Int. J. Man–Machine Studies (1993) **38**, 689–711

Evaluating user–computer interaction: a framework

M. Sweeney, M. Maguire and B. Shackel
HUSAT Research Institute, Loughborough University, UK

(Received 5 December 1990 and accepted in revised form 28 November 1991)

A framework is described, which classifies usability evaluations in terms of three dimensions; the approach to evaluation, the type of evaluation and the time of evaluation in the context of the product life cycle. The approaches described are user-based, theory-based and expert-based. The approach to evaluation reflects the source of the data which forms the basis of the evaluation. The types of evaluation are diagnostic, summative and metrication. These reflect the purpose of the evaluation and therefore the nature of the data and likely use of the results. The time of testing reflects the temporal location in the product life cycle at which the evaluation is conducted. This dictates the representation of the product which is available for evaluation.

The paper describes the relationship between these three framework dimensions. It also relates the methods of data capture, measurements and criteria which may be appropriately applied in various evaluation contexts. The latter part of the paper focuses on a more detailed review of methods which are associated with the most commonly applied and often most effective approach, i.e. the user-centred diagnostic evaluation.

Finally the paper considers the need to perform evaluations more effectively in the design of products and systems in the commercial world. The discussion addresses the need for computer support tools to facilitate the handling of resulting data from user trials.

1. Introduction

A framework for usability evaluation is presented in order to provide a methodological orientation which positions various approaches and types of evaluation within the context of a product development cycle. The framework initially proposes a definition of usability. From the definition it is possible to identify relevant usability indicators to be assessed, given a purpose for the results of the evaluation and a set of constraints within which the evaluation is to be conducted.

The paper focuses on the topic of the user-centred, iterative design-evaluation cycle in IT development, exploring the methods which may be applied and the data which they yield. Procedures and problems which are associated with conducting such evaluations are examined.

A usability evaluation framework

The framework proposed categorizes evaluations along three dimensions. The dimensions are the approach to evaluation, the type of evaluation and the time of evaluation. The first dimension classifies the approaches to evaluation as follows: user-, expert-, and theory-based. This classification reflects the underlying source of

0020-7373/93/040689 + 23$08.00/0

the data which forms the basis for the evaluation. The second dimension classifies evaluations into three basic types, diagnostic (or generative), summative and metrication (or certification). This reflects the underlying purpose of the evaluation. The third dimension, time of testing, reflects the temporal location in the product life cycle at which the evaluation is conducted. This dictates the representation of the product which is available for evaluation, i.e. specification, prototype or early mark.

Before discussing the dimensions upon which the framework for evaluation hangs, it is necessary to examine the definition of usability which forms the basis for the framework.

Definition of usability

Usability is an emergent quality of an [1*]optimum design, which is reflected in the [2*]effective and [3*]satisfying use of the IT.

As an emergent quality, usability is implicit in the design and manifests itself through interaction with the product. Although this definition implies that usability evaluation necessarily involves a user interaction, evaluation may also be conducted on the basis of the product's features and characteristics.

Usability may be measured by the extent to which the IT affords (or is deemed to be capable of affording) an effective and satisfying interaction to the intended [4*]users, performing the intended tasks within the intended environment at an acceptable [5*] cost.

Terms

[1*]*Optimum.* This refers to the fact that any design necessarily involves a cost-benefit analysis and compromises. These refer to the trade-off between providing sophisticated functionality and maintaining simplicity.

[2*]*Effective.* This refers to efficient and productive usage of the product which is reflected in the levels of speed, completeness and correctness which are achievable in the user's *performance* during interactions and/or of the output. It is also reflected in the following indicators, which, because they are context-dependent and require interpretation are difficult to specify in the abstract.

Efficient and productive interactions are manifest in solution-oriented activities which enable the user to achieve effective *behaviour.* They are also manifest in the contribution of mental models (e.g. knowledge of the task and the device) to effective *cognition* (or understanding).

User performance, behaviour and cognition are henceforth referred to as positive indicators of the usability of the product under evaluation.

[3*]*Satisfying.* This refers to the degree of general positive regard or emotion which the user attributes to the interaction with the IT. This is reflected in the level of positive *attitude/opinion* which is reported by the user.

[2*&3*]*Effective and satisfying use of IT.* Where evaluation is based on the data from user interaction, measures of effectiveness and satisfaction are sought. However, where the evaluation is based on the product's features, "guestimates" and predictions of these levels may be determined.

Effectiveness and satisfaction may be assessed against some level of expectation. This is called the criterion and it may be established with reference to the levels of user effectiveness and satisfaction which are currently attainable (e.g. by manual means), the levels attainable by using a competitive system or levels which are set by guidelines/standards or from the client's statement of business requirements.

[4*]*Intended user, task and environment.* For measurement of usability it is important to include characteristics underlying user, task and environment in the "usability evaluation equation". These may be added on a theoretical level by modelling them (this refers to our theory-based approach) at a practical level by simulating/emulating them (refers to our user-based approach) or at an abstract level by being conscious of their capabilities and limitations (expert-based approach).

