial web site. This phase was performed by iterations (with correction and advice from the analyst leading the study, at the end of each iteration; the analyst was an expert in the method) until their knowledge and application of CW was judged to be sufficient.
- An experimentation phase during which the four evaluators used CW successively and separately on four multimedia applications (fixed sequences as showed in Fig. 4), each one filling in the evaluation forms and the problem description forms.
- A final phase composed of the analysis of the results obtained.

For each software evaluated, a single task was chosen which was representative of the dominant scenational level and of the activities of typical users of the document. This task was broken down into sub-goals and then proposed to the four evaluators; unlike the definition of the method (Polson et al., 1992), the evaluators therefore did not participate in the evaluation preparation phase (which consists in determining the representative task and breaking it down). Moreover, as Wharton et al. (1994) indicate, the choice of one single task obviously does not

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

make it possible to produce a complete evaluation of the application, but on the other hand it is sufficient to find representative defects in the application (*Representative defects* can be defined as being those likely to be found at different places in the interactive application, and which are likely to suggest classes of defects which require correction).

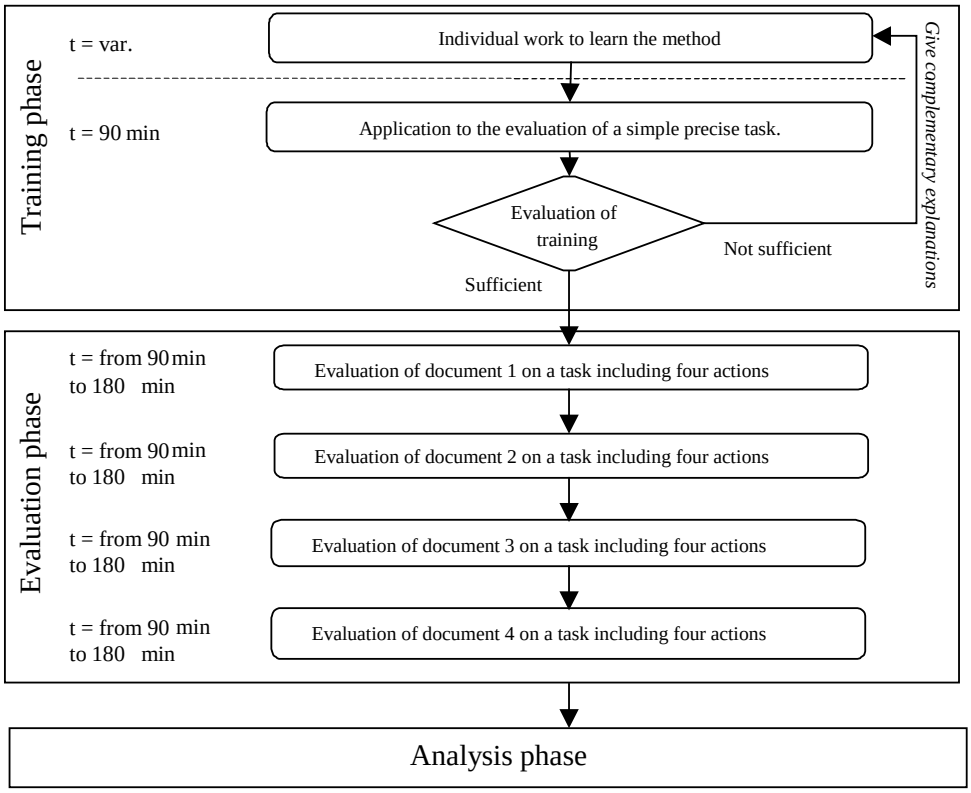

Fig. 4. Three major phases

5.4. Evaluation results[1]

5.4.1 Evaluation of document 1 (Consultation)

Document 1 proposes an arborescent type navigation, which means it has a degree #2 of interactivity. The representative task thus uses a downward path in the application structure. It consists of going to the page which explains the functioning of a “personal computer”. Four actions are necessary, with two screens of the application: one has to click on the image "machines", click on the "C" button in the index, click on "OK" to validate the choice and finally select and click on "Personal computer" in order to access the page required.

Over these four actions, 18 usability defects were found by experts 1 and 2, and 9 defects were found by experts 3 and 4, for a total of 20 defects shown up (two defects detected by experts 3 and 4 was not noted by experts 1 and 2). The defects mainly concern problems of coherence and self-evidence of the interface (in the sense of the type 2.i questions), even though some problems linked to the aims (type 1.i questions) and to feedback (type 3.i questions) are present. For example, for action 1 (click on “machines”), a child who wishes to see how a computer functions does not necessarily know that a computer is a machine: the child therefore does not have the goal of clicking on this image! Moreover, there is no feedback to indicate the possibility of an action. On the index screen, the child must click on “C” and then on “OK” to see the list of machines beginning with the letter “C”: here the interface reveals a problem of homogeneity as regards Windows-type interfaces, and

[1] In this study, the number of subjects (experts) was low (4), and the number of applications to be evaluated was limited (4); the number of defects shown up was also relatively low for each application evaluated. Consequently, there was no point in using statistical methods. Our aim was rather to enable an easy comparison of the results obtained by the evaluators; in this case, counts and percentages proved to be sufficient for this purpose.

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

also a problem of overload because several actions are used instead of just one. These problems can be generalised in this case to the whole of the application interface.

Table 3 presents the defects detected: the problems are organised in types which correspond to the questions on the evaluation forms. It should be remembered that the 1.i, 2.i and 3.i type questions concern respectively the current goals, the choice and execution of actions, and finally the system feedback (Fig. 2). The numbers in bold print correspond to the number of defects discovered simultaneously by at least two evaluators.

Table 3
Defects detected in a specific task in document 1 (based on CW)

	Action 1	**Action 2**	**Action 3**	**Action 4**	**Total**
Total defects	12	2	5	1	20
Total experts 1 and 2 (in-depth experience)	11	2	5	0	18
Total experts 3 and 4 (less experience)	3	2	3	1	9
Question 1.2	**1**	**1**	**1**	0	**3**
2.1	0	0	**1**	0	**1**
2.2	7+**1**	0	**1**	1	8+**2**
2.3	**1**	0	**1**	0	**2**
3.1	0	**1**	0	0	**1**
3.2	1	0	1	0	2
3.3	0	0	0	0	0
3.4	1	0	0	0	1

5.4.2 Evaluation of document 2 (Navigation)

Navigation in document 2 is made possible primarily by the map metaphor and an index. This index function was chosen by the evaluators: the task consists in accessing “The Mona Lisa” using the index. Three actions are necessary: clicking on “index” from the main menu; in the index, clicking on “M” to find the list of works beginning with the letter “M” and finally, clicking in the list on “Mona Lisa (the)”.

Experts 3 and 4 found 5 usability defects out of the 16 detected by experts 1 and 2. The defects detected are mainly concerned with the first action: when confronted with the application menu, the user perhaps does not know how to choose between the various possibilities described by the labels proposing “collections”, “guided tour”, “index”, “album”, … These defects will not necessarily be encountered by all users, but could be serious for novice computer users. The problems detected here also correspond to problems of goals and especially labels (cf. table 4).

