Preprint. Final version available as:

BLANDFORD, A., HYDE, J. K., GREEN, T. R. G. & CONNELL, I. BLANDFORD, A., HYDE, J. K., GREEN, T. R. G. & CONNELL, I. (2008) Scoping Usability Evaluation Methods: A Case Study. *Human Computer Interaction Journal.* 23.3. 278 – 327. DOI 10.1080/07370020802278254

This version is not definitive.

Scoping analytical usability evaluation methods: a case study

Ann E. Blandford

University College London

Joanne K. Hyde

Formerly at Middlesex University

Thomas R. G. Green

University of Leeds

Iain Connell

Formerly at University College London

Abstract

Analytical usability evaluation methods (UEMs) can complement empirical evaluation of systems: for example, they can often be used earlier in design, and can provide accounts of why users might experience difficulties, as well as what those difficulties are. However, their properties and value are only partially understood. One way to improve our understanding is by detailed comparisons using a single interface or system as a target for evaluation, but we need to look deeper than simple problem counts: we need to consider what *kinds* of accounts each UEM offers, and why. Here, we report on a detailed comparison of eight analytical UEMs. These eight methods were applied to a robotic arm interface, and the findings were systematically compared against video data of the arm in use. The usability issues that were identified could be grouped into five categories: system design; user misconceptions; conceptual fit between user and system; physical issues; and contextual ones. Other possible categories such as user experience did not emerge in this particular study. With the exception of Heuristic Evaluation, which supported a range of insights, each analytical method was found to focus attention on just one or two categories of issues. Two of the three 'home grown' methods (EMU and CASSM) were found to occupy particular niches in the space while the third (PUM) did not. This approach has identified commonalities and contrasts between methods and provided accounts of *why* a particular method yielded the insights it did. Rather than considering measures such as problem count or thoroughness, this approach has yielded insights into the scope of each method.

1 Introduction

Over the years, many analytical usability evaluation methods (UEMs) have been developed, each with a different theoretical basis, or addressing a particular class of usability problems. For example, TAG (Payne & Green, 1986) focuses on the consistency of syntax/semantics mappings, while FKS (Johnson & Hyde, 2003) focuses on task knowledge structures for collaborative working. Although UEMs have been developed from different theoretical perspectives, studies which have attempted to compare UEMs have tended to rely on usability problem count as the main dependent variable, rather than articulating in detail the different methods' scope and applicability. While some general trends can be deduced, it is difficult to

extract from these studies any firm conclusions about what issues they might identify. It is therefore difficult to assess the extent to which the various methods are complementary, contradictory, or overlapping.

Past studies which have compared Heuristic Evaluation (Nielsen, 1994) with other methods are in general agreement that while heuristic evaluation is good at finding a wide spread of general usability problems (Virzi, Sorce & Herbert, 1993; Cuomo & Bowen, 1994) at comparatively low cost (Jeffries, Miller, Wharton & Uyeda, 1991; Nielsen & Phillips 1993), other methods such as cognitive walkthrough may be necessary in order to focus on specific, task-related problems or re-design issues (Cuomo & Bowen, 1994; Desurvire, 1994; Dutt, Johnson & Johnson, 1994). However, the analyst must decide whether a problem is general, specific or task-related. There is stronger agreement that inspection methods (including heuristic evaluation and cognitive walkthrough) and user testing identify usability issues of different sorts and scope (Bailey, Allan & Raiello, 1992; Karat, 1994; Desurvire, 1994; Karat, 1997), with inspection being more effective in the earlier stages of the development cycle (Jeffries & Desurvire, 1992; Karat, Campbell & Fiegel, 1992; Desurvire, 1994; Karat, 1997). However, the precise nature of this difference is not well understood (Karat, 1997). Moreover, studies which have attempted to compare the predictive potential of heuristic evaluation and cognitive walkthrough (e.g. Desurvire, Kondziela & Atwood, 1992; Sears, 1997; Cuomo & Bowen, 1994) are in little agreement as to the proportion of empirical problems which might be successfully identified by the two methods.

A notable exception to the lack of attention given to method scope is the work of John and Kieras (1996a; 1996b), who present a clear account of what particular usability questions each of four variants of GOMS is suitable for addressing. In the work reported here, we take a similar perspective to that of John and Kieras, namely that one central consideration in selecting a UEM is what *kinds* of insights it will yield.

The motivation for conducting this study was to compare the scope of two novel methods, Evaluating Multimodal Usability (EMU: Hyde, 2002a) and Concept-based Analysis of Surface and Structural Misfits (CASSM: Blandford, Green, Furniss & Makri, forthcoming) with those of more established methods. The design of the study also made it possible to better understand existing approaches, and to reflect on the nature of craft skill in applying UEMs. These are themes to which we return in the Discussion.

1.1 Structure of this paper

The heart of this paper is a systematic review comparing the problem accounts furnished by the different UEMs. First, however, we describe the difficulties of evaluating UEMs, and set the scene by describing the target system to be evaluated by the chosen UEMS. We then report initial analyses, using each of the 8 UEMs; ESDA analysis of video extracts; and the systematic review.

The systematic review serves three functions. First, it records the hits, misses and false positives associated with each UEM (Figure 8). Second, by reorganising that data, it gives a better picture of what *kind* of issue each UEM can be expected to pick up (Figure 9). This will allow future analysts to choose a method suitable for their needs. Third, it allows us to separate out issues that were identified through the analyst's craft skill, rather than through simple and literal application of the UEM. The Discussion section considers the types of issues that we identified, the nature of craft skill, and the unusual but effective methodology employed in the present study.

2 Background: Evaluating UEMs

Evaluation of UEMs can take many different forms and address various questions. Ultimately, what matters is what the costs and benefits of applying any particular UEM are. Costs include the time and effort it takes to learn a UEM and then to apply it to a particular system; benefits include the insights obtained from applying a UEM. Other considerations might include how

well a UEM fits within ongoing design practice and how easy it is for different evaluators to apply the same method consistently.

Gray and Salzman (1998) criticise the earlier literature comparing UEMs against each other on two counts of validity. They specifically criticise the use of problem count as a measure of the effectiveness of a UEM, recommending that researchers limit both their expectations and their claims for UEM studies. Largely as a result of this critique, UEM practice has moved beyond simple head-to-head comparisons employing usability problem count as the sole measure, to consider criteria such as the following.

1. Reliability (also called internal validity) – the extent to which different analyses of the same system, using the same UEM, yield the same insights. Hertzum and Jacobsen (2001) report on studies of the evaluator effect, showing that different evaluators typically identify broadly different sets of problems, whether the method under study is the comparatively loose Heuristic Evaluation (Nielsen, 1994) or the more constrained Cognitive Walkthrough (Wharton, Rieman, Lewis & Polson, 1994) or even think-aloud protocols. Jacobsen, Hertzum and John (1998) focus particularly on how analysts working with the same UEM assessed the severity of problems and again found very little agreement between analysts.
2. External validity – the extent to which the findings from analyses conform to those identified when the system is used in the 'real world'. Cockton, Woolrych, Hall and Hindmarch (2003) report that encouraging analysts to reflect on their judgements when using Heuristic Evaluation can result in greatly improved validity of results, although the paper presents little detail of the empirical results against which the analytical findings are assessed for coming to this conclusion. Gray and Salzman (1998) and Lavery, Cockton and Atkinson (1997) adopt a distinction between what Lavery *et al.* term 'validity' (whether the UEM suggests observed problems or, conversely, 'false positives') and 'effectiveness'. Sears (1997) uses three ratio measures of UEM effectiveness ('thoroughness', 'validity' and 'reliability') to assess the differences between observed problems and predictions.
3. Thoroughness – defined by Sears (1997) and Hartson, Andre and Williges (2001) as the proportion of real problems that are found by a method. This draws on an analogy with information retrieval and the notion of recall (the proportion of documents that are retrieved on a topic compared to the proportion that should have been). This appears to correspond to what Lavery *et al.* (1997) term 'effectiveness'.
4. Effectiveness – defined by Hartson, Andre and Williges (2001) as the product of reliability and thoroughness. They also extend this definition to consider cost effectiveness, the effectiveness per unit cost. One of the practical difficulties with these definitions is that of obtaining accurate numbers to populate the formulae, so while these criteria are intuitively appealing, measuring them in practice is difficult.
5. Productivity – the number of problems a UEM identifies. This measure is probably the most widely discussed; for example, John and Marks (1997) present counts of the number of problems identified by each of six UEMs, each used by a single analyst, when assessing the same interface. Although the authors state clearly that this is a case study, not an experiment, the simple presentation of these figures in a table strongly suggests comparability of the UEMs on this dimension.
6. The practicalities – what is needed to integrate methods within design practice. This is the focus of work by, for example, Karat (1994). Spencer (2000) discusses the compromises that had to be made to integrate Cognitive Walkthrough with a design project.
7. Analyst activities – what analysts do when applying a UEM. To the best of our knowledge, no thorough treatment of this question has yet been conducted, but John and Packer (1995), John and Marks (1997), John and Mashyna (1997) and Jacobsen

and John (2000) present case studies that contribute to the picture of how people work with UEMs, with a particular focus on Cognitive Walkthrough. These studies include a consideration of how methods are effectively learnt. Of particular relevance to the study reported here is the finding of Jacobsen and John (2000) that the participant who had access to multiple descriptions of CW fared better with it than the participant who only had access to one publication on the method – although a comparison of just two individuals is not reliable. In a study of students learning Programmable User Modelling (PUM), Blandford, Buckingham Shum and Young (1998) found that students often had difficulty distinguishing between appropriate and inappropriate representations (e.g. when simplifying their description of a design), and that students appeared to get so focused on producing an appropriate representation that they sometimes lost sight of the fact that the representation was simply a tool to support reasoning.

8. Persuasive power – the ability of an analyst working with the UEM to persuade a developer to change the system as a consequence of problem identification. This was one focus of the John and Marks (1997) study. They went further to consider whether any resulting changes were ultimately beneficial to usability, although as a case study the findings were somewhat inconclusive, serving more to point to directions for further work than to give definitive answers to such complex questions.
9. Downstream utility – how useful the findings from an evaluation study are in informing redesign. This criterion is highlighted by Wixon (2003) and included in a list of criteria by Hartson, Andre and Williges (2001). While this criterion is similar to persuasive power, it implies a different relationship between analyst(s) and designer(s), including a suggestion of how to make design improvements. Hornbæk and Frokjær (2005) focus on this criterion in their study of how usability difficulties can inform system design.
10. Scope – what kinds of problems a method is and is not good for finding. As discussed above, we are aware of only one study which addresses this aspect of UEM effectiveness in detail, namely that of John and Kieras (1996a; 1996b) on the scope of four GOMS variants. Even Gray and Salzman (1998) appear to believe that UEMs should ideally have total coverage of the space of possible problems, stating that they are seeking "evidence that various analytic- and empirical-UEMs do indeed converge upon the same set of usability problems" (p243).

Methodologically, the comparison of UEMs is rife with traps. There are so many variables – from evaluator experience to the systems used in case studies – and so many possible questions that the landscape of possibilities is enormous, and any one study can only hope to map out a very small portion of the territory.

The study reported here circumvents the pitfalls identified by Gray and Salzman (1998) by adopting a clear focus on the *types* of usability problems and issues identified by eight methods, rather than comparing problem counts. The reanalyses are also inspectable (Blandford & Hyde, 2006), so that others can see how the conclusions are derived. However, there are still recognised limitations of the study, as discussed below (in the Discussion).

3 Setting the scene

3.1 Context of the work and methodology

The work reported here was not initially conceived as a single structured study, but evolved into its current form, as shown in Figure 1, over several years. The acronyms included in Figure 1 are all explained subsequently in this paper.

Figure 1. Roadmap of the research reported here

Dates (approx)	Activity	Purpose
1996 – 1998	Select device to study (robotic arm). This was chosen as being a simple device with a multimodal interface. Evaluate robotic arm using GOMS, CW, Z, STN, PUM. Task used was moving the arm to a selected position.	Overall goal: to develop and test novel usability evaluation method that focuses on multimodal issues ('Evaluating Multimodal Usability', or EMU). Objective 1: to gauge the scope of existing formal and semi-formal notations to assess whether there is a niche. Objective 2: to identify desirable properties and issues in learning to apply UEMs, to inform the design of EMU.
1998 - 2000	Robotic arm design was modified in response to evaluations. Develop and test EMU. Testing included applying EMU to the robotic arm. Task used was moving the arm to a selected position.	1) To gauge the learnability of EMU (not reported here). 2) To check whether EMU does indeed fill a niche (multimodal interaction).
2000	Robotic arm was destroyed in a flood, and development ceased. Complete Exploratory Sequential Data Analysis of limited video data of the arm in use. Task used was feeding. Complete preliminary systematic review of 6 methods.	To compare all analytical data against empirical data of the arm in use.
1996 – 2004	Develop and test CASSM.	This development was independent of the EMU development, and was intended to deliver a UEM that is less formal, and that focuses on conceptual misfits.
2004	Apply CASSM to the robotic arm.	To gather evidence on whether CASSM does fill the intended niche (conceptual misfits).
2004	Complete systematic review of all seven methods and compare against empirical data.	To develop understanding of scope of all methods in the context of the use of the robotic arm.
2006	Apply Heuristic Evaluation to the arm. Extend systematic review to 8 methods. Adapt it to the task represented in the video data.	To include a comparison with Heuristic Evaluation in the study.

