Intl. Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions

Thomas Mahatody, Mouldi Sagar & Christophe Kolski

To cite this article: Thomas Mahatody, Mouldi Sagar & Christophe Kolski (2010) State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions, Intl. Journal of Human–Computer Interaction, 26:8, 741-785, DOI: 10.1080/10447311003781409

To link to this article: https://doi.org/10.1080/10447311003781409

Published online: 27 Jul 2010.

INTL. JOURNAL OF HUMAN–COMPUTER INTERACTION, *26*(8), 741–785, 2010

ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447311003781409

Taylor & Francis
Taylor & Francis Group

State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions

Thomas Mahatody[1,2,3], Mouldi Sagar[1,2,3], and Christophe Kolski[1,2,3]
[1]Univ Lille Nord de France, Lille
[2]UVHC, LAMIH, Valenciennes, France
[3]CNRS, Valenciennes, France

This article discusses interactive system evaluation from the perspective of inspection methods, specifically the Cognitive Walkthrough (CW) method. The basic principles of CW are reviewed as proposed in the original version and the first two revisions. Then 11 significant extensions of CW are examined: Heuristic Walkthrough, The Norman Cognitive Walkthrough Method, Streamlined Cognitive Walkthrough, Cognitive Walkthrough for the Web, Groupware Walkthrough, Activity Walkthrough, Interaction Walkthrough, Cognitive Walkthrough with Users, Extended Cognitive Walkthrough, Distributed Cognitive Walkthrough, and Enhanced Cognitive Walkthrough. Four summaries are proposed: The first one concerns the conceptual, methododological, and technological aspects; the next two summaries deal with existing studies, first comparative and then noncomparative; and the last summary provides help for choosing a version or variant.

1. INTRODUCTION

The development of an interactive system is an iterative process, and evaluation is an important component of this process. For this reason, over the years, evaluation has been the subject of numerous studies and much research in the scientific community concerned with human–computer interaction (HCI). Today, many evaluation techniques and methods exist, and they have been reviewed and summarized to allow researchers an at-a-glance view of several methods (e.g., Dix, Finlay, Abowd, & Beale, 1993; Grislin & Kolski, 1996; Sears, 2003; Sweeney, Maguire, & Shackel, 1993). Usability inspection is the generic name for a set of methods that inspect (i.e., evaluate) the usability of the user interface (Bastien,

We thank the Nord-Pas de Calais region, the state, CISIT, and FEDER, which contributed to support this research. This research has been carried out under auspices of GDRE's "HumAn-MAchine SYstems in Transportation – HAMASYT." We thank the authors who sent us some of their articles in order to help us with our research, and the anonymous reviewers for their numerous constructive remarks.

Correspondence should be addressed to Christophe Kolski, LAMIH – FRE CNRS 3304, University of Valenciennes and Hainaut-Cambrésis, Le Mont Houy, F-59313 Valenciennes cedex 9, France. E-mail: Christophe.Kolski@univ-valenciennes.fr

2004; Cockton, Lavery, & Woolrych, 2003; Mack & Nielsen, 1994; Virzi, 1997). According to Nielsen and Mack (1994), "usability is a fairly broad concept that basically refers to how easy it is for users to learn a system, how efficiently they can use it once they have learned it, and how pleasant it is to use" (p. 3).

Cognitive Walkthrough (CW) is a usability inspection method that links the interface walkthrough to a cognitive model. The evaluator uses the interface to perform tasks that a typical interface user will need to accomplish. The actions and responses of the interface are evaluated according to the user's goals and knowledge through responses to questions related to the method's cognitive model, the differences between the user's expectations and the use reality (i.e., the steps really required by the interface). Like other common HCI evaluation methods, CW focuses on the basic principles of usability. However, unlike the other evaluation methods, CW also focuses on the cognitive activities of users, especially on their goals and knowledge when performing a specific task

CW was proposed by Lewis, Polson, Wharton, and Rieman in 1990, but it has since evolved. Later versions and extensions by other authors have been proposed. The basic principle remains the same in all versions and extensions: The method simulates the cognitive behavior of the user by responding to questions related to the user's cognitive model. Based on previous publications (Mahatody, Sagar, & Kolski, 2007a, 2007b), this article first describes the evolution of the CW method through its three versions: the first version proposed by Lewis et al. in 1990, their second version in 1992 (Polson, Lewis, Wharton, & Rieman, 1992), and finally the third version proposed 2 years later (Wharton, Rieman, Lewis, & Polson, 1994). Then, it reviews the various significant extensions of CW, providing a summary of the concepts, methods, and technology used in each: Heuristic Walkthrough (HW), The Norman Cognitive Walkthrough Method (NCW), Streamlined Cognitive Walkthrough (SCW), Cognitive Walkthrough for the Web (CWW), Groupware Walkthrough (GW), Activity Walkthrough (AW), Interaction Walkthrough (IW), Cognitive Walkthrough with Users (CWU), Extended Cognitive Walkthrough (Extended CW), Distributed Cognitive Walkthrough (DCW) and Enhanced Cognitive Walkthrough (Enhanced CW). Finally, four synthesis discussions are provided at the end of the article.

2. EVOLUTION OF THE ORIGINAL CW VERSIONS

2.1. First Version: CW1

The original version of CW (Lewis et al., 1990) was designed to evaluate walk-up-and-use interfaces (i.e., interfaces that can be used with little or no training). It is based on the theory of exploratory learning CE + (Polson & Lewis, 1990). The CE + model was the first cognitive learning theory to use HCI. Design guidelines, called "Design Principles for Successful Guessing," were derived from this theory to support the design of interactive systems requiring little or even no user training. The CE + model has three main components: a learning component, a problem-solving component, and an execution component. The model operates

as follows: A system user chooses an action among several alternatives based on the similarity between his or her goals and expected consequence of the action; after carrying out the action selected, the user evaluates the system response using the heuristics proposed by Lewis (1986, 1988). In this way, users evaluate their progress toward their goals. If the goal is achieved, the learning that occurs is registered by inscribing the steps taken by the system (i.e., the evaluated response) in the rule-based representation of procedural knowledge proposed by Kieras (1985). Otherwise, the problem-solving component is activated to discover appropriate action, and so forth. The execution component consists of triggering an applicable rule that matches the current context.

The first version of CW (called CW1 in this article) has two phases. In the preparation phase, first the evaluator, or a group of evaluators, specifies a series of tasks to be assessed. These evaluators may be designers, users, or usability inspection experts. Then, each task is broken down into action sequences. In the evaluation phase, each action is inspected by answering the questions that are generated by the simulation process of the CE+ model just described: selecting and performing the action and then evaluating the response of the system. Lewis et al. (1990) noted that the method was promising but not yet satisfactory because only 50% of observed errors were identified.

2.2. Second Version: CW2

To improve the first version of CW, a second version (called CW2 in this article) was proposed by the same authors in 1992 (Polson et al., 1992). It is based on an extension of the model CE +. The new model of this version is related to the theory of action (Norman, 1986) and the construction-integration model (Kintsch, 1988):

- Norman's action theory holds that task execution involves seven stages: establishing a goal, formulating an intention, specifying a sequence of actions, executing those actions, perceiving and interpreting the system state, and finally evaluating the system state in terms of the goal established in Stage 1. CW2 also uses this cyclic cognitive mechanism.
- Kintsch's model provides a framework for building two-phase connectionistic structures. In the first phase, called construction, all possible connections or relationships are built. During the second phase, called integration, inappropriate connections or relationships are eliminated based on the context of the field of study. In CW2, the Kintsch model is used to integrate the perceptual interface's representation of text or objects, taking the background and knowledge of the user into account.

In this second version of CW, the characteristics of the user, the context and the system use environment are all taken into account. In the new model adopted, in order to perform a task, the user must build a goal structure, which is similar to the hierarchical structure of the GOMS model (Card, Moran, & Newell, 1983): The main goal is broken down into intermediate goals corresponding to

the task breakdown, and the final goals correspond to actions. This goal structure incorporates the original goal by creating propositions representing the user's background knowledge, as well as the objects in the environment and the actions. An executable action is activated if there is a link between the original goal and this action. After the action is executed, each system response is interpreted, resulting in the deactivation of goals that have been reached and the construction of new propositions. These new propositions are linked to the current structure involving new actions, and so forth.

Like CW1, CW2 has two phases: preparation and evaluation. During the preparation phase, the tasks to be assessed are selected and broken down into sequences of actions and the user's goals are determined. During the evaluation phase, the evaluator uses three forms that help to guide the evaluation. These forms concern the problems that may arise and the failures that may occur while exploring the model. The first form assesses the relationship between the goals that are needed to manipulate the interface and common goals of the user, in order to determine whether there is an appropriate match between the two. Assuming that there is an appropriate match, the second form deals with the problems that arise when selecting the action. Taking the system response after the action has been carried out into consideration, the third form facilitates the development of user's goals.

According to its authors, this version is more complex and cumbersome to use than CW1 (Polson et al., 1992). Indeed, the evaluator must perform extremely detailed analyses, explicitly describing the user's goals and actions, analyzing how the user chooses an action, and finally explaining how the system feedback and its interpretation by the user changes the goals.

To mitigate the complexity and awkwardness of CW2, three of the original authors, plus three others, proposed an automated version (AutoCW2) implemented with Apple's Hypercard (Rieman et al., 1991). AutoCW2 allows the evaluator to complete the forms without using paper files and provides a simple editor for managing the goal structures, as well as for providing aid for each question.

2.3. Third Version: CW

The second version of CW has been the subject of various studies and reviews (Cuomo & Bowen, 1992; Desurvise, Kondziela, & Atwood, 1992; Rowley & Rhoades, 1992; Wharton, Bradford, Jeffrey, & Franzke, 1992;), which confirm the awkwardness of the method. Even using the automated version is still too time consuming. Two years after the second version, the same authors published a third version (called CW3 in this article) to address the criticisms (Wharton et al., 1994).

In CW3, the evaluator is invited to imagine a specific and credible scenario for each action that users must run to accomplish their task. To make the scenario credible, the evaluator must justify each action with respect to the user's background and knowledge and the feedback from the interface. To assess each action in the task, the evaluator must answer four questions related to various user thoughts and actions: (a) What is the user thinking at the beginning of the action (Q1: Will the user try to achieve the right effect?), (b) is the user able to locate the

command (Q2: Will the user notice that the correct action is available?), (c) is the user able to identify the command (Q3: Will the user associate the correct action with the effect that user is trying to achieve?), and (d) is the user able to interpret the feedback (Q4: If the correct action is performed, will the user see that progress is being made toward solution of the task?).

Studies and reviews were also done concerning this third version of CW (Huart, Kolski, & Sagar, 2004; Jacobsen & John, 2000; John & Packer, 1995; Riihiahoo, 2000; Sears & Hess, 1999). Most of the reviewing authors agreed that CW is tedious, despite being easy to learn and use.

3. VARIANTS OF THE ORIGINAL CW VERSIONS

Although still dependent on its theoretical foundations, the CW method has evolved mostly due to its practical value. This evolution has occurred to compensate for some limitations and to take into account the progress in the field of HMI or the different HMI types (e.g., Web, multimedia; Jeffries, Miller, Wharton, & Uyeda, 1991; John & Packer, 1995; Jacobsen & John, 2000; Huart et al., 2004). In this section, 11 extensions of CW are reviewed chronologically.

3.1. Heuristic Walkthrough

HW (Sears, 1997) is an evaluation method that combines two inspection processes—one scenario based and the other heuristic based—for use as a Usability Walkthrough method (Karat, Cambell, & Fiegel, 1992).

HW thus combines the Heuristic Evaluation (HE) method (Nielsen, 1992; Nielsen & Molich, 1990) and CW3 (Wharton et al., 1994). In HW, the evaluation is done in two phases. During Phase 1, the evaluator is guided by a prioritized task list and the list of questions from CW3. During Phase 2, however, the evaluator is free to explore any aspect of the system, guided by the knowledge obtained in Phase 1 (i.e., from the task list and the list of questions) and a list of heuristics, such as those provided by Nielsen (Nielsen, 1992; Nielsen & Molich, 1990).

Sears used the criteria proposed by Bastien and Scapin (1995) to compare the three methods: HE, CW3, and HW. These criteria are validity (ability of the method to focus on specific and relevant aspects of the interface), thoroughness (ability of the method to assess all aspects of the interface), and reliability (ability of the method to produce same results under the same conditions). Through this comparison, Sears showed that HW can find more usability problems than CW3 while producing fewer false usability problems than the HE method.