Table 4
Defects detected in a specific task in document 2 (based on CW)

	Action 1	**Action 2**	**Action 3**	**Action 4**	**Total**
Total defects	10	5	1	0	16
Total experts 1 and 2 (in-depth experience)	10	5	1	0	16
Total experts 3 and 4 (less experience)	4	1	0	0	5
Question 1.2	**1**	0	0	0	**1**
2.1	**1**	**1**	0	0	**2**
2.2	5+**1**	**1**	1	0	6+**2**
2.3	**1**	1	0	0	1+**1**
3.1	0	1	0	0	1
3.2	0	1	0	0	1
3.3	0	0	0	0	0
3.4	1	0	0	0	1

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

5.4.3 Evaluation of document 3 (Exploration)

The third software programme has two distinct parts: an "exercise" part (maths, French, …) and a "recreation" part representing the imaginary play world of the character who accompanies the child during the programme. In this universe representing a house and a garden, the movements and actions of the character are controlled by clicking on places or objects. The task chosen by the evaluators is that of planting a seed and then watering it; in order to do this, one has to "click on the garden", "click on the blue seeds", "click on the part of the garden in which one wants to plant the blue seed" and "click on the watering-can".

The distribution of the defects detected according to the actions is relatively homogeneous for experts 1 and 2 who detected 20 defects (which corresponds to the total number of defects detected for this application). Experts 3 and 4 found 5 defects concentrated on the first two actions. The interface is truly complex as practically nothing indicates what action is to be performed in order to achieve a task: this is perhaps a conscious decision on the part of the designers, but it remains debatable as regards the target. For action 1 therefore, defects regarding goals and labels were detected (Table 5). For action 2, the problem is identical. However, for actions 2, 3 and 4, problems of feedback were also detected.

The software therefore appears to be difficult for young children: each action can present problems when performed for the first time. However, this opinion can be qualified by comparing the interface to that of video games which are sometimes far more complex when first used, but then become truly enjoyable to use. Because of this, it can be assumed that certain usability defects were perhaps intentional on the part of the designers.

Table 5
Defects detected in a specific task in document 3 (based on CW)

	Action 1	**Action 2**	**Action 3**	**Action 4**	**Total**
Total defects	5	5	6	4	20
Total experts 1 and 2 (in-depth experience)	5	5	6	4	20
Total experts 3 and 4 (less experience)	3	2	0	0	5
Question 1.2	**1**	1	0	1	2+**1**
2.1	**1**	1	1	1	3+**1**
2.2	**1**	**1**	3	0	3+**2**
2.3	1+**1**	**1**	1	0	2+**2**
3.1	0	1	1	1	3
3.2	0	0	0	1	1
3.3	0	0	0	0	0
3.4	0	0	0	0	0

5.4.4 Evaluation of document 4 (Virtual visit)

Although the principle of virtual reality provides the user with a great range of manoeuvre possibilities, in order to progress through the game, the user (from the beginning to the end of the game) must perform a series of successive actions which have been practically predefined and ordered by the designers. At the beginning, the user does not know what the series of actions is, and so must guess at it, using the help of enigmas. The software therefore is aimed at posing problems of interface goal and self-evidence to its user: consequently, on the actions evaluated, the evaluators noted these defects (Table 6). In addition, the problems noted block the user: it is impossible to progress towards gains in the game unless a determined action has been performed; for example, in order to discover drawers, the user has to open curtains which, however, are part of the background scenery and which do not necessarily look as if they can be moved .

Experts 1 and 2 and experts 3 and 4 detected respectively 21 and 4 defects with 24 defects in total: these defects were considered critical from a usability point of view, although they do not exist in the context of the game. This raises the question of the efficiency of CW in the evaluation of software which is purposely based on a low level of usability.

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

Table 6
Defects detected in a specific task in document 4 (based on CW)

	Action 1	**Action 2**	**Action 3**	**Action 4**	**Total**
Total defects	12	2	5	5	24
Total experts 1 and 2 (in-depth experience)	9	2	5	5	21
Total experts 3 and 4 (less experience)	3	1	0	0	4
Question 1.2	3	0	1	0	4
2.1	**1**	1	0	0	1+**1**
2.2	6	1	1	1	9
2.3	1	0	1	0	2
3.1	0	0	1+**1**	2	3+**1**
3.2	0	0	0	1	1
3.3	0	0	0	0	0
3.4	1	0	0	1	2

5.4.5 Discussion

The use of CW led to different results according to the multimedia documents evaluated. Document 1 presents a considerable number of usability defects linked to goals and the choice of user actions when the user is looking for a specific piece of information: the interface is not easy and the choice of labels going with the buttons and images is not necessarily very explicit. In document 2 which is a very high quality product, the study of the index revealed some minor defects (bothersome but not blocking) but these defects do characterise the errors of a certain number of multimedia designers: indeed, the general problems of usability linked to indexes are well known in human-machine interfaces and a certain number of ergonomic recommendations exist on this subject (Vanderdonckt, 1994). In a more general manner, there are a certain number of errors which human-machine interface designers no longer commit but which are still to be found in multimedia documents. In document 3, in a simple task, practically each defect found by the evaluators is blocking, in relation to the target population of the software (which corresponds to children of age 4 minimum); this same problem is found in document 4. The majority of the usability problems detected therefore concern the self-evidence and the predictability of the interfaces. These problems are not necessarily blocking, but constitute a great hindrance in the use of the products, especially for novice users. This brings about problems of formulation and reformulation of goals for the users. However, this opinion should be qualified by the fact that the majority of multimedia applications are based on the difficulty of navigation. On the other hand, few problems of feedback were found, which demonstrates that the feedback rule has been correctly understood and applied at the professional level.

6. Efficiency of CW for the evaluation of multimedia documents

In this part of the article, we review the six questions put forward at the outset of the study (Cf. §5.2) and we attempt to provide answers, in the light of results obtained in terms of problems detected. Table 7 gives an overall view of the problems detected during the evaluation of each multimedia document for all of the evaluators. As in tables 3 to 6, the problems detected are situated in relation to the 1.i, 2.i and 3.i type questions given in the basic CW form (Fig. 2).

A version of this paper has been published in: *Interacting with Computers*, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

Table 7
The number and type of defects detected in the four multimedia applications (based on CW)

	Problems found in multimedia products							
	Experts 1 and 2	Experts 3 and 4	Experts 1 and 2	Experts 3 and 4	Experts 1 and 2	Experts 3 and 4	Experts 1 and 2	Experts 3 and 4
Question	**Document 1**		**Document 2**		**Document 3**		**Document 4**	
1.2	3	3	1	1	3	1	3	1
2.1	1	0	2	1	4	1	2	1
2.2	8	3	8	2	5	1	7	2
2.3	2	2	2	1	4	2	2	0
3.1	1	1	1	0	3	0	4	0
3.2	2	0	1	0	1	0	1	0
3.3	0	0	0	0	0	0	0	0
3.4	1	0	1	0	0	0	2	0
Total	**18**	**9**	**16***	**5**	**20**	**5**	**21**	**4**

** A same defect has been found by the experts 1 and 2, but using two different questions (2.1 and 2.2).*

[Q1] *Is it true that evaluators who use CW on multimedia applications have more difficulty in discerning the cognitive problems such as disorientation and the formulation of sub-goals to achieve a task than with self-evidence and predictability problems, and more generally classical usability problems (such as signifiance of codes, legibility, compatibility, and so on):*
A great majority of the problems detected are concerned with the choice and execution of actions (Fig. 5), and especially the question "Are there any other actions which might seem to be suitable for one of the current sub-goals?": the problems of comprehensibility and self-evidence of the interface are the only ones which are truly detected. Moreover, the questions linked to feedback and the reformulation of goals practically never led experts 3 and 4 (less experience) to fill in problem description forms, unlike the experts 1 and 2 (in-depth experience) who detected a certain number of problems for these same questions. Whereas cognitive problems such as disorientation and the formulation of user goals are often latent in hyperdocuments, it has to be noted that CW did not enable the evaluators to detect very many of them. A positive answer to this question is therefore supported by the results: firstly, CW is more adapted to the graphic aspects of the interface, and secondly, the evaluators who use CW have more difficulty in discerning cognitive problems.