The initial aim of the work was to develop and test a rigorous, analytical approach to usability evaluation that extended existing approaches to address multimodal usability issues such as modality clashes – e.g. a user being expected to read text while speaking different text. This approach was called Evaluating Multimodal Usability (EMU: Hyde, 2002a). Part of the preparation for this involved reviewing existing analytical evaluation methods that might form a basis for the new method.

Five methods were selected as a starting point; they were all formal or semi-formal, with a theoretical and/or representational basis. Some (most notably Z) focus primarily on the use of a notation to describe a system clearly; others (most notably Cognitive Walkthrough) focus

primarily on method, with a relatively informal description language. The methods were chosen as representing a range of formality, and having different base-line assumptions about users; for example, GOMS assumes experts, while Cognitive Walkthrough assumes novices learning through exploration. There were two system-oriented description methods (Z (Spivey, 1989) and STN (Dix *et al.*, 1993)), two established user-oriented methods (GOMS (Card, Moran & Newell, 1983) and Cognitive Walkthrough (Wharton *et al.*, 1994)) and one user-oriented method that had been developed locally (PUM: Young, Green and Simon, 1989; Blandford and Young, 1996). Z and STN are not standard usability modelling methods, being generally used in software engineering to describe the specification and functionality of a system. They were included to see what leverage well-known methods with no explicit usability analysis support could give to the understanding of the interface, against which other usability-specific methods could be compared.

These approaches represent some of the more formal modelling methods, but are not intended to be definitive. Indeed, other approaches, for instance Petri Nets (e.g. Bastide & Palanque, 1990), UAN (Hartson, Siochi & Hix, 1992) or Task-Action Grammar (e.g. Payne & Green, 1986), would have been equally applicable. Other methods such as syndetics (Duke *et al.*, 1998) and ICS Cognitive Task Analysis (Barnard & May, 1999) were not used because there is little published guidance on their application to interface analysis.

The five selected methods were all applied to the interface for a robotic arm, as described below. Following the STN analysis, feedback was given to the arm developer, who then implemented backtracking and consolidated 'continue' and 'go' into one operation (these were labelled usability issues 2 and 5 respectively — see Appendix A); for consistency, in the systematic review of all methods, all were assessed to establish whether they would have identified these issues or not.

The results of the work on a modality taxonomy and the experience of applying the five analysis methods formed the basis for the design of EMU. EMU was itself then subjected to evaluation, by teaching it to novice users and by applying it to the same interface as the five earlier methods.

As further validation of EMU, we compared the findings of all six UEMs (EMU plus the five applied earlier) to empirical data of the robotic arm in use. Unfortunately, during the development of EMU, the robotic arm system was destroyed in a flood and development was abandoned, so for this we had to rely on video data of the prototype system in use that had been collected before the flood.

By this point, it was very clear that some of the usability findings identified using each method could be attributed to the method, but that others were fortuitous, due to the general craft skill of the analysts or our growing understanding of the interface. Therefore, a systematic review was conducted to identify which insights could be attributed to the method, which to craft skill, etc., and also to identify which usability difficulties *should* have been identified using each method but were not. This was based on our judgement of how directly apparent the issue was from the representation.

Shortly after the completion of this study, we were developing a further evaluation method, CASSM (Concept-based Analysis of Surface and Structural Misfits, formerly known as OSM: Connell, Green & Blandford, 2003; 2004), and again the question of scoping arose. Therefore, a further analysis of the same robotic arm, based on the description presented below, was conducted. All the earlier data and analyses were revisited and expanded to include the new insights derived from the CASSM analysis.

Finally, in response to recommendations from referees of an earlier version of this paper, a Heuristic Evaluation (HE: Nielsen, 1994) of the arm was conducted, drawing on all the available information about the arm. The systematic review was redone to include the findings from HE, and also to frame it around the task featured in the video data. Although the final two analyses were conducted retrospectively, every effort was made to apply the same degree of rigour to them as to earlier ones.

3.2 Method

As discussed above, the work reported here was not originally conceived as a single, structured study. However, it can be understood as such. As a single study, the key steps of analysis were as follows:

1. Analysis of the robotic arm using the eight analytical evaluation methods introduced above.
2. Exploratory Sequential Data Analysis (ESDA: Sanderson & Fisher, 1994) of short video extracts of an individual using the robotic arm. This analysis focused on usability issues.
3. Systematic review of all eight analyses of the arm, taking the full list of usability issues compiled during steps (1) and (2) and constructing a careful account of why each method did, should have, should not have or did not identify each issue.

Before we describe each of these steps, we introduce the case study.

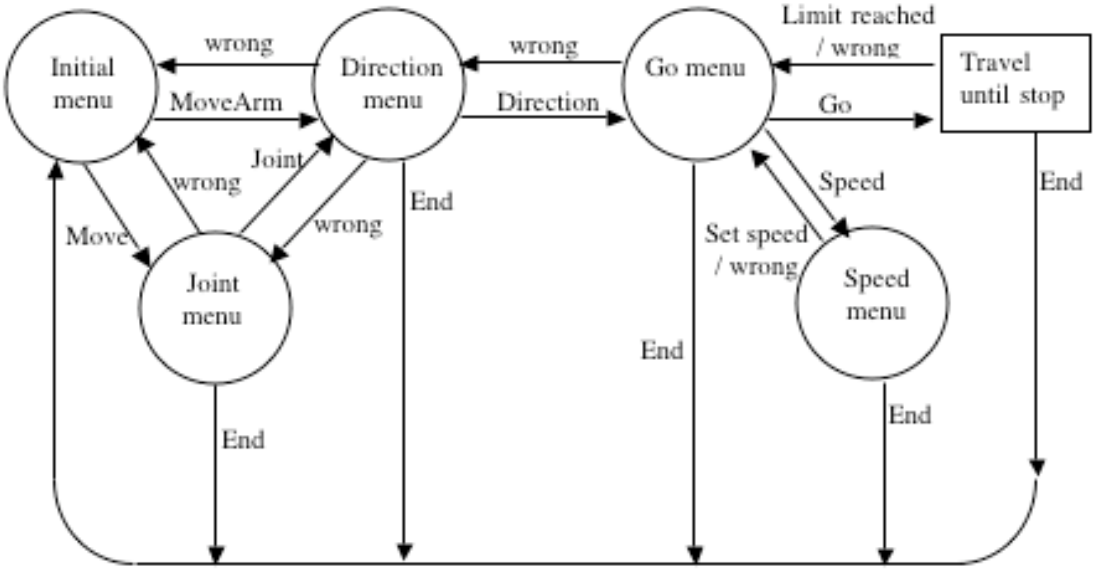
3.3 Case study: the robotic arm

The system chosen for analysis was a robotic manipulator for use by wheelchair-bound people (Parsons *et al.*, 1995; 1997). This was chosen because the interface was multimodal (and thus likely to test EMU well), the system was relatively simple (so that applying several evaluation methods was a tractable proposition) and the system was still under development (so that the analyses could actually inform design). The manipulator was intended to be used in a domestic context for everyday tasks such as feeding and grooming, and was developed primarily to prove that a sophisticated manipulator could be produced at a reasonable cost: usability issues were considered informally, if at all. The arm consisted of eight joints, powered by motors, which could move either individual joints or the whole arm at once, via the input devices. The user could either move joints explicitly (selecting the joint and direction of movement) or make use of pre-taught positions that were programmed in; in this study we focus on explicit movement.

The input devices interfaced to a Windows-based application which in turn sent motor control commands to a dedicated microprocessor that controlled the movement of the arm. The interface was based on menu selection. Three different devices could be used: a standard mouse; voice recognition; and a gesture-based interface. The voice recognition system allowed direct menu option selection simply by saying the menu option out loud. It was designed to be trained to individual voices. The gesture input system was based on a baseball cap with two sensors: one detecting movement forwards and backwards, the other detecting movement left and right. This allowed a variety of distinct gestures to form the gesture vocabulary. The gesture system was implemented so that a cursor moved along underneath the menu options cyclically, and an option was selected by making the correct gesture when the cursor was underneath that option.

For the purpose of analysis, only one task was considered, which exercised only part of the interface. However, the task is one that would be very common to all users, and would therefore give valuable information on the usability of the interface – namely, to move the robotic arm to a certain position without making use of any pre-taught positions. It is this kind of task that the developers of the arm consider to be a basic task, and that should be part of the core functionality of the interface. Figure 2 illustrates the possible states and transitions in the interaction between a user and the system interface while completing such a task.

Figure 2. The second STN diagram produced (including error correction)



3.4 Initial analyses

Each initial analysis was conducted by one of the first two authors and checked by at least one other author of this paper. At this stage, while every effort was made to conduct each analysis independent of all other analyses, there were inevitably learning and transfer effects. For some methods, it was necessary to invest substantial time on learning, by reading as many source documents as possible; for others, learning was negligible. There were also unavoidably effects due to the degree of familiarity with the device (familiarity grew throughout the study). These confounds are discussed at length by Gray and Salzman (1998); the systematic review aimed to take them into account explicitly.

An 'issue number' was used to index every usability issue identified. Appendix A presents a definition of each issue. There is no significance to the ordering of issues.

The eight analytical evaluation methods applied in this study are summarised in Figure 3. STN and Z are essentially device descriptions, whereas other approaches explicitly consider the user.

Figure 3. Overview of methods applied in this analysis

Method	Primary source of description	Developed locally?	Key features
State Transition Networks (STN)	Dix *et al.* (1993)	No	Diagrammatic
Z	Spivey (1989)	No	Formal notation based on set theory and first-order predicate logic
Heuristic Evaluation (HE)	Nielsen (1994)	No	Comparatively informal, based on heuristics.
Cognitive Walkthrough (CW)	Wharton *et al.* (1994)	No	Clearly defined method; natural language; goal-based
GOMS	John & Kieras (1996a)	No	Highly structured; hierarchical; goal-based
Programmable User Modelling (PUM)	Blandford, Good & Young (1998)	Yes	Highly structured; based on means-ends planning
EMU	Hyde (2002b)	Yes	Clearly defined method, focusing on multimodal issues
CASSM	Blandford, Connell & Green (2003)	Yes	Semi-formal, focusing on conceptual misfits between user and device

Here we very briefly summarise the six methods that are widely described in the literature, followed by more extensive descriptions of EMU and CASSM, which were the two approaches that motivated this study and are less well known.

STNs are a way of diagrammatically representing an interaction (Dix *et al.*, 1993) and can take various forms. For simple interaction sequences, STNs can clearly illustrate the flow of interaction and allow redundant cycles to be identified. The simplest type, as used here, has each state of the system represented by a circle, linked by lines, or transitions, which correspond to the actions necessary to move from that state to another. Figure 2 shows an STN diagram for the latest prototype of the arm controller.

While also being system-oriented, Z (Spivey, 1989) contrasts with STN in being a formal specification notation based on set theory and first order predicate logic. It makes use of schemas, which are collections of named objects with relationships specified by axioms. These schemas can be built up to define large specifications. Z focuses on the structures and relationships that are of importance, and allows the analyst to manipulate those relationships and examine the implications of change.

In contrast to the first two methods, Heuristic Evaluation (Nielsen, 1994) was developed specifically to support usability evaluation. It is also the least formal of the approaches considered here. Nielsen (1994, p.25) describes the motivation for developing HE as being to reduce the "intimidation barrier" to analysis. Completing a HE involves working through a checklist of ten questions and assessing the system against those questions. Optionally, severity ratings can be assigned to the problems identified. Nielsen advocates that 3-5 evaluators should be involved to achieve an appropriate cost-benefit trade-off in finding problems, reflecting the low inter-rater reliability of HE. He also notes that other heuristics can be added to the set as appropriate. In this study, HE was applied by one evaluator, since seven other approaches were also being used to generate usability issues and the focus was on the subsequent systematic review. Also, no severity ratings were proposed, since the study was focusing on what issues each approach identified rather than the craft skill of assigning severity ratings.

