Safety Science 62 (2014) 286–294

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/ssci

Comparing the comprehensiveness of three expert inspection methodologies for detecting errors in interactive systems

Christopher Power *, John Fox

Human Computer Interaction Research Group, Department of Computer Science, University of York, York YO10 5GH, United Kingdom

A R T I C L E I N F O

Article history:
Received 19 April 2013
Received in revised form 11 July 2013
Accepted 7 August 2013
Available online 8 October 2013

Keywords:
Expert inspection methods
Error inspection methods
Usability inspection methods
Error types
Behavioural deviation
Cognitive phase
SHERPA
HEHAZOP
Cognitive Walkthrough

A B S T R A C T

Expert inspection methods provide a means of evaluating interactive systems for error throughout the design lifecycle. Experts have a wide variety of methods available to them for detecting either potential user errors or usability problems that will provoke error; however, the data on what types of errors are detected by each method is very thin. This paper, presents the results of a study into the comprehensiveness of three expert inspection methods that were applied by nine evaluators across three devices. This study produced 350 errors that were analysed to identify, compare and contrast what types of errors were detected by each method. Of particular interest, the investigation revealed that a substantial number of distinct errors were detected by only one method. The paper closes with a discussion of the implications of these results on future practice for multi-method approaches as well as directions for future investigations.

1. Introduction

For any complex interactive system, it is essential that the designers have confidence that users can interact with it successfully. A key part of this is to ensure that interactive systems do not provoke errors that have users arrive in an undesirable state. This paper presents a study that compares and contrasts three different expert inspection methods for their comprehensiveness in uncovering potential errors in interactive systems.

In the broad literature of human error identification, there are three key types of evaluation methods that can be used to detect and correct potential errors during the design lifecycle (Baber and Stanton, 1996):

- Application of error inspection methods that focus experts in identifying specific errors that can occur in a system.
- Application of usability inspection methods that focus experts in uncovering usability problems in the interface that may provoke errors in users.
- Empirical user studies, through either true experiments or task-based usability studies, where users are observed interacting with a system.

Within this paper, we will refer to the first two types of evaluation as *expert inspection methods*. These expert inspection methods have the benefit that they can be applied at early stages of prototype development, without the final system being present. Further, they provide a more thorough and structured exploration of a system than is possible even with large user studies. Finally, in general, they are less expensive in terms of time and other resources than empirical user testing.

However, there are several potential problems that arise when working with different expert inspection methods. First, there are dozens of expert inspection methods available to evaluators of interactive systems, such as the 38 different methods surveyed by Kirwan (1998a, 1998b), many of which claim to provide comprehensive coverage of errors in systems. Second, there are serious problems regarding the validity of different methods. Of those same 38 methods, none of them satisfied the criteria set out by Kirwan for achieving a sound and valid accounting of potential errors in a system.

Indeed, even when validation data is available for methods it is usually on a small subset of the criteria. For example, several studies have looked at the *accuracy* of methods to identify errors that occur during real world use of a system. This includes studies that retrospectively investigated accidents with expert inspection methods from the HEI domain (Whalley and Kirwan, 1989 cited in Stanton et al., 2009; Kennedy, 1995) as well as direct compari-

* Corresponding author. Tel.: +44 1904325673.
E-mail address: Christopher.power@york.ac.uk (C. Power).

0925-7535/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ssci.2013.08.003

sons in live environments (Baber and Stanton, 1996). In other cases, methods have been compared regarding their accuracy, including false positive and false negative rates, in comparison to *in situ* environments. For example, Stanton and Baber (1998) compared two different methods, Systematic human error reduction and prediction approach (SHERPA) and Task analysis for error identification (TAFEI) and found that SHERPA yielded a consistently higher sensitivity to identifying errors that occurred in real world use of the evaluated system.

However, while some methods are providing reasonably good accuracy, it is not clear that they are all finding the same types of problems. There are very few studies looking at the *comprehensiveness* of methods regarding the types of problems that they find, as originally described by Kirwan (1992a, 1992b). For example, if two different methods A and B each find 75% of the errors in a system, are the two sets of problems found by methods A and B the same? Or is there a difference in the types of problems that are found and those that are missed? There is very little data on this particular point in the literature.

The issue of comprehensiveness becomes a very important one when looking at multi-method approaches to error identification. In the early 1990s Kirwan (1992b) proposed the combining of SHERPA with additional expert judgement to help provide comprehensiveness in coverage of errors. Then, in 1998, Kirwan suggested that a more modular, toolkit style approach was needed where multiple methods are applied to a system to eliminate the breadth of errors that occur (Kirwan, 1998a, 1998b).