[5*](*acceptable*) *Cost.* Refers to the level of user investment required to achieve and maintain high levels of the usability indicators described above. The cost to the user may be demonstrated in terms of level of *physical and/or mental effort* or *stress/anxiety* incurred. Cost indicators also cover the level of *organizational investment* required to attain effective and satisfying performance from its employees e.g. training time, user support documentation, incentives etc.

The goals of designing for usability are to maximize user effectiveness and satisfaction and to minimize the cost involved in achieving attainment criteria.

From this definition it is possible to abstract two main categories of indicator;

Positive indicators	**Cost indicators**
Direct positive indicator	*Direct cost indicator*
User performance	Physical and mental complexity
User attitude/opinion	Stress/anxiety
Expert opinion	
System conformance	
Indirect positive indicator	*Indirect cost indicator*
User cognition	Organizational investment
User behaviour	—economic cost
	—motivation and hygiene factors

The positive indicators may be employed to determine the levels of user effectiveness and satisfaction. The user performance, user and expert attitude/opinion and conformance measures provide a direct positive indicator of the effectiveness of the IT. The measures of users' behaviour and knowledge provide an indirect indicator of the effectiveness, as they need to be interpreted in the context of the evaluation. For example, models of the knowledge sources associated with performing the task using a particular system need to be developed, e.g. conceptual, technical and format knowledge or syntactic and semantic knowledge etc. *Cost indicators* may be employed to determine the investment required by the user to achieve and maintain effective performance levels. *Indirect cost indicators* may be determined by looking at the necessary organizational investments, such as levels of user support and training involved in bringing user effectiveness and satisfaction up to criteria levels.

Evaluations should focus on all of the classes of indicators, so as to cross-reference interpretation from the data. High levels of effective performance may be observed to be attained at a high price to the user in terms of stress or a high price to the organization in terms of training time.

The framework dimensions—approach, type and time of evaluation

Factors which determine the approach and type of evaluation adopted include constraints and limitations such as time scales, staffing and technical resources and access to users. The most significant determinant is the time of evaluation, as this dictates the representation of the product which is available for testing and hence the kinds of questions which may be posed in the evaluation.

Approach

The approach to evaluation defines the source of the data for the evaluation, i.e. user interaction, expert appraisal or theoretical prediction. The framework for human–computer evaluation describes three main approaches to evaluating human–computer interaction, user-based, theory-based and expert-based (see Figure 1). The scenario which characterizes a user-based approach involves one or more users completing one or more tasks, in an appropriate environment. In rigorous testing, relevant task, user and environment characteristics must match those for which the product is being designed. In simple terms, the scenario which characterizes a theory-based approach involves a designer or evaluator calculating the match between the task or user model and the system specification. This ultimately generates quantitative values for the learnability or usability of a system. The evaluation involves neither a user–computer interaction nor a tangible representation of the product. The scenario which characterizes an expert-based evaluation involves the evaluator (i.e. human factors practitioner) using the system in a more or less structured way in order to determine whether the system matches predefined design criteria. The evaluation results represent the evaluator's subjective judgement on the system's conformity to general human factors principles, approved guidelines (Brown, 1988; Smith & Mosier, 1984) and standards (BSI 91/40677, ISO 9241).

Each of the three approaches can be further described in terms of the indicators which they address. These indicators, when measured, reflect on usability and fitness for purpose of the system.

Figure 1 presents a basic taxonomy of the three evaluation approaches, indicating the relationship between the approaches described and the various usability indicators which may be assessed and the relationship between indicators and the kinds of measures which may be taken to sample them. Each indicator type and its related data inherently reflect the evaluation context in which they are most appropriately applied. The data types also imply the types of evaluation scenario and technology which need to be employed. For example, evaluating users' levels of understanding and knowledge of the system by application of verbal protocol analysis implies that the users must perform a set of predefined tasks within an environment which facilitates the recording of comments. Although this does not preclude field trials, exact control over the evaluation procedure is facilitated in a laboratory.

Approach	*Usability indicators*	*Data*
User-based evaluation	**Performance (user)**	Task times, % completed, Error rates Duration of time in HELP, Continuance of usage, Range of function used (*Objective*)
	Non-verbal behaviour	Eye movement, Orientation duration and Frequency of documentation access (*Objective*)
	Attitude (User's attitudes and opinions)	Questionnaire and survey responses, Comments from interviews and ratings Answers to comprehension questions (*Subjective*)
	Cognition (User's understanding and knowledge of system)	Verbal *protocols* Post-hoc comments (*Objective*)
	Stress	Galvanic skin response, Heart rate Event related brain potentials Electro-encephalograms Ratings or comments (*Objective & subjective*)
	Motivation	Enthusiasm, willingness and effort (*Subjective*)
Theory-based	**Performance (idealized)** (Predictions of usage)	Predictions of - task performance times - learning times - likely ease of understanding (*Objective*)
Expert-based evaluation	**Conformance** (Level of conformance with standards, guidelines and design criteria)	Level of adherence or conformance with - guidelines, principles and standards - design criteria (*Objective*)
	Attitude (expert) (Professional opinion)	Comments Rating of usability properties (*Subjective*)

FIGURE 1. Taxonomy of usability indicators (developed from Maguire & Sweeney, 1989).

