International Journal of Human-Computer Interaction
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653655

Comparing Inspections and User Testing for the Evaluation of Virtual Environments

Cedric Bach[a]; Dominique L. Scapin[b]
[a] IRIT, Toulouse, France, and Metapages, Toulouse, France [b] INRIA, Le Chesnay, France

Online publication date: 27 July 2010

To cite this Article Bach, Cedric and Scapin, Dominique L.(2010) 'Comparing Inspections and User Testing for the Evaluation of Virtual Environments', International Journal of Human-Computer Interaction, 26: 8, 786 — 824
To link to this Article: DOI: 10.1080/10447318.2010.487195
URL: http://dx.doi.org/10.1080/10447318.2010.487195

INTL. JOURNAL OF HUMAN–COMPUTER INTERACTION, *26*(8), 786–824, 2010

ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447318.2010.487195

Taylor & Francis
Taylor & Francis Group

Comparing Inspections and User Testing for the Evaluation of Virtual Environments

Cedric Bach[1] and Dominique L. Scapin[2]
[1]IRIT, Toulouse, France, and Metapages, Toulouse, France
[2]INRIA, Le Chesnay, France

This article describes an experiment comparing three Usability Evaluation Methods: User Testing (UT), Document-based Inspection (DI), and Expert Inspection (EI) for evaluating Virtual Environments (VEs). Twenty-nine individuals (10 end-users and 19 junior usability experts) participated during 1 hr each in the evaluation of two VEs (a training VE and a 3D map). Quantitative results of the comparison show that the *effectiveness* of UT and DI is significantly better than the *effectiveness* of EI. For each method, results show their *problem coverage*: DI- and UT-based diagnoses lead to more problem diversity than EI. The *overlap* of identified problems amounts to 22% between UT and DI, 20% between DI and EI, and 12% between EI and UT for both virtual environments. The *identification impact* of the whole set of usability problems is 60% for DI, 57% for UT, and only 36% for EI for both virtual environments. Also *reliability* of UT and DI is significantly better than *reliability* of EI. In addition, a qualitative analysis identified 35 classes describing the *profile of usability problems* found with each method. It shows that UT seems particularly efficient for the diagnosis of problems that require a particular state of interaction to be detectable. On the other hand, DI supports the identification of problems directly observable, often related to learnability and basic usability. This study shows that DI could be viewed as a "4-wheel drive SUV evaluation type" (less powerful under certain conditions but able to go everywhere, with any driver), whereas UT could be viewed as a "Formula 1 car evaluation type" (more powerful but requiring adequate road and a very skilled driver). EI is found (considering all metrics) to be not efficient enough to evaluate usability of VEs.

Thanks to the research team ETIC from University of Metz as well to the Laboratoire de Psychologie de Lorraine (LabPsyLor) for their active participation in this study. In particular, thanks to Pr. Éric Brangier, Vincent Burgun and Hervé Leguil, and to all the participants in the different phases of this study. Special thanks to Phil Gray for his extensive review of the initial draft paper. Some of this work, involving IRIT and Metapages, is part of the CARE project – Cultural experience: Augmented Reality and Emotion—partly funded by the French Research Agency (www.careproject.fr).

Correspondence should be addressed to Cedric Bach, Institut de Recherche en Informatique de Toulouse, Univerity of Toulouse, 118 Route de Narbonne, 31062 Toulouse, France. E-mail: cedric.bach@irit.fr

1. INTRODUCTION

The primary motivation for this study was the design and validation of Usability Evaluation Methods (UEMs) for Virtual Environments (VEs). Although many usability methods are offered in the literature (see, e.g., International Standards Organisation [ISO], 2001; Law et al., 2009; Nielsen, 1993; Scapin & Law, 2007; Shneiderman, 1998) the goal of this article is to report on a series of experimental comparisons focusing on three of the most widely used categories of methods: Expert Inspection (EI), Document-based Inspection (DI), and User Testing (UT). More specifically, this study aims at measuring, on two different VEs, the effectiveness of DI, compared with UT and EI. The main points of investigation concern the following: the nature, classification, and distribution of usability problems diagnosed with each method; the overall performance; and problem similarity (for reliability) within-method, between-method, and between-application effectiveness.

UT is probably the most well-known usability method. Generally, during the tests, a person participates either in the execution of a set of tasks representative of those for which the software or new technology was designed or through a free exploration of the software or new technology. The objective of these tests is to identify difficulties in use from the users' spontaneous comments and from various performance measures such as task execution time, accuracy of the results, number, and types of errors. Because the collection and analysis of such large sets of information is complex and difficult, there are a number of tools and technologies intended to facilitate detection (e.g., eye gaze), recording (e.g., videos, log files), event classification (video tracks analyzers, etc.), and, of course, statistical data processing, not to mention the complex installations necessary for testing new technologies (e.g., simulation, virtual environments, Wizard of Oz, quite an ancient technique [Chapanis, 1982] but still used for very recent technologies such as Mixed Reality Environments; Dow et al., 2005).

EI is another popular method, especially in industry. It is an informal evaluation based on the knowledge and experience of usually one (or several) usability expert(s). The expert diagnoses the problems "theoretically," that is, according to his or her knowledge of the most frequently reported problems in the current state-of-the-art. This can lead to the quick identification of potential usability problems and can make it possible to eliminate some of the causes of the problems. These evaluations are relatively inexpensive because they are relatively fast and can be carried out quite early in the design process. However, the performance of experts depends on their experience and knowledge of the state-of-the art, which implies strong variations in terms of the amount and type of detected design errors. Another limitation is, of course, that the diagnosis can relate only to relatively well-known problems and thus excludes new situations in terms of technologies, contexts, or users. In addition, the "theoretical" diagnosis arising from such evaluations can sometimes conceal surprises that can only be discovered during real use.

DI can supplement the evaluator's judgment: "In the Document-based methods (also called Document-based analysis), the usability specialist uses existing

checklists or other documents in addition to his own judgment" (ISO, 2000). There are several variations in these methods and various sources of documents, for example, cognitive inspection; conformity to manufacturers' style guides or custom "in-house" guides; conformity to handbooks or lists of recommendations; conformity to ergonomic dimensions (e.g., Ergonomic Criteria: Scapin & Bastien, 1997; Heuristics: Nielsen, 1993); and conformity to proprietary, national, or international standards (which can, in turn, correspond to general principles or specific recommendations). An important source of variation of these documents is the way in which they were constructed and validated: top-down, bottom-up, theoretically, bibliographically, empirically, via consensus, via compilations, and so on, and with complete, partial, or nonexistent validations. The contribution of these documents to the effectiveness of the diagnosis is obviously related to the more or less rigorous manner in which these documents were produced. In all cases, an important limitation of such methods is that they can only contribute to the detection of already known problems, starting from established knowledge. A strong constraint of such methods is that the evaluator must have sufficient experience to be able to use the available documents in a way that is appropriate to the context of use and efficient for design and evaluation.

This article introduces first the application domain on which the study has been conducted: VEs. Following a state-of-the-art section on usability methods comparisons, the third section discusses the main methodological orientations of the study. Then, the fourth section describes the experimental phases as well as the participants profiles, the VEs used, the material, and measurements. The article then provides the results, both quantitative and qualitative, and concludes by discussing the limits of the study and identifying further research directions.

2. APPLICATION DOMAIN: USABILITY OF VE

VEs correspond to an application domain for which the issue of usability methods is both important and novel. VEs are becoming widely used and have expanded to cover an extensive range of activities. An example of this expansion is the availability of applications such as Google Earth (http://earth.google.com) or Geoportail 3D (http://www.geoportail.fr) that allow computer-based access to 3D satellite maps. With such types of applications, users are able to handle large amounts of data. Although these applications have been adapted for office computers, in many new contexts of use their keyboard/mouse/screen-based interactions are not sufficient from a usability point of view. For instance, in a museum context where social dimensions and diverse user profiles are important issues, traditional interactive systems are not the answer (e.g., Hughes, Smith, Stapleton, & Hughes, 2004). Advanced, enriched, even ubiquitous interactions using large display screens with remote interaction devices (e.g., laser pointers, oriented sound flows, gesture recognition) are more likely to be used (Dubois, Truillet, & Bach, 2007). This may result in the interconnection of the Internet and Virtual/Augmented Reality (Stanney & Davies, 2005). Similarly, classical office applications are beginning to integrate satisfactorily (Agarawala & Balakrishnan,

2006) several interaction features usually incorporated in VEs (e.g., 3D, behaviors, collision and gravity management). The further dissemination of such advanced technologies requires a user-centered approach (Williams & Harrison, 2001) and, particularly, updated usability evaluation techniques for such systems (Durlach & Mavor, 1995). These systems, particularly certain "disposable" ones, designed for single use during marketing events (Stapleton & Hughes, 2005), must sometimes support immediate, fast learning, whereas, on the other hand, a more progressive learning curve is required in the context of specific professional applications (e.g., numerical models, surgery, maintenance, military applications) or for video games.

Actually, several studies have highlighted specific usability problems associated with VEs (Gabbard & Hix, 1997), whereas field studies of Virtual Reality designers have demonstrated a strong need for human–computer intereaction knowledge and methods (Kaur, Maiden, & Sutcliffe, 1996). Others have shown that the designers of VE systems cannot rely solely on the methods developed for standard 2D graphical user interfaces (GUIs) because their interaction styles and the use of 3D are radically different from standard GUIs (Bowman & Hodges, 1997; Poupyrev & Ichikawa, 1999; Stanney, Mollaghasemi, Reeves, Breaux, & Graeber, 2003). Thus, a great effort is needed to make available to designers and customers a set of usability evaluation methods for evaluation and design that are adapted to the diversity and complexity of VEs (Bowman, Gabbard, & Hix, 2002).

A few surveys have been published on existing knowledge and research results about the usability of Human Virtual Environment Interaction. These surveys concern different topics such as cognition (Wickens & Baker, 1995), usability characteristics (Gabbard & Hix, 1997), interaction techniques (Bowman, Kruijff, LaViola, & Poupyrev, 2005; Hand, 1997), usability-centered design (Kaur, 1998; Sutcliffe, 2003), Human Factors issues (Stanney, Mourant, & Kennedy, 1998), usability for collaborative VE (Tromp, 2001), state-of-the-art on VE (Stanney, 2002), UEMs for VE (Bowman et al., 2002), and guidelines compilation classified by Ergonomic Criteria (Bach, 2004). Most other studies are related to the development of methods and models dedicated to this type of interaction such as cybersickness questionnaire (Kennedy, Lane, Berbaum, & Lilienthal, 1993), framework for requirement analysis (Conkar, Noyes, & Kimble, 1999), model and notation of interaction (Dubois, Nedel, Dal Sasso Freitas, & Jacon, 2005; Kaur, Maiden, & Sutcliffe, 1999), and tools for usability engineering (Karempelas, Grammenos, Mourouzis, & Stephanidis, 2003; Stanney et al., 2003). Also, some studies are concerned with the adaptation of existing UEMs such as cognitive walkthrough (Sutcliffe & Kaur, 2000), usability questionnaire (Kalawski, 1999), User Testing (Tromp, Steed, & Wilson, 2003), Ergonomic Criteria (Bach & Scapin, 2003), and heuristic evaluation (Sutcliffe & Gault, 2004).

Conducting User Testing to evaluate VEs seems to be more difficult than testing GUIs or Web sites. Actually, Bowman et al. (2002) revealed a set of difficulties involved in designing User Testing dedicated to evaluating VEs. The authors assigned those difficulties to different categories: *physical environment issues*, *evaluator issues,* and *user issues*. This suggests that designing an efficient user testing to evaluate a VE, using complex interactive systems, is a challenge.

3. RELATED WORK

When done thoroughly, comparing UEMs is a very complex and lengthy process especially when researchers follow guidelines from Gray and Salzman (1998) or criteria from Hartson, Andre, and Willigies (2001). This is probably why there are not many studies published in the open literature, particularly comparing Inspection and User Testing: There are about 30 such papers, with several perspectives. Roughly speaking, about one third discuss the issues; another third perform method comparisons of several variations of the same method (different expertise, different application domain, etc.), and the last third actually perform comparisons of different usability methods.

A number of issues have been mentioned in the literature regarding the comparison of UEMs. The issues most often reported concern (see Cockton, Woolrych, Hall, & Hidemarch, 2003; Gray & Salzman, 1998; Hartson et al., 2001; Hornbæk & Frøkjær, 2005): using the appropriate metrics, the issue of problem severity, whether putative usability problems extracted in analysis are genuine usability problems, thoroughness, problem similarity, and usability problem interpretation.