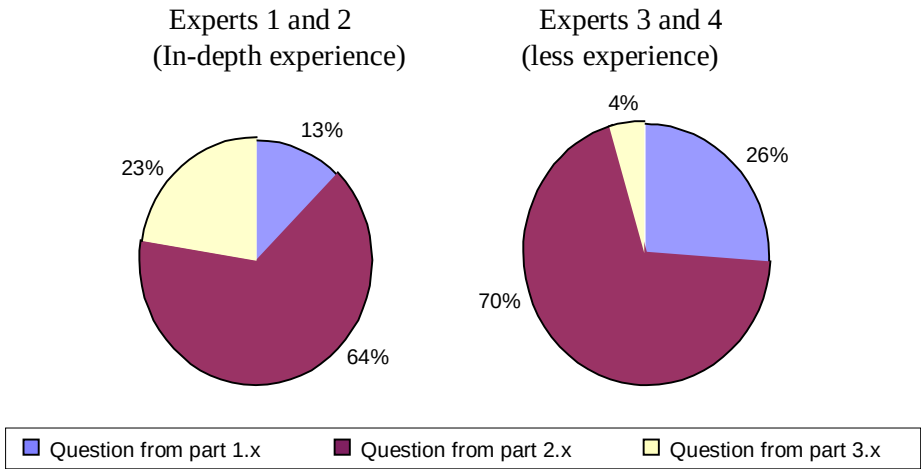

Fig. 5. Distribution of types of usability defects

- ***[Q2]*** *is it true that the method seems to be effective when the supposed user of the application has a specific task:*
  In documents 1 and 2, the tasks which the typical user performs are, on the whole, explicit, even though these products are designed to allow their users to move according to what they wish to discover. In documents 3 and 4, which are more exploratory in nature, the tasks performed by the user are potentially less explicit, the users elaborate goals as they move along in the background scenery provided as an interface: given that the

A version of this paper has been published in: *Interacting with Computers*, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

number of defects found for each of the four documents evaluated does not show any significant reduction in results according to the tasks, a positive answer to this question is not supported by the results.

- ***[Q3]*** *Is it true that the method does not seem to be adapted for the evaluation of software based on exploration (and which have a high degree of scenational interactivity):*
  When reading Table 7, it can be noticed that the degree of scenational interactivity does not seem to have any negative effect on the detection of defects by a set of evaluators: the number of defects detected is practically the same for each application. However, if the results are limited to those of experts 3 and 4 only, it appears that fewer usability defects were detected for the documents based on exploration (documents 3 and 4): the detection of usability defects by experts 3 and 4 therefore seems to be more difficult in documents with a high degree of scenational interactivity. A positive answer to this question is not supported by the results.

- ***[Q4]*** *Is it true that when software presents few ergonomic defects (in the sense of software ergonomics), the detection of defects is far more accurate? on the other hand, is it true that when software has many defects, the major (serious or critical) defects hide the minor (trivial or moderate) defects which are therefore not detected:*
  A positive answer to this question is not totally supported by the results, although it corresponds to the feelings expressed by experts 1 and 2. On the other hand, the estimated degree of severity of the defects seems to be linked to the phenomenon of habituation: indeed, in documents 3 and 4 which have a high degree of scenational interactivity, the final actions enabled experts 1 and 2 to detect defects which they estimated for the most part to be serious and/or critical, whereas on average (over four evaluators and four applications), the serious and critical defects represent less than a third of the total defects (Fig. 6). Nevertheless, Sears and Hess (1988) have shown that the degree of detail of a task has an effect on the number of defects found and their nature: thus, with a task which was not completely detailed, these authors noted a greater number of defects linked to the predictability of the interface; on the other hand, the more detailed tasks make it possible to find significantly more defects of type 1.i and 3.i.

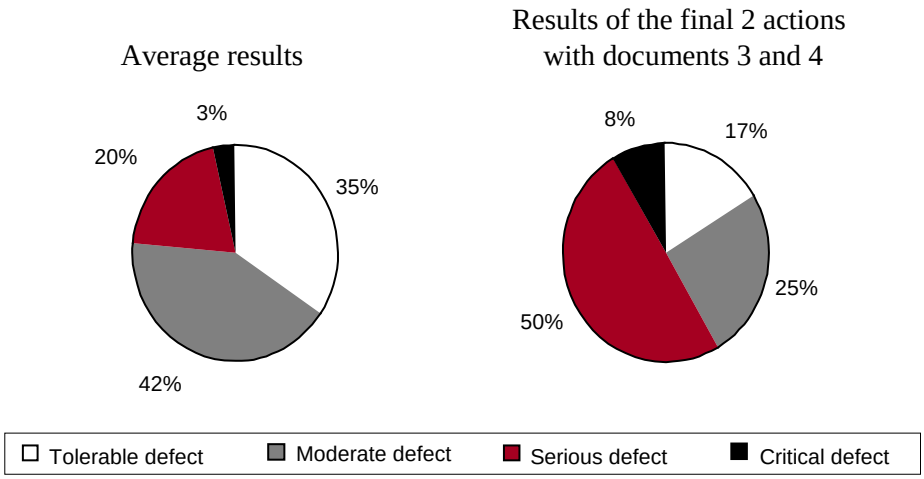

Fig. 6. Percentage of severity of the ergonomic problems detected on average and over the final two actions on the two documents with the highest degree of scenational interactivity (documents 3 and 4)

- ***[Q5]*** *Is it true that the number of defects detected is greater in the first action in the breakdown of a task and decreases as the evaluation progresses, independently from the defects truly present in the application:*
  An analysis of the detection of errors per action and per software shows in fact that in all of the multimedia documents tested, the number of usability problems detected is considerably greater in the first action and it decreases as the evaluation progresses (Fig. 7). This phenomenon of habituation and weariness when confronted with the extent of the evaluation work is greater with experts 3 and 4 (less experience). A positive answer to this question is supported by the results.

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

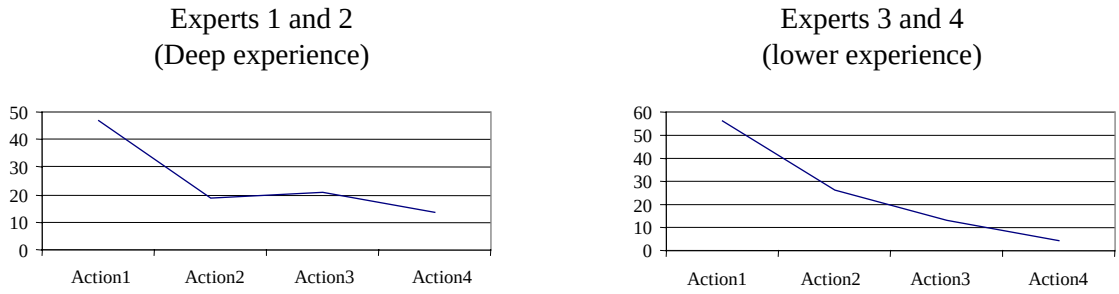

Fig. 7. Percentage of ergonomic problems detected per action composing the task to be evaluated (average over the four documents evaluated).

- ***[Q6]*** *Is it true that the more the degree of scenational interactivity increases, the greater the phenomenon described in the previous point (i.e. the detection of defects lessens as the evaluation progresses):*
  Whereas the results and comments of experts 3 and 4 could lead one to think that the previous point is dependent upon the degree of scenational interactivity of a multimedia document, the results of experts 1 and 2 contradict this statement: with experts 1 and 2, for document 3 for example (Fig. 8), the phenomenon of habituation in the application of the CW method is not dependent upon the degree of scenational interactivity. A positive answer to this question is not supported by the results.