3.2. Norman's Cognitive Walkthrough

As part of the project AVANTI (AdaptiVe and Adaptable iNteractions for multimedia Telecommunication applIcation), the NCW method addresses the particular problems raised by this project (e.g., the presence of design teams in different cities, even in different countries) and deals with interaction problems at a

high level of abstraction (Rizzo, Mandrigiani, & Andreadis, 1997). In fact, as the AVANTI team discovered, none of the initial three versions of CW is particularly appropriate for design teams composed of people from different cultures. In addition, all three version of CW focus on low-level interactions (e.g., typing on the keyboard, clicking with the mouse).

NCW is certainly based on the Norman model (Norman, 1986), but this model was amended by Rizzo et al. (1997) as shown in Figure 1. According to Hutchins, Hollan, and Norman (1985), the notion of cognitive distance refers to the amount and quality of information processing needed to fill the gap between states (e.g., intention, evaluation). This notion may be applied to both action execution and outcome evaluation. In the first case, the cognitive distance refers to the amount of information processing needed to reduce the gap between the intention to act and the physical actions through which the intention was communicated to the system, or in other words, the process of translating the user's thoughts and goals into the language of the system. In the second case (outcome evaluation), cognitive distance refers to the amount of mental effort needed to translate the information displayed by the system in the terms of the conceptual model adopted by the user.

The evaluation consists of testing the user's activities by running one or more tasks within a given scenario. The evaluator explores the system, looking for the actions that help to accomplish the task. The evaluator chooses actions whose descriptions correspond with what he or she is trying to do and then interprets the

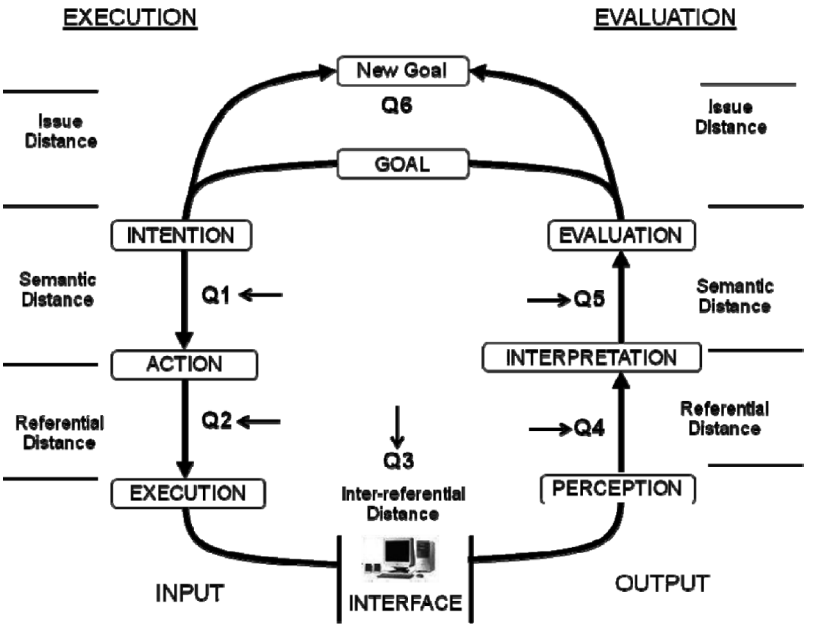

FIGURE 1 The modified Norman model (adapted from Rizzo, Mandrigiani, & Andreadis, 1997).

system's response to assess whether progress has been made toward accomplishing the task or if the goal should be reconsidered. This process allows the evaluator to, first, determine whether the user has correctly interpreted the meaning and form of the interface and, second, whether the user was able to set a feasible goal and execute the correct action on the adequate subject.

At each stage of the interaction, certain questions must be answered:

> Q1: Will the feasible and correct action be made sufficiently evident to the user and do the actions match the intention as stated by the user? (Intention-Action)
> Q2: Will the user connect the correct action description with what s/he is trying to do? (Action-Form)
> Q3: Will the user receive feedback in the same place and modality as where s/he has performed her/his action? (Action Input - Feedback Output).
> Q4: Will the user interpret the system's response to the chosen action correctly (i.e., will s/he know if s/he has made a right or wrong choice?) (Outcome - Form).
> Q5: Will the user properly evaluate the results (i.e., will s/he be able to assess if s/he got closer to her/his goal ?) (Form - Assessment).
> Q6: If the goal is wrong (or can be ameliorate), will the user understand that the intention s/he is trying to fulfil cannot be accomplished within the current state of the world (or will s/he find out alternative goals?) (Action/Outcome – Concern). (Rizzo et al., 1997, p. 308)

If the answer is not completely affirmative, the evaluator communicates the problems to design team members, with specific alternatives. This method allows the team members in different locations to communicate without ambiguity.

3.3. Streamlined Cognitive Walkthrough

Spencer (2000) found that CW3 did not produce consistently good results, and for this reason development teams may judge it to be inappropriate. In addition, in his opinion, CW3 was difficult to apply in large software development companies due to several constraints.

- Clearly, time is a major constraint for designers, who are often under pressure to complete their tasks on time. Any activity that is not immediately useful or that requires a lot of time is left out. Because a lot of time is needed not only to respond to CW3's four questions (see section 2.3), but also to process the voluminous responses produced, this method is consistently omitted.
- During the evaluation session, identifying problems often leads to a lengthy discussion of the design because the evaluation team is trying to solve the problems. This means that the evaluation time increases or is spent on the design process.
- Some designers become defensive during the evaluation session. Feeling that they have already invested so much time in designing the system, they try to defend their design in order to avoid going overtime on the project to take the problems raised by the evaluation into account.

To address these difficulties, Spencer proposed SCW. SCW is a variant of CW3 with five phases. The first phase defines the walkthrough input. In the second, the evaluation team is convened to define the role of each team member and to determine what should and should not be done during the evaluation, as well as what needs to be done to avoid making the designers defensive about their design. The third phase is the inspection phase, in which the evaluators proceed in the same manner as in CW3, responding to the following two questions:

1. Will the users know what to do at this step?
2. If the users do the right thing, will they know that they did the right thing and are making progress towards their goal? (Spencer, 2000, p. 355)

In the fourth phase, the problems detected are recorded, and in Phase 5, they are fixed. This method relies heavily on the rigor of the evaluation process.

3.4. Cognitive Walkthrough for the Web

CWW is a method for detecting and fixing errors that occur when browsing and searching for information on a Web site (Blackmon, Polson, Kitajima, & Lewis, 2002; Kitajima, 2006). This is an extension of CW3.

CWW is based on a theory called CoLiDeS (Comprehension-based Linked model of Deliberate Search; Kitajjma, Blackmon, & Polson, 2000). CoLiDeS is itself an extension of the model LICAI (Linked model of Comprehension-based Action planning and Instruction taking), which is an exploration model based on understanding (Kitajima & Polson, 1997). CWW allows users to simulate surfing on a Web page with a goal in mind. The exploration consists of selecting an action (e.g., click on an icon or a link) and then assessing the outcome in terms of the goal. CoLiDeS assumes that the selection of an action is a two-phase process: attention and action selection. During the attention phase, the user parses the page into a range of subregions, generates a brief description of each subregion based his or her background and knowledge about Web site page layout conventions, and then focuses on the subregion whose description matches his or her current goal. During the action selection phase, the user generates descriptions of all graphic widgets in the target subregion and acts on the one that is closest to his or her goal.

The questions for CWW evaluation are as follows:

Q1: Will the correct action be made sufficiently evident to the user?
Q2: Will the user connect the correct subregion of the page with the goal using heading information and her understanding of the sites page layout conventions?
Q3: Will the user connect the goal with the correct widget in the attended to subregion of the page using link, labels and other kinds of descriptive information?
Q4: Will the user interpret the system's response to the chosen action correctly? (Blackmon et al., 2002, p. 463)

CWW transforms the CoLiDeS approach using the Latent Semantic Analysis (LSA) (Landauer & Dumas, 1997): Instead of a subjective assessment of these questions by evaluators, LSA is used to estimate the semantic similarity between texts based on statistical analysis of a large corpus. In CWW, LSA is used to estimate the semantic similarity between the user's goal descriptions and descriptions of the page subregions, as well as the goals and descriptions of possible actions on the Web page.

The evaluation is done in four stages (Figure 2). In Stage 1, a set of realistic user goals is compiled; each goal described in 100 to 200 words, and then the correct selection for each goal is identified. In Stage 2, the semantic similarities between the goals, titles, and link labels are calculated. In Stage 3, problematic titles and link labels are identified, under CoLiDeS' assumptions. (These assumptions hold that a title may cause a problem if it is not familiar or if it can be confused with others; a title with an LSA value of less than 0.8 is assumed to be unfamiliar, whereas a pair of headings with a LSA value of more than 0.6 is assumed to be easily confused.) In Stage 4, goal-specific competition is inspected.

3.5. Groupware Walkthrough

A usability inspection method for groupware, GW is a "substantive modification of cognitive walkthrough" that allows the "complexities of teamwork" to

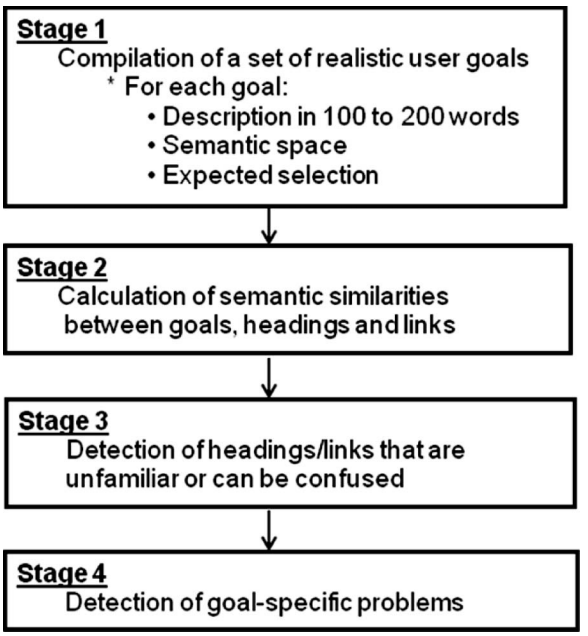

FIGURE 2 The stages of Cognitive Walkthrough for the Web (adapted from Blackmon, Polson, Kitajima, & Lewis, 2002).

be considered (Pinelle & Gutwin, 2002, p. 455). GW includes both a task model and a walkthrough process. The task model allows the identification and analysis of collaborative real-world tasks, as well as allowing the variability of the working group's actions to be expressed. The walkthrough process is designed to assess whether the system supports the tasks, which are modelled as a set of possible paths toward a desired result. Every possible path is explored during the walkthrough, and the support for each path in the interface is evaluated.

Collaboration involves two distinct types of work: *taskwork*, which includes the actions that must occur to accomplish the task, and *teamwork*, which refers to the actions that group members must perform as a group to accomplish a task. Traditional approaches to evaluation focus on taskwork. Teamwork can be analyzed at high levels and low levels of abstraction. High-level abstraction, which includes social and organizational factors, is analyzed using Groupware Task Analysis (GTA; Van der Veer, Lenting, & Bergevoet, 1996; Van der Veer & van Velie, 2000). The methods associated with GTA focus on task specification based on ethnographic observation. Although GTA is an effective way to gather information on the work context, its level of abstraction is too high to be used in a Walkthrough inspection situation. To capture the level of detail needed for the walkthrough, it is necessary to examine teamwork at the lower level of abstraction inherent to the mechanics of collaboration.

Nevertheless, the high-level analysis must precede the low-level analysis. The mechanics of collaboration (Gutwin & Greenberg, 2000) are the core activities of group work, the small-scale actions and interactions that the group must undertake to accomplish a task collaboratively. These activities are invariable through a variety of social and organizational factors. The mechanics of collaboration includes explicit communication, monitoring, action coordination, planning, assistance, and protection.

The main components of the task model are scenarios, individual and/or group tasks and subtasks, and actions. The scenario is a description of the high-level activities that must be accomplished to produce a specific result. Task analysis begins with the collection of observed data about the work and continues with the identification of episodes of collaborative interaction. Both scenarios and tasks can be specified, and teamwork can be analyzed using the mechanics of collaboration.