Compared to HE, Cognitive Walkthrough (Wharton *et al.*, 1994) is relatively structured. It is designed to uncover usability issues by following the sequence of actions a user would take to perform a set of tasks agreed by the analysts, and by analysing at each stage how successful the user would be in performing the action correctly. The method takes a task-oriented perspective, in that it considers the goal structure and the ways goals are addressed in completing the task. At every stage the interface is evaluated by answering set questions to determine whether or not it provides the necessary information for the user to successfully continue with the task, and what feedback the interface provides to the user. The analysis of user actions is done in terms of success and failure stories. Cognitive Walkthroughs concentrate on ease of learning; this perspective is justified by the fact that users tend to learn features of an interface as they need to, rather than all at once. Therefore, ease of learning is seen as essential to interface usability.

GOMS (Card *et al.*, 1983) is also a cognitively based method, but more formal than CW. It is based on the idea of the human as an information processor. GOMS stands for Goals, Operators, Methods, and Selection rules, and is based on the premise that a user's behaviour can be viewed as achieving goals by breaking them down into sub-goals which can then be separately achieved. The Operators are the ways available to accomplish the goals, Methods are defined sequences of operators and goals, and Selection rules determine how to choose between more than one method (John and Kieras, 1996a). The emphasis is not just on the physical aspects of interaction, but also on mental processes — for example, what the user has to know or remember. Varieties of GOMS address goal hierarchies, working memory load, schedule tasks, lists of operators, and production systems. The interface to the robotic arm was first analysed using CMN-GOMS (John and Kieras, 1996a). This version of GOMS was chosen as being comparatively easy to learn. It has a strict goal hierarchy, with each method represented as a series of steps that are performed in sequence. A further analysis was

conducted using CPM-GOMS (John and Kieras, 1996a) to examine more fully the cognitive, motor and perceptual aspects of the interaction.

PUM, a locally developed approach, was included in the study because features of PUM have informed the design of both EMU and CASSM. Like CW and GOMS, it has a cognitive basis. It focuses on user knowledge, and how that knowledge is used in the interaction to effect changes to the system state. A description of the knowledge that the user needs to operate the interface successfully is written in an Instruction Language (IL), which is then optionally compiled by a cognitive model to simulate predicted user behaviour (Blandford, Buckingham Shum and Young, 1998). Potential user difficulties can be identified both in the ease (or otherwise) with which the analyst can specify the required user knowledge in the IL and in observing the behaviour of the running model (if analysis is taken that far, which is was not in this case).

EMU was specifically developed to build on the strengths of existing methods while focusing particularly on multi-modal usability issues. To develop EMU, various existing approaches to assessing the usability of multi-modal and multi-media systems (Bernsen, 1995; Coutaz, Nigay & Salber, 1993; Dowell, Life & Salter, 1994; Coutaz, May, Young, Blandford, Nigay & Salber, 1995; Purchase, 1999) were investigated to derive critical properties of relevance to multi-modal usability. In addition, theories accounting for the mental and physical capabilities of users were invoked – notably the Human Information Processing Model (HIPM: Wickens, Sandry & Vidulich, 1983); the Model Human Processor (MHP: Card *et al.*, 1983); Interacting Cognitive Subsystems (ICS: Barnard and Teasdale, 1991); and the Executive Process-Interactive Control (EPIC: Kieras and Meyer, 1995); one particular concern was to identify cognitive restrictions applicable to multi-modal usability. From these theories, a focused definition of a modality was proposed: that a modality is a "temporally based instance of information perceived by a particular sensory channel".

A method to support reasoning about multimodal interaction, including a consideration of temporal issues and possible mismatches of modalities as well as processing constraints, was developed, drawing on experience of developing task-oriented descriptions using GOMS and CW. This method involves examining the interaction stage by stage, concentrating on the flow of modalities, and the conflicts and clashes between them. The task is defined, and the modalities are listed. The user, system and environment are profiled and compared to the modality listings in order to find any potential problems. The interaction sequence listing is completed using a notation that describes every step of the interaction in terms of the modalities expressed and received by user and system, and is examined for modality properties and clashes. An illustrative extract from the EMU analysis of the robotic arm is shown in Figure 4; this extract shows three (simultaneous) system output modalities and alternative user receptive modalities, depending on where the user is focusing their attention at the time.

Figure 4. Short extract from the EMU analysis describing the user looking at the display or the arm

[SE vis-lex-cont] and **menu display**	[SE vis-sym-dyn] and **moving cursor**	[SE hap-con-cont] **position of arm**
[UR vis-lex-cont] and **menu display** precon: [SE vis-lex-cont] *menu display* precon: looking at display	[UR vis-sym-dyn] **moving cursor** precon: [SE vis-sym-dyn] *moving cursor* precon: looking at display or [UR vis-con-cont] **position of arm** precon: [SE hap-con-cont] **position of arm** precon: looking at arm	

The system is displaying a menu, underneath which is a flashing cursor moving from one option in turn to another. The arm not moving and is at rest. The user is either looking at the menu with the cursor, or at the arm. A tutorial (Hyde, 2002b) on how to apply EMU was developed and tested by teaching the approach to a cohort of HCI students. The tutorial was used as the reference material for the systematic review reported here, and is available for download.

Finally, Concept-based Analysis of Surface and Structural Misfits (CASSM: Blandford *et al.*, 2003) focuses on structures rather than tasks or procedures. It draws on ideas of 'fit' that have been described previously by, for example, Moran (1983), Payne, Squibb and Howes (1990) and Norman (1986), and presents a methodology for reasoning about those ideas. Again, it has been developed to fill a perceived gap in the repertoire of usability evaluation methods. CASSM is designed to be applied in an iteratively deepening way, so that initial analysis can be quite sketchy, with thoroughness achieved through successive iterations (stopping as soon as the analyst judges that additional benefits are unlikely to be merited by the additional costs of going further). The analyst identifies the main concepts that the user works with, those represented at the interface, and those in the underlying system, and reasons about the quality of fit between the user, interface and system concepts. The full CASSM analysis of the robotic arm is reported by Blandford *et al.* (forthcoming) and an illustrative extract is shown in Figure 5. The systematic review of CASSM is based on the tutorial (Blandford *et al.*, 2003), which is available for download.

Figure 5. Short extract from the CASSM analysis showing the user concept of an 'object in the world' and its attributes. Such an object is meaningful for the user, but is not represented at ('absent' from) the interface and underlying system. According to this analysis, the user cannot create or delete objects, but can change their attributes indirectly (i.e. by moving the robotic arm when it is touching or holding the object)

	Name	User	Interface	System	Set / Create	Change / Delete	Notes
E	object in world	present	absent	absent	fixed	fixed	e.g. light switch, cup
A	position	present	absent	absent	fixed	indirect	any particular object may only have some attributes
A	configuration	present	absent	absent	fixed	indirect	may include orientation
A	speed	present	absent	absent	fixed	indirect	

The output from this first phase of analysis was a list of usability issues identified by each of the eight methods, including some duplicates where the same issue was identified by multiple methods. These usability issues are defined in Appendix A.

3.5 ESDA analysis of video extracts

A form of Exploratory Sequential Data Analysis (ESDA: Sanderson & Fisher, 1994) was used to analyse the only empirical data that was available for the robotic arm – namely 6 short episodes of use of the system by two individuals. ESDA techniques are observational and empirical, and include task analysis, protocol analysis and video analysis. In this particular case, the form of ESDA used was based on analysing the available video evidence in terms of where the user was looking, when the user made a selection (using whichever input device was featured in the episode), whether the arm was moving or not, whether or not there was any audible noise from the arm, and anything the user said. From this data, any perceptible user difficulties were identified.

The video data comprised six excerpts, each one showing a user performing a specific task and using a particular means of input, as shown in Figure 6. As indicated, most video excerpts used pre-taught positions, in which the user did not have to explicitly select and move individual arm joints. This makes them relatively uninteresting from a usability perspective, and means that excerpt 6, which involved a novice user (an individual with the kinds of movement difficulties that the device was intended to support) manipulating the arm at least partly manually, provided most data for this study.

Figure 6. Summary table of video data excerpts

EXCERPT:	One	Two	Three	Four	Five	Six
SECONDS:	123	83	89	50	89	525
INPUT:	Mouse	Voice	Gesture	Gesture	Voice	Mouse
TASK:	Feeding	Feeding	Feeding	Feeding	Feeding	Drinking
USER:	Expert	Expert	Expert	Novice	Novice	Novice
POSITIONS:	Pre-taught	Pre-taught	Pre-taught	Pre-taught	Pre-taught	Mixture

Since the robotic arm was no longer available, so that we could not tailor trials to closely match the rest of the study, ESDA was one of few approaches that could be used. However, as shown in Figure 6, all excerpts involved feeding or drinking tasks, rather than the task of moving the gripper to a particular position that was used for the analytical evaluations; this difference was taken into account in performing the systematic review of the methods (feeding and drinking involve a sequence of moves to particular positions, with additional opening and closing of the gripper, so are more complex than the analysed task).

Video evidence was found to corroborate thirteen of the usability issues identified, although in some cases the same behavioural phenomenon can be attributed to alternative usability problems, and it is not possible to disambiguate the attribution.

1. The video data shows four instances where all or part of the arm started to move in one direction, only for it to be stopped and moved in the opposite direction. This provides evidence that issues 12 ('problems of determining left and right, especially when arm contorted') and 13 ('user cannot check direction choice until arm starts to move') are real problems.
2. Video evidence shows various under- and over-shoots where the user had to subsequently correct the position of the arm, indicating that an error had occurred. This is indicative of user difficulties in judging arm movements and position (issues 14, 17, 23, 24, 25). On one occasion in excerpt 6, the gripper was poorly oriented for the task, and the user had difficulty seeing it (issue 25).
3. One of the users was heard to comment in excerpt 4: "I think it's on slow, innit?", indicating lack of display information about the current speed setting (issue 30).
4. The impoverished nature of the available video data means that there are issues for which there is inadequate or no video evidence. For example, there was no instance where the user paused in the middle of saying "move arm" (issue 9). There are also a few issues for which it is, in principle, not possible to have video evidence. For example, the redundancy of 'continue' (issue 5) would not appear in video data.

One additional usability issue was uncovered in the video data which is outside the scope of all the analytical evaluation approaches: it was found that the arm itself obscured the user's view at times. Twice in excerpt 6, the user had to move his head substantially to see around the arm.

All the usability issues are summarised in Figure 8 (see Results). The video evidence is classified as 'yes' (pretty clear video evidence of issue), 'poor' (some evidence, but not good), 'none' (no video evidence) or 'n/a' (not applicable: video evidence would not make this issue apparent). One of the surprising aspects of the very limited video data is the number of issues it highlighted (helped by the fact that the analyst was already aware of many of the possible difficulties of using the arm).

3.6 Systematic review

The analysis so far gave no insight into whether the issues identified analytically were identified according to the actual claims of the methods used or through the skill and insight of the analyst. Each of the analyses was therefore systematically re-examined using a single source of description for each particular method (as shown in Figure 1), asking the questions: should this method have supported the identification of this issue, and why (or why not)? This systematic review enabled us to consider whether the usability issues were identified due to the power of the method, the skill and knowledge of the analyst, or other factors. This, in turn, enabled us to assess the scope of each approach, at least in relation to the device and task used for this case study. Here, we have instantiated the idea of craft knowledge slightly differently from Long and Dowell (1990), defining craft skill as the analyst using their experiential knowledge *in conjunction with a method* to achieve insights that are informed by the method or notation being used, but not directly derivable from it.

It is important to understand the nature and status of the systematic review, which is central to this work: the systematic review involved creating a matrix of all the usability issues identified by any approach (analytical or empirical) and all the methods applied. For every cell of that matrix an account was generated of whether that issue should have been identified by that method and why (or why not).

Figure 7 shows the possible assessments made in the systematic review. In this Figure, 'A' issues straddle an ambiguous line between method and craft: had the problem been described differently, these issues would emerge through the method, but selecting the appropriate level of abstraction for the representation is itself a matter of craft skill. The nature of craft skill is discussed in more detail below.

Figure 7. Was an issue identified through the method or craft skill, or could it have been?

	Was identified	Was not identified
Should have been identified	M: Identified by method	O: Overlooked but should have been identified by method
Could have been found had the problem been described at a different level of abstraction	*[not applicable]*	A: Depends on abstraction level
Could have been identified (through craft skill)	C: Identified through craft skill of analyst	C?: Representation indirectly supports identification, but method does not explicitly
Should not have been identified	*[did not occur]*	[unlabelled]: outside scope of method and representation

Examples of extracts from the systematic review follow, exemplifying the different cells in Figure 7. The numbers are the indexes used to label each issue.