Stanton et al. (2009) observe that this type of multi-method approach has its own risks. First, pooling error data may increase the false alarm rate if the same problem is highlighted many times across different methods. This could be partially addressed by doing error matching between different methods to provide a set of unique error signatures. However, there is no data on what is the scale of overlaps that exist between different methods and how many errors are shared between different methods. Second, a multi-method approach will increase the amount of resources that is committed to evaluating any specific system. As a result, it is important to know what types of errors the methods find and whether methods are finding different problems or largely the same problems. With that information in hand, it can be quantified whether the increased resources are of value.

Once the decision has been made to take a multi-method approach, there is also a question as to whether methods should be drawn from the collection of error inspection methods that have arisen from human error identification research and practice (HEI), or if those should be mixed with usability inspection methods that have arisen from human computer interaction (HCI) research and practice. Thimbleby argues that usability inspections and user centred design practices are a necessary but insufficient means for detecting the robust set of errors that occur in systems (Thimbleby, 2007, 2008). However, it remains unclear what differences there are between HEI and HCI methods.

In this paper, we present a study that compares three different expert inspection methods in terms of their comprehensiveness of covering different types of errors. We present a methodology for collecting errors from expert evaluators and for analysing the results through different perspectives. The purpose of this study is to answer three questions:

1. Can we detect differences in the types of errors that different expert inspection methods find?
2. What is the level of overlap in the errors found between different expert inspection methods?
3. Are there substantial differences in the types of errors found by a usability inspection method when compared to error inspection methods?

2. Evaluation methods

We have chosen three evaluation methodologies to investigate the research questions proposed above. We chose two commonly used approaches for error inspection and one comparable, structured, usability inspection method.

2.1. Human Error Hazard and Operability analysis (HEHAZOP)

Originally developed by Kletz (1992, 2006) as a process for risk assessment in engineering domains, HAZOP (**Haz**ard and **Op**erability analysis) is a thorough method of assessing a system in even early design phases, and considers every element of the system based on engineering diagrams. The original HAZOP process has experts apply a set of structured guidewords to identify potential places where hazards could occur in the system. It is claimed to uncover potential for errors by humans when interacting with a system (Kirwan, 1992a). Whalley (1988 cited in Stanton et al., 2012) is produced a modified set of the original guidewords to be more targeted at uncovering previously observed human related errors.

In practice, the HEHAZOP methodology is applied by a team of experts from different domains working on an interactive system. This allows the broad range of experiences help assess the severity of potential hazards and possible design changes that can mitigate the effects (Stanton et al., 2012). For the purposes of this study, we have focussed on the just the stage of identifying errors in an interactive system.

We have chosen this method as it is commonly used as a comparator in the HEI literature (Kirwan, 1992b, 1998b; Stanton et al., 2009). Further, it is a very thorough method, looking at a broad range of potential errors that can occur. However, while there are claims that HEHAZOP is capable of identifying all errors in a system (Stanton et al., 2012), it is not clear what types of errors it most commonly uncovers.

2.2. Systematic Human Error Reduction and Prediction Approach (SHERPA)

The Systematic Human Error Reduction and Prediction Approach (SHERPA) was developed by Embrey (1986 cited in Stanton et al., 2012) in the mid-1980s for the nuclear reprocessing industry; however more recently it has been applied with success to a number of other domains (Baber and Stanton, 1996; Harris et al., 2005; Salmon et al., 2002). It is a taxonomic approach, where experts must first identify the error mode in which an action takes place, and then considers the operator behaviour that occurs within that mode to identify potential sources of error. SHERPA uses its own taxonomy of error type, which has its roots in Rasmussen's skills, rules and knowledge model (Rasmussen, 1983).

In contrast to HEHAZOP, one analyst applies SHERPA during an evaluation, with one of its major advantages being that it is claimed to require very little training to become a proficient user (Salmon et al., 2002).

Similar to HEHAZOP, SHERPA is a candidate for this study as it is a highly studied method that serves as a comparator for others methods. It is systematic method that allows the expert to identify errors within the tasks a user undertakes in a system and due to the robustness of the taxonomy it is claimed to identify the majority of errors that can occur in a system. There are two previously identified key shortcomings of the SHERPA method. First, it does not explicitly consider latent failures, and, second, it has been criticized for lacking the cognitive components in its error mechanisms (Stanton et al., 2005). However, there is little concrete data regarding if SHERPA, in fact, largely misses cognitive errors, or if they are revealed during evaluations.