Type of evaluation

The second dimension classifies evaluations into three basic types, diagnostic or generative, summative and metrication or certification.

This dimension is strongly related to the temporal dimension. Evaluations which are conducted earlier in the design cycle obviously have more potential to impact on the development of the product. The analysis of the results needs to indicate where redesign effort should be invested therefore the data will be diagnostic in nature.

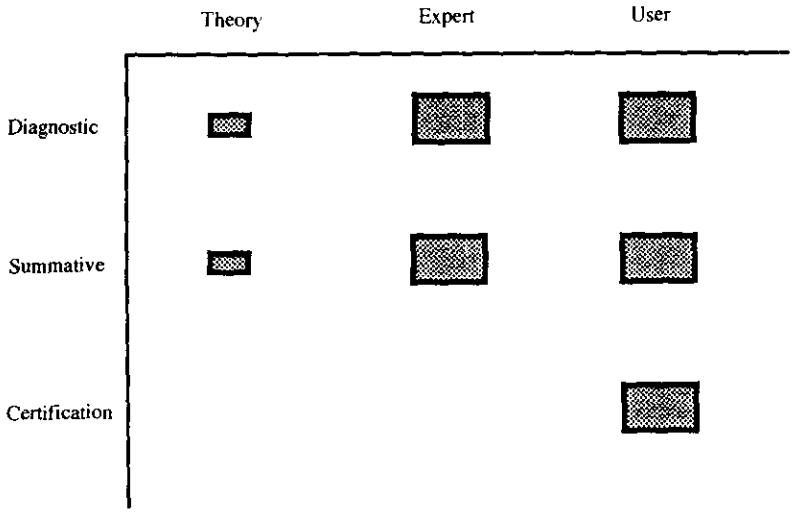

FIGURE 2. Relationship between approaches and types of usability evaluations. ▨ indicates an appropriate match, ■ indicates a possible match.

Later evaluations need to present a picture of the levels of effectiveness and satisfaction which can be achieved during interactions with the product.

Diagnostic evaluation seeks to identify short-comings in the design and recommend redesign solutions. This type of evaluation can be achieved by employing any of the approaches. Summative evaluations are placed somewhere in between diagnostic and certification, they seek to determine the extent to which the IT actually supports and is felt to support the user to complete the experimental task. This type of evaluation is achieved mainly by employing expert- and user-based approaches. However, it might be feasible, in some circumstances, to employ the theory-based approach. Certification tests seek to generate scores from an interaction which reflect the extent to which the product may be certified against certain criteria. Certification tests for any software may only be achieved by employing the user-based approach and by employing a (near) completed product design (although experts may be able to certify well-defined hardware aspects).

Figure 2 indicates the relationship between the approach (horizontal axis) and type (vertical axis) of evaluation.

The time of evaluation in the product life cycle

The third dimension proposed by the framework—time of evaluation—dictates the representation of the product which is available for evaluation. Four generic representations are characterized—specification, rapid prototype, high fidelity prototype and mark 1.

Earlier in the design cycle, evaluation tends to focus on the complexity and adequacy of the functionality of the system. The objective is to determine whether the specification for the proposed system meets the requirements. This may be achieved by using predictive theoretical models such as SANE (Bosser & Melchior, 1991) to test the specification against the task or user analysis data, or by having an expert human factors practitioner evaluate it against the user requirements.

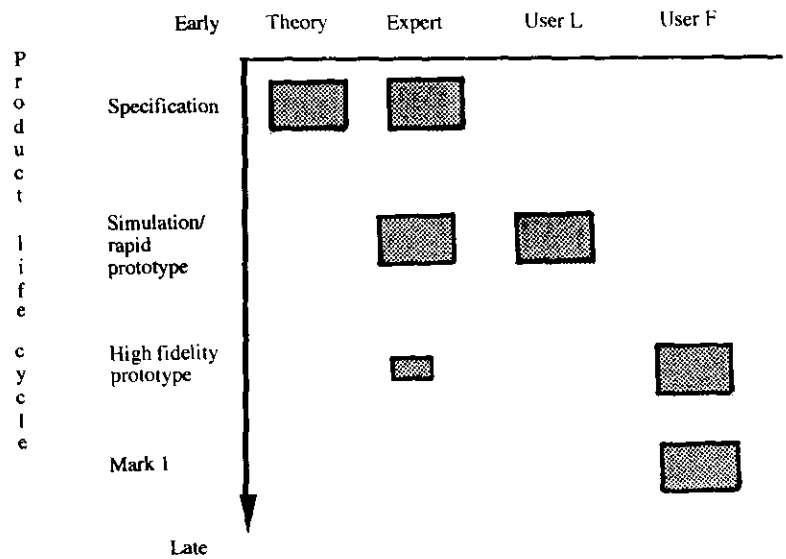

FIGURE 3. Relationship between approaches to evaluation and the product life cycle. L = laboratory trial, F = field trial, ▨ indicates an appropriate match, ■ indicates a possible match.