Once the teamwork analysis has been completed, GW can be conducted for each collaborative scenario as follows:

- Review the scenario to become familiar with the users, the intended outcome and the surrounding circumstances.
- For each task in the scenario:
  - Attempt to carry out each alternate subtask,
  - Record how each subtask was carried out, and
  - Record the problems encountered but assume they are resolved and continue.
- After each task, ask the following questions:
  - Can the task be performed effectively? (i.e., Does the interface supply the means to perform the task correctly?);

- Can the task be performed efficiently? (i.e., Would the group make the effort required to perform the task?); and
- Can the task be performed satisfaction? (i.e., Would the group be motivated to do this task, and would they be happy with the outcome?).

- After completing all tasks, determine whether the interface allows the group to produce the overall intended outcome.

3.6. Activity Walkthrough

AW (Bertelsen, 2004) is a modified version of CW3 based on activity theory. AW is powerful descriptive tool that focuses on understanding human activity, incorporating the concepts of intentionality, history, mediation, collaboration, and development (Nardi, 1996). According to activity theory, the unit of analysis is the activity itself. An activity is composed of a subject, an object, and the tools that serve as mediation. A subject is a person or group of persons engaged in an activity. An object is held by the subject and motivates the activity. Mediation can occur through the use of different types of tools: material tools and mental tools, including culture, ways of thinking and language. According to Kaptelinin (1996), the computer can be considered as a tool for mediation. The activity is executed through a conscious (intentional) goal directed by actions that are accomplished through unconscious (automatic) operations.

In AW, the evaluation is conducted in six phases. In Phase 1 (preparation), typical tasks to be analyzed are determined based on a needs specification. In Phase 2 (contextualization), the activities in which the application is used as a mediator are identified. For each activity, first the actions performed via the application and through which the activity is accomplished must be identified and then the objects and results of these actions. The other means of accomplishing the activity without the application (i.e., other artefacts used as activity mediators) must also be dealt with. In addition, user expectations, based on user experience with the application or similar tools, must be considered. In Phase 3 (task verification), the extent to which each task corresponds to purposeful actions in the activities in which the application will be embedded are assessed based on the contextualization of the application. In Phase 4 (task analysis), each task is broken down into sequences of atomic operations, as is done in CW3. In Phase 5 (evaluation) three questions about the perceptual context, the system's response and the user's learning, respectively, are asked and answered. In Phase 6 (verification), the task is analyzed based on the response given in the previous phase. Following these six phases, a report is prepared, summarizing the results of the evaluation process.

3.7. Interaction Walkthrough

IW (Ryu & Monk, 2004) is a modified version of CW2 based on the theory of cyclic interaction (Figure 3), initially introduced by Card et al. (1983) as the "recognise-act" cycle. Norman's seven-stage model (Norman, 1986) also imagined a cycle of interaction, but neither of these authors expressed explicitly how the environment

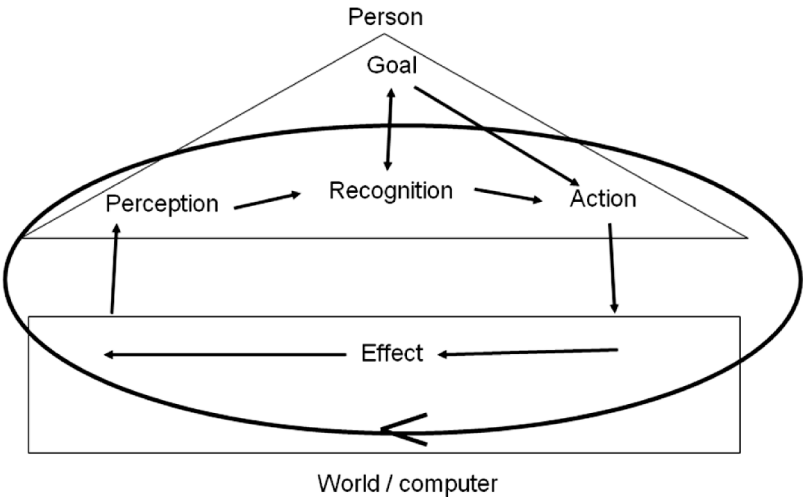

FIGURE 3 Cyclic interaction theory according to Monk (1998).

and context influence user interaction. The means available to the user to manipulate the system is focused on the action ("goal-action path"). An action triggers a system effect through the entry device ("action-effect path"). After the execution phase is complete, the system attains a new state within the environment. Then the "effect-goal path," which deals with perceived changes in the environment, continues the cycle for new goals in the interaction context. These three interaction paths lead to three classes of low-level problems: goal action, action effect, and effect goal.

Goal-action problems are often observed in cases of unpredictable actions designed to accomplish a goal. An action can be unpredictable due to the low affordance of a valid action or the high affordance of an invalid action. According to Norman (1988, 1999), affordance refers to the object attributes that allow human beings to know how to use them.

Action-effect problems occur when the same action produces different effects depending on the system mode. There are three types of problems provoked by mode ambiguity. The first type of problem is due to hidden modes. In this case, errors tend to occur when the user forgets that the system mode has changed. The second type is due to partially hidden modes. In this case, errors tend to occur when the user can't recognize the mode indicator in the information environment. The third type of problem is due to misleading mode signals. In this case, errors tend to occur when the user perceives an incorrect mode signal, which leads the user to believe that the correct action has been taken, when in fact the opposite is true.

Effect-goal problems are due to the goal reorganization process occasioned by the effect-goal path. Two types of problems are possible: goal construction problems and goal elimination problems. Goal construction problems occur when the goal is ambiguous or irrelevant, whereas goal elimination problems occur when the system cues are missing or misleading.

IW evaluation looks at the three paths of the interaction cycle for each of the problem categories just described: affordance (low & high), mode (hidden, partially hidden, & misleading), goal (construction & elimination). The evaluation takes place in three phases. In the preparation phase, the topics susceptible to have detectable problems are located. For example, for detecting a mode problem, the preparation phase identifies a same action that has different effects and lists the system effects that could inform the user of the current mode. In the walkthrough phase, questions are answered to allow to the different problems to be detected. For detecting a mode problem, the questions are

> Q1: Does the user recognise (rather than recall) the current mode from system effects?
> Q2: Are system effects sufficiently salient for the user to discriminate the mode change from the previous interaction?
> Q3: Is it possible that mode signals imply different modes? (Ryu & Monk, 2004, p. 313)

Finally, in the verification phase, the previously found problems are reviewed to verify whether they conform to the paths indicated by the system designer. (For more details, please consult Ryu & Monk, 2004.)

3.8. Cognitive Walkthrough With Users

CWU (Granollers & Lorés, 2006) explicitly integrates users into the walkthrough process. This version of CW takes place in three phases. In the first phase, CW is performed in the traditional manner. In the second, users are incorporated into the process as follows. First, representative system users must be recruited. After a brief introduction, these users are invited to perform all the tasks defined in the Walkthrough that correspond to their profile. During this interaction, the users are asked to express aloud their thoughts, feelings, and opinions on any aspect of the system or prototype. Users perform the tasks without any explanation other than the brief introduction; at the end of each task, they note the main deficiencies detected. Once the users have completed the tasks, they are invited to comment on the problems identified during the first phase in order to situate their point of view. In the third phase, experts review the doubts expressed by users during the second phase.

3.9. Extended Cognitive Walkthrough

Using CW often requires knowledge of cognitive psychology. Kato and Hori (2005) noted that some evaluators have difficulty understanding the differences between a correct object and a correct action, and also distinctions between the different questions. In an attempt to circumvent these problems, Kato and Hori (2005, 2006) suggest an extension of CW3, called Extended CW.

The theoretical model of the method is based on the Norman action model (Norman, 1986) but extended as shown in Figure 4. The changes in the model affect the third stage, splitting it up to allow data about the object and the action to be collected and interpreted separately. During the evaluation phase, nine questions must be answered (instead of the four in CW3); these questions are directly related to the extended Norman model (Figure 4):

(Q1) Will the user intend to achieve the right effect?
(Q2-a) Will the user notice that the correct object is available?
(Q2-b) Will the user know what the correct object refers to?
(Q2-c) Will the user notice that the correct action is available?
(Q2-d) Will the user know that the correct action should be applied to the correct object?
(Q3) Will the user be able to apply the correct action to the correct object without fail or difficulty?
(Q4) When the correct action is taken, will the user notice the physical change in the system state?
(Q5) Will the user know what exactly has happened to the system state?
(Q6) Will the user know that the current system state is nearer to completion of the task? (Kato & Hori, 2005)

Because information is accessible if it is easily perceptible and understood by system users, the authors of ECW think that information accessibility should be assessed according to its perceivabilty and its understandability. To make their method appropriate for detecting accessibility problems, they included questions

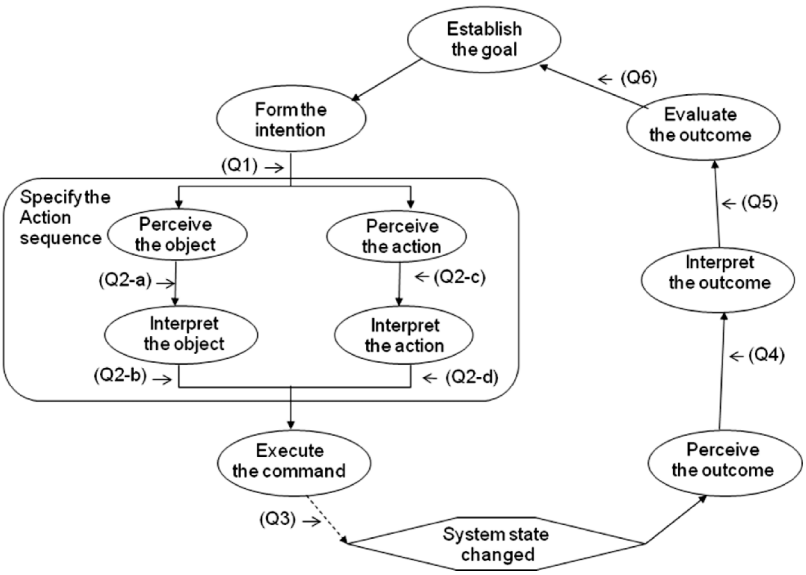

FIGURE 4 Extended Norman's action model (from Kato & Hori, 2005).

Q2-a and Q2-c for assessing perceptibility and Q2-b and Q2-d for assessing user understanding. They therefore find that ECW is better able to identify accessibility problems than CW3.

3.10. Distributed Cognitive Walkthrough

DCW (Eden, 2007, 2008) is a variant of CW3. This walkthrough method is based on two theories: distributed cognition (Hollan, Hutchins, & Kirsh, 2000) and distributed cognitive tasks (Zhang & Norman, 1994).

According to Hollan et al. (2000), the theory of distributed cognition, like any other cognitive theory, seeks to understand the organization of cognitive systems. However, these authors extend the definition of "cognitive" beyond the individual to include interactions between humans and objects in their environment. Applying distributed cognition to observed human activity involves three types of distribution processes: (a) Cognitive processes can be distributed through members of a social group, (b) cognitive processes may imply the coordination of internal and external information structures (see principle of distribution next), and (c) cognitive processes can be distributed over time so that the products of previous events can transform the nature of future events.

The representation of information affects the cognitive behavior of an individual for the same formal task structure. The theory of distributed cognitive tasks contributes to the problems of distributed representation of information, the interaction between internal and external representations, and the nature of the external representation.

The principle of distribution holds that part of a task is represented internally in mind and the other part is represented externally in the physical environment (Zhang & Norman, 1994). The internal representation may involve propositions, mental images, and production rules, whereas the external representation may involve physical symbols, external rules, external relationships embedded in physical configurations, and so on.

Like CW3, DCW includes four questions to which the evaluator must respond, with a negative response indicating a usability problem:

Q1: Will the way that information is represented provide all the knowledge required to carry out the task?
Q2: Will the way that information is represented show relevant previous progress towards completing the overall task?
Q3: Will the way that information is represented provide resources that keep the user from having to figure out or calculate anything in his/her head while carrying out the task?
Q4: If the current task is accomplished, will the way that information is represented be changed in any way so that the result of the task is accessible to current or future users at a later time or in a different place?

3.11. Enhanced Cognitive Walkthrough

Enhanced CW is an alternative version of CW3. It was proposed by Bligard and Osvalder (2007) and is used as an analytical approach for predicting and identifying use errors and usability problems. According to the IEC (2004), a use error is an act or omission of an act that has a different result than intended by the manufacturer or expected by the operator.