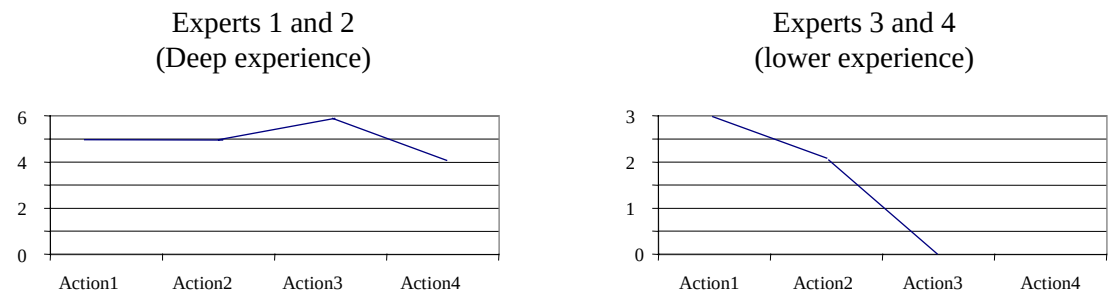

Fig. 8. Number of ergonomic defects detected per action composing the task to be evaluated for document 3.

6.1. Reactions of the users of the method

At the end of the study, each of the four evaluators (experts 1 to 4) were asked three questions. These questions and a synthesis of the evaluators' answers are presented below (the answers of experts 3 and 4 were given jointly).

6.1.1. About question 1: What did you think of the use of CW ?

Expert 1 thinks that CW is most adapted to preliminary studies, "quick and dirty" expert evaluations (as in Chignell et al., 1995; Jordan et al., 1996; Kolski et al., 2000), in order to obtain a rapid inventory but he feels it cannot lead to a pertinent ergonomic diagnosis. However, Expert 1 thinks that CW seems to be more useful than other methods used in software ergonomics such as the interview, the questionnaire or the extraction of knowledge away from the work situation, for example. Indeed, for him, CW makes it possible to guide the analysts better in their approach compared with the use of a classic questionnaire, since they are already in the action when they answer the questions posed and therefore they have to think about the problems encountered according to the current goal. But expert 1 also thinks that the use of CW depends upon three important variables: the support used (classical software, CD-ROM, web site…) and thus on the interaction; the expert (state, culture, experience, age…); the environment for the evaluation (work situation, research laboratory…). Finally, he found CW interesting, but sometimes annoying or even stressful, according to the degree of interactivity encountered.

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

Expert 2 is very satisfied with CW. He feels that CW makes it possible to ask systematic questions as regards each task to be evaluated, and that the problem forms make it possible to transmit improvement suggestions to the designers in a structured and justified form. He thinks that CW does not seem too cumbersome to assess, and that no more time is spent with CW than with another Heuristic Evaluation type method; on the contrary, it seems easier to draw up the evaluation report. However, in a specific field of application (multimedia applications, industrial process control, etc.), expert 2 suggests that CW must be completed using other questions or criteria.

Experts 3 and 4 (less knowledge) suggest that the level of computing skills of the evaluator plays an important role in the evaluation, which raises the question of the nature and the skills of the expert using CW. They suggest that the choice and the description of the representative task should be made or guided by the designer of the product used: the designer should also indicate which age group is targeted (indeed, in the evaluations concerning the document 3, an educational software programme for children, they found it difficult to express ergonomic problems). They also found that questions of the type 2.i were superfluous.

6.1.2. About question 2: How did you use CW ?

Expert 1 applied CW to the letter with no preliminary prospecting, and answered the questions as honestly as possible, even though he found that a certain diffuseness was sometimes tiresome.

Expert 2 tried to gain an overall view of each product to be evaluated, before going into the details of the tasks. Then, for each task, he used CW, at the same time keeping in mind his overall view of the product. He explained that he tried scrupulously to fill in each section, to answer each question, thinking that the questionnaires associated to CW exist primarily in order to facilitate and structure the task of the evaluator.

Experts 3 and 4 applied CW directly by answering the questions and filling in the problem forms when necessary. Nevertheless, the answers to the requests for explanations were often ignored as they found it difficult to answer this part.

6.1.3. About question 3: What improvement suggestions can you make for CW ?

For expert 1, CW could gain a lot by being smartened up a little (without losing its core) especially if it is to be applied to the field of multimedia. He also suggests a few elements to be developed:

- the aim of the evaluation, and thus the analysis of the demand;
- the status of the evaluators, their characteristics and skills;
- the role of a possible supervisor;
- the preparation and learning phase;
- the investment time and means;
- if one wishes to gain access to the reasoning of the evaluator, to the significance for the action and thus to the cognitive activity, is it necessary to record the verbalisations, and if this is the case, which verbalisations are to be used (simultaneous and/or provoked and/or self-confrontation verbalisations)?

Finally, expert 1 thinks that CW can be improved by introducing specific themes and by asking the evaluator of the tested product for improvement directions theme by theme and online. For example, concerning the current goal, the interface, the presentation of information, the feedback, the colours, the sound, the ease of use, and so on.

Expert 2 deems CW in itself to be a good method, but feels that the questionnaires can be refined, remembering that several authors have already proposed improvements (Abowd, 1995; Collins, 2000): in this case, it is necessary to ensure that the foundations of CW are not changed. He suggests studying CW in relation to a particular field of application (process control, web sites, educational software…): if this field has particular characteristics, specific constraints (e.g.: temporal constraints) linked to the tasks, to the needs of the users, to a particular targeted age group (e.g. children), he thinks it is necessary to remodel CW, or use it as a complement to other methods or criteria; in any case, the evaluator has to be aware of these specificities. He concludes by mentioning the fact that research should be carried out with specialists in each field.

Experts 3 and 4 propose testing the representativity of the task chosen by conducting an experiment involving users. They also suggest drawing up an explanatory text aimed at guiding the evaluators when using CW (with a view to increasing the understanding of questions). They think that there should be a limited number of well-targeted questions in order to make the evaluation procedure easier; in this case, each part could correspond to a theme (concerning usability), with the possibility of attributing a mark per theme.

Moreover, they propose adding a question to ask the evaluator if he or she has met any important problem with CW or can think of any improvements. In this way, CW would be improvable each time it is used.

Experts 3 and 4 have several ideas concerning the problem description form:

- provide three different problem forms for the three types of question (goal, action, goal structure);
- aim at quantifiable and easily usable answers (multiple choice questions, graduated scales) which would make it possible to give the software a mark, for example;

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

- insert the problem treatment questions into the evaluation form.

6.2. Discussion

The results show that using CW makes it possible especially to detect problems linked to the self-evidence and the predictability (and more generally to classical usability problems) of a multimedia interface; the problems linked to the cognitive functioning of the user are more difficult to define.