During an iterative design cycle rapid prototypes may be built which mirror (some of) the functionality as well as the look and feel of the proposed system. These enable the evaluation to focus on usability factors such as functionality, ease of use and learnability of the system. This may be achieved by having an expert "walkthrough" the prototype to test it against heuristics, previous experience or more formally against guidelines. Useful diagnostic data may also be generated by conducting laboratory-based user trials.

Later on, during the design cycle, high fidelity prototypes and early marks of the system enable (alpha and beta) evaluations in the field to address factors such as user acceptance in addition to functionality, ease of use and learnability. Such evaluation objectives may also be achieved by having an expert assess the fitness of the product for use.

Figure 3 indicates the relationship between the development stage in the product life cycle and the approach to evaluation which is appropriate.

Relationship between approach, type and time of evaluation

Predictive methods employed in theory-based approaches may be employed to diagnose faults in design. However, they are oriented towards a quantitative assessment of the product's usability. Usually conducted early in the design process, they yield a measure of the cognitive effort involved in using the system. The quantitative output from methods used in the theory based approach means that they are suitable for conducting summative assessments of IT products.

Expert-based evaluations are more qualitative and the results may be used to diagnose faults and make recommendations. More structured approaches can yield quantitative data which could be used in summative evaluations of IT.

User-based evaluations may yield a wide range of data which are suitable for application in diagnostic, summative and especially certification evaluations.

These relationships between types of evaluation reflect the basic distinction between the objectives which drive the evaluation. In reality the mappings between the three dimensions are not so clear cut. Obviously in generating metrics for certification the evaluator has access to a range of qualitative data which could be reported as a diagnostic.

Criteria setting

Most definitions of usability imply that usability may only be evaluated by conducting user trials and also that the interpretation of the data is always referenced to some criteria. Criteria are the expected levels or benchmarks against which observed levels of effectiveness and satisfaction are compared. Criteria generally apply to the performance indicators, although they may informally be set for many of the other indicators (e.g. the attitude/opinion and stress indicator). They are generally only applied in the user-based approach to evaluation as this yields performance data on the interaction.

Diagnostic evaluations do not require criteria to be established. Summative evaluations do not require precise criteria to be set, as relative relationships between scores or measures are more informative, (e.g. achievement levels across subjects, between conditions and over time). When certifying the usability of a product as part of a company's quality assurance procedures, relevant criteria are sourced to standards. However, many human factors standards are general and do not always provide precise criteria. In such instances the evaluator's judgement, the client procurer's operational requirements and results from previous tests are the basis for generating valid criteria.

Although some evaluations relate to a specific question which is being posed, some are rather more vague. In such cases the client may not be able to specify the relevant indicators or evaluation criteria without some human factors guidance. It is suggested that either of the framework matrices (see Figures 1, 4) might be a useful reference guide for identifying the evaluation objectives and approach.

2. A classification of usability methods

The framework (see Figure 4) attempts to match relevant usability indicators and their corresponding measures to the appropriate data capture methods.

Figure 4 is by no means complete. However, it presents the most accessible and frequently employed methods.

The *user-based* approach covers the following indicators:

user performance	e.g. speed (time) & accuracy scores (error)
behaviour	e.g. non-verbal gestures
knowledge	e.g. cognition and understanding
attitude/opinion	e.g. preferences and satisfaction
stress/anxiety	e.g. subjective or psychophysiological
physical and mental effort	e.g. subjective statement on complexity

Evaluation approach	*Usability indicator*	*Data*	Observation of user during interaction	Video-recording of user interaction	Audio-recording of user during interaction	System monitoring of user interaction	User's post-hoc comments	User interview	User questionnaire, survey or rating scale	Psycho-physical recording	Formal modelling	Expert "Walk-through"
User-based approach	Performance (Based on user interaction)	Task times, % completed Error rates *Duration of time in HELP* Range of functions used	✓	✓		✓						
	Non-verbal behaviour	*Eye movement* Duration and frequency of document usage	✓	✓								
	Attitude (User opinions)	Post-hoc comments Questionnaire, survey Interview comments and ratings			✓		✓	✓	✓			
	Cognition (understanding)	Verbal protocols Answers to comprehension questions, sorting task scores etc.			✓		✓	✓				
	Stress	Galvanic skin response *Heart rate*, Event related brain potentials *Electroencephalograms* Ratings of anxiety							✓	✓		
	Motivation	*Enthusiasm*, willingness and effort			✓		✓	✓	✓			
Theory-based approach	*Performance* (Usage prediction)	Predictions of - *task performance times* - learning times - ease of understanding									✓	
Expert-based approach	*Conformance* (Expert evaluation)	Comparisons with: - guidelines and standards - design criteria Technical measures										✓
	Attitude (Expert opinion)	Comments Rating of usability *properties*	✓									✓

FIGURE 4. Relationships between usability indicators and data capture methods. Ticks ✓ indicate possible data capture methods to support measure, Shaded areas ▩ indicate methods discussed in the present paper.