M: Our first example is taken from the review of Z, and illustrates an issue that is within the scope of the approach and was identified in the initial analysis:

> 5. Continue versus Go: Continue seen as redundant
>
> This issue was identified by the Z specification since both options share the same functionality, and were represented by different schemas with identical contents.

O: This example from the review of STN shows an issue that should have been identified but was not:

> 1. Long sequence of operators to move arm
>
> Since the STN shows the number of states that the user has to navigate through before the robotic arm can be moved, this issue should have been identified in the original analysis. That it was not identified shows the extent to which the analysis was dependent on the skill (or lack thereof) of the analyst.

A: An example from the review of PUM illustrates an issue that would have emerged had the problem been described in more detail:

> 7. Gesture input with twice as many operations as voice because dependent on cursor movement
>
> This did not come out in the original PUM analysis, because the analysis was not written at a low enough level of abstraction for this to be apparent: it was written in terms of conceptual operations rather than individual actions.

C: This example from the review of CASSM illustrates an issue that was identified due to the craft skill of the analyst, rather than directly from the approach:

> 12. Problems of determining left and right, especially when arm contorted
>
> The issue of judging directions when the arm is contorted emerged (with some craft skill) from looking at joints and what the user knows about the directions in which joints can move. It does not emerge directly from the CASSM representation.

C?: The fifth example is taken from the review of Cognitive Walkthrough, illustrating an issue that is outside the scope of the UEM but might be found by craft skill (C?):

> 2. Inability to backtrack
>
> CW does not deal with error in terms of its implications, therefore would not find this issue, although it might come out from the craft skill of the analyst through thinking about rectifying errors.

[unlabelled]: The final example is taken from the review of Heuristic Evaluation, illustrating an issue that was outside the scope of the method:

> 31. Arm obscuring user's view
>
> HE doesn't consider the context and details of system behaviour in this way, so this issue is outside scope.

The complete systematic review is presented by Blandford and Hyde (2006), and edited highlights are included as Appendix B of this paper.

4 Results

A full list of the issues identified by *any* method (including the video data) is presented as Figure 8, using the codes as summarised in Figure 7. The data is organised to highlight hits, misses and false positives (FPs) in the analyses. Here a Hit is a usability issue that was identified by using a UEM that was corroborated (in some cases weakly) by the video evidence; a Miss is a usability issue that emerges in the video data but was not identified by a particular method; and a False Positive is an issue that was predicted through analysis, but for which there is no supporting video evidence. This data must be viewed with caution (particularly the false positives) due to the limitations of the empirical data available. However, of particular concern is the number of issues that emerged in the video data (issues 4, 12 13, 29, 31, 32, 34) that were not found through *any* of the methods (although some were identified through the craft skill of the analyst); these issues are highlighted in Figure 8. These misses are discussed in more detail below.

Figure 9 shows an abstraction of the same data, focusing on issues that were or should have been identified through the method. In this Figure, the data has been restructured to visually highlight commonalities across methods by clustering. Again, these findings are discussed in more detail below.

Finally, Figure 10 focuses on craft skill. This Figure shows the issues that were or could have been readily identified through the craft skill of the analyst when applying each UEM. In contrast to the issues that could be found by the methods, there is no obvious pattern in the issues that might plausibly emerge through craft skill.

Figure 8. Hits, misses and false positives: data arranged to focus on video evidence. Shaded rows highlight issues from video data that were not reliably identified by any method.

	PROBLEM	GOMS	STN	Z	HE	CASSM	PUM	EMU	CW	video
17	Mismatch between way that arm works and way that user would move arm			C?		M	M	C	C?	yes
23	Difficulty of judging arm movements					M		M	C?	yes
24	Difficulty in judging speed and direction as getting close to target					M	A			yes
25	Difficulty in judging position, orientation and aperture of gripper as approaching target					M				yes
28	Voice recognition problems				M	M				yes
22	User looking one way, menu options in other direction					C		M	C?	yes
14	Time taken to interact with system to stop arm	M					C	C	C	yes
30	No display of speed				O	C?	M			yes
29	Speaking with mouth full…				C?			O		yes
12	Problems of determining left and right, especially when arm contorted					C?	C?		C	yes
13	User cannot check direction choice until arm starts to move				C?				C	yes
31	Arm obscuring user's view									yes
32	No arm reversing		C?		C?		C			yes
4	Lack of short cuts	C	C?	C?	C?		C?		C?	poor
34	Long time to recover from directional error				C					poor
1	Long sequence of operators to move arm	M	O	C?			C	C	*C*	poor
7	Gesture input with twice as many operations as voice	M	A	A			A			poor
3	Difficulty of choosing between Move Arm or Move	C		C?	C		M	M	M	none
6	Confusion over joint called Arm	C		C?	C?	O	M	M	M	none
11	Lack of feedback about selection				M		C?		M	none
18	Not clear that End returns user to main menu				O		M	O	M	none
19	End having two meanings						M	O	M	none
15	Similarity between moving joint and moving whole arm	M	M	M			C?		C?	none
16	Illegal options		C?	M					C?	none
26	Position and movement of most joints is of limited interest to the user					M			C?	none
9	If user pauses in middle of saying "Move arm"...				C?				C	none
10	If user engaged in conversation...				C?	C?		C?	C	none
27	Possible difficulty of timing gesture accurately as cursor moves between options	A			M	M			C?	none
2	Inability to backtrack	C	M	M	M		C		C?	n/a
5	Continue redundant	M	M	M	M		O		C?	n/a
8	Head moved to look at arm while gesture system operational may be interpreted as a command				C?	C?		M	C	n/a
20	Lighting conditions							M	C?	n/a
21	Difficulty for user to move field of vision							M	C?	n/a
33	Difficult to match names to joints				M	O	M		O	n/a

Hits/ Misses

FPs

Figure 9. Focus on methods: what *should* have been found by each method ('M's, 'O's and 'A's), and the types of issue associated with each UEM: The second column indicates issue type: S=System; M=user Misconceptions; C=Conceptual misfit; P=Physical misfit; X=contextual issue (see Discussion section).

		PROBLEM	GOMS	STN	Z	HE	CASSM	PUM	EMU	CW	video
1	S	Long sequence of operators to move arm	M	O							poor
5	S	Continue redundant	M	M	M	M		O			n/a
7	S	Gesture input with twice as many operations as voice	M	A	A			A			poor
15	S	Similarity between moving joint and moving whole arm	M	M	M						none
2	S	Inability to backtrack		M	M	M					n/a
16	S	Illegal options			M						none
14	P	Time taken to interact with system to stop arm	M								yes
27	P	Possible difficulty of timing gesture accurately as cursor moves between options	A			M	M				none
28	P	Voice recognition problems				M	M				yes
3	M	Difficulty of choosing between Move Arm and Move						M	M	M	none
6	M	Confusion over joint called Arm					O	M	M	M	none
18	M	Not clear that End returns user to main menu				O		M	O	M	none
19	M	End having two meanings						M	O	M	none
33	M	Difficult to match names to joints				M	O	M		O	n/a
30	M	No display of speed				O		M			yes
11	M	Lack of feedback about selection				M				M	none
20	X	Lighting conditions							M		n/a
21	X	Difficulty for user to move field of vision							M		n/a
8	P	Head moved to look at arm while gesture system operational may be interpreted as a command							M		n/a
22	P	User looking one way, menu options in other direction							M		yes
29	P	Speaking with mouth full…							O		yes
23	C	Difficulty of judging arm movements					M		M		yes
17	C	Mismatch between way that arm works and way that user would move arm					M	M			yes
24	C	Difficulty in judging speed and direction as getting close to target					M	A			yes
25	C	Difficulty in judging position, orientation and aperture of gripper as approaching target					M				yes
26	C	Position and movement of most joints is of limited interest to the user					M				none
9	P	If user pauses in middle of saying "Move arm"...									none
10	P	If user engaged in conversation...									none
13	M	User cannot check direction choice until arm starts to move									yes
12	X	Problems of determining left and right, especially when arm contorted									yes
31	X	Arm obscuring user's view									yes
32	S	No arm reversing									yes
4	S	Lack of short cuts									poor
34	S	Long time to recover from directional error									poor

Figure 10. Issues that either were (C) or could have been (C?, A) identified through craft skill.

	PROBLEM	GOMS	STN	Z	HE	*CASSM*	PUM	EMU	CW	video
1	Long sequence of operators to move arm			C?			C	C	*C*	poor
5	Continue redundant								C?	n/a
7	Gesture input with twice as many operations as voice		A	A			A			poor
15	Similarity between moving joint and moving whole arm						C?		C?	none
2	Inability to backtrack	C					C		C?	n/a
16	Illegal options		C?						C?	none
14	Time taken to interact with system to stop arm						C	C	C	yes
27	Possible difficulty of timing gesture accurately as cursor moves between options	A							C?	none
28	Voice recognition problems									yes
3	Difficulty of choosing between Move Arm or Move	C		C?	C					none
6	Confusion over joint called Arm	C		C?	C?					none
18	Not clear that End returns user to main menu									none
19	End having two meanings									none
33	Difficult to match names to joints					C?				n/a
30	No display of speed					C?				yes
11	Lack of feedback about selection						C?			none
20	Lighting conditions								C?	n/a
21	Difficulty for user to move field of vision								C?	n/a
8	Head moved to look at arm while gesture system operational may be interpreted as a command				C?	C?			C	n/a
22	User looking one way, menu options in other direction					C			C?	yes
29	Speaking with mouth full…				C?					yes
23	Difficulty of judging arm movements								C?	yes
17	Mismatch between way that arm works and way that user would move arm			C?				C	C?	yes
24	Difficulty in judging speed and direction as getting close to target						A			yes
25	Difficulty in judging position, orientation and aperture of gripper as approaching target									yes
26	Position and movement of most joints is of limited interest to the user								C?	none
9	If user pauses in middle of saying "Move arm"...				C?				C	none
10	If user engaged in conversation...				C?	C?		C?	C	none
13	User cannot check direction choice until arm starts to move				C?				C	yes
12	Problems of determining left and right, especially when arm contorted					C?	C?		C	yes
31	Arm obscuring user's view									yes
32	No arm reversing		C?		C?		C			yes
4	Lack of short cuts	C	C?	C?	C?		C?		C?	poor
34	Long time to recover from directional error				C					poor

5 Discussion

The case study was selected as being particularly suitable to the application of EMU, being a multimodal device with a simple task structure, but all eight methods highlighted important usability issues about the device.

Z and STN, although not designed to identify usability problems, were reasonably effective at supporting the identification of system-related problems such as the lack of an 'undo' facility, redundant operators, and long action sequences.

GOMS supported the identification of many of the same issues as Z and STN, plus some concerning the synchronisation of user actions with system behaviour. Apart from timing information, our GOMS analysis addresses all the kinds of issues outlined as being within scope by John and Kieras (1996b). Under the circumstances in which this study was conducted, it was not possible to include the detailed timing data that would enrich the GOMS analysis so that it should deliver more than the strictly device-centred Z and STN analyses. Lindegaard (2003) presents an argument that GOMS timing data is often irrelevant, since the aspects of interaction for which timings can be done are not the most significant in terms of total interaction times. A similar point is made by John and Kieras (1996b, p.299). Conversely, studies such as that of Gray, John and Atwood (1993) have shown the value of timing analyses in certain situations. It is not possible to be sure whether timing data would be informative in this case (where many of the timings are dictated by the device rather than mental processes and user actions). Nevertheless, it was a surprise to us that GOMS, as a cognitively based method, would have so much in common with system-oriented approaches and so little with other user-centred ones. The main explanation for this is likely to be that GOMS assumes users are experts, and therefore does not consider possible user misconceptions, focusing rather on user actions, which map directly onto device actions.

HE identified a range of issues, as defined by the particular set of ten heuristics applied (Nielsen, 1994). HE was (subjectively) the most difficult method for which to conduct the systematic review because the account of whether an issue was identified through the method or by craft skill was difficult to resolve: it depended very much on how the wording of the heuristic and its supporting text were interpreted. For example, the heuristic "match between the system and the real world" could be interpreted as covering issues such as "difficulty of judging arm movements". However, the supporting text is: "The system should speak the user's language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order." This gives a much narrower interpretation of the heuristic. These narrower interpretations were used in the systematic review. One subjective finding about HE was that it encouraged a more explicit consideration of the causes and consequences of errors (including, but not limited to, user misconceptions) than any of the other approaches tested.