Hornbæk (2009) provided recently an overview of UEMs assessments, which specified and extended the review by Gray and Salzman (1998), particularly on the issue of comparing UEMs. Hornbæk identified four main activities in studies assessing UEMs: evaluation, documenting, matching, and analysis. The first activity consists in applying a number of UEMs or, in different conditions, a single UEM to evaluate (usually) one single interactive application. Then, in the documenting activity, evaluators create a set of descriptions of usability problems, sometimes using a structured format, sometimes using a free-text format. In the third activity, a matching of usability problems occurs to identify duplicate problems and to compare sets of usability problems. Then, in the last activity, problems are counted both as problem instances (tokens) and, in relation to a classification, as problems classes concerning different areas of the interface (types or profiles).

According to Hornbæk (2009) these main activities do not support important dimensions to consider in UEMs assessment, particularly for formative evaluations. This is in agreement with Lindgaard's (2006) critique about the unnatural conditions, unrealistic interfaces and finally poor concern for "downstream utility."

The Hornbæk set of activities for UEMs assessments may include a few limits, for example, focus on problem count, little concern about procedures for matching problems, limited assumptions about the role of method prescriptions in evaluations, a main focus on how problems are taken up in design, a belief that a single best UEM exists, and the assumption that usability problems are all real. However, it is an interesting framework that we use in next section for positioning our own UEMs comparisons.

4. UEM COMPARISONS APPROACH

Along the four main activities just described, this section discusses the main methodological issues addressed in this study.

4.1. The Evaluation Step

The *evaluation* activity in UEMs comparisons involves mainly applying different UEMs, or a single UEM under different conditions. Usually individual evaluators (including end users) apply one UEM at a time to evaluate one interactive application (e.g., GUI, Web site). This step produces a set of usability problems, using different data-gathering techniques (e.g., thinking aloud, notes, video recording). It usually tries to answer to the following questions: (a) What is the difference between UEM_1 and UEM_2 and UEM_n? (b) What is the difference between their different conditions of use?

Current literature rarely mentions the *stability* of methods, which can be assessed through the evaluation of more than one application. Doing so on different applications that contain different usability problems allows to check whether the use of the methods differ from one application to the other. If it does not, it can be a fair indication of diagnosis *stability* across application types. To address this issue, the study reported here assessed the *stability* of methods through the evaluation of two quite different VEs (e.g., learning context, tourism context).

Current literature also rarely mentions that the *effectiveness* of an UEM Inspection can be compared with a control group. Actually, having a control group is not frequent in comparative studies, as mentioned by Chattratichart and Lindgaard (2008): Often only one method (or method variation) is compared to another. In the study reported here, the *effectiveness* of the DI was assessed by comparison with EI. As a matter of fact, EI, sometimes called Free Inspection, is used as a control group (Chattratichart & Lindgaard, 2008). This contributes to evaluating the *effectiveness* of the *Document*—in our case a document describing Ergonomic Criteria for VE (Bach & Scapin, 2005), but it also helps to highlight the usability problem coverage and variations depending on whether a *Document* was used for guidance.

Another issue concerns the realism of problems. Indeed, whatever the method used, the problems diagnosed can rarely be fully quoted as being real (except maybe through longitudinal testing studies, in context, with fully developed software). However, some care can be exerted in that direction. The realism of problems can be assumed by their source: Problems extracted from user testing are real in the sense that they arise from actual observable problems users have when using a system (e.g., comments, errors, undos, etc.), in "natural" conditions, that is, the ones that usability practitioners are confronted with while performing usability evaluations, and preferably with several experimenters analyzing data. Problems extracted from inspection are real in the sense that they are directly formulated by human factors specialists with minimal training. In our experiment, real user difficulties observed are used as a baseline, complemented with expert assessment; besides, having several methods being assessed allows problem comparisons, which also contributes to cross-checking the naturalness of problems.

One question remains: Where are the usability problems coming from? They can be explicitly generated, that is, programmed into the software for the sake of the experiment (see. e.g., Pollier, 1991; Scapin & Bastien, 1997) or just exist in freely available software applications. In our experiment, the choice has been to

carefully select applications rather than to design new ones, which might have been less realistic (and a very large investment in terms of VE development). One downside is though that thoroughness cannot be fully established, as usability problems have not been inserted explicitly.

4.2. The Documenting Step

This step corresponds to the individual description, by evaluators, of usability problems, using a structured format, or free-text. It sometimes also includes some levels of severity. Such a description differs depending on the UEM. For inspection (either Expert or Documented), it is somewhat less complex and much less time-consuming than for User Testing. It consists in data collection, organization, and homogenization of the problems diagnosed and documented directly by the participants. There, the issue is more to make sure the characterization of problems is explicit enough, which is addressed in the analysis step (see section 4.4).

For user testing, participants' interactions and comments (spontaneous thinking aloud) are recorded to facilitate direct and postexperiment interpretation. During the interpretation of the evaluation results, the problems are analyzed by experimenters as they were expressed in the context of their first appearance, by replaying the application and checking the participants' comments from recorded videos. There, the issue is to distinguish between *tokens* and false alarms. This was achieved through consensus, via multiexpert analysis.

Last, the issue of usability problem severity: Even though it is an important issue, it was not the goal of this study. Usability problems severity was not assessed in the course of the experiment also for the sake of the experiment duration; however, with the gathered material and data, it will be possible in the future to look into problem severity assessment, asking users and/or usability specialists to position usability problems on a severity scale.

4.3. The Matching Step

This step is usually conducted to compare sets of usability problems and to identify duplicate problems. For this, more or less structured formats can be used. In our experiment, both Ergonomic Criteria (Bach & Scapin, 2005) and the recommendations by Cockton and Lavery (1999) were used. While matching problems, special care was given to checking the equivalence in profile between inspection-based problems and user testing-based problems.

For each identified usability problem, an ergonomic criterion was assigned in order to build an organized map showing the distribution of the usability problems. This allowed an assessment of the diversity of the problems identified. This can be considered as a metric representing the *coverage* of the method considered and allowing to compare them.

Regarding the means used to extract usability problems, the procedure followed the recommendations by Cockton and Lavery (1999) for matching usability problems, and for judging whether two problems are similar. Problems have been

considered similar when the problem identification context, the interaction object concerned, and/or the interaction consequences (observable or inferable state changes) are similar.

4.4. The Analysis Step

Besides good experimental design and good statistics, which we do not discuss here (for details, see Chattratichart & Lindgaard, 2008; Hartson et al., 2001), the main item used to compare methods is the number of problems diagnosed, one way or another. Some may argue that such a plain problem count is not an appropriate way to assess the predictive power of a method. It is indeed an issue when considering the importance and priority of problems. However, such a measure, if carefully derived, makes sense in order to actually compare methods with respect to their capacity to generate usability diagnoses.

A plain count is indeed not enough; what is needed is careful identification and classification of problems using a common framework and a coherent format. Particular attention is required to assess the commonality versus specificity of usability problems (see the previous section). This will be achieved by iterative pairing of participants' results to differentiate problems diagnosed by over two subjects versus specific problems (i.e., reliability). The literature proposes different metrics to measure the reliability sometimes as a kappa (Hartson et al., 2001) or as the mean number of evaluators finding a problem (Chattratichart & Lindgaard, 2008). In the current study we used the mean number of evaluators finding a problem to measure reliability. Concerning the threshold for minimal reliability of a method, one can consider that a level of reliability of two (two evaluations discovering the same problem) is the minimal value to reveal a convergence of diagnoses and, of course, an effect of the method used to identify usability problems. Similarly, method-specific problems and method-shared problems (i.e., overlap) have to be distinguished.

In the case of UEM comparison of two applications, two types of analysis can be assess the various evaluation performances: a comparison of the average evaluation performance per evaluator/participant according to the applications and to the methods used. Another analysis concerned the *coverage* of methods by use of Ergonomic Criteria. A comparison of the distribution of the problems on the Ergonomic Criteria was also carried out.

The calculation of the proportions of similar versus specific problems helps to highlight the influence of each method on the overall problem identification. Such influence is further computed by adding the overlapping and the nonoverlapping proportions involving each method. Such an analysis is interesting as it shows the proportion of problems due to a particular method compared to the whole set of problems identified. An analysis is also carried out on the reliability of problem identification per method. A problem is considered to be identified, using a particular method, if at least two similar problem evaluations occur in the same experimental group.

Finally, as mentioned by Connell and Hammond (1999) an important issue is usability problem interpretation: tokens versus classes. During the gathering of

problems, all problem occurrences/tokens (individual problems. in a particular application context) are classified under classes of problems, that is, problems with a similar profile but applied to different contexts, different objects, widgets, and so on. While counting problems, special care was taken to clearly distinguish between problem classes and problem tokens.

4.5. Main Orientations for the Experiment

In this experiment, we followed the four main steps of UEMs comparison described in the previous paragraphs. In short, this study measures the effectiveness of the Ergonomic Criteria to document an inspection, compared with UT and an EI. This comparison, carried out on two different VEs, attempts to evaluate the following six main metrics:

- The distribution of usability problems in terms of Ergonomic Criteria classification (to evaluate coverage)
- The count of problems identified by each method (to measure evaluation performance)
- The within-method problem similarity (to evaluate reliability)
- The between-method problem similarity (to evaluate overlap)
- The between-application effectiveness of the methods (to evaluate stability)
- The profile of problems found with the various methods (to distinguish scope)

5. THE EXPERIMENT

5.1. Phases of the Experiment

Three UEMs (UT, DI, EI) were separately used to evaluate two VEs. Two VEs evaluated: an educational software (a 3D video game tutorial) and a 3D map of the Chamonix Valley (French Alps). Six experimental conditions conducted in three different experimental sessions (one session per UEM used to evaluate each of the two VEs). Each experimental session was 1 hr long (30 min to evaluate each VE). Each experimental condition produced a set of usability problems. Ten participants took part individually in the UT, and 19 junior experts took part in Inspections (10 in DI and 9 in EI). With this experimental design, 29 hr of usability evaluation activity performed in a laboratory context were analyzed. Figure 1 shows a synoptic view of the experimental phases for the three experimental sessions. The following sections detail the experimental sessions, tasks, participant profiles, VEs evaluated, material, data collection, and finally the six main metrics used for the processing of data.

UT. All participants in UT took part individually in only one experimental session. That experimental session consisted of two conditions allowing the

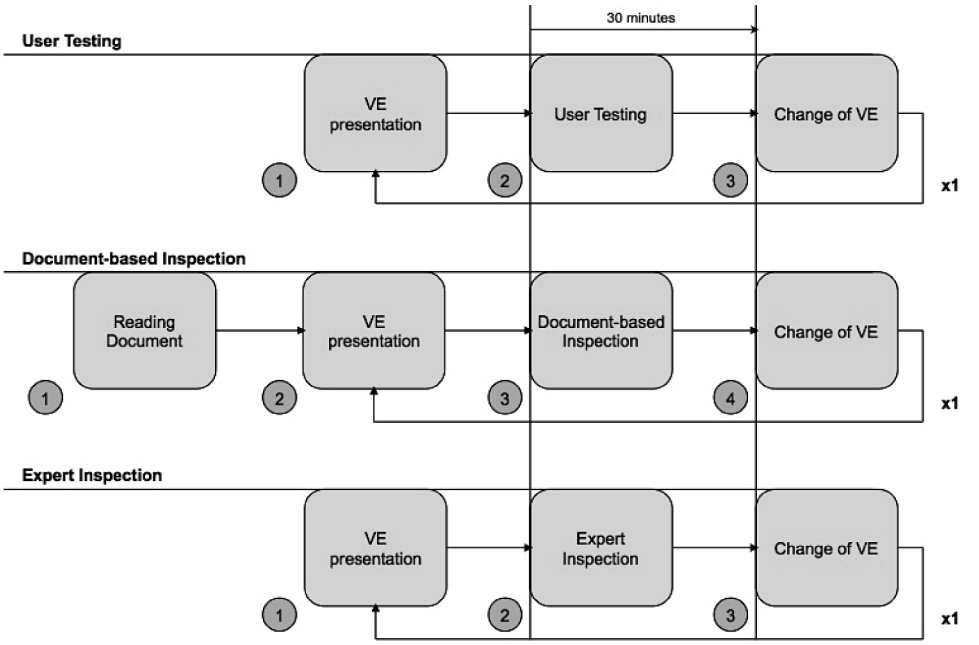

FIGURE 1 Synoptic view of experimental phases.
Note. VE = Virtual Environments.

evaluation of two VEs, in a counterbalanced order, using a rather broad spectrum of tasks for VEs functioning on a traditional computer. The instructions given to the participants were

- for Educational Software: "You have to complete the set of tasks asked by the virtual instructors." (Table 1)
- for 3D map: "You have to complete all the tasks described in this paper document." (Table 2)

The implemented task scenarios for the 3D educational software correspond to the four categories illustrated in Table 1. The completion of the whole set of tasks is possible in 30 min. Actually, an expert in using the system is able to perform the tasks within 15 min. Table 1 shows that the large majority of the tasks suggested by the trainer consist of pressing keys to trigger commands in a given context. There are in fact only four complex tasks. The difficulty of these tasks is not affected by the choice of difficulty level suggested by the system. Each participant interacts with the system default values but is free to tailor them.