The *theory-based* approach argues from "first principles" in order to make predictions about the complexity and hence learnability of a system.

performance (idealized user)	e.g. number of production rules to complete a task (indicating physical and cognitive complexity)

Finally the *expert-based* approach may provide measures relating to:

system conformance	e.g. indicating the level of conformance to guidelines, standards or design criteria (developed by human factors experts) e.g. expert opinions about usability based on heuristics or previous experience.

2.1. INDICATORS FOR USER-BASED EVALUATIONS

This section discusses user-based evaluations by considering both positive and cost indicators.

Positive indicators

Objective *performance* data are generated within the context of user-based evaluations, where the users are required to complete one or more (bench-mark) tasks. The performance is generally assessed in terms of measures such as time to completion, accuracy, or error levels and error types. Such indicators map easily on to operational definitions of the performance criteria to which system usage should be targeted. For example an operationally defined performance criterion may state that 90% of users should be able to perform a set of core tasks within a given time limit (see Shackel, 1990). The evaluation, therefore, is concerned with measuring performance times and comparing them with predefined criteria.

Data on *non-verbal behaviour* and gestures such as facial expression and visual direction (eye movements) may also be generated from video-recordings of the interaction. *Cognitive* data on the user's knowledge or understanding of how a product or system works may be elicited by recording and analysing users' verbal comments made concurrently during the interaction. This method was first popularized by Ericsson and Simon (1980, 1983) and is known as verbal protocol analysis. Similar methods such as post-hoc comments or talk-throughs involve asking the user to comment after usage of the system. Comprehension questions or sorting exercises are also a useful means of gathering cognitive data.

Subjective data such as user *opinions or attitudes* towards the system may be assessed in a number of ways (e.g. interview, survey questionnaire and rating scale). Such methods are valuable tools for obtaining users' reactions to the technology, e.g. whether it is satisfying and enjoyable to use (see Chin, Diehl & Norman, 1988; Oppenheim, 1966). These data are important since they provide a measure of acceptability (a broader concept than usability), which in turn reflects on whether the bespoke system will be organizationally effective or whether a generic product is likely to succeed in the market place.

Cost indicators

Methods of recording psychophysiological data yield objective measures which may reflect on cognitive workload and *stress* levels which the user is experiencing during system usage. These include event-related brain potentials (ERPs), electroencephalograms (EEG), ocular responses, heart rate (ECG) and GSR (galvanic skin response). For a review of these measures see Gale and Christie (1987) and for physiological measurement devices see Kak (1981). Stress may also be subjectively assessed by asking users to rate the levels of stress and anxiety which they experienced during the interaction (Hart & Staveland, 1988).

Data on the degree of *enthusiasm* which the user requires in order to maintain effective levels of performance may be assessed by using interview, rating scales and user comments (post-hoc or concurrent).

User-based evaluation enables the system to be assessed in circumstances (i.e. user, task and environment) which closely resemble those for which it was intended. The representation of the system may range from a prototype to the finished system, depending on when in the life cycle the evaluation is conducted. The results of the "early" evaluation may therefore be the input to an iterative product design process, whereas the results from the finished product may be used to shed light on how the system might operate in practice. Nevertheless, the user-based approach can prove costly in terms of resources required for conducting user trials and analysing the resulting data.

2.2. INDICATORS FOR THEORY-BASED EVALUATIONS

Positive indicators

This approach involves the application of *formal methods*—a catch-all term for those analytical evaluation techniques which measure key attributes of the system interface, such as ease of use, ease of learning and consistency, by applying models of expert user behaviour. Examples include CLG—command language grammar (Moran, 1981), Reisner's Formal Grammar (Reisner, 1981), TAG—task action grammar (Payne 1984), GOMS—goals, operators, methods and selection rules (Card, Moran & Newell, 1983) and CCT—cognitive complexity theory (Kieras & Polson, 1985). Models of user–system interaction can provide a predictive evaluation of proposed designs without involving users directly. In terms of preparation, some form of task analysis needs to be conducted in order to compile the set of production rules or keystrokes which are involved in performing the task. The evaluation determines a measure of the physical and/or cognitive complexity involved in using the system.

Cost indicators

Methods which yield measures of the complexity of a system may also be used to generate data on the requirements for user support, training and learning time. This enables the evaluator to guestimate the level of organizational investment required to introduce the new technology.

Although a very active field of research, formal modelling is still very limited in its application and it places assumptions on the interaction (e.g. error-free usage). Bellotti (1988) states that many of these models are not successfully applied in the world of commercial design projects due to constraints which lead to problems such

as poor communication between designers and users, uncertainty about requirements, exclusion of users in the evaluation, expanding task outlines, designers unfamiliarity with the task domain etc. For a review of the implementation of formal methods in evaluating commercial designs see Bellotti (1988).

2.3. INDICATORS FOR EXPERT-BASED EVALUATIONS

Positive indicators

This approach involves an expert (human factors, system design, software engineering) making an assessment of the system. This may firstly be an appraisal of *system conformance* where the expert compares the system with established human factors standards, guide-lines or principles (e.g. see Brown, 1988).