The approach suggested by Bligard and Osvalder (2007) comprises four phases: (a) the definition of the evaluation, (b) the description of the system, (c) the analysis of the interaction, and (d) the compilation of the results and reflection. Enhanced CW intervenes in the third phase (analysis of the interaction), in parallel with the predictive analysis of use errors (Bligard, 2007).

Enhanced CW proceeds in two stages. The first aims to predict usability problems; it consists of inspecting the sequence of user actions, using the following questions:

Q1: Will the user know that the evaluated function is available?
Q2: Will the user try to achieve the right effect?
Q3: Will the user interface give clues that show that the function is available?
Q4: Will the user be able to notice that the correct action is available?
Q5: Will the user associate the right clue with the desired function?
Q6: Will the user associate the correct action with the desired effect?
Q7: Will the user get sufficient feedback to understand that the desired function has been chosen?
Q8: If the correct action is performed, will the user see that progress being made towards the solution of the task?
Q9: Will the user get sufficient feedback to understand that the desired function has been performed?

The answer to each question is a number ranging between 1 and 5, where 1 means *no*, 2 means *probably no*, 3 means that the answer is unknown, 4 means *probably yes*, and 5 means *yes*. Like the third CW version, each answer must be justified. Any answer other than 5 suggests a potential usability problem.

The second phase of Enhanced CW is the identification of the problems; this phase consists of determining the category of each problem. The categorization is based on the justification and the description of the problem. Depending on the type of the HCI and the task concerned, there can be several problem categories. Bligard and Osvalder (2007) proposed five categories of utilisability problem:

- problems related to user knowledge and background
- problems related to information posting (i.e., the interface does not give any indication on the functions available or the way of using them)
- problems related to the way information is illustrated (i.e., the interface's location, appearance and contents can be badly interpreted or not understood)
- the problems related to the sequence of actions (i.e., the actions are probably performed in an unnatural way)

- the problems related to information feedback (i.e., the interface does not give enough indications of what the user should do)

4. SUMMARIES

We propose four complementary summaries, which are each described in succession. The first deals with the conceptual, methodological, and technological aspects of each version or variant. The second presents comparative studies of at least one version or variant of CW and at least one other evaluation method. The third focuses on noncomparative studies (i.e., the methods depend only on CW). The fourth summary is intended to facilitate the choice between the versions and variants of CW.

4.1. Summary of the Conceptual, Methodological, and Technological Aspects

At the beginning, CW was a method for assessing usability problems in "walk-up-and-use" systems (i.e., systems that can be used with little or no training). Due to its success, practitioners and researchers interested in this method tried to improve it or to adapt it for other specific types of application. Eleven modified or extended versions of CW have been thus far proposed. In most cases, the modifications and extensions concern theoretical or conceptual aspects; however, some concern methodological aspects. Only two versions deal with the technological aspects of the method.

CW1 (Lewis et al., 1990) requires an implicit knowledge of cognitive science to understand the terminology used in the forms that the evaluators fill out. For example, to succeed in using the method, the evaluators must understand the distinction between a goal and an action. To address this problem, CW2 (Polson et al., 1992) is more formal, guiding evaluators through detailed analyses to avoid ambiguities. The addition of Norman's action model and Kintsch's construction-integration model allows detailed analyses, structured according to the goals and actions. But the increased number of forms and lists of questions makes the method more complex and cumbersome to apply. An automated tool was developed (Rieman et al., 1991) to help evaluators fill out the different forms, thus helping them to focus on the evaluation. But even with this help, the authors of CW2 found the method costly in terms of time and energy. CW3 (Wharton et al., 1994) was developed in an effort to reduce the method's complexity and operating costs by introducing methodological changes. Nevertheless, in our opinion, even this third version of CW is far from perfect. In fact, most of the authors of the various extensions presented in this article have made comments, criticisms, and suggestions, ending up proposing an extension that would address the problems they identified.

In HW, changes were made on the methodological level. Because the first phase of HW includes the whole CW process, the second phase could be seen as redundant in terms of evaluation. Clearly, HW—with its second system assessment— is

cumbersome. Still, Sears's (1997) tests show that this redundancy allows HW to find more usability problems than CW. This is a promising result, which calls for further study.

The NCW method adopts the action model as a theoretical foundation, which allows the problem to be dealt with at a high level of abstraction, thus permitting the source of problems to be located during the evaluation. The methodology focuses on the level of cognitive distance, thereby facilitating the location of the problems. However, the reference document for this method (Rizzo et al., 1997) provides no information that would allow us to judge the validity of this method.

SCW makes several important methodological changes, involving the preparation of the evaluation team, the imposition of background rules for supervising the CW process, and the reformulation of the initial four questions into only two questions. These changes affect the time constraint, eliminate the lengthy design discussions during the evaluation, and prevent designers from going on the defensive (i.e., defending the mistaken views that are influenced by their design experience). We think that making the questions more general instead of more detailed could compromise the method's effectiveness for novice evaluators, but this needs to be verified in future experiments.

CWW makes changes on the conceptual level, adopting the CoLiDeS cognitive model. In addition, this is the only extension that makes changes in the technological aspects. This extension was designed specifically for assessing the Web site usability, with four questions that focus on navigation and information retrieval. To prevent subjective judgements of the answers to these four questions, the method employs LSA. LSA uses rich goal descriptions to incorporate more information about the users' understanding of the tasks, which results in more realistic goals. The choice of semantic space knowledge applied in LSA allows CWW to take a target population (e.g., chosen by age) into account during the evaluation. In our opinion, the effectiveness of this method will depend on the quality of the semantic space knowledge chosen and its interpretation.

GW is the only extension designed to evaluate Computer Supported Cooperative Work applications. Based on a model of the mechanics of collaboration, GW addresses the complexity and particularities of these applications, among other usability problems related to collaboration. Although its use of GTA (Van der Veer et al., 1996; Van der Veer & van Velie, 2000) does facilitate the detailed examination of the context of each task connected to the evaluation, GW is not able to detect classic usability problems and it would have to be improved in order to detect and take into account the different socio-organizational uses in real collaborative systems (Pinelle & Gutwin, 2002).

Based on the activity theory, AW allows evaluators to include more context in the evaluation and the dynamic commitment of actions, because the basic element of the analysis is the real activity of users. However, we believe that tackling CW through the activity theory—which requires that many factors be taken into account, including intra- and interindividual variability, but also use contexts—will certainly complicate the evaluation, making it more cumbersome and increasing the difficulty of results interpretation, which already requires extensive expertise. Both the method's author (Bertelsen, 2004) and others (e.g., Ryu, 2008) appear to share this point of view, because they emphasize that the method is complex and cumbersome

IW is based on the cyclic interaction model and considers low-level problems. In contrast to other methods, which are supposed to detect fairly complicated problems, locating them and explaining their origins, IW detects problems more easily because they are low level. However, Ryu and Monk (2004) believe that user behavior is dependent on the technology used and that technology must be seen in the context of a state of activity that has meaning for the user. Given the complexity of the interaction analysis, the validity of this method's results may be doubted.

CWU is a CW method that embraces real user involvement in the broadest sense of the term. It is an interesting option, in terms of ergonomics, but because it involves both experts and users, the end result seems to remove the method from the "for experts only" category of inspection methods. Given this change in category, we wonder whether the chosen models and methods for intervention and analysis are the most appropriate (e.g., use of concurrent "think-aloud" verbalization).

ECW modifies CW conceptually, adopting an extended model of the Norman action theory. ECW helps to evaluate information accessibility, taking into account all aspects from cognitive perception to understanding to action implementation. Like in the NCW method, the questions used in this method focus on cognitive distances. After trying the method, Kato and Hori (2006) concluded that ECW detects more usability problems than CW3 in a comparable time. Although the theoretical foundations of the method seem solid, the experimental conditions (e.g., type of participants, compensation, types of HMI evaluated) described by Kato and Hori (2005) do not allow us to draw conclusions about the effectiveness of ECW in other conditions, in terms of the validity criteria proposed by Gray and Salzman (1998).

DCW is a variant of CW3 that adopts cognitive distribution theories. Theoretically, this method can identify usability problems through dimensions, such as time, space, and social structure. However, the author of the method did not provide results of experiments using the method.

Last, Bligard and Osvalder (2007) underlined the effectiveness of Enhanced CW within the framework of an analytical approach for predicting and identifying usability problems and use errors. It seems to us that it is now necessary to show its effectiveness if used outside of this framework.

Table 1 summarizes the salient features of all the CW versions and extensions. Three columns contain the theoretical/conceptual aspects, the methodological aspects, and the technological aspects.

To conclude this overview, we would like to emphasize that, although they are not homogeneous, the variations of CW show what happens when users evaluate a particular system by performing tasks with this system (conceptual aspect), providing specific questions (methodological aspect), and using various tools and techniques (technological aspect). Due to space constraints, we cannot detail the questions asked in each method, but in general the questions are designed to evaluate, for each interaction, whether users manage to (a) formulate goals, (b) perform the appropriate actions that will help them reach their goals, (c) interpret the system responses correctly, and finally (d) accomplish their goals properly.

Table 1: Summary of All the Cognitive Walkthrough (CW) Versions and Their Evolution

Name	*Theoretical and Conceptual Aspects*	*Methodological Aspects*	*Technological Aspects*
CW1 (Lewis et al., 1990)	CE+ learning by exploration model	1. Preparation: choosing the task to be evaluated, breaking up the task into atomic actions 2. Evaluation using one form	- Without proposed material
CW2 (Polson et al., 1992)	CE+ learning by exploration model, Norman's model of action, Construction-intégration model	1. Preparation: choosing and describing the task to be evaluated, breaking up the task into atomic actions, identifying the target user of the system, describing the initial goals structures 2. Evaluation using three forms	- Evaluation forms: goal structure, choosing and executing action, modification of goal structure - Proposition of automated version (Rieman et al., 1991)
CW3 (Wharton et al., 1994)	Learning by exploration model CE+	1. Preparation: choosing the task to be evaluated, describing in detail each scenario 2. Evaluation taking into account the detail of scenario	- Proposition of guide
Heuristic Walkthrough (Sears, 1997)	CE+ learning by exploration model	Combining CW, Heuristic Evaluation and Usability Walkthroughs	- List of heuristics
Norman CW (Rizzo et al., 1997)	Norman's model of action	As CW3 with questionnaire relying Norman's model	- Questionnaires focusing on the action
Streamlined Cognitive Walkthrough (Spencer, 2000)	Like optimized CW	As CW3 with reinforcement of the evaluation control	- Proposition of questionnaires - Proposition of ground rules for conducting the evaluation
Cognitive Walkthrough for the Web (Blackmon et al., 2002; Kitajima, 2006)	CoLiDeS model	Four stages: compiling a set of realistic user goal and intended selection, assessing semantic similarity with Latent Semantic Analysis, identifying problematic heading/link labels, finding goal-specific problems	- Analysis tools avalaible at http://autocww.colorado.edu/HomePage.html
Groupware Walkthrough (Pinelle & Gutwin, 2002)	Dedicated to Groupware (collaboration processes)	- Task modeling (scenario, task, subtask) - Task analysis with Groupware Task Analysis - Evaluation of task scenario	- Proposition of questionnaires

Activity Walkthrough (Bertelsen, 2004)	Activity theory	Seven stages: identifying typical task to be analysed, contextualization, task vérification, task analysis, evaluation, global verification, reporting	- Proposition of questionnaires - Synthesis report
Interaction Walkthrough (Ryu & Monk, 2004)	Cyclic interaction theory (Monk, 1998)	- Finding the usability problem relying the interaction(about goal-action, action-effect, goal construction and goal elimination) - Four evaluations consisting of three stages for each (preparation, evaluation, verification)	- Proposition of questionnaires
Cognitive Walkthrough with User (Granoller & Lorés, 2005)	- CW involving the user - User verbalization	- CW with real user intervention. - Run in two stages: CW by expert, user intervention	- Proposition of questionnaires
Extended Cognitive Walkthrough (Kato & Hori, 2005)	Extended Norman's model of action	As CW3 but the questionnaires are focusing to the cognitive semantic distance identified by the Extended Norman's model	- Without additional tools proposed
Distributed Cognitive Walkthrough (Eden, 2007)	- Distributed Cognition - Distributed Cognitive Task	As CW3 but the questionnaires are focusing on the distributed cognition	- Without additional tools proposed
Enhanced Cognitive Walkthrough (Bligard & Osvalder, 2007)	Learning by exploration model CE+	Considered as a part of an analytical approach for predicting and identifying use errors and usability problems	- Proposition of (a) questionnaires as CW3 and (b) list of problem types

4.2. Summary of the Comparative Studies

In this section, we look at the diverse studies that evaluate the performance of the various versions and variants of CW. Table 2 summarizes all the comparative studies examined. In these studies, one of the three CW versions is compared with another version or a variant of CW or another evaluation technique. The evaluated system is specified, including the types of evaluators taking part in the study. Then, the main quantified results related to usability problems are provided. To provide more details about the various elements in the table, the conditions of each study and the corresponding conclusions are summarized next.[1]

Lewis et al. (1990)

Conditions of the study. These authors evaluated two tasks for each of the four user interface designs in a mail messaging system. Four evaluators, including three who were conversant with the CE+ theory, took part in the evaluation. Each evaluator independently evaluated the two tasks for each of the four interfaces. The results were compared with the previously available data, which came from empirical evaluations of the same tasks carried out on the four interfaces with at least 15 subjects.