For the 3D map task, scenarios were presented to participants in a paper document (unlike the educational software, which had a virtual instructor). In the experiments, the participants were asked to carry out a scenario of 10 tasks. That scenario was aimed at covering several aspects and was divided into five categories representing 20 subtasks. Table 2 presents the various subtask categories and their number of appearances within the scenario.

Table 1: Categories of Tasks Implemented in the Educational 3D Video Game

Task Categories	*N*	*Description*	*Examples*
Access to the simulator	3	The user must use the mouse to reach mission 1.	The participant must click on one of the three icons which will give access to mission 1.
To operate a control	20	The system indicates to the user, using an instruction posted on the screen, which key of the keyboard must be pressed.	The system displays the message: press on the countermeasures key "H"
To operate an unknown control	3	The system requires the user to activate a command, but it does not indicate which key to press	The system requires the participant to activate the hyperpropulsion to be able to continue the course of the scenario. It concerns key "J," but the user must discover it.
Complex tasks	4	The system orally requires the user to carry out a complex task requiring the satisfaction of several prerequisites.	During the 17th task, the system requires the user to destroy a large vessel by destroying 5 specific subtargets from a set of 15 subtargets.

Table 2: Various Categories of Tasks Associated With the User Testing Using Chart 3D

Task Categories	*N*	*Description*	*Examples*
Configuring the interface, use of tools	5	Aims to simulate the use of certain tools, to check if the interface parameter setting is efficient.	Identifying information from a panel can be done using a screen capture tool. Or, set field of vision to 90°.
Controlling automatic visits	2	Aims to check if the participant manages to control the automatic visits.	In task 1 of the scenario the user is asked to stop the automatic visit to consult information on a panel.
Finding an object, a specific piece of information	5	Aims to evaluate system effectiveness in highlighting information or objects useful to a task.	Several tasks of the scenario require the user to find panels of information associated with various ski resorts
Seeking a particular place	5	Aims to evaluate system assistance in a navigation task.	The participant is asked to return to the top of Mount Blanc or the needle of Argentière.
Controlling movement	3	Aims to evaluate the efficiency of the means of movement available in the VE.	The participant is requested to turn him(her)self around or to move around an object.

Note. VE = Virtual Environments.

The completion of the whole set of tasks is possible in 30 min. Actually, a system expert can carry them out in about 15 min. Table 2 shows that the distributions of the various categories of task are nearly equivalent, even though navigation tasks (search for places and/or objects) have been highlighted, particularly at the beginning of the scenarios. That aspect is one of the difficulties most often associated with VE interaction.

DI. In this experimental condition, no particular inspection strategy was suggested. Each one of the 10 participants carried out the entire protocol individually. The DI using the Ergonomic Criteria consisted of four phases.

During the first phase, the experimenter asked the participants to read in its entirety the document presenting the Ergonomic Criteria, with their definitions, justifications, examples of recommendations, and applications. There was no time constraint for reading the document. Previous work (Fischoff, 1982) indicated that a reading phase, which can be considered here as making the reader aware of a choice model (Zachary, 1986), can be an approach for producing a debiasing effect that may improve the diagnosis activity.

In the second phase, the participants were asked to read the experimental instructions indicating the steps to be followed:

> The experimenter will introduce you a 3D application to evaluate. You have to detect in your own way usability problems in this application. You can refer anytime to the document describing Ergonomic Criteria you read before. When you find a usability defect, you have to describe it aloud and write it down. After 30 minutes you will evaluate another 3D application.

Then the experimenter provided a general presentation of the first application. After making sure the instructions were well understood, the experimenter returned to the technical zone of the usability laboratory to monitor and record the participants' activity.

During the third phase, which consisted in the DI of the first application, the participants had to briefly (during 30 min) describe, vocally, and/or in writing, the usability problems they identified.

During the fourth phase, the experimenter presented the second application to be inspected, the procedure being the same as in Phase 3. At the end of that second 30-min phase, the inspection was stopped and the experiment ended.

EI. Such an inspection can be carried out at various levels of expertise. In this experiment, the available expertise was *junior* expertise. In these experimental inspections, as with the others, no particular inspection strategy was suggested. All nine participants carried out the entire protocol individually, through the same three phases as the three last phases of the DI (i.e., instructions, evaluation of the first VE then evaluation of the second VE).

5.2. Participants

There were two different categories of participants: end-users who took part in UT and advanced students in ergonomics who took part in the two types of inspections.

The group of participants in UT consisted of five men and five women, 19 to 24 years old, the average being 21.8 years ($SD = 1.5$). All participants' sight and hearing abilities were normal or corrected-to-normal. Six of the university

students were 2nd-year psychology students, one was a 3rd-year English student, one a master's student in ecology, and one a master's student in communication. All participants used a traditional computer (i.e., GUI, screen, keyboard, mouse) regularly at the university.

Initially, participants sought for this study were those familiar with the use of traditional computer equipment but not with 3D applications. This profile was required for two reasons: (a) The evaluated applications, although being in 3D, operate on a traditional computer, and (b) the aim was to limit the role of the "gender" factor by recruiting an equal number of women and men. Men are known to be more likely to be 3D video games users than women (Adamo-Villani, Wibur & Wasburn, 2008; Green & Bavelier, 2006); therefore, we thought it would have been easier to recruit both men and women who are not frequent 3D video game players. But, despite a strict recruitment process conducted with more than 200 people, the men recruited were slightly more familiar with this type of computerized environment. On the other hand, it would have been harder to look for an equal number of women familiar with 3D applications. Besides, recruiting people who are very familiar with 3D applications would have considerably increased the chances of recruiting people having already played with the educational software video game evaluated in this study.

The group of 19 participants in the two types of inspections (DI and EI) was all 5th-year students in work psychology; all trained in software ergonomics. The training was mainly theoretical; they did not have practical experience in usability inspection and they had not attended a course on the Ergonomic Criteria for GUIs. Their knowledge of VEs was almost nonexistent; none of them had experience with the VE applications evaluated in the experiment. The participants were randomly affected to the two inspections conditions: 10 students for the DI (5 men, 5 women; M age $= 24.5$ years, $SD = 2.5$) and 9 students for the EI (3 men, 6 women; M age $= 26$ years, $SD = 7$).

5.3. The Two Virtual Environments

The two VEs selected for the experiment do not use sophisticated VE platforms to simplify the *physical issues* mentioned by Bowman et al. (2002): "In VEs, nontraditional input and output devices are used, which can preclude the use of some types of evaluation. Users may be standing rather than sitting, and they may be moving about a large space, using whole-body movements" (p. 405). The training VE was a desktop educational video game (Microsoft Game Studios, 2000) that reproduces the cockpit of a spaceship. It is an entirely simulated environment. The interaction with the system was performed using the keyboard (the mouse is not usable in the game itself but only in the system requirements area). It is possible to choose the level of complexity of the game (three choices: "easy," "medium," "difficult"). In the game scenario, the VE was a flight simulator in which a participant had to learn how to use a spaceship in order to use it effectively in game missions. The mission evaluated in the experiment was "Mission 1," supposedly the simplest. In this training mission, a virtual and automatic trainer explains how to use the basic commands of the spaceship and asks the players to carry out a

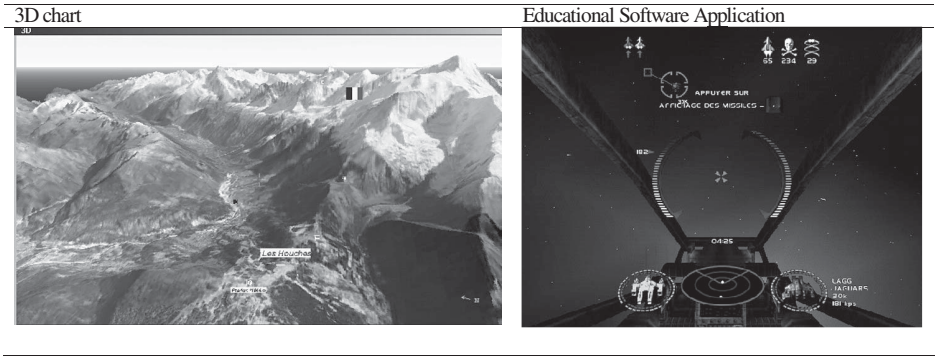

FIGURE 2 Screen pages of the two desktop Virtual Environments evaluated.

certain number of exercises. Guidance is supported by two sensory channels as outputs: the audio channel (the virtual trainer explains orally) and the visual channel (the system posts messages on the screen, reveals targets, etc.). The point of view of the user, by default, is egocentric from the cockpit of the spaceship (see Figure 2), the hands of its avatar being visualized. The user can modify this point of view. The educational software follows a constrained scenario, which requires carrying out the tasks progressively in order to move from one task to the next. The scenario provides 35 tasks at various levels of difficulty. The system can simply require the participant to press a key or to carry out a complex task requiring planning, subobjectives to reach and movements.

The tourism VE uses Terra Explorer 4.0.0 (Skyline Software Systems, Chantilly, VA), which requires a high bandwidth Internet connection. Interaction with the system is mainly performed via a mouse. This VE uses only visual output; it does not include sound. The users' point of view is egocentric (Figure 2); there is no representation of an avatar. Movements are carried out according to several interaction metaphors:

- As with a helicopter, the user must control his or her movements either directly on the 3D map or with a virtual flight interface.
- When moving a map, the user's point of view remains fixed; he or she can move the 3D map using three modes. A panoramic mode (the map turns around the virtual position of the participant), a rotation mode (the map turn around its own center), and a slip mode (the map moves along a plan).
- By jumps: The user jumps from a starting point to a point of arrival, seeing the movement which is entirely automated.
- By teleportations: The user goes from the starting point to the point of arrival without seeing the movement, which is instantaneous.

This application allows a user to visit a 3D map of the Chamonix Valley generated from geographical data (aerial pictures and/or satellites). It allows the user to collect tourist information about the valley through information panels or links to Web sites.

5.4. Material

In this section we present the experimental apparatus used in this study: computer equipment, the experimental document describing the Ergonomic Criteria, and the laboratory in which the experiments were carried out.

Computer equipment. The two VEs operate on same computer equipment. It is a rather traditional hardware configuration corresponding to the majority of current computers with screen, mouse, keyboard, and speakers.

Document. The document used in this study is an adaptation of the Ergonomic Criteria to VEs that followed the three phases proposed by Scapin (1990) to organize human factors knowledge: the collection and organization of experimental results, followed by an experimental validation of the dimensions obtained (see Bastien & Scapin, 1992), and concluding with an experimental evaluation of the utility of Ergonomic Criteria-based inspection methods (for GUIs, see Bastien & Scapin, 1995; and for the Web, see Bastien, Scapin, & Leulier, 1999; Leulier, Bastien, & Scapin, 1998).

Similarly, the first phase of the adaptation of Ergonomic Criteria to VEs consisted of a thorough analysis of the literature, leading to the compilation of 170 ergonomic recommendations dedicated to VEs, classified by a series of 20 Ergonomic Criteria and by a corresponding set of 73 interactive elements to which these recommendations applied. This initial work led to the creation of two new Ergonomic Criteria: *Grouping/Distinguishing by the Behaviour* and *Physical Workload*, and to the modification of the criterion *Significance of codes and denominations* into *Significance of codes, denominations, and behaviors*.