In a more subjective evaluation based on *expert opinion,* the expert relies on experience and heuristics in order to make a judgement on whether the system will succeed in the particular application area, with given users and tasks. Aids such as check-lists (Brown, 1988) and inventories (Ravden & Johnson 1989) may be used in order to systematize the representation of the results of the system conformance evaluation, whereas rating scales may be used to represent the results based on the experts' subjective assessment.

Expert assessment has the advantage of being both quick and relatively cheap to conduct. Experts can draw upon their knowledge of what design features make a product usable with certain contexts (e.g. level of user's skill, organizational culture). However, even experts may only have limited knowledge of the task and may have difficulty in putting themselves in the users' shoes, and personal biases may also colour their judgement.

2.4. SUMMARY

So far the paper has concentrated on presenting the framework which establishes the context for approaches and types of evaluations and the relationship between relevant indicators and the methods which address them.

The remainder of the paper focuses on specific issues which relate to conducting such user-based diagnostic evaluation of early designs. The topics which are discussed include the design of such evaluations, a review of the methods which are commonly employed, and a discussion of how the data is analysed and fed back into the design cycle.

3. A discussion of usability methods

The following sections discuss a set of data capture methods which are highlighted in Figure 4. The methods which are discussed are typical though not exhaustive of those employed in laboratory-oriented, user-based evaluations. These include:

Observation of user	3.1
Video-recording of user interaction	3.2

The discussion in each section will describe how the data gathering technique may be implemented, the advantages and disadvantages of each method, how the data might be analysed and the possibilities for computer-based support for evaluation.

3.1. OBSERVATION OF USER DURING INTERACTION

Typically with this method, the evaluator watches "over the user's shoulder", as it were, noting problematic aspects of the interaction. Objective features of user interaction which may be recorded manually include time to complete task, points of apparent user difficulty, number, frequency and approximate duration of relevant events, e.g. errors made, general demeanour of user, approaches to using the system etc. (see Figure 4).

Advantages:

1. Can be conducted quickly or tailored to suit the circumstances as necessary.
2. Can be flexible since no special technology is necessary.
3. Provides first hand feedback of user interaction, thus assisting correct interpretation of objective measures.
4. Does not curtail movement and interaction between the user at the workstation and co-workers, a particularly relevant attribute when the evaluation is being conducted in the field and where the public may be involved in the interaction (e.g. welfare service offices, banks or travel agencies).
5. Observation can potentially yield as much insight as a lengthy and costly experimental evaluation. There may be additional facets to the interaction which may require attention (e.g. use of documentation, the need for pen and paper, supplementary information etc.), which methods such as automatic monitoring omit.

Disadvantages:

1. There is no permanent record of the interaction to be reviewed later, hence relevant incidents may be missed or forgotten.
2. The user may find it intrusive and uncomfortable to have someone hovering behind them. This may affect the subject's performance.
3. Observation may yield less accurate results than automatic recording methods, as the process of recording may cause the observer to miss some actions or inaccurately record the timing of events.
4. Observers may produce varying amounts of data as a result of the level of detail to which they attend within the interaction.

Capture and analysis

Observations vary in the degree to which they are structured. In a structured recording the observer records information on to a paper-based form containing a check-list template which lists categories of action types. A tick and/or the time may be noted on the form when the action is seen to occur. Free-hand notes describing the context may also be added to a "general comments" column. For unstructured recording, the observer records events perceived to be significant, in free format as and when they arise. A form of short-hand devised specifically for the experiment is useful for recording notes more rapidly. More structured observations are appropriate when the focus is restricted and where the issues being addressed are explicitly outlined in advance. Less structure is appropriate when the evaluation is exploratory and where less preconception of valid issues is possible.

Comments

In order to ensure objectivity, the goals and issues being addressed should be refined as far as possible. This should help the evaluator to focus on relevant target variables and make the observation task less strenuous.

Being observed can be unnerving and in the context of an evaluative study the observer needs to re-assure the subjects that it is the product which is being assessed and not themselves. It is important to spend some time before the experiment, briefing the subjects and discussing the project. The more aware the subjects are of the objectives of the study, the more comfortable they feel and the more they can contribute to the data gathering.

Computer support

The process of live observation is enhanced when the evaluator is equipped with a portable and unintrusive input device for noting events with a time stamp. A flexible approach might be the use of a computer linked touch pad, where each key denotes a specific "event marker." Manual recording of events is thus performed via the pad. Researchers have written simple programs to accept event marker entries and then automatically calculate the duration and frequency of the events on PCs and "Psion Organizers", thus saving effort.

3.2. VIDEO-RECORDING OF USER'S INTERACTIONS

The following sections discuss audio-, and video-recordings as separate capture methods although video generally implies audio-recording too.