One of the missions of the evaluators was to note any malfunctions occurring during the use of CW and to judge the method. They suggest the following analysis. Experts 3 and 4 considered that the writing of problem description forms was long and difficult. Expert 1 confirmed that the forms were tedious to fill in, sometimes superfluous and that the phenomenon of habituation is a consequence of this; expert 2 tempered this judgement, indicating that the formulation and organisation of the questions had enabled him, on the contrary, to structure his ideas into a coherent form. For example, the problem description form requests a description of the problem, followed by an estimation of the circumstances and the severity of the problem, with justification required for these elements. This important demand for information (which is already a handicap for the method in the case of human-machine interfaces), although it is necessary for the estimation of the severity of the ergonomic defect, was not appreciated in most cases by the evaluators, especially because their degree of knowledge of multimedia products did not allow them to perform the estimations in depth and above all did not allow them to give a systematic justification. This raises the problem of the nature of the expert using CW: an expert in the field of human-machine interfaces can perform a relatively correct evaluation of the interface, but may possibly detect defects which are not actually defects, and will probably miss certain defects linked to the characteristics of the multimedia field. There are also currently very few true experts in multimedia interfaces and multimedia applications and, moreover, few rules exist. In addition, the evaluators often felt the need to express the explanation of the problems detected in the "other comments" section, which shows that the forms are not well adapted. Finally, superfluous elements were noted in the drafting of the problem description forms: for example, type 2 questions sometimes led the evaluators to fill in several corresponding forms for the same problem, because they had difficulty in defining the exact nature of the problem, and the question concerning it. Such superfluous elements have also been noted in the case of the evaluation of human-machine interfaces (John and Packer, 1995).

The evaluators, and especially the experts 3 and 4, also pointed out the hindrance caused by questions 3.2, 3.3, and 3.4. Two reasons can explain this problem. Firstly, as the breakdown of the task had been imposed on them, the evaluators had great difficulty in envisaging modifications to the structure of the user goal: however, this problem is linked to the study. Secondly, many users of multimedia documents have no precise goal for their use, and the designers take advantage of this: the exploratory nature of multimedia documents therefore tends to limit the interest of the questions linked to the user goals.

This problem is also linked to the difficulty in targeting the users of the documents evaluated. It is extremely difficult to envisage the characteristics and level (even just in computing skills) of multimedia document users. This problem, which is general to multimedia, constitutes a considerable weakness for CW which requires a good knowledge of the user. For example, for the evaluation of document 3, an educational game software intended for children aged 4 to 7 years, the evaluators had a great deal of trouble in imagining the perceptive and cognitive behaviour of children.

6.3. Directions for the adaptation of CW

CW is an expert's method. In the case of human-machine interfaces, the number of experts is fairly high (Wixon et al., 1994): ergonomists, interface designers, cognitive science specialists. Besides this, in the majority of cases when an interface is designed, the potential users are relatively well known: the method is applied strictly and systematically; the estimation of the degree of severity of the defects is objective and rational. In the case of multimedia, few experts are available: the field of multimedia is vast and includes many techniques and much knowledge. In addition, the knowledge of the user is often limited and the diversity of the public targeted greatly hinders the generalisation of reasoning on the potential defects in multimedia interfaces. In particular, the estimation of the degree of severity of the defects detected proved to be more or less impossible for the evaluators.

CW therefore would appear to present certain shortcomings at this level, and the difficulty found by the evaluators in envisaging positive developments of the method for multimedia is a consequence of this. More globally, as far as usability is concerned, the results of the study can be compared with those of other study projects.

Firstly, of the 80 usability errors detected in the four multimedia documents, 22 were noticed by at least 2 evaluators and only 3 defects were detected by all of the evaluators (Fig. 9): Nielsen and Mack (1994) and Pollier (1991) have already highlighted the need to entrust the study of software usability to several analysts in order to increase efficiency. The evaluation of usability using inspection methods is dependent upon the culture

A version of this paper has been published in: *Interacting with Computers*, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

of the expert: in our case, each expert made it possible to detect defects linked more specifically to his or her own field of expert knowledge. In terms of individual differences, this result points in the same direction as that of the research performed by Hertzum and Jacobsen (1999).

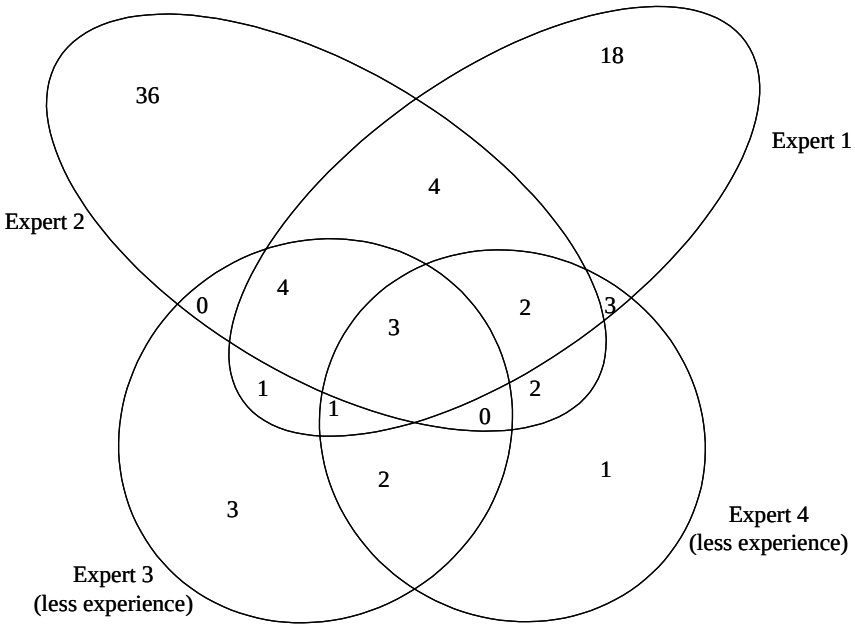

Fig. 9. Intersections between defects found.

It should be noted that 82% of the 22 defects detected by at least two evaluators were detected by at least one of the experts 3 and 4 (Fig. 9), whereas both of these experts (less experience) found only 28% of the total defects simultaneously. This makes it possible to situate the performance of such experts using the method compared with that of the experts 1 and 2 (with in-depth experience). The degree of knowledge of the experts with less experience in the evaluation of human-machine interfaces is sufficient to detect latent usability defects but does not necessarily make it possible to detect a majority of defects, especially in the case of this study.

The defects detected during this study mostly concern the self-evidence and also classical usability problems of the interface: Wharton et al. (1994) indicate that CW only seeks to evaluate the overall facility of learning. Thus, the evaluation of an interface designed to increase productivity can be negative. When transposed to the field of multimedia, the problem becomes a major one: a great number of hyperdocuments are purposely based on exploration and the lack of self-evidence of the interface (for example, in some applications, the buttons are hidden in order to oblige the user to explore the screen).

The complexity and cumbersome nature of implementing CW constitute a problem which is equally important: the evaluation of a web site (with a task associated to 12 actions) by an experienced user of CW lasted more than three hours. John and Packer (1993) report the case of an evaluator who spent ten weeks analysing two tasks associated to 36 and 104 actions and who decided during the evaluation to create macro-actions in order to limit the evaluation time. The use of CW can therefore be costly (Mack and Nielsen, 1994). The complexity of CW can be lightened, however: the four evaluators responsible for the study suggest limiting the number of questions and targeting them better in order to avoid the current complexity of establishing the reports. Whilst keeping their significance, the three types of question (current goals, choice and execution of actions and system feedback) would gain from being transformed into three single questions; the evaluators also indicate that the justifications and the degree of severity of the errors should be given, either in the answer to each question, or on a form regarding the type of problem. Some authors have already suggested a simplification which goes in this direction: Abowd (1995) and Collins (2000) limit the number of questions to four, whilst including the major points of CW: Have the users got the correct goal? Will they be capable of seeing that the correct action is a possible choice? From the moment when the users find the correct action in the interface, will they know it is the correct one for the goal they wish to attain? Once the action has been performed, will they understand the system feedback? However, their opinions diverge regarding the way of drawing up the problem form: Collins puts the problem descriptions on the action evaluation form, whereas Abowd insists on the need to

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

record the errors detected via the creation of separate problem forms. Suzuki et al. (1999) suggest a simpler CW method (with an easier way of describing user goals, a simplified rating of severity, a reduced number of questions): many experiments are required to prove its efficiency as regards multimedia applications with varying degrees of interactivity. Blackmon et al. (2002) also propose a transformation of the CW method for evaluating web sites.