Based on this work, a document was compiled that presented the definitions, justifications and examples of applications of the Ergonomic Criteria adapted to VEs. This document was used in an initial experimental validation. The second phase of the adaptation of the Ergonomic Criteria to VEs consisted of an intrinsic validation (Bach & Scapin, 2003) which led to an edited version of the Ergonomic Criteria adapted to VEs (Bach & Scapin, 2005). The modifications, which attempted to limit the very few wrong assignments and overgeneralizations observed, were addition of new examples, addition of comments allowing a better distinction between criteria, and the refinement of some definitions to ensure good independence and distinctness.

The Document itself consists of two parts. The first part is the list of the Ergonomic Criteria (see Figure 3). This list is made up of three levels of criteria. The first level consists of eight main criteria. Five of these are subdivided in subcriteria (second-level criteria) from which some are subdivided into sub-subcriteria (third-level criteria). The term *elementary criterion* refers to the criterion or subcriterion that is not further divided. Overall, the list contains 20 elementary criteria (bold in Figure 3).

The second part of the document presents each criterion individually, one criterion per page, including its definition, its justification, and examples of ergonomic recommendations. On average, the whole document is read in 25 min ($SD = 5$) by

1. **Compatibility**
2. Guidance
   - 2.1. **Legibility**
   - 2.2. **Prompting**
   - 2.3. Grouping / distinguishing items
     - 2.3.1. **Grouping/ distinguishing by Location**
     - 2.3.2. **Grouping/ distinguishing by Format**
     - 2.3.3. **Grouping/ distinguish. by Behavior**
   - 2.4. **Immediate Feed-Back**
3. Explicit Control
   - 3.1. **Explicit User Actions**
   - 3.2. **User Control**
4. **Significance of codes and behavior**
5. Workload
   - 5.1. **Physical Workload**
   - 5.2. Brevity
     - 5.2.1. **Minimal Actions**
     - 5.2.2. **Conciseness**
   - 5.3. **Information Density**
6. Adaptability
   - 6.1. **Users' Experience**
   - 6.2. **Flexibility**
7. **Consistency**
8. Error Management
   - 8.1 **Error Protection**
   - 8.2 **Quality of error messages**
   - 8.3 **Error Correction**

FIGURE 3 List of ergonomic criteria for Virtual Environments.

the experts having taken part in the intrinsic validation of the Ergonomic Criteria (Bach & Scapin, 2003).

The laboratory. The experiment was conducted at the Pergolab platform of Metz University (France). This usability laboratory is equipped with a video unit, sound recording devices, and one-way mirrors. In this study, the three video/sound recordings (behavior of the participant) were synchronized on the same channel, with the video output of the computer, the keyboard, and the mouse.

5.5. Data Collection and Analysis

To assess methods, both quantitative and qualitative analyses were required. As previously described in the sections titled Related Work and UEMs Comparison

Approach, quantitative analyses show the distribution of problems in term of Ergonomic Criteria classification, the evaluation performance, the reliability, the overlap, and the stability of each method. Qualitative analysis shows which profile of problems are common versus specific to each method.

Time was recorded continually for all experimental conditions. Time has been mainly used for calibrating the session length and for task duration in User Testing.

The next sections describe the various metrics used to evaluate the methods individually and to compare them. A global qualitative analysis is then presented with the goal of illustrating the differences in problem profiles, depending on the evaluation methods used in the experiment.

Metrics used to evaluate different aspects of each evaluation method. Three different main metrics was used to evaluate different aspects of each method. As mentioned in the previous section, we are interested to evaluate more dimensions of UEMs than plain count:

- The evaluation of *coverage* by classification of usability problems using Ergonomic Criteria. The assignment of each problem to Ergonomic Criteria was performed by experimenters also for the DI group because the participants did not have time to perform it themselves. This is understandable when comparing the time available to discover and carry out the inspection of an application (30 min) to the time that was necessary (48 min, $SD = 9$) for usability experts to classify 40 problems for a single assignment task (Bach & Scapin, 2003). Obviously 30 min was not sufficient to both identify and characterize the problems with the Ergonomic Criteria.
- The *evaluation performance* was calculated from the average number of problems highlighted in the evaluations as well as the group performance; moreover, *t* tests were conducted to check the effect of *gender* and *application* factors.
- The *reliability* was evaluated as the mean number of evaluators finding a problem (Chattratichart & Lindgaard, 2008).

Metrics used to compare evaluation methods. As indicated in the UEM Comparisons Approach section, different metrics are used to compare methods such as reliability (by use of Kruskal-Wallis analysis of variance [ANOVA] and *U* tests) and evaluation performance (by use of ANOVA and *t* tests). This section describes more precisely first the comparisons in terms of *coverage*, then *the qualitative analyses.*

First, the analysis concerned the *coverage* of methods by use of Ergonomic Criteria. A comparison of the distribution of the problems on the Ergonomic Criteria was carried out. Through the use of 3D histograms presenting in the X-coordinate (X) the Ergonomic Criteria, in ordinate (Y) the number of the identified problems, and in depth (Z) the type of method used. This resulted in two graphs for each application. This type of graph allows the visualization of the variations according to the methods used.

A *qualitative analysis* concerned the assignment of problem *tokens* (i.e., individual occurrences of usability defects) identified with each UEM. This was based on the Ergonomic Criteria assignment performed in a previous stage. Two classifications were conducted.

The first classification attempts to show problem specificity and *overlap* between UEMs. Seven classes of problem have been used: problems common to all methods, problems common to UT and DI, problems common to DI and EI, problems common to UT and EI, problems specific to UT, problems specific to DI, problems specific to EI. These classes are named "Overlap Classes," even though they include classes that are specific but from the point of view of overlapping.

The second classification, attempts to cover the main profiles of *tokens*. These classes are named "Problem Profile Classes." This qualitative classification followed a strategy of aggregating Problem Profile Classes initially classified by Overlap Classes. The focus has been on a characterization of the problem profile by looking at the problem identification context, the interaction object concerned, and/or the interaction consequences (observable or inferable state changes).

6. RESULTS

This section describes the results of the experiment based on the various analyses previously described. The results first concern each individual method, then the comparison of methods, and finally the qualitative analysis.

6.1. Problems Diagnosed With Each Method

Diversity of problems identified with UT. Figures 4 and 5 show the distribution (assigned to the Ergonomic Criteria) of the problems identified with all methods. The focus here is only on user testing. There is no strong interapplication difference in terms of problem distribution. Two Ergonomic Criteria were not used to characterize the problems: *Informational Density* and *Physical Workload*. The latter is explained by the fact that physical problems were avoided in the choice of applications to preserve the well-being of the participants.

The capability of problem assignment to Ergonomic Criteria seems satisfactory, as all problems were allocated to the Ergonomic Criteria. For this reason, the *coverage* of Ergonomic Criteria seems satisfactory and sufficient to assign the actual problems identified during UT. In terms of application diversity, for 3D educational software (Figure 4) problems are distributed along 17 *elementary criteria*, while for the 3D map (Figure 5), the problems are distributed along 15 *elementary criteria*. Overall, UT led to a wide diversity of problems.

The evaluation performance with UT. The evaluation performance is on average 29.40 ($SD = 9.59$) problems identified by participant for the 3D educational software and 25.60 ($SD = 4.70$) for the 3D chart.

A student's t test was calculated to determine the effect of first application type (20 measures) and secondly gender of the participants (20 measures) on

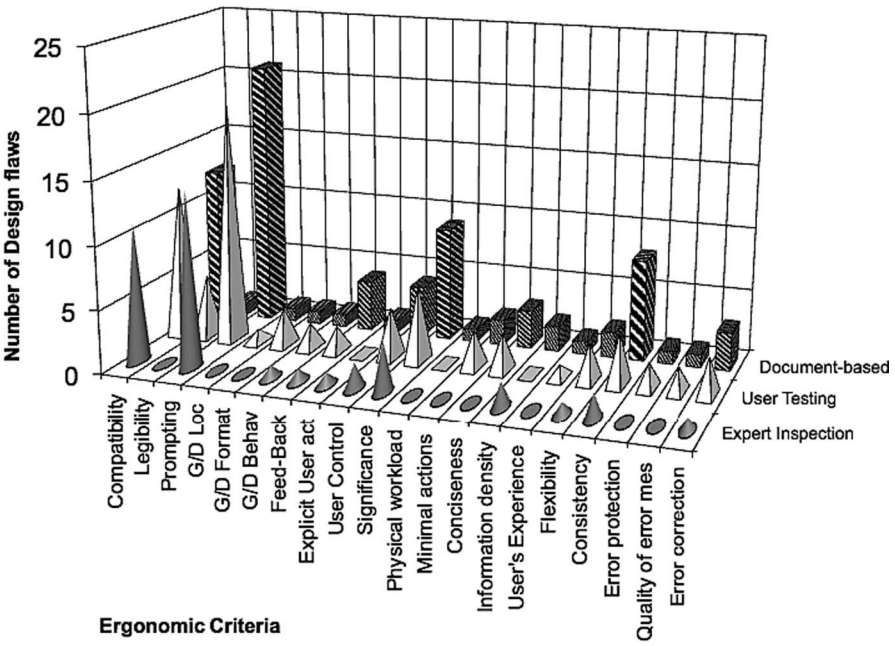

FIGURE 4 Comparison of design flaws per method in the 3D educational software.

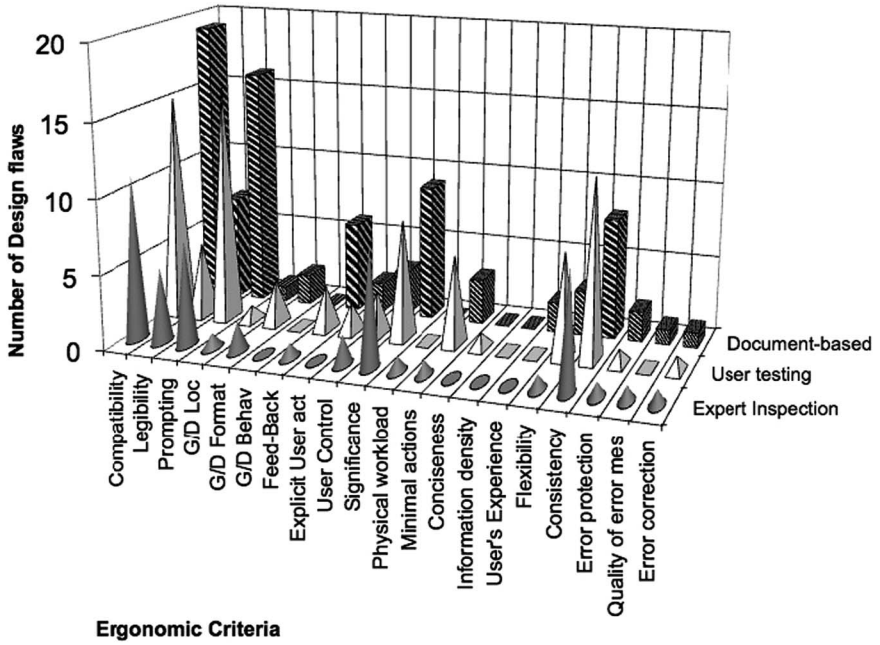

FIGURE 5 Comparison of designs flaws per method in the 3D map.

evaluation performance. There is no significant difference between the two applications, $t(18) = 1.01$, $p = .327$, two-tailed; on the other hand there is a significant effect of the participants' gender, $t(18) = 2.814$, $p = .011$, two-tailed. This difference may be explained by the fact that women are generally less exposed to interactive 3D applications, particularly 3D video game or other factors as mentioned in Adamo-Villani et al. (2008).

Reliability of the results with UT. Figure 6 shows the percentage of design flaws identified in each VE. Figure 6 shows also the number of user testing sessions identifying similar problems. In other words, Figure 6 shows the distribution of percentage of design flaws by a reliability degree for each VE evaluated. A problem is considered to be specific when only one UT identified it. A problem is considered to be similar as soon as it is identified in at least two cases. There is a large difference between the applications in terms of specific problems: 42.8% of the problems on the 3D map are specific, against 25% for the 3D educational software. As mentioned before, this variation can be explained by the constraint of the 3D educational software scenario on the exploration of the application. The 3D map leaves much more freedom for exploring new contexts of use, particular manipulation trials ($M = 3.12$). Also, similar problems (from three degrees of reliability, the same problem diagnosed by at least three different tests) are overall more numerous for the 3D educational software ($M = 3.96$).