The use of video-cameras to record human–computer interaction is becoming increasingly popular. Recording facilities range from a simple hand-held camera which may "roam", but tends to focus on the system display, to a fully fitted usability laboratory, featuring remotely controlled cameras, an interaction area (the simulated working environment containing system, furniture etc.) and an observation area housing the camera control units and video-editing suite. This is separated from the interaction area by a one-way mirrored window. The following large-scale computer manufacturers and users all have usability laboratories in the UK—ICL, IBM, Hewlett Packard, DEC and British Telecom.

Video-recordings of the user interaction can provide views of documentation, facial expression or the workstation (eg. input devices), in order to provide data

relating to measures of behaviour. Such data present a rich view of the context in which the events occur. However, the video-camera is generally used to capture the screen output for an analysis of the product interface.

Audio-recordings (one microphone for the subject and another for the evaluator's comments) provide information on the subject's rationalization of what he or she is doing, or reaction to the system. Data on the user's intentions are complementary to that which is generated from objective performance assessment.

Advantages:

1. A comprehensive record of a session can be captured.
2. A session can be replayed allowing a fuller analysis of the interaction than is permitted by direct observation.
3. Reliability of data analysis can be increased by having a number of evaluators analyse the same recording (this is very difficult with direct observation).
4. Tapes can be edited and a compilation tape may be presented to the design and marketing teams to illustrate problems that occurred.

Disadvantages:

1. Recording may interfere with the users' performance as it may cause the subject to feel self-conscious and unable to talk to colleagues or ask for help (the intrusiveness may be minimized when gantry-mounted cameras are used).
2. Analysis of tapes is very time consuming (as a rule of thumb, previous experience suggests that 10 h for coding/analysis must be allowed for every 1 h of taped interaction).
3. Video-recording of the subject's screen may yield a poor quality replay unless a video camera with an electronic shutter is used together with a box called "The Shutter". This reduces the flicker on the video tape by synchronizing the shutter speed with the scan rate of the computer monitor being recorded. Obviously technical solutions such as recording the screen via a hardware device such as a "mediator" enable perfect recordings to be made. However this adds to the cost of the set-up.

Capture and analysis

The evaluator uses video to record the users' interaction with the system. Generally, one camera is used to record the system display as seen by the user, and another camera is used to capture non-verbal behaviour (e.g. reading manuals, using the keyboard or other input device). A video-timer is used to time-stamp recordings so that data from other sources (e.g. commentary, keystroke records) may be replayed in synchronization and accurate time scores can be taken. Although the evaluator's task is dictated by the question being addressed, it is generally concerned with reviewing the tape to determine problematic aspects of the dialogue or insufficient aspects of the functionality afforded by the product. These may be demonstrated by problems, errors and circular attempts at completing a task, or events such as the user resorting to paper-based methods of processing the information.

Comments

The use of video has revolutionized the process of studying human–computer interaction. However, in some respects it encourages evaluators to skimp on

preparatory work such as planning exactly what aspects of the interaction are to be measured, prior to user trials. Thus, there is a tendency when using video to record absolutely everything. However, it must be said that sometimes important interaction patterns only become apparent after looking at the data more than once or by comparing data across many subjects.

Computer support

Analysis of the video-tape, is time consuming. Techniques for digital video-recording in the future will offer random access which should speed up analysis. As described in the section on live observation, a simple computer-based event recording device would be of value so that recorded events could be counted or timed automatically. Even more powerful, would be a means of displaying the recorded events on a time-line on screen. The evaluator could then view the data and, using a pointing device, select and replay sections of the video which are of particular interest. For instance, one might look back to an event label "problems occurred here" and issue a command for the video to rewind to that part of the interactive session and replay it. By marking and locating periods of interest in this way, the time-consuming process of video-analysis can be reduced. A number of workbenches have now been developed to facilitate such video-capture and analysis (see Maguire & Sweeney, 1989; Theaker *et al.*, 1989, Walsh & Laws, 1990; Al-Isa, 1990).

3.3. AUDIO-RECORDING OF USER DURING INTERACTION

Audio-recordings may be conducted within the framework of verbal protocol analysis (VPA) where the recording is made of the users' concurrent rationalization of their performance. Audio-recordings may also be made in the context of post-hoc comments during play back, where users' comments are made after completing the task. The data captured and the reduction and analysis procedures are the same as for VPA.

3.3.1. Verbal protocols

Verbal Protocol Analysis (VPA), otherwise known as "think-aloud" or "talk-aloud" protocol, involves eliciting and recording a verbal explanation by the user of what they are doing and why, *during* the execution of a task. The objective is to gather information on the user's intentions during the performance. The data may be analysed in order to shed some light on the strategies, plans or rationale adopted by the individual. The scenario usually involves a complete system or sophisticated interactive prototype, where the task is structured so that common patterns across users can be observed.

The main advantage of this method is that it may be the only source of data on the cognitive processes involved in using the system. These data are very useful for determining the source of users' misconceptions and for determining how sophisticated their conceptual models of the system are.

Disadvantages:

1. It is intrusive—the user has to be asked to think aloud initially and subsequently prompted.

2. The intrusion may completely change the nature of the interaction.
3. Subjects vary in the depth and amount of information which they provide.
4. Large volumes of verbal data require a great deal of effort in analysis.