The evaluators also insisted on the need they felt to include the intervention of users in the preparation of tasks and the evaluation process. John and Packer (1995) indicate that during their experiments, the CW user regretted not having been able to meet users in order to achieve a truly representative task selection. Although this does not come into the framework of the CW method at all, many studies have shown that the association of empirical and informal methods optimise the evaluation of usability (McCall et al., 1997): this is especially the case with the SUE method (Garzotto et al., 1995; Garzotto and Matesa, 1997) which combines a heuristic evaluation and a user test; the evaluations performed by Parlangeli et al. (1999) on multimedia systems also point in this same direction.

Finally, having experienced a problem with the presentation of the evaluation results for each multimedia product, the evaluators suggest the weighting of questions and problems in order to be able to determine the degree of severity of these problems more easily. However, it appears to be obvious to us that the degree of severity of a problem can vary considerably from one type of application to another

7. Conclusion and perspectives

The multimedia documents studied include a certain number of usability defects. During this study, the current formulation of CW only made it possible on average to detect the defects linked to the self-evidence and the predictability of the interface. It was particularly difficult to evaluate the management of user goals. It should be noted, however, that amongst the defects detected in the multimedia documents are defects which are no longer found in classic interactive systems because ergonomic rules are used in the design of these systems. Therefore from a general point of view, it can be said that multimedia applications are slightly behind as regards usability compared with classic human-machine interfaces, except concerning the feedback for which the basic ergonomic rules are used. Nevertheless, these conclusions must be tempered by recognising that the use of CW led the evaluators to detect usability defects which had, however, been purposely created by the designers (in the case of exploration-based software). Concerning CW, the study appears to show that the efficiency of the method decreases along with the amount of interactivity of the product evaluated: for applications of the exploratory type (virtual reality, for example), the method would appear not to be adapted at all.

CW is currently too complex, both in its form and its use, to allow an efficient and low cost evaluation of multimedia documents. Nevertheless, provided that relatively important modifications are made in the form of the questionnaires on which the method is based, it could be used for the evaluation of usability of certain multimedia documents with low or average degrees of interactivity. However, the evaluation of multimedia documents is not limited to the detection of usability defects, and must also deal with problems regarding the significance and format of messages. In this perspective, the CW user would appear to be unprepared. In any case, our study has not made it possible to envisage the use of the method for this, and even less to envisage its development in this direction.

Different solutions can therefore be considered: to restrict the use of CW to the evaluation of usability alone on applications with low amount of interactivity (which is already a major objective in itself) and to use other tools in order to measure the specific parameters of multimedia products; Wharton et al. (1993) indicate that the method can indeed be used for this purpose. Otherwise, the solution would be to adapt CW further still to the specificities of multimedia, by integrating criteria specific to multimedia communication into the questionnaire and/or to the thoughts of the evaluator (cf. on this subject Leulier et al., 1998); progress can be made by finding out more about what the relevant criteria for multimedia applications are - for instance concerning the risk of perceived disorientation (Smith, 1996, McDonald and Stephenson, 1998, Ahuja and Webster, 2001) - and after that by reconsidering the method. A great deal of research therefore remains to be done for an efficient use of CW, whatever the degree of interactivity of the multimedia application evaluated.

References

G. Abowd, 1995. Performing a Cognitive Walkthrough. Available at: http://www.cc.gatech.edu/computing/classes/cs3302/documents/cog.walk.html

J.S. Ahuja, J. Webster, 2001. Perceived disorientation: an examination of a new measure to assess web design effectiveness. Interacting With Computers 14, 15-29.

J.P. Balpe, 1997. Hypertexte et interactivité. Hypertextes et Hypermédias 1 (1), 11-22.

J.M.C. Bastien, C. Leulier, D.L. Scapin, 1998. L'ergonomie des sites web. In J.C. Le Moal, B. Hidoine (Eds.), Créer et maintenir un service web, Paris : ADBS, pp. 111-173.

J.M.C. Bastien, D.L. Scapin, 1993. Ergonomic criteria for the evaluation of human-computer interface. Research report, INRIA, Paris.

J.M.C. Bastien, D.L. Scapin, 2001. Évaluation des systèmes d'information et Critères Ergonomiques. In Kolski C. (ed.), Environnements évolués et évaluation de l'IHM, Interaction homme-machine pour les SI 2, Hermes, Paris, pp. 53-80, 2001.

J.M.C. Bastien, D.L. Scapin, C. Leulier, 1999. The ergonomic criteria and the ISO/DIS 9241-10 dialogue principles: a pilot comparison in an evaluation task. Interacting With Computers 11, 299-322.

R.G. Bias, 1994. The Pluralistic Usability Walkthrough: Coordinated Empathies. In J. Nielsen, R.L. Mack (Eds.). Usability inspection methods, Elsevier, pp. 63-76.

M.H. Blackmon, P.G. Polson, M. Kitajima, C. Lewis, 2002. Cognitive Walkthrough for the Web. 2002 ACM conference on human factors in computing systems (CHI'2002), 463-470, ACM Press.

N. Bowers, 1996. Weblint: Quality Assurance for the World Wide Web. Proceedings of the fifth International World Wide Web Conference, Paris.

M. Carretier, S. Nicot, E. Benoît, 1999. Utilisation de la méthode Cognitive Walkthrough pour l'évaluation de produits multimédias. Master Thesis in Industrial Ergonomics, LAMIH, University of Valenciennes (unpublished), February.

H.M. Chignell, T. Motoyama, V. Melo, 1995. Discount video analysis for usability engineering. In Y. Anzai, K. Ogawa, H. Mori (Eds.). Symbiosis of Human and Artifact, Social aspects of Human-Computer Interaction, Elsevier Science B.V., Amsterdam, pp. 323-328.

P. Collins, 2000. MetaWeb: A plan for Cognitive Walkthrough. Available at: http://c2000.gatech.edu/c2000/cs6751_96_fall/projects/glass/walkthrough.html.

J. Conklin, 1987. Hypertext: An Introduction and Survey. IEEE Computer 20 (9), 17-41.

M. Cooper, 1999. Evaluating accessibility and usability of web pages. In J. Vanderdonckt and A. Puerta (Eds.). Computer-Aided Design of User Interfaces II, Kluwer Academic Publishers, pp. 33-42.

R.R. Dholakia, M. Zhao, N. Dholakia, D.R. Fortin, 2000. Interactivity and revisits to websites: a theoretical framework. Working paper, RITIM, accessible at: http://ritim.cba.uri.edu/wp/

H. Dix, J. Finlay, G. Abowd, R. Beale, 1993. Human-Computer Interaction, Prentice Hall.

T. Duffy, J. Lowyck, D. Jonassen, 1993. Designing Environments for Constructive Learning. In R. Scott Grabinger, Hypermedia, Taylor Graham, pp. 144-149.

A. Durand, 1997. Modélisation moléculaire, vers un nouvel outil d'aide à la conception multimedia. Ph.D. Thesis, University of Valenciennes.

A. Durand, J. Huart, S. Leleu-Merviel, 1997a. Vers un modèle de programme pour la conception de document. Hypertextes et Hypermédias, 1(1), 79-101.