Diversity of problems identified by DI. Figures 4 and 5 show that the problems identified are distributed across all the criteria without exception. This shows

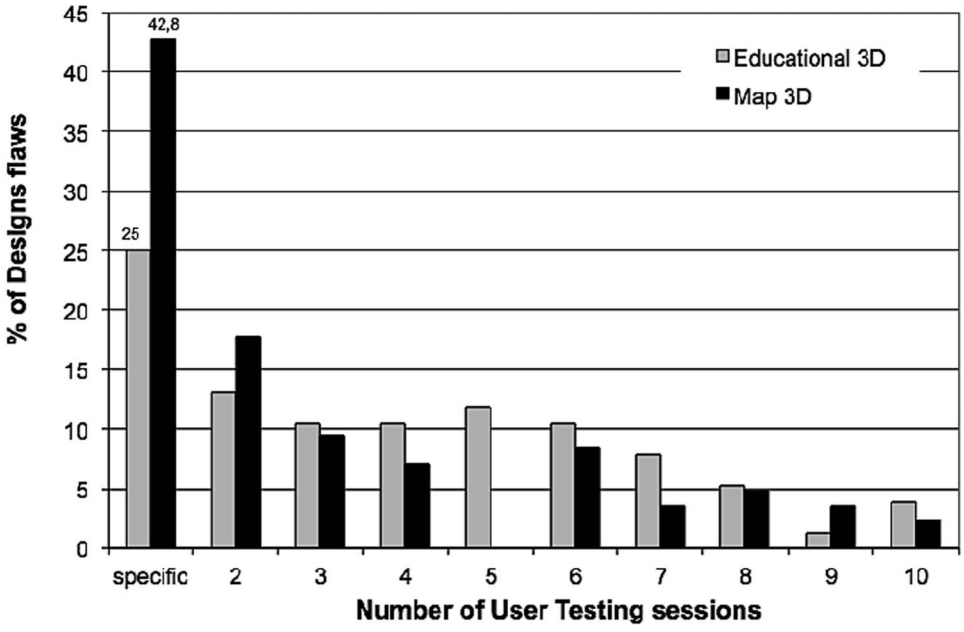

FIGURE 6 Similarity of problems identified during User Testing.

that the group of participants used all the Ergonomic Criteria presented in the document. Concerning the diversity of problems identified by application, the evaluation of the 3D educational software (Figure 4) leads to problems distributed across the 20 *elementary criteria*, which corresponds theoretically to a maximum *coverage*. For the 3D map (Figure 5), 16 *elementary criteria* are used to assign the identified problems. This result shows that the DI using the Ergonomic Criteria allows an evaluation covering a large number of different ergonomic dimensions and leads to the identification of a large variety of usability problems.

The evaluation performance of DI. The evaluation performance averages 25.10 ($SD = 7.37$) problems identified by participant on the 3D educational software and 21.60 ($SD = 7.99$) problems for the 3D map. A student's t test was calculated to test the effect of application type and gender on the evaluation performance. There is no effect of the application, $t(18) = 1.018$, $p = .322$, two-tailed, or gender, $t(18) = 0.836$, $p = .414$, two-tailed, on the performance of evaluation.

Reliability of the results concerning DI. Figure 7 shows the percentage of design flaws identified in each VE and also the number of Document-based Inspections sessions identifying similar problems (reliability). There is a variation of 6% between the specific problems of the 3D educational software (33%) reliability, which is on average 3.19, and those of the 3D map (39%) reliability, which is on average 2.53. This highlights the homogenization effect of the

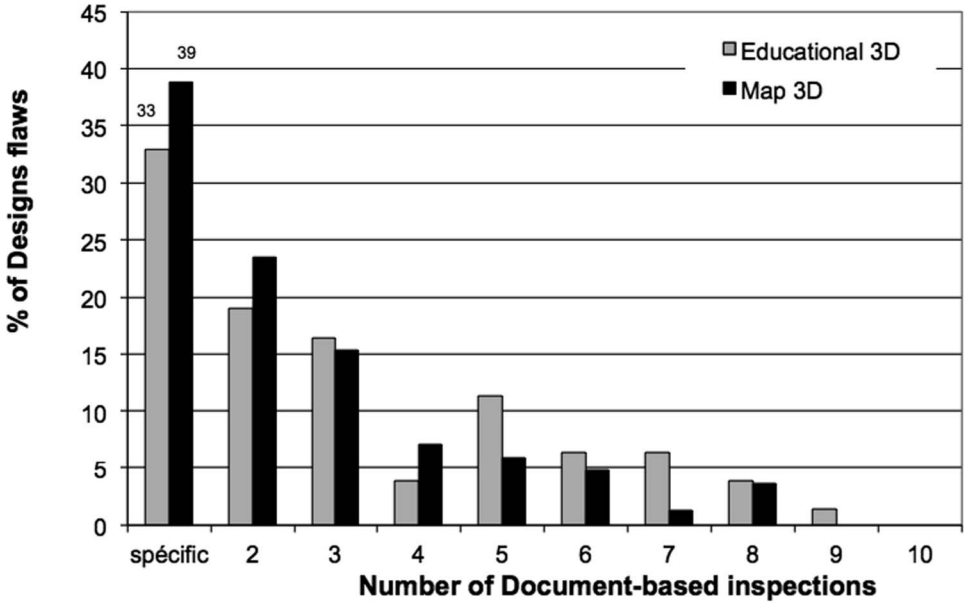

FIGURE 7 Similarity of problems identified during Document-based Inspections.

Ergonomic Criteria on the problems diagnosed for both VE evaluated. The percentage of problems common to at least two participants rates between 61% and 67%.

Diversity of the problems in the EI group. Figures 4 and 5 show that the problems identified with EI are distributed differently according to the applications. For example, no participant diagnosed any problem related to the criteria *Conciseness* and *Taking into account the user's experience* for either application. Problems diagnosed on the 3D educational software (Figure 4) were distributed among 11 *elementary criteria*, particularly the criteria *Compatibility* and *Guidance*. The problems diagnosed on the 3D map (Figure 5) are distributed among 15 *elementary criteria*. This shows, at least for the 3D educational software, that the problems are not very diversified compared to what is possible to obtain theoretically.

Evaluation performance with EI. The evaluation performance of the EI group is, on average, 7.56 ($SD = 3.43$) problems identified by each participant for the 3D educational software and 11.25 ($SD = 4.98$) problems identified for the 3D map. A student's t test was calculated to test the effect of application type and gender on the evaluation performance. There is no significant effect of the application type, $t(15) = -1.800$, $p = .092$, two-tailed, or gender, $t(15) = 0.026$, $p = .980$, two-tailed, on the evaluation performance. It also shows that the evaluation performance is quite similar in the two applications: For educational software, 8 problems were identified on average; for the 3D map, 11 problems were identified on average.

Reliability of the results concerning EI. Figure 8 shows the percentage of design flaws identified in each VE and the number of DI sessions identifying similar problems (reliability). There are no problems common to more than 6 participants. With regard to specific problems, there is a variation of 15% between the 3D educational software (70%) reliability, which is on average 1.69, and the 3D map (55%) reliability, which is on average 1.76. Overall, EI leads to a larger proportion of specific problems.

6.2. Quantitative Comparison Between the Various Methods

Distribution of the problems identified by use of Ergonomic Criteria. Figures 4 and 5 show the distribution of the number of problems using the various methods (assigned to the Ergonomic Criteria). Figure 4 presents the results of the 3D educational software, and Figure 5 presents the results of the 3D map.

Comparing the diversity of results obtained with each method, one can observe in Table 3 that the DI and the UT show less problem diversity on the 3D map: They are distributed on a maximum diversity of 18 criteria, as this application does not include problems associated with *Informational Density* and

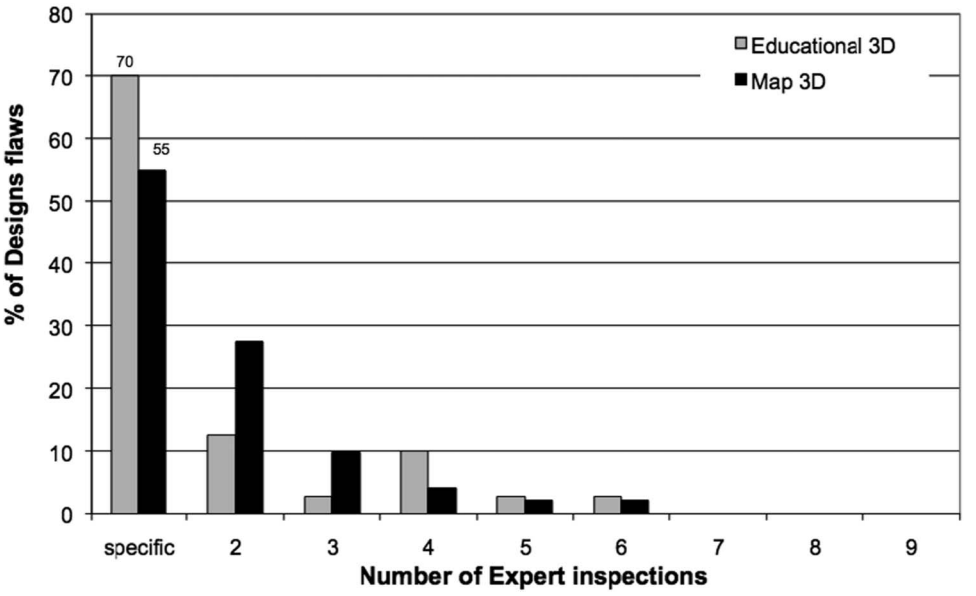

FIGURE 8 Similarity of problems identified during Expert Inspections.

Table 3: Usability Coverage for Each Method

	User Testing	*DI*	*EI*	*Maximum Coverage*
Educational 3D	17	20	10	20
Map 3D	15	16	15	18

Note. DI = Document-based Inspection; EI = Expert Inspection.

Grouping/Distinguishing by Behavior (Figure 5). By comparing the diversity of the results for each method compared to the maximum diversity, one can notice that the DI allowed the identification of a larger diversity of problems than UT and EI.

A quite interesting result in Figures 4 and 5 is that the same criteria distribution appears, whatever the method. However, there are scaling effects for the EI group where the distribution is similar, but where maximum values are lower (this reflects the fact that the overall performance of that EI group is lower). On some criteria, the distribution is even equal in terms of problem numbers: for example, for *Compatibility* with the educational software 3D (Figure 4) or for *Guidance* with DI and UT on the 3D map (Figure 5). This shows that there is a between-method consistency with respect to the types of usability problems present in the applications.

Average evaluation performance per participant. As mentioned before, two one-way ANOVAs were performed. The first shows that there is no effect

of the application type on the evaluation performance $F(1, 55) = 0.118$, $p = .733$. The second shows a significant effect of the method on the evaluation performance, $F(2, 53) = 34.325$, $p = .000$. An additional analysis using two students' t tests shows that there is no significant difference, $t(36) = 1.638$, $p = .1096$, two-tailed, between the evaluation performance of the DI group and the UT group. On the other hand, there is a significant difference, $t(35) = 6.62$, $p = .000$, two-tailed, between the evaluation performance of the DI group and the EI group. Another significant difference, $t(35) = 8.600$, $p = .000$, two-tailed, is found between the UT group and the EI group.

This result shows that DI leads to finding almost as many problems as with the UT method. It also shows that the document describing the Ergonomic Criteria has a significant effect on the evaluation performance in inspection situations. In other words, inspections carried out using the document describing the Ergonomic Criteria allow identification of, over a given time, a more significant number of usability problems than during inspections based solely on the evaluators' knowledge. This result is very encouraging, taking into account the fact that the participants are beginners in terms of usability inspection on VEs and that they only discovered the Ergonomic Criteria during the experiment.

Overlap of problems between the various methods. Here we focus on the proportion of overlap between-methods as well as the proportion of the problems specific to each method. For the 3D educational software 24% of the problems found are specific to the UT group, 19% are specific to the DI group, and 5% to the EI group. On the other hand, 23% of the problems were identified by both UT and DI groups and 11% by both UT and EI group. Eighteen percent of the problems were identified by both DI and EI groups.

Concerning the 3D map, the proportion of problems specific and similar to each method are overall identical to those of the 3D educational software: 22% of the problems are specific to the UT group, 18% to the DI group, and 5% to the EI group. Twenty-two percent of the problems were identified by both UT and DI groups, and 12 % by both UT and EI group. Twenty-one percent of the flaws were identified by both DI and EI groups.

Looking now at the impact of each method on the identification of the whole set of problems by application, that is, to the proportion of problems identified by each method, the results show that the UT group found 58% of the whole set of educational software problems and 56% of the 3D map problems. The DI group found 60% of the educational software and 61% of 3D map problems. Finally the EI group identified 34% of the educational software and 39% of 3D map problems.