Conclusions of the study. The authors highlighted 20 problems, 18 of which were found by at least two of the four evaluators. Then they compared the results from the CW1 evaluations with the empirical data, which led them to conclude that using CW1 allowed the detection of 50% of the usability problems detected using the empirical method.

Polson et al. (1992)

Conditions of the study. These authors reported the results of tests using CW2 on student projects. Three interfaces were evaluated: a voice mail directory, a text editor, and a document router. The first interface was tested both with CW2 and with a thinking-aloud user test. The second interface was tested using CW2 on about 50 varied tasks, and then the results were compared with those obtained by an empirical method (i.e., field trouble reports and user interviews). Finally, the last interface was evaluated using CW2 on a rather simple task, and the results were compared with those from subsequent usability tests.

[1]Insofar as several of these studies are rich in results and analyses, it was necessary to summarize them. The three authors provide the summaries and conclusions in this article for the convenience of the readers. For more details, it is necessary to refer to the original articles (see References section).

Table 2: Summary of the Comparative Studies

	CW			CW Variants											Other Means of Evaluation												
Study	*CW1*	*CW2*	*CW3*	*HW*	*NCW*	*SCW*	*CWW*	*GW*	*AW*	*IW*	*CWU*	*TCW*	*DCW*	*HCW*	*EM*	*TA*	*HE*	*UT*	*GL*	*CA*	*GOMS*	*UAN*	*RS*	*Type of System Evaluated*	*Type and No. of Evaluators*	*Numerical Results Concerning the Detection of Usability Problems*
Lewis et al. (1990)	X														X									Mail messaging system	4 experts	CW1 detected 50% of the problems found using the empirical method
Polson et al. (1992)		X														X								Voice mail directory,	Students (no. not provided)	CW2 detected 50% of the problems found using the think aloud method
		X													X									Text editor	Students (no. not provided)	CW2 detected 30% of the problems found using other methods (e.g., user interview, field trouble report)
		X																X						Document routing system	Students (no. not provided)	CW2 detected 100% of the problems found by subsequent user tests
Sears (1997)			X	X													X							System for learning the visual effects of rendering algorithms	20 students	For Heuristic Evaluation: - on average 1.5 to 3 serious problems, depending on the number of evaluators. - on average 2.5 to 5.8 problems of intermediate severity, depending on the number of evaluators. - on average 3 to 10.7 problems of minor severity, depending on the number of evaluators. - on average 1.7 to 7 nonproblems, depending on the number of evaluators

(*Continued*)

Table 2: (Continued)

Study	CW			CW Variants											Other Means of Evaluation									Type of System Evaluated	Type and No. of Evaluators	Numerical Results Concerning the Detection of Usability Problems
	CW1	CW2	CW3	HW	NCW	SCW	CWW	GW	AW	IW	CWU	TCW	DCW	HCW	EM	TA	HE	UT	GL	CA	GOMS	UAN	RS			
																								System for learning the visual effects of rendering algorithms	20 students	For CW: - on average 1.3 to 3 serious problems, depending on the number of evaluators. - on average 2.9 to 5 problems of intermediate severity, depending on the number of evaluators. - on average 2 to 6 problems of minor severity, depending on the number of appraiser. - on average 0.3 to 1.4 nonproblems, depending on the number of evaluatorrs
Sears (1997)			X	X													X							System for learning the visual effects of rendering algorithms	20 students	For Heuristic Walkthrough: - on average 2.1 to 3 serious problems, depending on the number of evaluators. - on average 3.3 to 5,9 problems of intermediate severity, depending on the number of evaluators. - on average 3.4 to 9.6 problems of minor severity, depending on the number of aevaluators. - on average 0.3 to 1.4 nonproblems, depending on the number of evaluators

Granoller & Lorés (2004)		X	X						Microscope simulator	One expert and 10 students	CW3 detected 31% of the problems found using CW with Users
									Conference manage-ment system	One expert and 3 users	CW3 detected 37% of the problems found using CW with users
Kato & Hori (2005)		X		X					Digital camera interface	20 students	CW3 detected 72% of the problems found using Extended CW
									Paint software		Extended CW detected 88% of the problems found using CW3
Eden (2008)		X			X				Restaurant system	42 computer science students	Average severity rating =1.68 for DCW compared to 0.82 for CW; actionability rating =2.38 for DCW compared to 1.71 for CW, usability relevance rating =2.76 for DCW compared to 1.81 for CW
Jeffries et al (1991)	X					X	X	X	Graphical user interface for an operating system	3 data- processing specialists (using CW2), 4 HCI specialists (using Heuristic evaluation), one human factors specialist and 6 users (using Usability testing), 3 other data- processing specialists (using Guidelines)	CW2 detected 35 problems, compared to 105 with the heuristic evaluation, 31 with Usability testing and 35 with Guidelines

(*Continued*)

Table 2: (Continued)

Study	CW			CW Variants										Other Means of Evaluation									Type of System Evaluated	Type and No. of Evaluators	Numerical Results Concerning the Detection of Usability Problems
	CW1	CW2	CW3	HW	NCW	SCW	CWW	GW	AW	IW	CWU	TCW	DCW	HCW	EM	TA	HE	UT	GL	CA	GOMS	UAN	RS			
John & Marks (1997)			X														X			X	X	X	X	Multimedia authoring system	Data-processing specialist (master's degree)	CW3 detected 42 problems including 2 nonproblems, 13 problems concerning the non-implemented aspects of the system and 11 problems that led the developers to modify the code.
			X														X			X	X	X	X	Multimedia authoring system	Evaluator with a master's degree in English and skill in 2 computer programming languages	Heuristic Evaluation detected 88 problems including 24 nonproblems, 3 problems concerning the non implemented aspects of the system and 7 problems that led the developers to modify the code.
			X														X			X	X	X	X	Multimedia authoring system	Evaluator with a master's degree in architecture and skill in 3 computer programming languages	GOMS detected 44 problems including 6 nonproblems, 17 problems concerning the non-implemented aspects of the system and 3 problems that led the developers to modify the code.
			X														X			X	X	X	X	Multimedia authoring system	Evaluator with a bachelor's degree in electrical engineering and skill in six computer programming languages	Claims Analysis detected 25 problems including 6 nonproblems, 5 problems concerning the non implemented aspect of the system and no problem that led developers to modify the code.

	X			X		X	X	X	X	Multimedia authoring system	Data-processing specialist (bachelor's degree)	Using Action Notation detected 17 problems including 14 nonproblems, 3 problems concerning the non implemented aspects of the system and no problem that led the developers to modify the code.
	X			X		X	X	X	X	Multimedia authoring system	Evaluator with skill in one computer programming language	Reading Specification detected 69 problems including 4 nonproblems, 20 problems concerning the nonimplemented aspects of the system and no problem that led the developers to modify the code.

Note. CW1 = Cognitive Walkthrough, first version; CW2 = Cognitive Walkthrough, second version; CW3 = Cognitive Walkthrough, third version; HW = Heuristic Walkthrough; NCW = Norman's Cognitive Walkthrough; SCW = Streamlined Cognitive Walkthrough; CWW = Cognitive Walkthrough for the Web; GW = Groupware Walkthrough; AW = Activity Walkthrough; IW = Interaction Walkthrough; CWU = Cognitive Walkthrough with Users; TCW = exTended Cognitive Walkthrough; DCW = Distributed Cognitive Walkthrough; HCW = enHanced Cognitive Walkthrough; EM = Empirical Method; TA = Think Aloud; UT = Usability testing; GL = Guidelines; CA = Claims Analysis; GOMS = Goals, Operators, Methods and Selection rules; UAN = User Action Notation; RS = Reading specification.

Conclusions of the study. The test on the first interface showed that the CW2 method allowed the detection of 50% of the usability problems found with the thinking-aloud user test. The comparison data for the second interface were not matched specifically to the tasks evaluated during the CW2 tests; nonetheless, CW2 detected 30% of the usability problems, including 70% of the problems considered to be serious. Finally, all the usability problems detected for the third interface with the subsequent usability tests were also detected by CW2.

Sears (1997)

Conditions of the study. This study compared the performances obtained with the following methods: CW3, HW, and HE. Twenty computer science graduate students, who had taken courses in user interface design and evaluation, took part in the evaluation of a system for learning the visual effects of rendering algorithms. The evaluations were carried out by groups of evaluators, comprising 1 to 5 participants.

Conclusions of the study. The results of the study showed that, for the groups of 2 to 5 participants, HW allowed the identification of significantly more minor and intermediate-level problems compared to CW3, whereas HE allowed the identification of significantly more minor problems than CW3 (see Table 2). In addition, HW and CW3 allowed the identification of significantly more false positive problems than HE for the groups of 2 to 5 participants.

Granoller and Lorés (2006)

Conditions of the study. CWU was tested on a microscope simulator and a conference system management. For the first system, an expert evaluator evaluated the application with CW3. Then 10 medical students individually carried out the second phase of CWU. For the second system, an expert evaluator evaluated the application. Then 3 users with different profiles representing the end users of the conference system management carried out independently the second phase of CWU.

Conclusions of the study. The results showed that with CW3, the expert evaluator found five usability problems on the first system. By using CWU, 11 other usability problems were found. For the second system, the expert evaluator found 9 problems in the first phase. In the second the three users found 14 other usability problems.

Kato and Hori (2005)

Conditions of the study. This study compared CW3 and Extended CW. Twenty students with prior experience using CW3 took part in the study. Two interactive systems (a digital camera's user interface and a paint software) were evaluated. The participants were asked to use a 7-point scale to rate the ease with which they understood and were able to answer the questions in the two versions

of CW. Then, they compared the two versions. Finally, they were asked to pick the version they would most like to use in the future.

Conclusions of the study. There were 43 potential usability problems found on the digital camera interface using CW3 compared to 59 with Extended CW. For the paint software, CW3 allowed the detection of 70 usability problems compared to 62 for Extended CW. The results also showed that the Extended CW questions were easier to understand than those of CW3. Moreover, 70% of the participants judged that it was easier to answer the Extended CW questions than the CW3 questions. For this reason, 70% of them preferred to use Extended CW in the future.

Eden (2008)

Conditions of the study. The objective of this study was to compare DCW and CW3 in terms of several criteria: average severity rating, usability relevance rating, and actionability rating. The first criterion (average severity rating) measures the usability of all the questions used during a complete evaluation session; according to Eden (2008), "this measure of the average severity rating provides the most direct assessment of the usability of a design for a task, with respect to the evaluation method used" (p. 62). The second (usability relevance rating) provides an overall rating in terms of how well each complete participant evaluation makes relevant usability statements. This criterion allows the assessment of the relevance of the evaluation results as carried out by an unspecified evaluator using a given method. The third criterion (actionability rating) is used to assess up to what point the usability problem descriptions indicate as explicitly as possible what should be changed to resolve the problems.

A scenario drawn from the hospitality industry (specifically, a restaurant scenario) was studied. For this scenario, the waiters must take the various drink orders, which then must be tracked until the drinks have been prepared and served to the customers. The participants were undergraduate students in computer science who did not have experience with usability evaluation methods.

Conclusions of the study. According to the three criteria, both DCW and CW3 provided interesting results. For the first criterion, the average value was 1.68 for DCW compared to 0.95 for CW3, which suggests that DCW seems to encourage the evaluators to be more severe than CW3. For the second criterion, the average value was 2.76 for DCW compared to1.81 for CW3, which indicates that DCW led to more relevant results than CW3. For the third criterion, the average value was 2.38 for DCW compared to 1.71 for CW3, which means that DCW led the evaluators to better describe the solutions to the problems found than CW3.