Like HE, Cognitive Walkthrough supported the identification of additional issues through craft skill well, compared to other approaches. The issues identified through the method could all be classified as relating to possible user misconceptions when interacting with the system (consistent with the theory underpinning CW as concerning learning through exploration). We are unsure why CW seemed to encourage the identification of more issues than are strictly within its scope, when compared to the other approaches tested, but surmise that this may be because it is a comparatively discursive and unconstrained approach – a property it shares with HE – and also because it encourages empathy with the user, since the analyst is actually 'walking through' the task.

PUM supported the identification of similar issues to CW. Subjectively, this was at greater cost of analysis. However, it also supported the identification of some issues that emerged naturally by applying CASSM. This is not surprising, since the development of CASSM was informed by earlier experience of working with PUM.

EMU also covered some of the same territory as CW and PUM – again, not surprising, since the development of EMU drew extensively on these task-oriented approaches. In addition, it supported the identification of various issues relating to the modalities of interaction – e.g. that disabled users might have difficulty shifting their attention between the arm and the display in a timely way. These were the kinds of issues that EMU was designed to identify.

Finally, CASSM also covered some of the same territory as CW, PUM and EMU, but also raised issues relating to the conceptual fit between how to operate the arm controller and what the user would want to do with the arm 'in the world' (e.g. concerning how easily the user could judge arm movements). These are the kinds of issues that CASSM is designed to identify, but this device was not selected as being particularly well suited to a CASSM analysis, so it is encouraging to find evidence that CASSM fills the intended niche.

The systematic reanalysis made it possible for some important qualitative issues to emerge. First, the *scope* of each method has become apparent, including some unexpected overlaps and disjuncts between the findings of different UEMs, and also some perturbing omissions (usability issues that emerged in the video data – even with impoverished video data – that were not found by any analytical method). Second, issues about the nature of craft skill in usability evaluation have emerged, particularly through the reflective process imposed by the systematic review. Finally, we reflect briefly on the methodology applied.

5.1 The scoping of methods

Using the nine approaches (eight UEMs plus video data), the usability issues can be classed into groups, according to what the primary focus of the issue is. The classification that emerged comprised five classes:

- (S) System design,
- (M) user Misconception,
- (C) quality of the Conceptual fit between user and system,
- (P) Physical issues or
- (X) conteXtual ones

The second column of Figure 9 indicates the type of each issue.

System design

This concerns the logical design of the system itself, and issues that might make it difficult for the user to work with. In this particular case study, it includes issues such as number of task steps and redundant commands. In other cases, it might include safety and reachability concerns. Broadly, the system-oriented and expert-focused methods (STN, Z and GOMS) were the strongest at identifying these kinds of issues. All of these approaches focus on procedural aspects of system design. Other system-oriented approaches such as ERMIA (Green and Benyon, 1996) that focus on structure rather than procedures might complement these methods, but further investigation lies outside the scope of the present study.

User misconceptions

There is a set of issues that all relate to possible user misconceptions about the state of the system. In practice, few of these issues emerged in the video data. In this particular study, this is partly explained by the fact that only excerpt 6 included non-pre-taught moves, and was therefore the only video data that included any requirement on the user to apply their understanding of the system. Thus, the poor quality of the video evidence leaves some unanswered questions about the value of these kinds of analysis methods that should be the focus of further study. The methods included in this study that consider user misconceptions are CW, PUM and, to a lesser extent, EMU and CASSM.

Conceptual fit

In contrast to the user misconception issues, and perhaps surprisingly, a relatively high proportion of the usability issues for which there is video evidence relate to the conceptual fit between the user's perspective (what they are trying to achieve 'in the world' – e.g. eating) and the system implementation (how they use the interface – e.g. moving the arm in a particular

direction). This illustrates well the difference between users' conceptions as represented within CASSM and potential user misconceptions, as represented within CW or PUM. Unsurprisingly (since this was the intention in developing it), CASSM provides the most support in identifying issues regarding conceptual fit.

Physical fit

For a device such as a robotic arm and its interface, physical considerations – for example, concerning timing and the interpretation of multimodal commands – are important. In particular, there is scope for system misinterpretation of user intentions so that the command issued by the user is not that received by the system. Consistent with the motivation for developing EMU to consider multimodal issues, this method proved the strongest for identifying these issues in the interaction.

Use in context

Finally, there were issues that emerged due to the physical nature of the device and the way it is used in context. Some methods (particularly the more established approaches that consider system design or user knowledge) encouraged the analyst to focus on the interaction with the system controller and consequently pay little attention to the arm being controlled. Even the newer approaches, which were developed to address broader usability concerns, missed some of the most important issues such as the arm itself obscuring the user's view of the gripper at times.

One might reasonably argue that this is the kind of domain knowledge that is more properly the focus of broader domain analysis methods, and therefore outside the legitimate concern of HCI. Nevertheless, this study illustrates that overall usability includes such context-specific factors, and that they need to be accommodated within a total usability analysis.

Horses for courses: choosing the appropriate UEM

This grouping of usability issues is not exhaustive for all systems – for example, it does not include user experience (Norman, 2004; Hassenzahl & Tractinsky, 2006). Nevertheless, it represents the groupings identified for this particular kind of system. Given these groupings, we see that most UEMs have their main strengths within one particular group, and that some important usability issues are missed by all the analytical methods evaluated. Hornbæk (2006), in a comprehensive review of usability studies, structures his discussion around the ISO categories of efficiency, effectiveness and satisfaction; against this categorisation, the System issues typically relate to efficiency and all the other categories to effectiveness: none of the methods investigated here relate to satisfaction or other subjective measures.

In considering the overlaps between the findings of the different methods (Figure 9), the groupings that emerge are as follows.

1. At the levels of abstraction at which these analyses were conducted, STN, Z and GOMS identified very similar issues. These related to the *system design* and, to a lesser extent, the *physical fit* between user and system (notably synchronisation issues).
2. In contrast, the other user-oriented methods consider *user misconceptions*, leading to another clear grouping of issues, for which PUM, CW, EMU and CASSM all have a high degree of overlap.
3. A third set of issues was identified only by EMU – all concerned with the *physical relationship* between the user and the device and the *context* (for example, concerning lighting conditions and hence the user's ability to perceive information correctly from the system).
4. A fourth set of issues was identified only by CASSM; appropriately, these were issues that could be classed as *conceptual misfits* between user and system – for example, that the user might have difficulty judging the position, orientation and aperture of the

gripper as it approached a target. Although in this case many of these concern an understanding of the physical orientation of the gripper, because of its role in achieving the user's real-world goals, they are classed as conceptual rather than physical misfits because they relate to the user's understanding of how the system work rather than the directly physical constraints.

5. Finally, there were issues that emerged in the video data that had not been anticipated by any of the analytical evaluation methods. Some of these had been identified through craft skill while applying a UEM, but others were missed completely. These were in various categories, but include issues about use in context – for example, that the physical arm obscured the user's view of the target at some points in the interaction.

HE was the only approach that did not result in a cluster of issues being found: the HE findings were spread across themes, and are more difficult to classify than those that emerged from approaches with a clearer semantics or theoretical basis.

5.2 The nature of craft skill

In our systematic review, we considered whether issues 'should' have been identified by a particular UEM. In practice, this was a more complex question than anticipated, as illustrated by the extracts from the reviews included in Appendix B. Some cases were fairly clear-cut: either the issue was within the scope of the method or it was not. However, others were less so. Aspects of this were:

1. Task or scenario generation. On several occasions, we could see that had the task been described slightly differently, or had the scenario been embellished more, then the issue would have emerged from the analysis, and been naturally credited to the method rather than craft skill (or being missed completely).
2. Level of abstraction. For four methods (STN, Z, GOMS and PUM), we could see that there were issues that would have emerged had the problem been described at a different (but equally appropriate) level of abstraction or, conversely, that some issues were identified because of the level of abstraction adopted, which might not have emerged had a different representation been chosen. For example, issue 27 (concerning timing gestures accurately as the cursor moves between options) should have emerged in a GOMS analysis if the user—system interaction description had included the detail of every wait and mental operator of assessing whether the cursor was in the right place yet, but our analysis simply said, in effect, "gesture when the cursor is in the right place".
3. Source materials. Tutorial and explanatory materials routinely make use of examples to communicate and illustrate points more effectively. Occasionally, it was apparent that the particular example used in tutorial material helped in issue identification, and that the issue might not have emerged otherwise. One obvious example of this was identifying the 'inability to backtrack' issue of the first version of the system using STN. The role of source materials was also central to the application of Heuristic Evaluation: the explanatory text accompanying each heuristic led the analyst to consider particular features of the design.
4. Representation. As is widely recognised (e.g. Cheng, 1999; Cockayne *et al.*, 1999), representations can serve an important role in helping the problem solver 'see' the problem in a particular way, which makes particular issues apparent and (conversely) hides others. Thus, even notations such as STN and Z, which are not traditionally used as evaluation methods, made certain issues apparent, but did not highlight others. This is also true, though not as starkly, of the user-oriented approaches. This matter of representation is the main determinant of whether or not we classified an issue as 'findable by craft skill': this was based on our judgement of whether or not the representation made an issue reasonably apparent.

5. Skill with notation. The analyst's skill in working with a notation or applying a method appeared, at least subjectively, to influence the quality of insights obtained through applying that UEM. Although this was more obvious with the more formal representations (such as Z), it was also an issue with the more discursive approaches (such as CW). In some cases, we were aware in conducting the initial analyses that the demands of the notation – requiring that the representation be consistent and complete – dominated the analysis, drawing attention away from the system being analysed towards the notation being used for describing it.

As this list illustrates, there are several important factors that influence the efficacy of applying any UEM to a particular interface. These factors contribute to the 'evaluator effect' (Hertzum & Jacobsen, 2001). They also contribute to overall strengths and weaknesses of studies (including this one) that compare UEMs.

5.3 Methodology

Finally, we reflect briefly on the methodology applied in this study. The approach adopted has circumvented many of the pitfalls identified by Gray and Salzman (1998) as discussed above (see Background) but introduced other confounds that we have aimed to identify and account for in the analysis.

The findings presented above have some clear limitations. They are confined to one interface and one task, and the initial analyses were all performed by one of two people. The second of these factors has influenced the results in some particular ways:

1. The differences between 'M's (found by method) and 'O's (should have been found by method but were not) relate directly to the skill and experience of the analyst.
2. Similarly, the differences between 'C's and 'C?'s (were and could have been identified by craft skill) can be attributed to an analyst effect.
3. Finally, the way the problem was represented (including the level of abstraction for the analysis) was chosen by the analyst.

The evaluator effect (Hertzum and Jacobsen, 2001) is reduced by using only one, very small, team of evaluators to perform all analyses; we have made real efforts to 'level the playing field' by systematically revisiting all analyses multiple times to make them consistent with each other. This has, however, introduced a converse problem, which is that the analysts have different depths of understanding of the different approaches – most obviously, between the approaches that are 'home grown' (PUM, EMU, CASSM) and those from elsewhere, but also between the approaches that we have used extensively ourselves (HE, CW, STN) and those that were learnt for conducting this study (Z, GOMS). It might be that analysts with more skill in these approaches would have identified additional issues, or could have accounted for how their approaches would support the identification of more of the issues in the current issue set. We have tried to account for these concerns by relating our findings back to independent descriptions of methods (e.g. John & Kieras, 1996b), and by making the systematic reviews publicly available, so that they are open to inspection and criticism.

Gray and Salzman (1998) discuss cause-effect issues – that a correlation between a problem being identified and a method being applied does not necessarily mean that the application of the method resulted in the identification of the problem. The systematic review has presented an account of why each UEM does or does not support the identification of each usability issue in our set; this also identified issues that might plausibly be found through analyst craft skill, extending slightly beyond the scope of the method. While there is some inevitable subjectivity in these assessments, we have made the assessments inspectable. The method that we found most difficult to assess in terms of cause-effect was Heuristic Evaluation, which has no theoretical basis on which to ground assessments.

Another concern raised by Gray and Salzman (1998) is that of 'method shift' (in which the definitions of UEMs change over time); this was addressed by basing the reanalyses on single, defined sources of description for each method.

This study has not addressed other possible criteria such as what it takes to learn a new UEM (we started with different levels of expertise in the different approaches); what the costs of applying a UEM are (it depends on many factors, including depth of analysis, number of evaluators, expertise of evaluators); persuasive power or downstream utility (after arm development was abandoned, there were no developers to interact with). However, it has resulted in a grounded account of the scope of eight analytical UEMs.