This result shows that the DI group identified a proportion of problems almost twice as large as the EI group. It also shows that the proportion of problems found with the DI is slightly higher than for the UT group.

Comparison of the reliability of the problems identified by the various methods. Two types of comparison were performed, first between methods to test the level of reliability and second between applications to test its stability.

The distribution was not normal in all cases hence non-parametric statistics were used. Concerning the effect of method on reliability, a Kruskal-Wallis one-way ANOVA was performed. The test reveals a significant effect of methods, $H(2) = 30.488$, $p = .0000$. Further investigations with Mann-Whitney U test show several interesting results about the level of reliability. First, UT and DI allow a significantly better reliability level than EI, for both VE evaluated (UT vs. EI: U-test $z = 5.235$, $p = .0000$; DI vs. EI: U-test $z = 4.353$, $p = .0000$; global means for each methods are UT = 3.51, DI = 2.85, EI = 1.73). In all cases, the level of reliability of UT and DI is above the threshold fixed at 2 and the EI is below this threshold. A marginal difference was found between UT and DI only on the educational software (U-test $z = 1.705$, $p = .0882$; means are UT = 3.96, DI = 3.19) but not on the Chart 3D (U-test $z = 0.7106$, $p = .4773$; means are UT = 3.12, DI = 2.53). This result shows that UT seems to have a highest level of reliability than DI in the evaluation of the educational software, an application binding itself his exploration by users or evaluators.

A second set of tests was used for the stability of reliability across applications. A Kruskal-Wallis one-way ANOVA was performed and reveals a significant effect of the applications on reliability, $H(1) = 5.827$, $p = .0158$. U tests show a significant difference of the stability of reliability on UT ($z = 2.243$, $p = .0249$; means are educational software = 3.96, Chart 3D = 3.12). A marginal effect was found with the DI ($z = 1.766$, $p = .0774$; averages are educational software = 3.19, Chart 3D = 2.53) and no effect with EI ($z = 1.026$, $p = 0.3049$; averages are educational software = 1.69, Chart 3D = 1.76). This result reveals that the DI group is much more stable than the UT group concerning the stability of reliability across applications. Following this rationale, one can state that EI is very stable across applications but with a poor level of reliability (under the threshold of 2). One can note a convergence of the reliability level of this EI group (1.7) with the reliability level that Chattratichart and Lindgaard (2008) have reported for their Heuristic evaluation group (1.5).

6.3. Qualitative Comparisons

This section describes the classification procedure and coverage, then the classification results, which are mainly illustrated in Table 4.

Table 4: Distribution of Classes of Problem Profiles Over "Overlap Classes"

	Overlap Classes						
Profiles of Problems	*Com. All*	*UT/DI*	*DI/EI*	*UT/EI*	*UT*	*DI*	*EI*
Lack of guidance on interactive objects	1 (51)	3 (29)	2 (36)	3 (5)	3 (20)	2 (9)	
Guidance unsuited to the interaction context	2 (46)	1 (42)			4 (14)	10 (5)	
Lack of guidance on task in progress	4 (36)	(6)			7 (11)		
Lack of guidance on task goals					9 (9)		

(*Continued*)

Table 4: (*Continued*)

	Overlap Classes						
Profiles of Problems	*Com. All*	*UT/DI*	*DI/EI*	*UT/EI*	*UT*	*DI*	*EI*
Lack of guidance on task action sequence					10 (8)		
Lack of guidance toward a previous state			6 (18)			(2)	
Guidance unsuited to the user's virtual position	(14)				(6)		
VE learnability – error in scenario progression	3 (43)						
VE learnability – information timing		(7)	3 (27)				
VE learnability – lack of reminders			4 (21)				
VE learnability – lack of error correction	(7)		9 (7)		(5)		
VE learnability – lack of free discovery					8 (10)	(3)	
Lack of significance leading to an misunderstanding	5 (31)				(5)	3 (8)	
Lack of significance leading to a mistake	7 (21)						
Doubts about the meaning of the wording		(5)	1 (37)		(4)	(2)	
Lack of control over motion or moving	6 (24)	9 (8)	7 (15)		5 (12)		
Problems when selecting objects while moving	(8)						
Overall legibility problem	(14)		5(21)		(3)	(4)	
Legibility problems with information		5 (14)					
Legibility problems from a particular standpoint		6 (12)			(6)		
Lack of compatibility with users' profiles			8 (13)		(2)	5 (7)	
Lack of compatibility with users' expectations	9 (18)		(5)		6 (11)	(5)	
Recommendations for solving problematic situations	10 (16)					8 (6)	
Doubts about help effectiveness		(4)		1 (7)			
Doubts about overall VE efficiency	(15)	7 (10)	10 (7)		2(26)	(3)	
Doubts about specific command efficiency						7 (6)	
Doubts about feedback quality			(3)			6 (7)	
Doubts about function utility						9 (5)	

(*Continued*)

Table 4: (*Continued*)

	Overlap Classes						
Profiles of Problems	*Com. All*	*UT/DI*	*DI/EI*	*UT/EI*	*UT*	*DI*	*EI*
Criticisms about object absence, position, or inadequate size	(14)		(5)		(5)	1 (12)	1 (3)
Lack of presentation conciseness		4 (18)			(4)	(4)	
Lack of consistency between different VE contexts			(6)		1 (32)	4 (7)	2 (3)
Lack of consistency between different VE modalities	8 (20)	8 (9)			(3)		3 (2)
Wayfinding or orientation problems				2 (6)	(4)	(4)	4 (2)
Interaction technique problems		2 (38)			(5)	(2)	
Problems about emotional design		10 (8)			(2)		

Note. UT = User Testing; DI = Document-based Inspection; EI = Expert Inspection; VE = Virtual Environments.

Classification procedure and coverage. In a first step, a set of 46 Problem Profile Classes was identified. From these Problem Profile Classes and Overlap Classes, a table was designed to illustrate the sorted problem *tokens* (rows: Problem Profile Classes; columns: Overlap Classes). From this table, an ordered list of problem *token* frequency (number of *tokens* per Problem Profile Class) was extracted for the seven Overlap Classes. For each Overlap Class, the 10 most often identified problem *tokens* were selected to characterize each class. That way, a list of the 33 most frequent Problem Profile Classes was obtained.

The table was then completed with all other problem *tokens* not selected based on frequency. Also, in the same table, two additional Problem Profile Classes have been included. These classes did not produce a high-enough frequency to be part of the top 10 classes but were selected due to their very specific profile to virtual environments: (a) problems related to inaccurate prompting about the user position in the VE; (b) problems related to object selection during a user move in the VE.

Therefore, the current classification (Table 4) of Problem Profile Classes contains 35 classes. Two types of information are provided: in bold, from 1 to 10, the frequency associated with Problem Profile Classes; in parentheses, the frequency of each problem *token* for each Problem Profile Class within the Overlap Classes. This allows the further qualitative interpretation of the problem profile differences depending on the usability method used.

The 35 Problem Profile Classes cover 93% of the 1,225 problem *tokens* identified in all experimental conditions, by either one of the methods, as assigned to the 7 Overlap Classes in Table 4. That table contains two types of information: in bold, from 1 to 10, the order of frequency associated with Problem Profile Classes; in parentheses, the frequency of each problem *token* of each Problem Profile Class within the Overlap Classes. In other words, the first digit refers to frequency ranking and the second one refers to the number of instances; for example, "1 (51)"

means that it is the highest frequency ranking, corresponding to 51 instances of the same problem, identified with each method, with the following profile *Lack of guidance on interactive objects*. This allows the further qualitative interpretation of the *tokens* profile differences depending on the usability method used.

The following sections presents the results interpreted from Table 4. The classification results concern the profiles of problems specific to each method, the problems identified only with inspections, the problems identified only with UT and DI, and problems identified by all three methods, including four level of consistency of profiles problem classes.

Problems specific to each method. The focus is here on the three Overlap Classes to the right of Table 4: EI, UT group, and DI group. These three Overlap Classes show a set of problem situations that are specific to them. These classes are named Specific Overlap Classes.

For the EI group, problem tokens have been related to four Problem Profile Classes (mainly for consistency issues). However, the identification performance (in parentheses) of the EI group compared to the other methods is quite low. Therefore, one cannot consider that the EI group did identify classes of problems that are specific to their profile, unlike the other methods.

For the UT group, problem tokens have been related to three Problem Profile Classes (mainly for guidance issues, both on task goals and on action sequence). In addition, a particular Problem Profile Class has been identified both by this UT group and by the DI group, even though in a less efficient manner. This category covers the problem tokens related to the *lack of opportunistic discovery of the 3D training* (in other words, the application does not allow the user to explore).

For the DI group, problem tokens have not been, strictly speaking, related to the main Problem Profile Classes but correspond to problem-prone situations as expressed by the participants. Such situations concern the efficiency of certain commands or functions and their utility.

Problems identified only with inspections (Document & Expert). A set of seven Problem Profile Classes have been identified in inspection situations only (DI & EI). Three of these classes have also been identified by the User Testing group (2, 5, 8, respectively). However, inspections have been more efficient on these classes (higher frequency of problem *tokens*).

The heuristic convention was the following: Overlap Classes are not considered efficient if their maximum frequency of problem *tokens* is inferior to the third of the Problem Profile Class with the highest problem *tokens* frequency. For instance, for the category Overall legibility problem, four Overlap Classes are represented: All, DI/EI, UT, DI. The most efficient Overlap Class is DI/EI with 21 problem *tokens*, the second is All (UT = 5_DI = 5_EI = 4), the third is DI = 4, and the last is UT = 3. The rejection threshold is $21/3 = 7$ problem *tokens*. The condition rejected is therefore UT, which is always inferior to 7.

For these seven Problem Profile Classes, more problem *tokens* have been identified in the DI condition. It can therefore be recommended to use that evaluation technique for identifying such problems. These categories are

- *User guidance to return to a previous system state*
- *Problems related to the learnability of the VE*, for example, lack of redundancy in operational instructions
- *Compatibility with the user profile*
- *Quality of feedback*
- *Timing of information presentation* (rate of prompting, reading time allowed, etc.).
- *Overall legibility problems*
- *Doubts about the meaning of the wording*

Profiles of problems identified only with UT and DI. Five Problem Profile Classes have been identified by both UT and DI:

- *Problems associated with interaction devices*
- *Problems of legibility identified from specific virtual positions*
- *Problems related to Emotional Design* (e.g., not pretty, not realistic, . . .)
- *Conciseness of specific presentations*
- *Legibility of specific information*

Problems identified preferentially by UT and EI. Only one Problem Profile Class is common to UT and EI: Problems with the efficiency of the help system. This means probably that the extra effort of inspecting such help features meant that the features were not much used by the DI group.

Problems identified by all methods. In this section, classes of problem profile are presented as a function of their problem *token* assignment into the Overlap Classes. Four levels are distinguished, according to their between-method problem *token* consistency:

- *No* consistency: a set of problem profile classes for which the problem *tokens* have been identified only by specific Overlap Classes
- Weak consistency: a set of problem profile classes for which the problem *tokens* have been identified at least by four Overlap Classes
- Average consistency: a set of problem profile classes for which the problem *tokens* have been identified partly in specific Overlap Classes (actually only for UT & DI) and partly in regular overlap classes (at least four of them)
- High consistency: a set of problem profile classes for which the problem *tokens* have been identified solely in the Overlap Class All

The first level is no consistency. Two problem profile classes belong to the no consistency category:

- *Problems with consistency among the various locations of the VE*; for instance, various entry points do not present the same information; information presentation is sometimes visual, sometimes auditory, sometimes both.

- *Problems with orientation or geolocalization*; for instance boundaries between virtual territories are not represented, or difficulties in spatial positioning.

The lack of consistency between methods for these two classes could be explained by their intrinsic profile. Indeed these classes refer to issues associated with places, positions, specific geographical movements, and so on. Besides, they are not directly identified through task performance but are mentioned through recall ("it was like that there ..."). These types of problem generate lots of variations in information presentation, precisely because there is an inherent lack of consistency between and within applications.