Jeffries et al. (1991)

Conditions of the study. In this study, CW2 was tested by a group of evaluators and then the results were compared to other evaluation methods (HE, usability testing, and guidelines). The interface evaluated was the graphical interface HP-VUE of an operating system UNIX for a Hewlett-Packard workstation.

This interface was not a walk-up-and-use interface. Four variable-size groups of evaluators, each assessing a different method, took part in the study. The group using the heuristic evaluation method was composed of four HCI specialists, the one using the usability testing method was composed of six regular PC users who were not familiar with the UNIX system and led by a human factor specialist. Finally, the groups using the guidelines and CW2 methods were each composed of three computer scientists, who were not members of the system design team but were members of the same laboratory.

Conclusions of the study. The heuristic evaluation method allowed 105 usability problems to be identified, whereas usability testing identified 31, guidelines identified 35, and CW2 identified 35. The authors noted that even though the heuristic evaluation method allowed the identification of many problems, particularly serious problems, it required the participation of HCI specialists, whereas CW2 can be used by developers who are not necessarily HCI specialists.

John and Marks (1997)

Conditions of the study. The goal of this study was to assess the predictive and persuasive power of six evaluation methods—CW3, HE, Claims Analysis, GOMS, User Action Notation, and Reading Specification—by comparing their predictions with that of the user test, Think Aloud). Six evaluators evaluated a multimedia authoring system. They each used a different evaluation method. The evaluators who used Claims Analysis had a bachelor's degree in electrical engineering. Those who used CW3 had a master's in computer science. Those who used GOMS had a master's in architecture. Those who used HE had a master's in English. Those who used User Action Notation had a bachelor's in computer science. Last, those who used Reading Specification were in the junior year of an undergraduate computer science program. Twenty undergraduate business students used the Think Aloud method.

Conclusions of the study. CW3 allowed the detection of 42 problems compared to 88 for HE, 24 for Claims Analysis, 44 for GOMS, 17 for User Action Notation and 69 for Reading Specification. However, 2 (5%) of the 42 problems found by CW3 were false alarms (i.e., which are not usability problems) compared to 24 (17%) for HE, 6 (24%) for Claims Analysis, 6 (13%) for GOMS, 14 (82%) for User Action Notation, and 4 (6%) for Reading Specification.

Among the 40 problems found using CW3, 13 (33%) were related to aspects of the system that had not yet been implemented, compared to 3 of 64 (5%) for HE, 5 of 18 (26%) for Claims Analysis, 17 of 38 (44%) for GOMS, 3 of 3 (100%) User Action Notation, and 20 (31%) for Reading Specification. These results confirm the conclusions of Wharton et al. (1994), who stated that CW3 could be used early in the development process. Among the 15 problems found by CW3 related to the aspects of the system that had already been implemented, 11 of 15 (73%) led the developers to change the code, compared to 7 of 41 (17%) for HE, 0 of 11 (0%) for Claims Analysis, 3 of 10 (30%) for GOMS, 0 of 0 (0%) for User Action Notation, and 9 of 23 (39%) for Reading Specification.

By combining similar problems, the number of problems found by the six evaluators was reduced to 54. Among these 54 problems, 26 were observed in the user tests; among these 26 problems, 13 led to code changes.

4.3. Summary of Noncomparative Studies

Table 3 summarizes all of the noncomparative studies undertaken by various authors. For each study, the system evaluated is described, as well as the type of participant taking part in the study. When numerical results are available, they are provided.

Rizzo et al. (1997)

Conditions of the study. This study did not have numerical results. The authors simply described an example implementation of the method. The scenario studied involved a person with wheelchair who lived in Rome and decided to plan a one-day visit to Siena, in Italy. For each activity in the task, the members of design and evaluation team used the questions of the action model.

Conclusions of the study. Related to the NCW method, this study concluded, "It allowed the design team working in different laboratories to avoid ambiguity in communicating the problems discovered during the evaluation by pointing out the questions that received a negative answer and the factors involved (e.g. Intention-Action or Action/Outcome - Concern) with an indication of both the cognitive and the physical aspect of the problem" (p. 39).

Spencer (2000)

Conditions of the study. This study highlighted the effectiveness of the SCW method in a project management context with social constraints (see section 3.3). The Integrated Development Environment under development was evaluated by a eight-member team (three usability specialists, one graphic designer, and four project managers responsible for various aspects of the Integrated Development Environment specifications). Only the team leader, a utilisability specialist, was familiar with CW method. The test was conducted in two sessions, separated by 1 week. The first session lasted 90 min, including 20 min devoted to the preparation phase and covered 32 action sequences (only the results of the first session are discussed in their publication).

Conclusions of the study. The first session allowed 24 potential problems to be found. The participants agreed that 14 of the 24 usability problems found were due to a lack of user knowledge, and the 10 remaining problems were due to a lack of system feedback when the correct action was carried out. The results show that the team members did not waste time defending their design choices.

Table 3: Summary of the Noncomparative Studies

Study	CW			CW Variants											Type of System Evaluated	Type and No. of Evaluators	Conclusions of the Study
	CW1	CW2	CW3	HW	NCW	SCW	CWW	GW	AW	IW	CWU	TCW	DCW	HCW			
Rizzo et al. (1997)					X										Planning system for trips for the elderly and handicapped	Members of the design and evaluation team	The method allowed the team working in various laboratories to avoid ambiguity when communicating the problems found during the evaluation.
Spencer (2000)						X									Integrated Development Environment system	3 usability specialists, 1 graphic designer, and 4 project managers	Detection of 24 potential problems in 90 min.
Blackmon et al. (2002)							X								Web site (online encyclopedia)	Students (no. varied depending on the experiment)	Experiment 1: 41% "first click" success rate for the goals affected by usability problems detected by CWW compared to 70% for the goals not affected by problems. Experiment 2: 68% "first click" success rate for the goals affected by usability problems detected by CWW compared to 89% for the goals not affected by problems. Experiment 3: 38% "first click" success rate for the goals affected by usability problems detected by CWW compared to 62% for the goals not affected by problems.
Pinelle & Gutwin (2002)								X							Home care system	The two authors	Five main types of problems detected
Huart et al. (2004)		X													Multimedia application: encyclopedia for children	2 experts (1 usability specialist, 1data-processing specialist) and 2 nonexpert ergonomists	The expert evaluators found 18 problems compared to 9 for the nonexpert evaluators.

				Multimedia application: interactive museum visit		The expert evaluators found 16 problems compared to 5 for the nonexpert evaluators.
				Multi-media application: educational software system intended for children 4 years old and older		The expert evaluators found 20 problems compared to 5 for the nonexpert evaluators.
				Multi-media application: virtual reality game		The expert evaluators found 21 problems compared to 4 for the nonexpert evaluators.
Bertelsen (2004)		X		Student project	Students	Method considered too complicated by the students
Ryu & Monk (2004)		X		None (only illustrations are provided in the article)	Not applicable	No numerical data is provided, but interesting paths are discussed in terms of the helpfulness of the method for the design process.
Bligard & Osvalder (2007)			X	Dialysis machine	Not communicated	No numerical data are provided allowing conclusions to be drawn about the validity of Enhanced CW; nevertheless, the global approach seems effective for the evaluation of complex systems (of the type evaluated).

Note. CW = Cognitive Walkthrough; HW = Heuristic Walkthrough; NCW = Norman's Cognitive Walkthrough; SCW = Streamlined Cognitive Walkthrough; CWW = Cognitive Walkthrough for the Web; GW = Groupware Walkthrough; AW = Activity Walkthrough; IW = Interaction Walkthrough; CWU = Cognitive Walkthrough with Users; TCW = exTended Cognitive Walkthrough; DCW = Distributed Cognitive Walkthrough; HCW = enHanced Cognitive Walkthrough.

Blackmon et al. (2002)

Conditions of the study. This study compared the the "first-click" success percentage of the participants on headings or links with usability problems and headings or links without problems. The authors carried out three experiments simulating the search function of an online encyclopedia. The participants in these three experiments were college educated. Experiment 1 was carried out on a initial online encyclopedia. Experiment 2 was carried out on the same encyclopedia, but with an additional post (a long description of the current goal). Experiment 3 was carried out on a second online encyclopedia, using the same procedure as Experiment 2 (i.e., with an additional post describing the current goal).

Conclusions of the study. Some of the analysis results are shown in Table 3. The success rate of Experiment 2 was higher than that of Experiment 1 by approximately 23 points, and the success rate of Experiment 3 is lower than that of Experiment 2 by approximately 29 points. According to the authors, the results show that "the CWW can identify characteristics of Web pages that differentially affect to user performance. (p. 468)" In addition, CWW provides specific diagnostics, which can guide the resolution of the problems that it identifies.

Pinelle and Gutwin (2002)

Conditions of the study. The authors published a case study to show how to use GW. They evaluated an initial prototype of a home care system. The system allowed the members of the medical team (e.g., physicians, nurses) to share documents and to communicate via a chat tool.

Conclusions of the study. Five main usability problems were detected: the person receiving a message might not see it, the variable typing skills in the group could generate difficulties, the system does not provide enough information on a person's availability, the system does not allow multiple meetings of subgroups within the home care team, and the identification of an element on a shared document can cause problems. They noted that GW let them to revise their design easily and quickly.

Huart et al. (2004)

Conditions of the study. This study used CW2 to evaluate four multimedia systems with various degrees of interactivity. The first system, an encyclopedia for children from 8 to 14 years old, had lowest degree of interactivity in the study. The second system was a museum exploration tool intended for any public. The third system was an educational software system intended for children 4 years old and older. The fourth system, a virtual reality game, had the highest degree of interactivity in the study. Four evaluators—including a ergonomist, a interactive system design, and evaluation specialist (both called expert evaluators) and two master's students in industrial ergonomics (called nonexpert evaluators)—evaluated each of the four systems in turn.

Conclusions of the study. By the end of the evaluations, 80 problems had been detected in the four systems, 22 of which were detected by at least two evaluators. Among the problems detected by the experts, 58 were not detected by the nonexperts; among those detected by the nonexperts, 6 were not detected by the experts. The authors concluded that CW2 can be used to evaluate multimedia systems with various degrees of interactivity. Nevertheless, the effectiveness of the method depends greatly on the expertise of the evaluator, with the expert evaluators finding many more problems than the nonexpert evaluators.

Bertelsen (2004)

Conditions of the study. The objective of this preliminary study was to obtain information about the applicability and usefulness of AW. The participants were students (level not specified in the study) taking the author's HCI course. The students formed groups of three, with each group evaluating an HCI interface of its choice.

Conclusions of the study. This preliminary study indicates that AW is very complicated to use in its current form. Nevertheless, several students concluded that the contextualisation in Phase 3 is an important advance over state-of-the-art inspection.

Ryu and Monk (2004)

Conditions of the study. The authors described the stages and the necessary questions for using the IW method. Although not proposing a complete study, they do illustrate their explanations with examples of HCI.

Conclusions of the study. The authors concluded that IW method is a promising alternative for evaluating HCI with low-level interaction.

Bligard and Osvalder (2007)

Conditions of the study. In this article, Enhanced CW was integrated into a broader analytical approach for predicting and identifying usability problems and use errors. Two studies were conducted in the medical field on a dialysis machine interface. In the first study, the suggested approach was used to compare three different design solutions. In the second study, the objective was to improve a design.

Conclusions of the study. The authors considered that the evaluations provided the company with more specific information about the usability problems and use errors related to the dialysis machine interface prototype. However, the results do not provide any information concerning the performance of Enhanced

CW if it is used separately from the analytical approach for predicting and identifying usability problems and use errors.

4.4. Summary of the Studies Intended to Help Make the Choice of a Version or Variant

We did not find any studies intended to help the evaluators (both practitioners and researchers) to choose a method among the various CW versions and variants presented in the literature. A priori, this seems an extremely difficult, or even impossible, task. There are very important differences between the methods in terms of structure, experimental methodology, expected results, type of the study, evolution of the versions, and variants. The various summaries (sections 4.1, 4.2, and 4.3) that we have proposed may provide some assistance in making this choice, but of course more needs to be done.

Moving in this direction, we offer Table 4, which completes the information given in the preceding summaries. We highlight key quotations from the original articles that we think characterize each method's specificity ans we give an informal opinion for each version and variant.