Ultimately, it may not be possible to conduct a methodologically clean comparative study of UEMs. If multiple evaluators are employed then inter-individual differences will confound results, whereas if only one evaluator is used then their familiarity with different UEMs and their growing familiarity with the system being analysed are confounds. There are problems over the attribution of observed user difficulties to underlying causes. And the kinds of difficulties identified will depend on the system and tasks chosen for analysis. However, if no studies are conducted then our understanding of the strengths and limitations of UEMs will not advance.

6 Conclusion

Although this study has focused on one interface and task, the findings are not about that particular interface. This focus inevitably means that there are issues that have not emerged in this analysis that might have, had a different kind of system been used, or a broader set of tasks and contexts of use considered; nevertheless, the findings from this study contribute a piece to the jigsaw of understanding the scope and properties of analytical UEMs.

Similarly, although this study has focused on eight UEMs, it is not just about the features of those particular methods: it has also forced reflection on the nature of analytical evaluation, its strengths and limitations. We have identified several factors that contribute to the quality of an analysis, including the appropriateness of tasks selected, the details of how scenarios of use are described, the level of abstraction used in modelling (applicable to some methods but not all) and the analyst's expertise in the method. As others have reported (e.g. Karat, 1997), analytical methods and user testing yield results of different kinds and scope: analytical methods yield greater insight into why users might have difficulties with an interface, and hence should provide better support for redesign than empirical approaches which focus more on behaviour and subjective assessments. On the other hand, empirical approaches cover a broader spread of possible issues than any individual analytical approach, and reveal issues that are outside the scope of any of the analytical methods we tested.

Some of the UEMs included in this study encourage a focus on the control interface rather than on the arm or other aspects of the domain and context of use; others have broader scope; some (notably CW) encourage a focus on local issues (about *this step* in the interaction) so that the broader picture tends to get lost. For a novice analyst, the more difficult methods encouraged a focus on the notation and getting the representation 'right' rather than using the notation to gain insights about usability. John and Marks (1997) suggest that unstructured consideration of a design description can be just as insightful as the use of a particular analysis method; however, they say little about the precise skills of the individual doing the inspecting. Ultimately, it may be that UEMs provide structure to help the analyst get going and to ensure coverage of issues within the scope of the approach, but their limitations also need to be recognised.

This work has presented a systematic approach to comparing UEMs and validating the findings against empirical evidence. There have been two limitations to comparing analytical findings against empirical data in this study. The first is particular to this study and relates to the poor quality of the video evidence available, which made it difficult to be confident about some of the false positives (was it just that issues did not emerge because the interactions were

too short or too undemanding?). Considering how limited the data was, the findings from it were surprisingly rich. The second issue is more general: it concerns the difficulty of relating behavioural observations to underlying causes. Hollnagel (1998) refers to this as the difference between genotypes (underlying causes) and phenotypes (surface manifestation); this is a difficulty that will continue to plague HCI, and remains a strong argument in favour of analytical methods: observation of surface behaviour can highlight user difficulties, but does not directly point to the possible sources of those difficulties, and hence to design solutions that will remove them. Also, although false positives are often considered undesirable (e.g. Cockton *et al.* 2003), there may be usability difficulties that do not emerge in finite empirical data – whether because they are rare but critical difficulties or because they cause unnecessary mental workload but no obvious physical manifestation.

This study has reinforced some earlier findings, such as the ability of HE to find a wide spread of general usability problems and the strength of CW as focusing on task-related problems. The findings of the GOMS analysis are consistent with those of earlier studies, except that we were unable to conduct a timing analysis. To the best of our knowledge, Z and STN have not previously been used for evaluation in the way presented here, but these methods were found to have similar (though more restricted) scope to GOMS. There has not been a previous scoping study of PUM, though it is not a surprise to find that it covers similar territory to CW.

One of the purposes of this study was to check whether EMU and CASSM did indeed fill the niches that they were designed to. Both methods delivered as designed. In the case of EMU, the device selected for analysis was particularly suited to the approach (EMU would deliver few useful insights that are not more easily acquired by other approaches for systems that are not multimodal). Although EMU fills the intended niche (delivering a method that encapsulates theory about multimodal usability), a parallel (unpublished) study of the usability of the method has shown that the current method is difficult to learn and tedious to apply, so the next step in the development of EMU will be to refine the approach to make it more learnable and usable.

In the case of CASSM, the device was not chosen particularly to test the method, but CASSM supported the identification of some important usability issues; to our surprise, most of these issues were corroborated by empirical evidence (we had thought that CASSM's focus on conceptual misfit would mean that it would reveal issues that increased mental workload but did not affect perceptible behaviours). So far, the development and testing of CASSM has taken place within a research context; future work will focus on downstream utility and fit with design practice.

As noted above, some issues were identified in the video data that were not covered by any of the evaluation methods tested, most notably problems concerning the physical context of use. We are not aware of any techniques that would readily identify such issues, indicating that there is a niche here for a new evaluation approach that focuses on physical context issues. A method such as that adopted in this study could be applied to different kinds of interactive systems to identify other niches for which evaluation support is needed. For example, we can imagine that interactive multi-user games would highlight issues concerning experience, communication and mutual awareness that did not emerge in this study, and for which there are as yet no validated analytical evaluation techniques.

There are many criteria on which UEMs can be assessed; in this paper we have focused on scope. Much of the earlier work on comparing UEMs has assumed that problem count (or similar measures such as thoroughness or the number of hits and false positives) is one of the central considerations. In this paper, we have shown that different UEMs do not simply deliver different numbers of problems: they also support the analyst in identifying different kinds of problems. The question should not be which UEM delivers more, but what kinds of insights each UEM delivers.

NOTES

Acknowledgments. We are grateful to reviewers of earlier versions of this paper for constructive criticisms that have greatly improved it, and to Stephen Payne for encouraging us in those improvements.

Support. Joanne Hyde was funded by a studentship from the School of Computing Science, Middlesex University. Work on CASSM was funded by EPSRC grant GR/R39018.

Authors' Present Addresses. Ann Blandford, UCLIC, University College London, United Kingdom, E-mail A.Blandford@ucl.ac.uk. Joanne Hyde, E-mail jo@jkhyde.co.uk. Thomas Green, Department of Computer Science, University of Leeds, United Kingdom, E-mail greenery@ntlworld.com. Iain Connell, E-mail iain_connell@btinternet.com.

HCI Editorial Record. (supplied by Editor)

REFERENCES

Bailey, R.W., Allan, R.W. & Raiello, P. (1992). Usability testing vs. heuristic evaluation: a head-to-head comparison. *Proceedings of the Human Factors Society 36th Annual Meeting, 1992, **1**, 409-413*. Human Factors Society.

Barnard, P. J. & May, J. (1999) Representing cognitive activity in complex tasks. *Human Computer Interaction*, 14, 93-158.

Barnard, P.J. & Teasdale, J. (1991). Interacting Cognitive Subsystems: A systemic approach to cognitive affective interaction and change. *Cognition and Emotion*, 5, 1-39.

Bastide, R. & Palanque, P. (1990) Petri net objects for the design, validation and prototyping of user-driven interfaces. In Diaper, D. Gilmore, D., Cockton, G. & Shackel, B. (Eds.): Human-Computer Interaction - INTERACT'90. pp.625-631, Elsevier Science Publications, North Holland, Netherlands.

Bernsen, N. O. (1995) A revised generation of the taxonomy of output modalities. In Bernsen, N. O., Jensager, F., Lu, S., Verjans, S. (eds): *Information theory and information mapping*, Amodeus project deliverable D15

Blandford, A. E., Buckingham Shum, S. & Young, R. M. (1998) Training software engineers in a novel usability evaluation technique. *International Journal of Human-Computer Studies*, 45(3), pp. 245-279.

Blandford, A., Connell, I. & Green, T. (2003) CASSM Tutorial. Working paper available from http://www.uclic.ucl.ac.uk/annb/CASSMpapers.html

Blandford, A. Good, J. & Young, R. M. (1998) Programmable user modelling analysis for usability evaluation. Tutorial available as Working Paper WP11a from www.cs.mdx.ac.uk/puma/ (mirrored at http://www.uclic.ucl.ac.uk/annb/CASSMpapers.html)

Blandford, A., Green, T. R. G., Furniss, D. & Makri, S. (forthcoming) User-centred evaluation using CASSM: focusing on concepts and structures. To appear in *International Journal of Human–Computer Studies.*

Blandford, A. & Hyde, J. (2006) Rational Reanalyses of a Robotic Arm. Working paper available from http://www.uclic.ucl.ac.uk/annb/CASSMpapers.html

Blandford, A. E. & Young, R. M. (1996) Specifying user knowledge for the design of interactive systems. *Software Engineering Journal*. 11.6, pp. 323-333.

Card, S. K., Moran, T. P. & Newell, A. (1983) *The Psychology of Human Computer Interaction*, Hillsdale NJ: Lawrence Erlbaum.

Cheng, P. C-H. (1999) Unlocking conceptual learning in mathematics and science with effective representational systems. *Computers and Education*, 33(2-3), 109-130.

Cockayne, A., Wright, P.C. & Fields, B. (1999). Supporting Interaction Strategies Through the Externalization of Strategy Concepts. In: Sasse, M.A. and Johnson, C. (eds.) *Proceedings INTERACT'99*, pp. 582-588. IOP Press.

Cockton, G., Woolrych, A., Hall, L. & Hindmarch, M. (2003) Changing Analysts' Tunes: the Surprising Impact of a New Instrument for Usability Inspection Method Assessment. In: People and Computers XVII: Proceedings of HCI'03, Springer, pp 145-161.

Connell, I., Blandford, A. & Green, T. (2004) CASSM and cognitive walkthrough: usability issues with ticket vending machines. *Behaviour and Information Technology*. 23(5). 307-320.

Connell, I., Green, T. & Blandford, A. (2003) Ontological Sketch Models: Highlighting User-System Misfits. In E. O'Neill, P. Palanque & P. Johnson (Eds.) *People and Computers XVII, Proc. HCI'03*. 163-178. Springer.

Coutaz, J., May, J., Young, R., Blandford, A., Nigay, L., Salber, D. (1995) Integrating system and user modelling through abstraction: the CARE properties for reasoning about multimodality. In: Nordby, K., Helmersen, P., Gilmore, D. J., and Arnesen, S. (eds): *Human-Computer Interaction: Interact'95*. Chapman and Hall, pp. 115-120

Coutaz, J., Nigay, L. & Salber, D. (1993) The MSM Framework: A Design Space for Multi-Sensory-

Motor Systems. In L. Bass, J. Gornostaev, & C. Unger (eds): *Lecture Notes in Computer Science 753 (EWHCI'93) Selected Papers*. pp. 231-241. Springer-Verlag.

Cuomo, D.L. & Bowen, C.D. (1994). Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers, 1994,* ***6(1)****, 86-108*.

Desurvire, H.W. (1994). Faster, cheaper !! Are usability inspection methods as effective as empirical testing ? In J. Nielsen and R.L. Mack (eds.), *Usability Inspection Methods*, pp. 173-201. New York: John Wiley & Sons.

Desurvire, H.W., Kondziela, J.M. & Atwood, M.E. (1992). What is gained and lost when using evaluation methods other than empirical testing. In A. Monk, D. Diaper and M.D. Harrison (eds.), *People and Computers VII.* Proceedings of HCI '92, York, September 1992, pp. 173-201. British Computer Society Conference Series 5. Cambridge: Cambridge University Press.

Dix, A. J., Finlay, J., Abowd, G. & Beale, R. (1993) *Human-Computer Interaction,* Hemel Hempstead: Prentice Hall International.

Dowell, J., Life, A. & Salter, I. (1994) The design space for a multimodal multimedia travel facility. In: *Proceedings of ECCE7, European Society for Cognitive Ergonomics*, September 5-8 1994, Bonn

Duke, D. J., Barnard, P. J., Duce, D. A. & May, J. (1998) Syndetic Modelling. *Human-Computer Interaction*. 13, 337-394

Dutt, A., Johnson, H. & Johnson, P. (1994). Evaluating evaluation methods. In G. Cockton, S.W. Draper and G.R.S. Weir (eds.), *People and Computers IX. Proceedings of HCI '94*, Glasgow, August 1994. Cambridge: Cambridge University Press.

Gray, W., John, B & Atwood, M. (1993) Project Ernestine: Validating a GOMS Analysis for Predicting and Explaining Real-World Task Performance. *Human-Computer Interaction*, 8. 237-309.