The second level is weak consistency. This set of classes corresponds to most of the problems identified during the three experimental conditions. Seven problem profile classes were identified:

Guidance problems amount to 52.5% of this problem set. These two Problem Profile Classes mostly contain problems common to all methods but also a subset of specific problems:

- *Guidance towards interactive objects*
- *Improper guidance considering the context of use*

Other Problem Profile Classes also mostly contain problems common to all methods but also a subset of specific problems (19%):

- *Problems with movement control*
- *Multimodal inconsistencies*

Some Problem Profile Classes contain an equivalent number of method-specific problems (leading method per problem class in parentheses) and problems common to several methods (28%):

- *Problems with position or size of objects* (DI)
- *Doubts about the overall efficiency of the application* (UT)
- *Lack of compatibility with user expectations* (UT)

Two thirds of the Problem Profile Classes can be considered with an overlap of weak consistency only because it is impossible to overlook problems found with specific methods. Indeed, the assignment of problems to overlap and specific classes (see Table 4 for distribution details) is not equivalent for the first four Problem Profile Classes (see earlier). Only the last three Problem Profile Classes show an equivalent distribution of problems between overlap and specific classes. In other words, these three classes are the only ones to show a clear overlap of weak consistency.

Third is average consistency. Two categories have been extracted. Both categories correspond to problems related to task considerations:

Problems identified by UT:

- *The lack of guidance on the state of task achievement*; for instance, how long will a VE visit last; how many steps are needed to destroy a spaceship, and so on.
- *Inappropriate guidance on the user's virtual position*; for instance, not being able to move efficiently to locate a target, not knowing an interesting tourist spot is nearby, and so on.

Problems identified by DI:

- *Problems of significance leading to misunderstanding*; for instance, unclear statements from the virtual trainer, wrong units on a scale.
- *Issues with ways of resolving dead-end situations*; for instance, guidance following an unsolved task, guidance toward a point outside the field of view.

Finally, there is high consistency. These classes correspond to the easiest types of problems, with the best discoverability, that are found in each experimental condition:

- *Errors in the scenario sequence* (educational software scenario only). For example, the educational software shows how to move the spaceship after a pursuit exercise. Logically, the method of movement should have been taught before the more complex task of pursuit (following and then shooting a target).
- *Misleading vocabulary and icons*
- *Problems when selecting objects during automatic moves* (very frequent in the 3D map)

This categorization allows for the characterization of overall tendencies in terms of problem profile as they are identified by either method. For UT, usability problems are mainly related to tasks and activities, whereas for DI, the usability problems are mainly related to the understandability and learnability of the software systems.

7. CONCLUSION

This study was conducted to compare experimentally a UEM based on a document describing Ergonomic Criteria adapted specifically to VEs. The full definitions, justifications, and examples of recommendations can be found in Bach and Scapin (2005). A series of comparisons of evaluation performance, mainly following the criteria from Gray and Salzman (1998) and Hartson et al. (2001), were conducted using UT and EI. All forms of evaluation were performed on two different VEs (a 3D Educational software and a 3D map) in order to estimate the evaluation *stability* of the various methods. The comparisons were first carried out using the problem classification based on Ergonomic Criteria, which has already been demonstrated to be effective (Bach & Scapin, 2003). Ten participants took part

in UT. The experimenters themselves then analyzed and classified the problems the participants had encountered in their usability laboratory sessions. Ten other participants looked for usability problems using a Document, whereas 9 other participants did the same, simply based on their own knowledge. In both cases, the experimenters carried out the final assignment of the various problems identified by the participants, by use of the Ergonomic Criteria.

The main result of this comparison is that there is a significant difference between using the DI and EI, in terms of number of usability problems found over time. This result is encouraging as compared to the results from Chattratichart and Lindgaard (2008), who showed that using heuristics for Web site evaluation leads to a lower performance than using EI.

In addition, the performance for UT and DI is rather similar, although the results did not show a significant effect of the applications type on method performance. This result is quite similar to results from Molich and Dumas (2008) obtained through a study crossing different methods with or without users to evaluate a Web site. This shows that DI is a potential method for short-circuiting the difficulties related to UT (Bowman et al., 2002) to evaluate VEs more complex technically. This method, which is less costly, could help alleviate the most obvious usability problems.

Concerning the reliability of the methods, the results showed both variations due to the evaluated application and disparities between the methods. Indeed, interapplication variability as observed in the UT was attenuated in DI group. The results show a tendency for the DI to result in better stability than UT. However, for UT, reliability is more important for the training application. The interappraisers variability has already been highlighted in the 2D world, for UT (Faulkner, 2002) and for Free Inspection (Pollier, 1991). However, for interapplication variability, no such studies have been identified; most experiments have dealt with only one type of application at a time.

The results concerning the diversity of problems identified with each method show that the DI allowed the identification, in both applications, of more problem diversity than the two other methods. This result is important in terms of problem identification *coverage* of the usability methods. The trade-off is, of course, between a very narrow field of diagnoses and an important dispersion with a large scope. On this issue, it seems that DIs are a good compromise compared to the two other methods.

The results concerning the intermethod overlap show a relative stability interapplication for the problems, both common and specific. The proportion of problems shared by inspection and UT is twice more if the DI are used (approximately 22%). This result supports the idea of using the DI to help characterize the real difficulties in use observed during UT.

Another interesting result relates to the diagnosis power of each method, assessed with the complete set of problems identified in the whole experiment, regardless of the method. The DI demonstrates the highest power of diagnosis, together with the largest identification *coverage*. Overall, these characteristics lead to a high and stable evaluation performance (about 60% of all problems). These results, using several metrics, corroborate the recommendations of Hartson et al.

(2001) concerning the necessary composite approach (multicriteria) for methods evaluation as well as systems in general.

Two important issues concerning the variables used for quantitative evaluation of methods are worth mentioning:

- First, on all evaluation metrics used, the DI were found to be significantly more powerful than the EI group, which tends to validate the usefulness of usability DI of VEs.
- Second, on certain evaluation metrics, UT were shown to be more powerful than the DI but relatively less stable when looking at interapplication and intersubject performance (e.g., gender effect for UT not for DI). This leads to the conclusion that the DI offers a greater stability in the overall diagnosis performance regardless of application or evaluator.

Using a motor vehicle metaphor, one could say that the DI could be viewed as "an SUV evaluation type" (less powerful under certain conditions but can go everywhere with any driver) whereas UT could be viewed as "Formula 1 car evaluation type" (more powerful but requiring adequate road and a very skilled driver).

The results obtained with quantitative analyses provide other interesting conclusions. Qualitative data show that the different usability evaluation methods do not tend to identify the same profile of problems. UT seems particularly efficient for the diagnosis of problems that require a particular state of interaction to be "detectable." On the other hand, DI supports the identification of problems "directly observable," often related to learnability and basic usability.

These results should lead to additional research work, as many questions remain open. For instance:

This experiment made the assumption that reading a document before inspecting has a debiasing effect. Indeed the results tend to show such an effect, mainly when comparing interviews with the EI group to those from the DI group: For the latter group, participants suggested they memorized a sort of analysis grid. Of course, this remains to be confirmed in detail and related to other literature results (Fischoff, 1982; Zachary, 1986) on this phenomenon, in other application areas.

Other interview results show also that the evaluators (DI and EI) mention having difficulties in understanding the VE structure in terms of interactive elements. If confirmed, these results would lead to the design of inspection grids that would illustrate (e.g., in terms of inspection order) the VE structure. Actually, for another study, the basic elements of such a grid have been designed under a form of 73 interactive elements used to classify our ergonomic guidelines. In addition, this set was used to complement a metamodel of the ASUR notation (Dubois, Abou Moussa, Bach, & Bonnefoy, 2008), aimed at modeling Mixed Systems (environments that are both physical and digital).

Similarly, it would be interesting to reinforce the *profile problem classification* obtained from this experiment by further analysis dedicated to its intrinsic validity, for example, by use of an adapted statistical model such as the one developed by Schmettow and Vietze (2008) or with other classification schemes, such as user

action framework, or UAF (Andre, Hartson, Belz, & McCreavy, 2001). Augmenting the quality of such a problem classification would be useful as it has been shown (Chattratichart & Lindgaard 2008) that a list of problem profiles helps Web site inspection performance. Overall, a promising direction for further work would be to improve the efficiency of DI through the joint use of the Ergonomic Criteria, the set of interactive elements, and a problem classification scheme. This could generate results comparable to those obtained in a more mature and more specialized area, namely, Web accessibility for the blind (Mankoff, Fait, & Tran, 2005).

Another perspective concerns the tools for inspection. Investigating such tools could provide some flexibility to the inspection and better accommodate the various evaluator strategies, including continuous reading of the structure and content of the inspection document; mental association of problems diagnosed through document search; free problem search, but a posteriori classification of the problems with the document. In addition, a flexible, points-of-view based document could help in inspection reporting, which may be beneficial for downstream utility (Hartson et al., 2001). Some tools have already been proposed by Stanney et al. (2003) and by Karampelas et al. (2003).

It would be also useful to work on the relationship between DI methods with other methods dedicated to specific aspects of user interaction with VEs, such as cybersickness and presence. As stated by Bowman et al. (2002), it is difficult to envision inspection methods for those topics, whereas other methods are good candidates, such as questionnaires (Kalawsky, 1999; Kennedy et al., 1993; Witmer & Singer, 1999).

Finally, research is also needed to limit the difficulties in UT protocols for evaluating VEs. One aspect relates to the gender effect. Such type of effect has been identified by Green and Bavelier (2006) who only selected male individuals to participate in their video game experiments. To control such aspects, other work (Griffiths, Sharples, & Wilson, 2006) was carried out using a series of predictive tests for user performance in VEs to support appropriate subject selection for UT.

The perspectives just mentioned, as well as many others, are only examples of the large research effort needed to support the extensive dissemination of VEs in our future professional and personal lives.

7.1. Limitations

This extensive quantitative and qualitative comparison of UEMs led to interesting results regarding the role and coverage of three different UEMs. However, a number of limits must be acknowledged, including the following.

The allocated time of 30 min per application for inspection is obviously lower than the time necessary for a full VE evaluation, under normal conditions. For example, Wilson, Eastgate, and D'Cruz (2002) observed that time necessary to carry out an inspection was approximately a day to which additional time (about a week) must be added for problem description and report production. However, this can vary according to the inspection goals, their level of exhaustiveness, and the profile of the VE.

Obviously, the time measurements focus on the experimental sessions, on the participants' performance. It does not include the time required for extraction and analysis of problems identified by the various methods. Future work should incorporate techniques for independent evaluation of such activities. Nevertheless, it seems obvious that the processing time for UT analysis is much longer than for inspection reports.

The issue of problem severity (e.g., how important they are; how difficult it is to fix them) is of importance but not answered here; it would need a follow-up study.

Related to this, there is the issue problem equivalence. As long as no appropriate priority or weights have been established, no overall effectiveness can be stated when assessing individual methods. But again, for comparing methods, it is perfectly legitimate to use such numbers in terms of coverage and thoroughness.

The application systems selected for the experiments were only desktop VE. It would be interesting to carry out usability evaluations with more technologically complex VEs (e.g., immersive VE, stereoscopic visualization, haptic feedback), particularly in terms of interaction techniques.

This study does not answer the issue of cost/benefit nor does it highlight the possible method asymptote threshold. This would require focusing on the number of newly identified problems compared to the previous evaluations, as well as a reordering of the participants' performance. Such an automatic data reorganization will be worked on in the future.

It would also be useful to carry out additional analyses to complement current results, by using additional qualitative information collected during postexperimental sessions. Such data, which were not described in this article, concern for instance the evaluation of satisfaction during UT or inspection.

Finally, the participants conducting the evaluations were rather novice in terms of usability inspection. However, this may not be such a limitation considering past results showing quite limited differences in evaluation performance between experienced analysts and novices (Bastien & Scapin, 1992; Chattratichart & Lindgaard, 2008). Because such differences may be mostly qualitative, namely in terms of problem description, it is quite difficult to shed light on such differences in rather short experiments. In any case, because the VE technology is quite recent, not many evaluators can be considered experts at this time.

REFERENCES

Adamo-Villani, N., Wilbur, R., & Wasburn, M. (2008). Gender differences in usability and enjoyment of VR educational games: A study of SMILE™. In *Proceedings of the 2008 International Conference* (pp. 114–119). Washington, DC: IEEE Computer Society.

Agarawala, A., & Balakrishnan, R. (2006). Keepin' it real: Pushing the desktop metaphor with physics, pile and the pen. In *Proceedings of the ACM CHI 2006* (pp. 1283–1292). New York: ACM Press.

Andre, T. S., Hartson, H. R., Belz, S. M, & McCreavy, F. A. (2001). The user action framework: a reliable foundation for usability engineering support tools. *International Journal of Human-Computer Studies*, *54*, 107–136.