A. Durand, J.M. Laubin, S. Leleu-Merviel, 1997b. Vers une classification des procédés d'interactivité par niveaux corrélés aux données. In J.P. Balpe (Ed.). Actes de la conférence $H^2PTM'97$ (September 1997, St Denis), Hypertextes et Hypermédias, Volume 1 (2-3-4), Hermès, Paris, pp. 367-382.

S.M. Eisenstein, 1929. Dramaturgie de la forme filmique, Dunod, Paris.

C. Farenc, M.F. Barthet, V. Liberati, 1996. Automatic ergonomic evaluation: which are the limits. Proceedings CADUI'96, 2nd International Workshop on Computer-Aided Design of User Interfaces, Namur, Belgium, 5-7 june.

T. Forse, 1989. Qualimétrie des systèmes complexes, mesure de la qualité du logiciel. Les éditions d'organisation, Paris.

F. Garzotto, L. Mainetti, P. Paolini, 1995. Hypermedia design analysis, and evaluation issues. Communications of the ACM 38(8), 74-86.

F. Garzotto, M. Matesa, 1997. A Systematic Method for Hypermedia Usability Inspection, The new review of hypermedia and multimedia (3), 39-65.

W. Gibbs, 1994. Software's Chronic Crisis. Scientific American, September, 72-81.

M. Grislin, C. Kolski, 1996. Evaluation des interfaces homme-machine lors du développement de système interactif. Technique et Science Informatiques (TSI) 15 (3), 265-296.

J. Grudin. Utility and usability: research issues and development contexts, Interacting With Computers, 4 (2), 209-217.

J.F. Head, N. Archer, Y. Yuan, 2000. World wide web navigation aid. International Journal of Human-Computer Studies 53, 301-330.

M. Hertzum, N.E. Jacobsen, 1999. The evaluator effect during first-time use of the cognitive-walkthrough technique. In H.J. Bullinger, J. Ziegler (Eds.). Proceedings of the 8th Human-Computer Interaction International Conference 1999 (HCI'99, Munich, Germany, August 22-26, 1999), Volume 2, Lawrence Erlbaum Associates, London, pp. 1063-1067.

O. Hû, P. Trigano, 1999. A tool for evaluation using dynamic navigation in a set of questions. In J. Vanderdonckt, C. Farenc (Eds.). Tools for Working With Guidelines TFWWG'2000, Springer Verlag, pp. 273-281.

J. Huart, 2000. Mieux concevoir pour mieux communiquer à l'ère des nouveaux médias, vers des méthodes de conduite de projets et d'évaluation qualité de documents multimédias. Ph.D. Thesis, University of Valenciennes, France.

J. Huart, C. Kolski, S. Leleu-Merviel, 1998. Vers la correction et la prévention des erreurs méthodologiques dans le cycle de vie d'applications multimédias. Proceedings 6ème Colloque Ergonomie et Informatique Avancée ERGO'IA (4 to 6 Nov., Biarritz), ESTIA/ILS, Bayonne, pp. 59-68.

INA, 1994. Facteurs-clés de succès des Produits Multimédias Interactifs, étude guide. Institut National de l'Audiovisuel Bry-sur-Marne, France.

ISO DIS 14915,1999. Multimedia user interface design – Software ergonomic requirements. ISO.

B.E. John, H. Packer, 1995. Learning and using the Cognitive Walkthrough method: a case study approach. Proceedings of CHI'95 (May 7-11, Denver), ACM Press, pp. 429-436.

P.W. Jordan, B. Thomas, B. Weardmuster, I. McClelland (Eds.) (1996). Usability evaluation in industry. Taylor & Francis, London.

M.J. Kahn, A. Prail, 1994. Formal Usability Inspections. In J. Nielsen, R.L. Mack (Eds.). Usability inspection methods, Elsevier, pp. 141-171.

C.M. Karat, 1994. A Comparison of User Interface Evaluation Methods. In J. Nielsen, R.L. Mack (Eds.). Usability Inspection Methods, Elsevier, pp. 203-233.

C.M. Karat, R. Campbell, T. Fiegel, 1992. Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. Proceedings of CHI'92, ACM Press, pp. 397-404.

T. Kelley, L. Allender, 1995. Why choose ? A Process approach to usability testing. In Y. Anzai, K. Ogawa, H. Mori (Eds). Symbiosis of human and artifact: human and social aspects of Human-Computer Interaction, Elsevier Science B.V, pp. 393-398.

B. Kemp, K. Buckner, 1999. A taxonomy of design guidance for hypermedia design. Interacting With Computers 12, 143-160.

J. Kirakowski, B. Cierlik, 1998. Measuring the Usability of Web Sites. Proceedings Human Factors and Ergonomics Society Annual Conference, Chicago.

C. Kolski, 1997. Interfaces homme-machine, application aux systèmes industriels complexes. Hermes Science Publications, Paris.

C. Kolski, P. Millot, 1991. A rule-based approach for the ergonomic evaluation of man-machine graphic interface. International Journal of Man-Machine Studies 35, 657-674.

C. Kolski, B. Riera, T. Berger, 2000. A questionnaire-based discount evaluation mathod using guidelines for process control interactive applications. In J. Vanderdonckt, C. Farenc (Eds.). Tools for Working With Guidelines TFWWG'2000, Springer, London, pp. 127-138.

G. Kouroupetroglou, C. Viglas, C. Metaxaki, 1994. A Generic Methodology and Instrument for Evaluating Interactive Multimedia. Proceedings of Telematics for Education and Training Conference, november, Dusseldorf.

N. Layaïda, C. Keramane, 1995. Maintaining Temporal Constistency of Multimedia Documents. Proceedings of the ACM Workshop on Effective Abstractions In Multimedia, November.

S. Leleu-Merviel, 1997. La conception en communication. Hermes Science Publications, Paris.

C. Leulier, C. Bastien, C., D. Scapin, 1998. Compilation of Ergonomic Guidelines for the Design and Evaluation of Web Sites. Research report, Esprit Project "Commerce & Interactions" (EP 22287).

C. Lewis, 1997. Cognitive Walkthroughs. In M. Helander, T.K. Landauer, P. Prablhu (Eds.). Handbook of Human-Computer Interaction, Elsevier, Amsterdam, pp. 717-732.

C. Lewis, J. Rieman, 1994. Task-centered User Interface Design – a practical introduction. Complete text on-line. Also available through anonymous ftp: cs.colorado.edu (Directory: /pub/cs/distribs/clewis/HCI-Design-Book/.

J. Löwgren, T. Nordqvist, 1992. Knowledge-based evaluation as design support for graphical user interfaces. In Bauersfeld P., Bennett J., Lynch G. (Eds.). CHI'92 ACM Conference on Human factors in Computing Systems, Monterey, California, pp. 181-188, May 3-7, New York: ACM Press.

R.L. Mack, R., J. Nielsen, 1994. Executive Summary. In Nielsen J., R.L. Mack (Eds.). Usability Inspection Methods, Elsevier, pp. 4-23.

J.A. McCall, P.K. Richards, G.F. Walters, 1997. Factors in Software Quality, Volumes I, II and III. US Rome Air Development Center, Reports NTIS ADIA-049014,015,055, National Technical Information Services, US Department.

S. McDonald, R.J. Stephenson, 1998. Effects of text structure and prior knowledge of the learner on navigation in hypertext. Human Factors, 40 (1), 18-27.

M.E. Mendes, W. Hall, R. Harrison, 1998. Applying Metrics to the Evaluation of Educational Hypermedia Applications. Journal of Universal Computer Science 4(4), 382-403.