Gray, W. D. & Salzman, M. C. (1998) Damaged merchandise? A review of experiments that compare usability evaluation methods. *HCI Journal*. pp. 203-261

Green, T. R. G. & Benyon, D. (1996) The skull beneath the skin: entity-relationship models of information artifacts. *International Journal of Human-Computer Studies*, 44(6) pp. 801-828

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001) Evaluating usability evaluation methods. *International Journal of Human-Computer Interaction* 13 (4), 373-410.

Hartson, H. R., Siochi, A. C. & Hix, D. (1992) The UAN: A user-oriented representation for direct manipulation interface designs. *ACM Transactions on Office Information Systems,* 8, 181-203.

Hassenzahl, M., & Tractinsky, N. (2006). User Experience - a research agenda. *Behaviour & Information Technology*, 25(2), 91-97.

Hertzum, M., and Jacobsen, N.E. (2001). The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13(4), 421-443.

Hollnagel, E. (1998) Cognitive Reliability and Error Analysis Method (CREAM) Oxford : Elsevier Science.

Hornbæk, K. (2006) Current practice in measuring usability: challenges to usability studies and research. *International Journal of Human–Computer Studies*. 64. 79-102.

Hornbæk, K. & Frokjær, E. (2005) Comparing usability problems and redesign proposals as input to practical systems development. *Proc ACM CHI 2005*. 391-400.

Hyde, J. K. (2002a) *Multi-Modal Usability Evaluation*. PhD thesis. Middlesex University.

Hyde, J. K. (2002b) Tutorial notes on applying EMU available from http://www.uclic.ucl.ac.uk/annb/CASSMpapers.html.

Jacobsen, N., Hertzum, M. & John, B. (1998) The evaluator effect in usability studies: problem detection and severity judgements. *Proc. HFES 42nd Annual Meeting.* 1336-1340.

Jacobsen, N. E., & John, B. E. (2000). Two case studies in using cognitive walkthrough for interface evaluation. School of Computer Science Technical Report CMU-CS-00-132. Pittsburgh, PA: Carnegie Mellon University. PDF: http://reports-archive.adm.cs.cmu.edu/anon/2000/CMU-CS-00-132.pdf

Jeffries, R. & Desurvire, H. (1992). Usability testing vs. heuristic evaluation: was there a contest ? *SIGCHI Bulletin, October 1992, **24(4)**, 39-41.*

Jeffries, R., Miller, J.R., Wharton, C. & Uyeda, K.M. (1991). User interface evaluation in the real world: a comparison of four techniques. In S.P. Robertson, G.M. Olson and J.S. Olson (eds.), *Reaching Through Technology: CHI '91 conference proceedings.* ACM conference on human factors in computing systems, New Orleans, April-May 1991. New York: Addison-Wesley.

John, B. & Kieras, D. (1996a) The GOMS family of user interface analysis techniques: comparison and contrast. *ACM Transactions on CHI*, 3, 320-351.

John, B. & Kieras, D. E. (1996b) Using GOMS for user interface design and evaluation: which technique? *ACM ToCHI* 3.4. 287-319.

John, B. & Marks, S. (1997) Tracking the effectiveness of usability evaluation methods. *Behaviour and Information Technology* 16, No. 4/5, 188-202.

John, B. E., & Mashyna, M. M. (1997) Evaluating a Multimedia Authoring Tool with Cognitive Walkthrough and Think-Aloud User Studies. *Journal of the American Society of Information Science, 48* (9) 1004-1022.

John, B. E. & Packer, H. (1995) Learning and using the Cognitive Walkthrough method: A case study approach. In *Proceedings of CHI'95.* 429-436. ACM Press: New York.

Johnson, H. & Hyde, J. (2003) Towards modeling individual and collaborative construction of jigsaws using task knowledge structures (TKS) *ACM Transactions on Computer-Human Interaction* 10.4, 339 - 387 .

Karat, C. M. (1994) A comparison of user interface evaluation methods. In J. Nielsen & R. Mack (Eds.), *Usability Inspection Methods.* 203-233. New York: John Wiley.

Karat, C-M., Campbell, R. & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In P. Bauersfeld, J. Bennett and G. Lynch (eds.), *CHI '92 Conference Proceedings: striking a balance.* ACM conference on human factors in computing systems, Monterey, California, May 1992. Reading, MA: Addison-Wesley.

Karat, J. (1997). User-centered software evaluation methodologies. In M. Helander, T.K. Landauer and P. Prabhu (eds.), *Handbook of Human-Computer Interaction*, 2nd edition, pp. 689-704. Amsterdam: Elsevier Science B.V.

Kieras, D. E., Meyer, D. E. (1995) An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, 12, 391-438

Lavery, D., Cockton, G. & Atkinson, M. P. (1997) Comparison of evaluation methods using structured usability problem reports. *Behaviour and Information Technology*, 16 (4/5), 246-266.

Lindegaard, G. (2003) The Misapplication of Engineering Models to Business Decisions. In M. Rauterberg, M. Menozzi & J. Wesson (Eds.) *Proc. Interact 2003*. IOS Press. 367-374.

Long, J. & Dowell, J. (1990) Conceptions of the discipline of HCI: craft, applied science, and engineering. In A. Sutcliffe and L. Macaulay (Eds.) *People and Computers V.* 9-32 Cambridge: Cambridge University Press.

Moran, T. P. (1983) Getting into a system: external-internal task mapping analysis. In A. Janda (ed.), *Human Factors in Computing Systems.* 45-49.

Nielsen, J. (1994) Heuristic Evaluation. In J. Nielsen & R. Mack (Eds.), *Usability Inspection Methods* (pp. 25-62) New York: John Wiley.

Nielsen, J. & Phillips, V.L. (1993). Estimating the relative usability of two interfaces: heuristic, formal and empirical methods compared. In S. Ashlund, A. Henderson, E. Hollnagel, K. Mullet and T. White (eds.), *Human Factors in Computing Systems: INTERCHI '93.* Proceedings of INTERCHI '93. Amsterdam: IOS Press.

Norman, D. A. (1986) Cognitive engineering. In D. A. Norman and S. W. Draper, Eds. *User Centered System Design,* Hillsdale NJ: Lawrence Erlbaum. 31-62

Norman, D. (2004) Emotional Design: why we love (or hate) everyday things. Basic Books.

Parsons, B., Prior, S. D. & Warner, P. R. (1995) An Approach to Designing Manipulator Controller Software Employing Context Dependent Command Interpretation. In: *European Conference on Assistive and Rehabilitation Technology*, Lisbon, October 1995

Parsons, B., Warner, P., White, A.. & Gill, R. (1997) An Approach to the Development of Adaptable Manipulator Controller Software. In: *Proc. International Conference on Rehabilitation Robotics (ICORR97).*

Payne, S. J. & Green, T. R. G. (1986) Task-Action Grammars: The Model of the Mental Representation of Task Languages *Human-Computer Interaction*, 2, 2, 93-133

Payne, S. J., Squibb, H. R. & Howes, A. (1990) The nature of device models: the yoked state space hypothesis, and some experiments with text editors. *Human-Computer Interaction*, 5. 415-444.

Purchase, H. (1999) A semiotic definition of multimedia communication. In: *Semiotica*, 123-3/4, 247-259

Sanderson, P. M. & Fisher, C. (1994) Exploratory sequential data analysis: foundations. Human-computer Interaction 9, 215-317.

Sears, A. (1997). Heuristic walkthroughs: finding the problems without the noise. *International Journal of Human-Computer Interaction, 1997, **9(3)**, 213-234.*

Spencer, R. (2000) The Streamlined Cognitive Walkthrough Method, Working Around Social Constraints Encountered in a Software Development Company, *Proc. CHI'2000*. pp. 353-359.

Spivey, J. M. (1989) *The Z Notation: A Reference Manual.,* Prentice-Hall International.

Virzi, R.A., Sorce, J.F. And Herbert, L.B. (1993). A comparison of three usability evaluation methods: heuristic, think-aloud and performance testing. *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting, 1993, **1**, 309-313.* Human Factors and Ergonomics Society.

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

Wickens, C. D., Sandry, D. L., Vidulich, M. (1983) Compatibility and resource competition between modalities of input, central processing, and output. In: Human Factors, 25, 2, pp 227-248

Wixon, D. 2003. Evaluating usability methods: why the current literature fails the practitioner. *interactions* 10, 4 (Jul. 2003), 28-34.

Young, R. M., Green, T. R. G. & Simon, T. (1989) Programmable User Models for Predictive Evaluation of Interface Designs, in K. Bice. & C. Lewis (eds.), *Wings for the Mind: CHI '89 Conference Proceedings*, pp.15-19. ACM conference on human factors in computing systems, Austin, Texas, April-May 1989. Reading, MA: Addison-Wesley.

Appendix A: definitions of the usability issues identified

1. Long sequence of (mental) operators to move arm

The number of decision and action steps needed by the user to get the arm going is greater than necessary (does not apply to pre-taught positions). This is particularly so if the user wishes to move an individual joint, or to change the speed of arm movement.

2. Inability to backtrack.

In the first version of the system analysed, there was no 'undo' option. As shown in the 'second STN' (Figure 2), this omission was soon corrected.

3. Difficulty of choosing between Move Arm or Move

The user's first decision is between MoveArm (which moves the whole arm) and Move, which then allows the user to select an individual joint to move. The semantics of this choice may be difficult for novice users to grasp.

4. Lack of short cuts

There is no quick way to return to the direction menu, which might be required if the arm overshoots or if the whole arm is being moved and needs a change of direction.

5. Continue serves same function as Go, and is redundant

There was originally an option called 'continue', which served exactly the same function as 'go' and was therefore eliminated in an early redesign of the interface.

6. Confusion over joint called Arm

The term 'Arm' is used to refer to both the whole arm and an individual joint called 'arm'.

7. Gesture input with twice as many operations as voice because dependent on cursor movement

This refers to mental operations, not physical ones. The number of physical operations is the same in both cases, but it takes more mental effort to spot the gesture option and maintain attention on it until the cursor is in the correct place to select that option.

8. Problem if head moved to look at arm while gesture system operational may be interpreted as a command

Since gestures may be part of user's normal repertoire of head movements, it is possible that the user might move their head in a way that is interpreted by the system as a gesture when it was not intended as such.

9. If user pauses in middle of saying "Move arm"…

Because "move arm" is made up of "move" and "arm, and "move" and "arm" are also valid commands, a pause in the middle could cause misinterpretation by the system.

10. If user engaged in conversation…

If the user of the speech controlled system is also engaged in another conversation, it is possible that some conversational words might be interpreted as commands by the system.

11. Lack of feedback about selection

This arose in the CW analysis specifically in relation to MoveArm. This reflects a broad concern that the system as analysed did not give feedback on selections at the time of analysis, although the gestural and voice input mechanisms did request user confirmation of choice.

12. Problems of determining left and right, especially when arm contorted

If the arm is contorted then 'its' right and left may be different from right and left (or indeed up and down or in and out) as perceived by the user.

13. User cannot check direction choice until arm starts to move

This is really a combination of 11 and 12: that the user neither gets feedback on what they have selected nor can anticipate which actual direction corresponds to the command for a contorted arm until the arm starts to move.

14. Time taken to interact with system to stop arm

The user has to anticipate how long it will take the system to respond to 'stop' and issue the command at the right time.

15. Similarity between moving joint and moving whole arm

Both moving the joint and moving the arm follow a similar pattern of states and transitions. The interaction could be made more efficient and maybe clearer by combining these options into a single menu.

16. Illegal options

When the arm has reached its limit of movement, it is possible to issue command that would, in principle, send it beyond its limit. The only feedback to the user is that the arm does not move.

17. Mismatch between way that arm works and way that user would move arm

The way the user conceptualises what they are doing 'in the world' does not map readily on to the way the user has to program the arm to work.

18. Not clear that End returns user to main menu

This is about labelling: firstly, 'end' is semantically confusable with 'stop'; secondly, 'end' does not mean 'return to initial menu', although that is the effect of this action.

19. End having two meanings

Under all circumstances, 'end' returns the user to the initial menu. Other than at the end of the overall interaction, the user has a motivation to complete this step; right at the end of the interaction the user has no reason to restore the interface to its initial state, and may therefore omit the 'end'. This is unlikely to cause substantive user difficulties in the circumstances.

20. Lighting conditions

If lighting is poor, the user may have difficulty seeing options or seeing the arm's current position.

21. Difficulty for user to move field of vision

Disabled users may have difficulty shifting their visual attention from the display to the arm and vice versa.

22. User looking one way, menu options in other direction

The user has to divide their visual attention between the arm position or movements and the display that controls the arm.

23. Difficulty of judging arm movements

For novice users, it is likely to be difficult to judge exactly how the arm is moving and where it currently is. This issue is expanded below as more detailed issues.