Table 4: Quotations and Perspectives on the Specificity of the Various Versions and Variants

CW, Variants, and Their Evolution	*Quotations*	*Our Point of View*
CW1	"The walkthrough with a very limited investment in resources, approximately an hour per task per interface, can detect almost 50 percent of the problems encountered by users of the design." (Lewis et al., 1990, p. 240)	Intended for walk-up-and-use interfaces with at least a mockup For each user action, rates the percentage of potential users expected to have problems
CW2	"The cognitive walkthrough is an evaluation methodology that focuses on ease of learning. It is especially appropriate for the development of applications where users must (or prefer) to master a new application or function by learning through exploration." (Polson et al., 1992, p. 742)	Intended for walk-up-and-use interface with at least a mockup For each walkthrough question, rates the percentage of potential users expected to have problems
CW3	"The cognitive walkthrough method promises to be a valuable addition to the designer's suite of tools. The new version of a method is flexible enough to fit into given software development process. The method identifies problems with a design early in the process and, by describing the reasons for those problems, it suggests design changes early on." (Wharton et al., 1994, p. 139)	Uses a success/failure story to encourage the choice of an action

(*Continued*)

Table 4: (Continued)

CW, Variants, and Their Evolution	*Quotations*	*Our Point of View*
HW	"A new technique is described that combines the benefits of heuristics evaluation, cognitive walkthroughs, and usability walkthrough." (Sears, 1997, p. 243)	This method that associates the approaches well known by usability professionals, particularly heuristic evaluation. This makes it possible to potentially detect more usability problems.
NCW	"The design issues are: the problems deriving from distribution of the teams collaborating to the project in several cities (sometimes different European countries); and the need to face high-level interaction problems in the evaluation process. One important action taken to face these issues was the development of a variation of the Cognitive Walkthrough based on the Norman's model of action." (Rizzo et al., 1997, p. 305)	This method allows high-level problems to be solved (unlike IW, see below). It is interesting as a first approach to the evaluation of an interactive system, without going into detail about its realization.
SCW	"Managers, developers, and other team members are pressured for time, tend to lapse into lengthy design discussions, and are sometimes defensive about their user-interface designs. By enforcing four ground rules, explicitly defusing defensiveness, and streamlining the cognitive walkthrough method and data collection procedures, these social constraints can be overcome, and useful, valid data can be obtained. This paper describes a modified cognitive walkthrough process that accomplishes these goals." (Spencer, 2000, p. 353)	This variant has a social dimension that does not exist in the other variants.
CWW	"The new cognitive walkthrough for the Web (CWW) is superior for evaluating how well websites support users' navigation and information search task." (Blackmon et al., 2002, p. 463)	This method allows the design of Web pages under construction to be critiqued as they are completed or the pages of an online site to be evaluated. It is an automated method, which is not the case for all the others.
GW	"The technique is a substantive modification of cognitive walkthrough to include consideration for the complexities of teamwork. The two components of groupware walkthrough are a task model for identifying and analysing real-world collaborative tasks, and a walkthrough process for assessing a system's support for those tasks." (Pinelle & Gutwin, 2002, p. 455)	This task inspection method uses GTA. It is suitable for evaluation using collaborative scenarios.

(*Continued*)

Table 4: (Continued)

CW, Variants, and Their Evolution	*Quotations*	*Our Point of View*
AW	"The method is a modified version of the cognitive walkthrough, and is aimed to systematically include the context and history of use." (Bertelsen, 2004, p. 251)	This method is based on activity theory. It systematically integrates the use context, which is not the case for the other variants.The advantage of this method is that it allows a compromise to be reached: the contexts are taken into account, without requiring an in-depth description.
IW	"The method is a modified version of cognitive walkthrough and the analysis focuses on the issue of direct concern to the practitioner who intends to identify low-level interaction problems in their design specification." (Ryu & Monk, 2004, p. 304)	This method is an alternative for analyzing low-level HCI.
CWU	"The new variant which can be regarded as a new method incorporates users in way that combines the advantages of the initial method with those contributed by the availability of end users in the evaluation of interactive systems." (Granollers & Lorés, 2006, p. 254)	This method combines expert and user interventions, unlike all the other versions and variants. It is an interesting approach that integrates "think-aloud" verbalization.
Extended CW	"Applied to a Web design evaluation study, the extended CW was shown to be more effective in identifying accessibility and usability problems while remaining as efficient as the currently-practiced CW." (Kato & Hori, 2007, p. 1)	This is an interesting method for the evaluators without a lot of knowledge about cognitive psychology. It is intended to facilitate detection of accessibility problems.
DCW	"The Distributed Cognitive Walkthrough (DCW) method is useful for the identification of aspects of usability issues related to interaction between people, artifacts, and information, across dimensions such as time, space, and social structures." (Eden, 2008, p. 2)"DCW method uses concepts from distributed cognitive theory to view interaction between people and information as transcending interactions with graphical user interfaces, allowing the DCW method to be useful for evaluation of design ideas in almost all areas of interaction design; for example, evaluation of ubiquitous computing, service design (e.g. Starbucks customer/worker experience), and mathematical notations (e.g. Newton versus Leibniz Calculus notation)." (Eden, 2007, p. 1)	This method offers interesting prospects in terms of general interactions (i.e., not only between one human and one interactive system). It is based on distributed cognition, which is not the case for the other versions and variants.

(*Continued*)

Table 4: (Continued)

CW, Variants, and Their Evolution	*Quotations*	*Our Point of View*
Enhanced CW	"The approach is based on the methods Hierarchical Task Analysis (HTA), Enhanced Cognitive Walkthrough (ECW) and Predictive Use Error Analysis (PUEA)." (Bligard, 2007, p. 427)"ECW employs a clearly detailed procedure for simulating the user's problem-solving process in each step of the interaction. Throughout the interaction, it is checked whether the supposed user's established goal and previous experience will lead to the next correct action." (Bligard, 2007, p. 432)	This method classifies the problems according to their severity, proposing 5 problem categories that allow a better identification. It is integrated into more global approach to complex systems.

Note. CW1 = Cognitive Walkthrough, first version; CW2 = Cognitive Walkthrough, second version; CW3 = Cognitive Walkthrough, third version; HW = Heuristic Walkthrough; NCW = Norman's Cognitive Walkthrough; SCW = Streamlined Cognitive Walkthrough; CWW = Cognitive Walkthrough for the Web; GTA = Groupware Task Analysis; GW = Groupware Walkthrough; AW = Activity Walkthrough; IW = Interaction Walkthrough; HCI = human–computer interaction; CWU = Cognitive Walkthrough with Users; TCW = exTended Cognitive Walkthrough.

5. CONCLUSIONS AND PERSPECTIVES

Generally speaking, CW allows the evaluation of the ease with which a user completes a task with minimal system knowledge and the ease of interface exploration/learning. Initially, it was based on a learning model (Polson et al., 1992) that was itself inspired by the Norman action theory (Norman, 1986). Users integrate a perceptual entry into their background knowledge to build a representation that will enable them to perform a task. Through the interface, CW aims to simulate the user's cognitive process to accomplish the task. Forms with specific questions that must be answered guide the user through the walkthrouth. CW can be applied in the system design and development phase to identify system usability problems as soon as the system is modelled and its features specified. It can also be used retroactively to determine the difficulties in using a system through executing specific scenarios and special testing. Representing the cognitive processes employed by the user can help to improve system usability.

CW is a widely recognized evaluation method in the HMI scientific community. Its theoretical foundations and its practical interest are such that, since it appeared on the scene, it has been the subject of numerous studies and conceptual, methodological, and technological extensions. To demonstrate the efficiency and performance of each method mentioned in this article, the authors of the methods often perform a limited number of tests (sometimes as few as a single test), and their conclusions are too often subjective. To determine which method is the best, the literature provides only a few studies that make partial comparisons, such as Eden (2008), Granoller and Lorés (2005), Kato and Hori (2005, 2006),

and Sears (1997). See Table 2. The authors of the various versions and extensions often adopt different methodologies and techniques for judging the validity of the evaluation results. No comprehensive study has ever been made to compare the different versions and extensions to determine which method is better using the same techniques to judge the results.

Interestingly enough, some studies have been done to compare CW with other types of evaluation methods (e.g., Desurvire et al., 1992; Jeffries et al., 1991; Karat et al., 1992). However, Gray and Salzman (1998) examined these comparative studies and have observed that they have many anomalies. Several studies have shown that an evaluation method may be effective for a given system but not necessarily for others. For example, Huart et al. (2004) noted that CW is not well suited for evaluating certain types of multimedia applications.

Considering the various steps adopted by the authors to show the effectiveness and performance of the methods, we think that a global comparison of the various techniques would constitute a particularly interesting direction for further research.

It would be important to involve various types of evaluators (e.g., expert, beginners) and various types of systems, taking inspiration, for example, from Huart et al. (2004), who considered in their study several multimedia systems with different levels of interactivity and in diverse contexts (e.g., the DCW method that requires an approach involving distributed cognition).

Our investigations suggest that CW has evolved significantly, particularly in terms of its conceptual aspects. However, we were not able to systematically prove the validity and effectiveness of each change or extension. With respect to the methodological and technological aspects of CW, we find that there are still many imperfections. Future research should thus go in this direction. Many methodological weaknesses remain in terms of preparing for the assessment (e.g., evaluator training, task analysis, context considerations) and using the evaluation results. On the technological front, we believe that the partial automation of the method is necessary because it would help to make the method easier to use and reduce the time needed for the evaluation. We expect to explore these two aspects in future research. Moreover, an audio-video recording could be integrated explicitly in one or more versions or variants of CW to facilitate the collection, analysis and interpretation of the data (toward what John and Marks proposed in 1997).

We have shown that an evaluation method can be effective for a given system, but not inevitably for other systems. This observation has led some researchers to adapt certain versions to particular system types and application domains. For example, Bligard and Osvalder (2007); Edwards, Moloney, Jacko, and Sainfort (2008); Jasper (2008); and Niculescu (2008) adapted certain versions and variants of CW for evaluating HCI in the medical domain. Antona, Mourouzis, and Stephanisis (2007) adapted CW2 for universal access HCI evaluation. Gonzáles, Lorés, and Granollers (2007) adapted CW3 for evaluating Latin-American Web sites. Ruttkay and Akker (2008) also adapted CW3 for evaluating the interaction between humans and virtual humans.[2] And finally Wang (2008) also adapted CW3

[2]According to Cassell, Sullivan, Prevost, and Churchill (2000; as cited in Ruttlay & Akker, 2008), "Virtual humans (VHs) are synthetic characters produced by computational means, which are intended to look like and communicate as real people do" (p. 90).

for designing and evaluating dynamic intelligent menus. Given the start made by these researchers, we think that carrying out further experiments with CW in various domains would also be an interesting direction for research.

As for us, our detailed analysis of the literature, our application of the CW method in student practical work in the domain of the Web and interactive software for 15 years, the various experiments with multimedia systems, as well as the observations made earlier, have led us to study and design an evaluation assistance environment, exploiting a set of variants of the CW method. The first mockup of this environment is described in Mahatody, Kolski, and Sagar (2009). From this platform, we hope to generalize the use of CW for various types of HCI and application domains.

REFERENCES

Antona, M., Mourouzis, A., & Stephanisis, C. (2007). Towards a walkthrough method for universal access evaluation, universal access in human computer interaction. In *Coping with Diversity, Lecture Notes in Computer Science*, *4554*, 325–334.

Bertelsen, O. W. (2004). The activity walkthrough: An expert review method based on activity theory. *Proceedings of NordiCHI*. pp. 251–254.

Bastien, J. M. C. (2004). L'inspection ergonomique des logiciels interactifs: Intérêts et limites. [Ergonomic inspection of interactive software: interests and limits] In J. M. Hoc & F. Darses (Eds.), *Psychologie ergonomique: tendances actuelles* [Ergonomic Psycology: Current Texdexies]. Paris: PUF.

Bastien, J. M., & Scapin, D. L. (1995). Evaluating a user interface with ergonomic criteria, *International of Human–Computer Interaction*, *7*, 105–121.

Blackmon, M., Polson, P., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the Web. *Proceeding of CHI*, 463–470.

Bligard, L. O. (2007). *Prediction of medical device usability problems and use errors—An improved analytical methodical approach*. Chalmers University of Technology, Göteborg.