J. Mullin, A.H. Anderson, L. Smallwood, M. Jackson, E. Katsavras, 2001. Eye-tracking explorations in multimedia communications. In Blanford A., Vanderdonckt J., Gray P. (Eds.). People and Computer XV - Interaction without Frontiers - Joint Proceedings of HCI 2001 and IHM 2001, pp. 367-382, London: Springer.

J. Nielsen, 1993. Usability Engineering, Academic Press.

J. Nielsen, 1994a. Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier. Useit Papers and Essays. Available at: http://www.useit.com/papers/guerrilla_hci.html

J. Nielsen, 1994b. Heuristic evaluation. In J. Nielsen, R.L. Mack (Eds.). Usability inspection methods, Elsevier, pp. 25-62.

J. Nielsen, 1997. Loyalty on the Web. Useit Alertbox (August 1997). Available at: http://www.useit.com/alertbox/9708a.html

J. Nielsen, 1998. The Web Usage Paradox: Why Do People Use Something This Bad ? Useit Alertbox (August 1998). Available at: http://www.useit.com/alertbox/980809.html

J. Nielsen, R.L. Mack (Eds.), 1994. Usability inspection methods, Elsevier.

J. Nielsen, R. Molich, 1990. Heuristic evaluation of user interfaces. Proceedings CHI'90 Conference, Seattle, ACM Press, pp. 349-356.

D.A. Norman, 1986. Cognitive Engineering. In D.A. Norman, S.W. Draper (Eds.). User centred system design: new perspectives on human computer interaction, Erlbaum, Hillsdale, NJ, pp. 31-61.

O. Parlangeli, E. Marchigiani, S. Bagnara, 1999. Multimedia systems in distance education: effects of usability on learning. Interacting With Computers 12, 37-49.

A. Pollier, 1991. Evaluation d'une interface par des ergonomes: diagnostics et stratégies. Research report, INRIA, n°1391, France.

P.G. Polson, C.H. Lewis, J. Rieman, C. Wharton, 1992. Cognitive Walkthroughs: a method for theory-based evaluation of use interfaces. International Journal of Man-Machine Studies 36, 741-773.

J. Prümper, 1993. Software Evaluation based upon ISO 9241 Part 10. In T. Greching, M. Tscheligi (Eds.). Human Computer Interaction Vienna Conference, VCHI '93 Proceedings, Berlin, Springer, pp. 255-265.

H. Ruokamo, S. Pohjolainen, 1998. Pedagogical Principles for Evaluation of Hypermedia-Based Learning Environments in Mathematics. Journal of Universal Computer Science 4(3), 292-307.

D.L. Scapin, J.M.C. Bastien, 1997. Ergonomic criteria for evaluating the ergonomic quality of interactive systems, Behaviour and Information Technology 16, 220-231.

J. Scholtz, L. Downey, 1998. Methods for Identifying Usability Problems with Web Sites. Proceedings of the 7th International Confe. on Engineering for Human-Computer Interaction ECHT'98 (Sept 14-18, Crète).

J. Scholtz, S. Laskowski, 1999. Developing Usability Tools and Techniques for Designing and Testing Web Site. National Institute of Standards and Technology (NIST). Available at: http://www-09.nist.gov/div894/vvrg/jpaper/hf_and_web.htm.

A. Sears, D.J. Hess, 1998. The Effect of Task Description Detail on Evaluator Performance with Cognitive Walkthroughs. In C.M. Karat, A. Lund, J. Coutaz, J. Karat (Eds.). Proceedings of the Conference on Human Factors in Computing Systems (CHI'98, Los Angeles, CA, 18-23 April 1998), ACM Press, New York, pp. 259-260.

B. Senach, 1990. Evaluation ergonomique des interfaces homme-machine: une revue de la littérature. Research report, INRIA, n°1180, Sophia Antipolis, France, March.

B. Shackel, 1991. Usability, context, Framework, Definition, Design and Evaluation. In Shackel B. and Richardson S., Human Factors for Informatics Usability, Cambridge University Press, pp. 21-37.

N. Simpson, 1999. Managing the use of style guides in an organisational setting: practical lessons in ensuring UI consistency. Interacting With Computers 11, 323-351.

P.A. Smith, 1996. Towards a practical measure of hypertext usability. Interacting With Computers 8, 365-381.

A version of this paper has been published in: *Interacting with Computer*s, 16, 183-215, 2004.
The final version is available at:
http://www.sciencedirect.com/science?_ob=IssueURL&_tockey=%23TOC%235644%232004%23999839997%23484145%23FLA%23&_auth=y&view=c&_acct=C000036578&_version=1&_urlVersion=0&_userid=674936&md5=32ab4d10aa605e4dc3ee80e32add4ff5

S. Suzuki, A. Kaneko, K. Ohmura, 1999. QUIS: applying a new walkthrough method to a product design process. In H.J. Bullinger, J. Ziegler (Eds.). Proceedings of the 8th Human-Computer Interaction International Conference 1999 (HCI'99, Munich, Germany, August 22-26, 1999), Volume 2, Lawrence Erlbaum Associates, London, pp. 933-937.

M. Sweeney, M. Maguire, B. Shackel B., 1993. Evaluating user-computer interaction: a framework. International Journal of Man-Machine Studies 38, 689-711.

P. Trigano, 1997. Evaluation de l'interface homme-machine des logiciels éducatifs. Le journal du multimédia, 18, 12-15.

T.S. Tullis, 1998. A System for Evaluating Screen Formats: Research and Applications. In H.R. Hartson, D. Dix (Eds.). Advances in Human-Computer Interactions, Volume 2, Ablex Norwood, N.J., pp. 214-286.

J. Vanderdonckt, 1994. Guide ergonomique de la présentation des applications hautement interactives. Presses Universitaires de Namur, Namur, Belgium.

J. Vanderdonckt, 1998. Conception ergonomique de pages WEB. Vesale.

J. Vanderdonckt, 1999. Development milestones towards a tool for working with guidelines. Interacting With Computers 12, 81-118.

R.A. Virzi, 1997. Usability inspection methods. In M. Helander, T.K. Landauer, P. Prablhu (Eds.). Handbook of Human-Computer Interaction, Elsevier, Amsterdam, pp. 705-715.

C. Wharton, L. Rieman, C. Lewis, P. Polson, 1994. The Cognitive Walkthrough Method: a Practioner's Guide. In J. Nielsen, R.L. Mack (Eds.). Usability Inspection Methods, Elsevier, pp. 105-140.

D. Wildman, 1995. Getting the Most from Paired-User Testing. Interactions 2(3), 21-27.

S. Wills, 1994. Beyond browsing: making interactive multimedia interactive. Proceedings EdTech94 Conference, Rethinking the Role of Education in the Technological Age, Singapore, pp. 58-68, May.

D. Wixon, S. Jones, L. Tse, G. Casaday, 1994. Inspections and Design Reviews: Framework, History, and Reflection. In J. Nielsen, R.L. Mack (Eds.). Usability Inspection Methods, Elsevier, pp. 77-103.

P. Wright, A. Monk, 1991. A Cost-Effective Evaluation Method for Use by Designers. International Journal of Man-Machine Studies 35(6), 891-912.

J. Ziegler, 1999. Standards for Multimedia User Interfaces – Opportunities and Issues, in Human-Computer Interaction: Communication, Cooperation, and Application Design. In H.J. Bullinger, J. Ziegler (Eds.). Proceedings of the 8th Human-Computer Interaction International Conference 1999 (HCI'99, Munich, Germany, August 22-26, 1999), Volume 2, Lawrence Erlbaum Associates, London, pp. 858-862.