24. Difficulty in judging speed and direction as getting close to target

As the gripper gets close to the target, it needs to reach it without overshooting or colliding. Depending on the direction of approach, the user may find this very difficult to judge.

25. Difficulty in judging position, orientation and aperture of gripper as approaching target

Similarly, the position of the gripper may be difficult to ascertain.

26. Position and movement of most joints is of limited interest to the user

Since the user's main concern is with the position of objects in the world, which can only be manipulated by the gripper, the main concern is about getting the gripper in the right place, i.e. by moving the whole arm. Exceptions might be when fine-tuning the angle of the gripper on approach, and if avoiding other obstacles in the room.

27. Possible difficulty of timing gesture accurately as cursor moves between options

The user of the gestural interface has to time their gesture to select the correct option. This timing may be difficult for novices.

28. Voice recognition problems

If the user does not speak clearly, their words may not be interpreted correctly by the voice recognition system.

29. Speaking with mouth full…

If the user of a voice recognition system tries speaking while eating, there are likely to be voice recognition problems.

30. No display of speed

There is no feedback (other than the perceived speed of the arm while actually moving) of the current speed setting.

31. Arm obscuring user's view

The arm itself may get in the way of the user's view of the target object in the world.

32. No arm reversing.

It is not possible to reverse direction of the arm without going all the way through the set-up procedure again. This matters in cases where the user overshoots.

33. Difficult to match names to joints

For the novice user, it may take a while to learn the names of all the joints.

34. Long sequence of operators to recover from directional error

This is a combination of issues 1 & 32 plus an extra consideration, which is that if the user selects any wrong parameter (joint, direction, speed), it takes many steps to recover from that error.

Appendix B: extracts from the systematic reviews

The full systematic reviews are available from Blandford and Hyde (2006). Here, the most interesting issues are presented – typically those that illustrate points made in the discussion section.

STN reviewed

1. Long sequence of operators to move arm

Since the STN shows the number of states that the user has to navigate through before the robotic arm can be moved, this issue should have been identified in the original analysis. However, STN deals with only physical state changes, and does not consider mental operations, so the effect is less marked for STN than it was for GOMS. That it was not identified shows the extent to which the analysis was dependent on the craft skill (or lack thereof) of the analyst.

2. Inability to backtrack [STN]

This issue is apparent from the STN, and was identified as a problem. However, the identification of this issue was possibly influenced by the explicit mention of this kind of problem in a discussion on 'undo' in the source materials (Dix *et al.*, 1993, p.291). This shows the effect that the source materials of a method has on the application of a method.

3. Difficulty of choosing between Move Arm or Move

The STN concentrates on the actual choice between the system states, rather than on the difficulties the user has in choosing between them. It is therefore not an issue that the STN on its own would be expected to identify, but might have been identified through craft skill – looking at the problem with particular questions in mind.

4. Lack of short cuts

Since the STN explicitly shows the possible path of the interaction through the various states, the lack of short-cuts was an issue that might have become apparent if the analyst had been looking for it. This is therefore an issue that is a combination of craft skill and representation. That it was not noticed was possibly because the analyst's attention was more on obtaining the correct representation of the system states.

6. Confusion over joint called Arm

The STN did not go into the detail of the individual options, so this issue did not arise. If the STN had been done at a different level of abstraction, this issue might have been identified through the craft skill of the analyst. It is not something that the STN would identify directly however, since it is concerned more with the user understanding of what a particular option choice means rather than with the option choice itself. Thus this issue highlights questions associated with both craft skill and appropriate levels of abstraction.

7. Gesture input with twice as many operations as voice

The STN was not written at the level of abstraction which would identify this issue. If it had been, this issue would probably have been identified, since it would be concerned with the number of states and transitions. For the gesture input, there are a series of states and transitions between them as opposed to the voice input which has one state with multiple transitions coming from it. This raises questions concerning the appropriate level of abstraction of an analysis.

16. Illegal options

This issue was not represented on the STN. There was no state showing that the arm had reached its limit of movement, nor was there an end option leading from the travel until stop state which might also represent it. This shows how difficult it is to draw STNs correctly, and relates to the level of skill of the analyst in determining how the system states should be represented. However, even if the STN diagram had been correctly drawn, it is still unlikely that this issue would have been identified without explicitly checking for illegal options.

27. Possible difficulty of timing gesture accurately as cursor moves between options

STN does not explicitly consider timing. With a more detailed STN (level of abstraction), this issue might have been spotted through craft skill. In the event, it was not.

32. No arm reversing.

Because the STN focuses on the device states, and the direction of motion is simply a parameter on that state, the domain requirement to make it easy to reverse does not appear through the STN. It would have required a very different kind of STN to allow this issue to emerge.

CW reviewed

2. Inability to backtrack

CW does not deal with error in terms of its implications, therefore would not find this issue, although it might come out from the craft skill of the analyst through thinking about rectifying errors.

11. Lack of feedback about selection

For the purposes of the original analysis, this was not relevant, since the feedback had not been implemented, but it was an important issue raised by the method that would have to be addressed once feedback had been implemented.

17. Mismatch between way that arm works and way that user would move arm

One of the aims of the method is to uncover this kind of issue, however there is not much support within the questions for this to be identified at a high level, because of the method's concentration on the step-by-step nature of the task. This is more likely to be uncovered by craft skill therefore.

30. No display of speed

Because the display of speed (or the lack of it) is outside the essential task definition (unless the task were to be to move the arm at a particular speed, which would involve craft skill in perceiving the need for such a task), this would not naturally emerge from a CW analysis.

CMN and CPM GOMS reviewed

The CPM GOMS analysis was unable to identify many issues over and above those identified by CMN GOMS, other than the difference between the use of voice and gesture operators. Thus only issue seven was able to be identified, and this was the only issue that can be considered to be within the bounds of the method. This re-analysis consequently focused on the use of CMN GOMS. A different CPM GOMS analysis that indicated where the user would want to look at the arm to check its position or movement would have raised more issues.

4. Lack of short cuts

By writing out the methods, the long sequence showed that this would take a long time and that there were no short cuts. Whether this emerges from the analysis or is derived through craft skill is a moot point.

10. if user engaged in conversation…

This issue is outside the scope of CMN GOMS and was not identified.

If the task description included reference to another conversation, this issue should be identified through CPM GOMS; however, this depends on analyst insight in specifying such a task.

26. Position and movement of most joints is of limited interest to the user

This issue did not emerge. Indeed, a task definition would include a specification of which joints to move, so this issue is more strongly excluded from the set of possible issues than most.

27. Possible difficulty of timing gesture accurately as cursor moves between options

CMN GOMS does not consider timing issues such as this.

CPM GOMS should have spotted this issue, had the interface been described at the appropriate level of abstraction.

29. Speaking with mouth full…

This issue is outside the scope of CMN GOMS and was not identified.

It would only be identified by CPM GOMS with a very inspired choice of tasks.

PUM reviewed

1. long sequence of operators to move arm

This issue was mentioned in the original analysis, but not in a strong enough way for it to be apparent as an issue of consequence. It was identified from looking at the heavy ordering identified by the analysis, and was therefore dependent upon the craft skill of the analyst.

2. inability to backtrack

The original analysis found a heavy ordering, which is within the bounds of the PUM method. However, from this was derived the lack of backtracking provision, which is therefore identified by the craft skill of the analyst, based on the representation provided by the method.

7. Gesture input with twice as many operations as voice because dependent on cursor movement

This did not come out in the original PUM analysis, because the analysis was not written at a low enough level of abstraction for this to be apparent.

11. Lack of feedback about selection

The output was not included in the original analysis. If it had been then the PUM analysis might have picked up on this issue, in the modelling of the user knowledge, because the user would not know that the option had been selected.

15. similarity between moving joint and moving whole arm

The way that the PUM analysis was conducted meant that this issue was not identified, although it would probably have been recognised if the analyst was looking for it. Therefore, although the PUM analysis represented the operations, it would take the craft skill of the analyst to identify their similarity.

24. Difficulty in judging speed and direction as getting close to target

If a much more detailed PUM model had been constructed, it is possible that this issue might have been identified, through the process of describing a 'monitoring' activity more detailed than the 'wait and then stop' implemented in the current model. This is therefore both a level of abstraction and a craft skill issue.

26. Position and movement of most joints is of limited interest to the user

Because PUM doesn't encourage the analyst to 'step back' in this way, it is unlikely that this issue would fall inside the scope of a PUM analysis.

27. Possible difficulty of timing gesture accurately as cursor moves between options

It would be necessary to construct a PUM model at a much finer grain of detail for this issue to emerge. This is not a level at which PUM naturally works, so it is unlikely that this issue would be spotted.

Z reviewed

1. long sequence of operators to move arm

This issue was not apparent because of the way that the specification was constructed, although the specification did represent it. This issue therefore highlights the important difference between an issue being represented and identified. It would take a certain amount of craft skill on the part of the analyst to identify this issue.

4. Lack of short cuts

The Z specification represented the lack of backtracking opportunities, due to its concentration on the ordering of the interaction. The lack of short-cuts was therefore also represented. However, the issue, although represented, was not identified, which again illustrates the difference between an issue being represented and identified, and the importance of the craft skill of the analyst in identifying significant issues.

7. Gesture input with twice as many operations as voice because dependent on cursor movement

This issue was not identified by the Z specification because the specification was not written at a low enough level of detail to represent the cursor movement. This illustrates the need for the appropriate level of abstraction of the representation.

EMU reviewed

10. if user engaged in conversation…

This was not an issue identified by the method since it was not indicated in the initial scenario that the user would be engaged in conversation. If that information had been included in the environment profile, then this issue would have been identified by EMU.

19. End having two meanings

This issue was not identified in the original analysis and should have been, since it is a potential mismatch. This demonstrates how the identification of any issue is dependent upon the analyst, and that mistakes and omissions can occur.

23. difficulty of judging arm movements

This is a clash unless expert issue, and the method instructs the analyst to look for these clashes.

29. Speaking with mouth full…

This issue should have been identified by EMU had a different task been considered – i.e. one that included feeding.

31. Arm obscuring user's view

Paradoxically, this issue is outside the scope of EMU, unless it were identified through craft skill, because the bulk of the rest of the arm (other than the gripper) is not represented.

CASSM reviewed

6. Confusion over joint called Arm

With a slightly expanded CASSM description that includes the concept of the whole arm as being made up of joints, this issue should have emerged. This issue *should* have been identified.

12. problems of determining left and right, especially when arm contorted

The issue of judging directions when the arm is contorted emerged (with some craft skill) from looking at joints and what the user knows about the directions in which joints can move. It does not emerge directly from the CASSM representation.

30. No display of speed

This probably should have emerged through the consideration that there is a difference (misfit?) between the perceived speed of the arm as moving and the speed setting as determined (but not displayed) through the interface. This one's a bit marginal…

Heuristic Evaluation reviewed

2. Inability to backtrack

This issue was identified through the 3rd heuristic ("user control and freedom") which includes the consideration of undo and redo.

3. Difficulty of choosing between move arm or move

This issue arose through craft skill while considering the names of joints, because the similar names match different real-world terms (heuristic 2).

6. Confusion over joint called arm

This might have been discovered by considering match between system and real world, but would have relied on substantial craft skill.

10. If user engaged in conversation…

If the analyst were very familiar with voice input systems, this might be identified under heuristic 5 (error prevention).

12. Problems of determining left and right, especially when arm contorted

While heuristic 2 concerns the match between the system and the real world, there are no cues to make this kind of high level match, so this is outside the scope of HE.

13. User cannot check direction choice until arm starts to move

This might be spotted through craft skill if the analyst is familiar with this kind of system (visibility of system status).

17. Mismatch between way that arm works and way that user would move arm

This kind of high level mismatch between the system and the real world would be unlikely to emerge from a HE unless the analyst were looking out for it specifically (under heuristic 2).

18. Not clear that end returns user to main menu

This should have emerged from heuristic 2: that 'end' is the wrong term for this meaning.

26. Position and movement of most joints is of limited interest to the user

The heuristics are too general to focus on issues like this.

27. Possible difficulty of timing gesture accurately as cursor moves between options

This emerged by considering possible causes of user error (heuristic 5).

29. Speaking with mouth full…

If the analyst were very familiar with this kind of system this might emerge while considering heuristic 5.

32. No arm reversing.

This is part of error recovery, and is a specific example of issue 34, which emerged in this analysis, and so is covered as a special case of that.

33. Difficult to match names to joints

This emerged while considering match between the system and the real world.

34. Long sequence of operators to recover from directional error

This was identified through craft skill while considering error recovery (heuristic 9).