Bligard, L. O., & Osvalder, A. L. (2007). An analytical approach for predicting and identifying use error and usability problem. *HCI and Usability for Medicine and Health Care, Lecture Notes in Computer Science*, *4799*, 427–440.

Card, S. K., Moran, T. P., Newell, A. (1983). *The psychology of human–computer interaction*. Hillsdale, NJ: Erlbaum.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied conversational agents*. Cambridge, MA: MIT Press.

Cockton, G., Lavery, D., & Woolrych, A. (2003). Inspection-based evaluations. In J. A. Jacko & A. Sears (Eds.), *The human–computer interaction handbook* (pp. 1118–1138). Mahwah, NJ: Erlbaum.

Cuomo, D. L., & Bowen, C. D. (1992). Stages of user activity model as basis for user-centered interface evaluation. *Proceedings of the Human Factor and Ergonomics Society 41st annual meeting*, 1254–1258.

Desurvise, H. W., Kondziela, J. M., & Atwood, M. E. (1992). What is gained and lost when using evaluation methods other than empirical testing. In A. Monk, D. Diaper, & M. D. Harrison (Eds), *People and computers VII: proceedings of HCI92, York, September 1992* (pp. 89–102). Cambridge University Press.

Dix, H., Finlay, J., Abowd, G., & Beale, R. (1993). *Human–computer interaction*. Englewood Cliffs, NJ: Prentice Hall.

Eden, J. (2007, June). Distributed cognitive walkthrough (DCW): A walkthrough-style usability evaluation method based on theories of distributed cognition. *Proceedings of the 6th ACM SIGCHI conference on Creativity & Cognition*. p. 283.

Eden, J (2008). *The Distributed Cognitive Walkthrough: The impact of differences in cognitive theory on usability evaluation*. Unpublished doctoral dissertation, Drexel University, Philadelphia, Pennsylvania.

Edwards, P. J., Moloney, K. P., Jacko, J. A., & Sainfort, F. (2008). Evaluating usability of a commercial electronic health record: A case study. *International Journal of Human–Computer Studies*, *66*, 718–728.

González, M. P., Lorés, J., & Granollers, A. (2007). Assessing usability problems in Latin-American academic Web pages with Cognitive Walkthroughs and datamining techniques, usability and internationalization. *HCI and Culture, Lecture Notes in Computer Science*, *4559*, 306–316.

Granollers, T., & Lorés, J. (2006). Incorporation of users in the evaluation of usability by Cognitive Walkthrough. In R. Navarro-Prieto & J. L. Vidal (Eds.), *HCI related papers of interacción 2004*, 243–255.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human–Computer Interaction*, *11*, 203–261.

Grislin, M., & Kolski, C. (1996). Evaluation des interfaces homme-machine lors du développement des systèmes interactifs [Human–machine interface evaluation during interactive system development]. *Technique et Science Informatiques*, *15*(3), 265–296.

Gutwin, C., & Greenberg, S. (2000). The mechanics of collaboration: developping low cost usability evaluation methods for shared workspaces. *Proceedings of 9th IEEE WETICE 2000*, 98–103.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human–computer interaction research. *ACM Transactions on Computer–Human Interaction*, *7*(2), 174–196.

Huart, J., Kolski, C., & Sagar, M. (2004). Evaluation of multimedia applications using inspection methods: The cognitive walkthrough case. *Interacting with Computer*, *16*, 183–215.

Hutchins, E. L., Hollan, J. D., & Norman, D. A. (1985). Direct manipulation interface. *Human–Computer Interaction*, *1*, 311–388.

IEC. (2004). *Medical electrical equipment—Part 1–6: General requirements for safety—Collateral standard: Usability. International Electrotechnical Commission (EIC)*.

Jacobsen, N. E., & John, B. E. (2000). Two case studies in using Cognitive Walkthrough for interface evaluation (Tech. Rep. CMU–CS-00-132/CMU-HCI-00-100)., Pittsburgh, PA: Carnegie Mellon University.

Jaspers, M. W. M. (2008). A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. *International Journal of Medical Informatics*, *78*(5), 340–353.

Jeffries, M., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User interface evaluation in real world: A comparison of four techniques. *Proceeding of ACM Computer–Human Interaction*, 119–124.

John, B. E. & Packer, H. (1995). Learning and using the Cognitive Walkthrough Method: A case study approach. Proceedings of CHI, 1995 (Denver, Colorado, May 7–11, 1995) ACM, New York. pp. 429–436.

John, E. B., & Marks, S. J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, *16*(4/5), 188–202.

Kaptelinin, V. (1996). Computer-mediated activity : Functional organs in social and developmental contexts. In B. Nardi (Ed.), *Context and consciousness: Activity theory and human–computer interaction* (pp. 45–68). Cambridge, MA: MIT Press.

Karat, C. M., Cambell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. *Proceedings of ACM CHI*, 397–404.

Kato, T. & Hori, M. (2005). Articulating the cognitive walkthrough based on an extended model of HCI. *Proceedings of HCI International*, Las Vegas, July, 2005.

Kato, T., & Hori, M. (2006). Beyond perceivability: Critical requirements for universal design of information. *8th Annual ACM Conference on Assistive Technologies*, 287–288.

Kieras, D. (1985). An approach to the formal analysis of user complexity. *International Journal of Human–Computer Studies*, *22*, 365–394.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 63–182.

Kitajima, M. (2006). Cognitive Walkthrough for the Web. In *International encyclopedia of ergonomics and human factors* (2nd ed., chap. 216, pp. 1044–1046). CRC Press.

Kitajjma, M., Blackmon, M. H., & Polson, P. (2000). A comprehension-based model of Web navigation and its application to Web usability analysis. *People and Computer XIV*, 357–373.

Kitajima, M., & Polson, P. G. (1997). A comprehension based model of comprehension. *Human–Computer Interaction*, *12*, 345–389.

Landauer, T. K., & Dumas, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lewis, C. (1986). A model of mental model construction. *Proceeding of ACM Computer–Human Interaction*, 306–313.

Lewis, C. (1988). Why and how to learn why: Analysis-based generalization of procedure. *Cognitive Science*, *12*, 211–225.

Lewis, C., Polson, P., Wharton, C., & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up and-use interface. *Proceeding of ACM Computer-Human Interaction*, 235–242.

Mack, R. L., & Nielsen, J. (1994). Executive summary. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 4–23). Amsterdam: Elsevier.

Mahatody, T., Kolski, C., & Sagar, M. (2009). CWE: Assistance environment for evaluation operating a set of variations of cognitive walkthrough ergonomic inspection method, In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics, HCII 2009, LNAI 5639* (pp. 52–61). Springer-Verlag.

Mahatody, T., Sagar, M., & Kolski, C. (2007a). Cognitive Walkthrough for HCI evaluation: Basic concepts, evolutions and variants, research issues. Proceedings EAM'07 European Annual Conference on Human-Decision Making and Manual Control June, 2007. Technical University of Danemark, Lyngby.

Mahatody, T., Sagar, M., & Kolski, C. (2007b). Cognitive Walkthrough pour l'évaluation des IHM: Synthèse des extensions et évolutions conceptuelles, méthodologiques et technologiques [Cognitive walkthrough for HCI evaluation: Synthesis of extensions and conceptual, methodological and technological extensions]. *Proceedings of IHM 2007, 19ème Conférence de l'Association Francophone d'Interaction Homme–Machine*, 143–150. International Conference Proceedings Series, ACM Press, Paris.

Monk, A. (1998). Cyclic interaction: a unitary approach to intention, action and the environment. *Cognition*, *68*(2), 95–110.

Nardi, B. A. (1996). *Context and consciousness: Activity theory and human–computer interaction*. Cambridge, MA: MIT Press.

Niculescu, A. (2008). Affordances in conversational interactions with multimodal QA systems. *HCI and Usability for Education and Work, Lecture Notes in Computer Science*, *5298*, 221–236.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. *Proceedings of ACM CHI*, 373–380.

Nielsen, J., & Mack, R. (1994). *Usability inspection methods*. New York: Wiley & Sons.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of CHI Conference on Human Factors in Computing System*, 249–256.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman & S. W. Draper (Eds.), *User centered systems design: New perspectives in human–computer interaction* (pp. 31–61). Hillsdale, NJ: Erlbaum.

Norman, D.A. (1988). *The psychology of everyday things*. New York: Basic Books.

Norman, D.A. (1999). Affordances conventions and design. *interactions*, *6*(3), 38–42.

Pinelle, D., & Gutwin, C. (2002). Groupware walkthrough: Adding context to groupware usability. *Proceedings of ACM CHI*, 455–462.

Polson, P. G., & Lewis, C. H. (1990). Theory-based design for easily learned interfaces. *Human-Computer Interaction*, *5*(5), 191–220.

Polson, P., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive Walkthrough: A method for theory-based evaluation of user interface. *International Journal of Man-Machine Studies*, *36*, 741–773.

Rieman, J., Davies, S., Hair, D. C., Esemplare, M., Polson, P., & Lewis, C. (1991). An automated walkthrough: Description and evaluation. *Proceeding of ACM Computer–Human Interaction*, 427–428.

Riihiaho, S. (2000). *Experiencies with usability evaluation methods*. Licentiate's thesis, Helsinki University of Technology, Laboratory of Information Processing Science, Helsinki, Finland.

Rizzo, A., Mandrigiani, E., & Andreadis, A. (1997). The AVANTI Project: Prototyping and evaluation with a Cognitive Walkthrough based on the Norman's Model of Action. *Proceeding of Designing Interactive Systems: Processes, Practices, Methods, & Techniques*, 305–309.

Rowley, D. E., & Rhoades, D. G. (1992). The Cognitive Jogthrough : A fast-paced user interface evaluation procedure. *Proceeding of ACM Computer–Human Interaction*, 395–398.

Ruttkay, Z., & Akker, R. (2008). Affordances and Cognitive Walkthrough for analyzing human–virtual human interaction. *Verbal and Nonverbal Features of Human–Human and Human–Machine Interaction, Lecture Notes in Computer Science*, *5042*, 90–106.

Ryu, H. (2008). Collective Web Usability analysis: Cognitive and activity walkthroughs. *International Journal of Web Engineering and Technology*, *4*(3), 286–312.

Ryu, H., & Monk, A. (2004). Analysing interaction problems with cyclic interaction theory: Low-level Interaction Walkthrough. *Psychology Journal*, *2*(3), 304–330.

Sears, A. (1997). Heuristic Walkthroughs: Finding the problems without the noise. *International Journal of Human–Computer Interaction*, *9*, 213–234.

Sears, A. (2003). Testing and evaluation. In J. A. Jacko & A. Sears (Eds.), *The human–computer interaction handbook* (pp. 1091–1092). Mahwah, NJ: Erlbaum.

Sears, A., & Hess, D. J. (1999). Cognitive Walkthroughs: Understanding the effect of task-description detail on evaluator performance. *International Journal of Human–Computer Interaction*, *11*(3), 185–200.

Spencer, R (2000). The Streamlined Cognitive Walkthrough method, working around social constraints encountered in a software development company. *Proceeding of ACM CHI*, 353–359.

Sweeney, M., Maguire, M., & Shackel, B. (1993). Evaluating user–computer interaction: A framework. *International Journal of Man–Machine Studies*, *38*, 689–711.

Van der Veer, G. C., Lenting, B. F., & Bergevoet, B. A. J (1996). GTA: Groupware Task Analysis—Modeling complexity. *Acta Psychologica*, *91*, 297–322.

Van der Veer, G. C., & van Velie, M. (2000). Task-based groupware design: Putting theory into practice. *Proceedings of DIS'2000*, 326–337.

Virzi, R. A. (1997). Usability inspection methods. In M. Helander, T. K. Landauer, & P. Prablhu (Eds.), *Handbook of human–computer interaction* (pp. 705–715). Amsterdam: Elsevier.

Wang, X. (2008). Design and evaluation of intelligent menu interface through Cognitive Walkthrough procedure and automated logging for management information system. *Computer Supported Cooperative Work in Design IV, Lecture Notes in Computer Science*, *5236*, 408–418.

Wharton, C., Bradford, J., Jeffrey, R., & Franzke, M. (1992). Applying Cognitive Walkthrough to more complex user interfaces: Experiences, issues, and recommendations. *Proceedings of Computer–Human Interaction*, 381–388.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The Cognitive Walkthrough method: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 105–140). New York: Wiley & Sons.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, *18*, 87–122.