Bach, C. (2004). *Élaboration et validation de Critères Ergonomiques pour les Interactions Homme-Environnements Virtuels* [Development and validation of ergonomic criteria for human virtual-environments interactions]. Unpublished doctoral thesis, Université Paul Verlaine, Metz, France.

Bach, C., & Scapin, D. L. (2003). Adaptation of ergonomic criteria to human-virtual environments interactions. In *Proceedings of Interact'03* (pp. 880–883). Amsterdam: IOS Press.

Bach, C., & Scapin, D. L. (2005). *Critères ergonomiques pour les interactions homme-environnements virtuels: Définitions, justifications et exemples* [Ergonomic criteria for human-virtual environments interactions: definitions, justifications, and examples] (Tech. Rep. No. 5531). Rocquencourt, France: French National Institute for Research in Computer Science and Control (INRIA).

Bastien, J. M. C., & Scapin, D. L. (1992). A validation of ergonomic criteria for the evaluation of human–computer interfaces. *International Journal of Human-Computer Interaction*, *4*, 183–196.

Bastien, J. M. C., & Scapin, D. L. (1993). *Ergonomic criteria for the evaluation of human–computer interfaces* (Tech. Rep. No. 156). Rocquencourt, France: French National Institute for Research in Computer Science and Control (INRIA).

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

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

Bowman, D., Gabbard, J. L., & Hix, D. (2002). A survey of usability evaluation in virtual environments : classification of methods. *Presence: Teleoperators and Virtual Environments*, *11*, 435–455.

Bowman, D. A., & Hodges, L. F. (1997). An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In *Proceedings of the 1997 symposium on Interactive 3D graphics* (pp. 35–38). New York: ACM Press.

Bowman, D., Kruijff, E., LaViola, J., & Poupyrev, I. (2005). *3D user interfaces: Theory and practice*. Boston: Addison-Wesley.

Chapanis, A. (1982). Man/computer research at Johns Hopkins. In R. A. Kasschau, R. Lachman, & K. R. Laughery (Eds.), *Information Technology and Psychology: Prospects for the future* (pp. 238–249). New York: Praeger.

Chattratichart, J., & Lindgaard, G. (2008) A comparative evaluation of heuristic-based usability inspection methods. In *Proceedings of the ACM CHI '08 extended abstracts on Human factors in computing systems* (pp. 2213–2220). New York: ACM Press.

Cockton, G., & Lavery, D. (1999). A framework for usability problem extraction. In *Proceedings of Interact'99* (pp. 344–352). Amsterdam: IOS Press.

Cockton, G., Woolrych, A., Hall, L., & Hidemarch, M. (2003). Changing analysts' tunes: The surprising impact of a new instrument for usability inspection method assessment. In *Proceedings of People and Computers XVII: Designing for Society* (pp. 145–162). New York: Springer Verlag.

Conkar, T., Noyes, J. M., & Kimble, C. (1999). CLIMATE: A framework for developing holistic requirement analysis in Virtual Environments. *Interacting with Computers*, *11*, 387–402.

Connell, I. W., & Hammond, N. V. (1999). Comparing usability evaluation principles with heuristics: problem instances vs. problem types. In *Proceedings of Interact'99* (pp. 621–629). Amsterdam: IOS Press.

Dow, S., MacIntyre, B., Lee, J., Oezbek, C., Bolter, J. D., & Gandy, M. (2005). Wizard of Oz support throughout an iterative design process. *IEEE Pervasive Computing*, *4*(4), 18–26.

Dubois, E, Abou Moussa, W., Bach, C., & Bonnefoy, N. (2008). Modelling and simulation of mobile mixed systems. In J. Lumsden (Ed.), *Handbook of research on user interface design and evaluation for mobile technology* (pp. 346–363). Hershey: IGI Global.

Dubois, E., Truillet, P., & Bach, C. (2007). Evaluating advanced interaction techniques for navigating Google Earth . In *Proceedings of the 21st BCS HCI Group Conference HCI 2007* (Vol. 2, pp. 4–7). London: BSC.

Dubois, E., Nedel, L. P., Dal Sasso Freitas, C. M., & Jacon, L. (2005). Beyond user experimentation: Notational-based systematic evaluation of interaction techniques in virtual reality environments. *Virtual Reality*, *8*, 118–128.

Durlach, B. N. I., & Mavor, A. S. (1995). *Virtual reality: Scientific and technological challenges*. Washington, DC: National Academy Press.

Faulkner, L. (2002, July). Reducing variability—Research into structured approaches to usability testing and evaluation. In *Proceedings of Annual Conference of the Usability Professionals Association*.

Fischoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, UK: Cambridge University Press.

Gabbard, J. L., & Hix, D. (1997). *A taxonomy of usability characteristics in virtual environments*. Unpublished master's thesis, Virginia Tech, Blacksburg, VA.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, *13*, 203–261.

Green, C. S., & Bavelier, D. (2006). Effect of action video games on the spatial distribution of visuospatial attention. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 1465–1478.

Griffiths, G., Sharples, S., & Wilson, J. R. (2006). Performance of new participants in virtual environments: The Nottingham tool for assessment of interaction in virtual environments (NAIVE). *International Journal of Human-Computer Studies*, *64*(3), 240–250.

Hand, C. (1997). A survey of 3D interactions techniques. *Computer Graphics Forum*, *16*, 269–281.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*, 373–410.

Hornbaek, K. (2009). Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology*, pp. 1–15.

Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. In *Proceedings of ACM Conference on Human Factors in Computing Systems* (pp. 391–400). New York: ACM Press.

Hughes, C. E., Smith, E., Stapleton, C. B., & Hughes, D. E. (2004). Augmenting Museum Experiences with Mixed Reality. In *Proceedings of Knowledge Sharing and Collaborative Engineering 2004*. Calgary, Canada: Acta Press.

International Standards Organisation. (2000). *ISO / TS 16982 Ergonomics of Human-System Interaction–Usability methods supporting Human Centred Design*. Geneva: Author.

Kalawski, R. S. (1999). VRUSE: A computerised diagnostic tool for usability evaluation of virtual/synthetic environment systems. *Applied Ergonomics*, *30*(1), 11–25.

Karampelas, P., Grammenos, D., Mourouzis, A., & Stephanidis, C. (2003). Towards I-Dove, an interactive support tool for building and using virtual environments with guidelines. In *Proceedings of the 10th HCI International* (Vol. 3, pp. 1411–1415). Mahwah, NJ: Erlbaum.

Kaur, K. (1998). *Designing virtual environments for usability*. Unpublished doctoral thesis, City University, London.

Kaur, K., Maiden, N., & Sutcliffe, A. (1996). Design practice and usability problems with virtual environments. In *Proceedings of Virtual reality World'96 conference*. Stuttgart, Germany: IDG conferences.

Kaur, K., Maiden, N., & Sutcliffe, A. (1999). Interacting with virtual environments: an evaluation of a model of interaction. *Interacting with Computers*, *11*, 403–426.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire an enhanced method for quantifying simulator sickness. *International Journal of Aviation Psychology*, *3*, 203–220.

Law, E., Scapin, D. L., Cockton, G., Springett, M., Stary, C., & Winckler, M. (2009). Maturation of Usability Evaluation Methods: Retrospect and prospect. *Proceedings of COST294-MAUSE Closing Conference*. Toulouse, France: IRIT Press.

Leulier, C., Bastien, J. M. C., & Scapin, D. L. (1998). *Compilation of ergonomic guidelines for the design and evaluation of Web sites* (Contract report). Rocquencourt, France: French National Institute for Research in Computer Science and Control (INRIA).

Lindgaard, G. (2006). Notions of thoroughness, efficiency, and validity: Are they valid in HCI practice? *International Journal of Industrial Ergonomics*, *36*, 1069–1074.

Mankoff, J., Fait, H., & Tran, T. (2005). Is your web page accessible?: a comparative study of methods for assessing web page accessibility for the blind. In *Proceedings of the ACM CHI '05* (pp. 41–50). New York: ACM Press.

Molich, R., & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, *27*, 263–281.

Microsoft Game Studios. (2000). *StarLancer*. Redmond, WA: Author. Retrieved June 12, 2008, from http://www.microsoft.com/games/da/starlancer/

Nielsen, J. (1993). *Usability engineering*. San Diego, CA: Academic.

Pollier, A. (1991). *Evaluation d'une interface par des ergonomes: Diagnostics et stratégies*. [User interface evaluation by ergonomists: diagnoses and strategies] (Research Rep. No. 1391). Rocquencourt, France: French National Institute for Research in Computer Science and Control (INRIA).

Poupyrev, I., & Ichikawa, T. (1999). Manipulating objects in virtual worlds: Categorization and empirical evaluation of interaction techniques. *Journal of Visual Languages and Computing*, *10*(1), 19–35.

Scapin, D. L. (1990). Organizing human factors knowledge for the evaluation and design of interfaces. *International Journal of Human-Computer Interaction*, *2*, 203–229.

Scapin, D. L., & Bastien, J. M. C. (1997). Ergonomic criteria for evaluating the ergonomic quality of interactive systems. *Behaviour & Information Technology*, *16*, 220–231.

Scapin, D. L., & Law, E. (2007). Review, report and refine Usability Evaluation Methods (R3 UEMs). *Proceedings of the COST294-MAUSE 3rd International Workshop*. Athens, Greece.

Schmettow, M., & Vietze, W. (2008). Introducing item response thery for measuring usability inspection processes. In *Proceedings of the ACM CHI '08* (pp. 893–902). New York: ACM Press.

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Human-Computer Interaction*. MA: Addison-Wesley.

Stanney, K. M. (2002). *Handbook of Virtual Environments: Design, implementation, and applications*. Mahwah, NJ: Erlbaum.

Stanney, K. M., & Davies, R. C. (2005). Augmented reality in Internet applications. In R. W. Proctor & K-P L. Vu (Eds.), *Handbook of human factors in Web design* (pp. 647–657). Mahwah, NJ: Erlbaum.

Stanney, K. M., Mollaghasemi, M., Reeves, L., Breaux, R., & Graeber, D. A. (2003). Usability engineering of virtual environment (VEs): Identifying multiple criteria that drive effective VE system design. *International Journal of Human-Computer Studies*, *58*, 447–481.

Stanney, K. M., Mourant, R. R., & Kennedy, R.S. (1998). Human Factors issues in Virtual Environments: A review of the literature. *Presence: Teleoperators and Virtual Environments*, *7*, 327–351.

Stapleton, C. B., & Hughes, C. E. (2005). Mixed reality and experiential movie trailers: Combining emotions and immersion to innovate entertainment marketing. In *Proceedings of 2005 International Conference on Human-Computer Interface Advances in Modeling and Simulation* (pp. 40–48). New Orleans.

Sutcliffe, A. (2003). *Multimedia and virtual reality : designing multisensory user interfaces*. Mahwah, NJ: Erlbaum.

Sutcliffe, A., & Gault, B. (2004). Heuristic evaluation of Virtual Reality applications. *Interacting With Computers*, *16*, 831–849.

Sutcliffe, A., & Kaur, K. (2000). Evaluating the usability of virtual reality user interfaces. *Behaviour & Information Technology*, *19*(6), 415–426.

Tromp, J. G. (2001). *Systematic usability design and evaluation for collaborative virtual environments*. Unpublished doctoral thesis, University of Nottingham, Nottingham, United Kingdom.

Tromp, J. G., Steed, A., & Wilson, J. R. (2003). Systematic Usability evaluation and design issues for collaborative Virtual Environments. *Presence: Teleoperators and Virtual Environments*, *12*, 241–267.

Wickens, C. D., & Baker, P. (1995). Cognitives issues in virtual reality. In W. Barfield & T. A. Furness III (Eds.), *Virtual environments and advanced interface design* (pp. 514–541). New York: Oxford University Press.

Williams J. S., & Harrison M. D. (2001). A toolset supported approach for designing and testing virtual environment interaction techniques. *International Journal of Human-Computer Studies*, *55*, 145–165.

Wilson, J. R., Eastgate, R. M., & D'Cruz, M. (2002). Structured development of Virtual Environments. In K. M. Stanney (Ed.), *Handbook of Virtual Environments. Design, implementation, and applications* (pp. 353–378). Mahwah, NJ: Erlbaum.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, *7*, 225–240.

Zachary, W. (1986). Cognitively based functional taxonomy of decision support techniques. *Human–Computer Interaction*, *2*, 25–63.