However, accepting the fact that the subject is being more conscientious and taking time to talk through actions, the wealth of data about their cognitive activity may prove invaluable in certain evaluations. The trade-off which must be made for such a rich source of data is the high cost required for analysis.

Capture and analysis

In terms of preparation for analysis, the evaluator should select at least two subjects' protocols to listen to, in order to get an idea of the type of events and trends which are evidenced. On the basis of this, a classification scheme detailing event categories can be devised for systematically analysing the data. The classification scheme should label and describe the events which are relevant to the evaluation. It can be expanded during analysis as more undefined event categories become apparent.

In order to ensure reliability and validity of the results, it is recommended that two independent and experienced evaluators conduct the analysis. They should both use the scheme as their guide to interpreting the data. This is done by simply ticking the event category once it is encountered in the data. A quick measure of inter-rater reliability may then be determined by calculating the number of agreed ticks (where both raters similarly identify the presence of a particular event or characteristic), and then dividing by the total number of categories in order to get the level of agreement as a percentage of the possible agreement. If this score is less than 80% it indicates that there is a problem with the interpretation of the data. This may mean that the various categories need to be defined more clearly.

The results of VPA are descriptively presented rather than statistically represented, but it is important to determine how reliable the conclusions are by checking the inter-rater reliability. Similarly, a rule of thumb for determining the validity of an event category is that it should be observed in the protocols of more than 20% of the subject pool. If the event has not been referenced frequently enough it means that it is probably an idiosyncratic observation.

3.3.2. Post-hoc comment

This method again involves recording a subject or user's verbalizations of their experience of using the product. The scenario might involve asking the subject, whilst viewing a video play-back of their activity using a system, to describe what they were doing. It is a good alternative to VPA, although criticisms include the fact that it relies on the subject's memory; it is easier to rationalize problems after they have been solved and subjects may say things which make them appear more intelligent (Halo effect) or which they believe the experimenter wants to hear. For these reasons it is more productive to record post-hoc comments in order to gather subjective data on the user's opinion, and VPA in order to gather cognitive data on their mental processes.

Advantages:

1. Provides valuable feedback on the user's opinions and cognitive processes.
2. Avoids interference problems which are encountered when subject is required to talk aloud during the interaction (c.f. concurrent verbal protocols).

Disadvantages:

1. Users may forget their original thoughts.
2. Users may be embarrassed to admit to errors in their thinking.
3. Users may feel that they must provide logical explanations for illogical behaviour.

Capture and analysis
The method is generally conducted by playing the video-recorded interaction and possibly dubbing an audio-track of the users' comments on to the video at this time. This makes future reviews of the tape more convenient. The methods of coding data and analysing it are the same as those outlined in the above section on verbal protocol analysis (VPA).

Comments
The addition of post-hoc comments to video data offers tangible benefits to the evaluator. It provides information on the underlying reasons for the user's interaction strategy, and offers an insight into the user's conceptual model of how the system operates and impressions of the ease of use and learning, flexibility, transparency etc. In order to ensure that the data obtained from the analysis of the post-hoc commentaries is reliable and valid, it is important to establish a rapport with the subject, in a non-judgemental atmosphere where the subject is aware that it is the system which is being tested and not the user. It is also important to prompt the subject on target issues but not to pose leading questions.

3.3.3. Content analysis
Content analysis is a knowledge elicitation technique, first popularized by Krippendorff (1980). The main difference between VPA and content analysis is that the former is concerned with examining what the subject is thinking while performing some operation, whereas the latter is concerned with determining the elements of a concept as it is understood by the subject. VPA is based on the user's commentary on the *process* of using "it", whereas content analysis is based on the user's conceptualization or *description* of what "it" is.

The scenario need not involve human–computer interaction and it might be based on the subject's memory of past experiences with the task (e.g. driving), technology (e.g. in-car technology) or concept (e.g. navigator, negotiation, communication) which is under examination.

Capture and analysis
The implementation of content analysis involves recording the subject's comments on a given concept or issue. The subject is required to relate all they know about the issue under investigation. The representation may actually be offered by the subject in written or verbal form. It is suggested that the written form is appropriate where the purpose of the study is to generate as much descriptive detail as possible. Verbal representations will be more spontaneous and will be likely to reflect a structure (priority) in terms of the order of discussion.

Comment
Content analysis is normally used to produce information which may form the basis of a survey or questionnaire. The term "content" refers to the nature of the factors or dimensions which are involved in the concept. In the context of HCI it might be employed in a preliminary study of mental models (e.g. where people are asked to write or talk about what they understand by the concept of an electronic desktop). The data yielded by this could then be used as a basis on which to formulate a survey or questionnaire on the effective usage of electronic desktops.

3.4. SYSTEM MONITORING OF USER INTERACTION

This is a performance assessment method, which yields system derived records of the interaction.

System monitoring may be technically implemented in a number of ways in order to gather interaction data when video is not available. One mechanism is to record all user inputs such as keystrokes and mouse movements. The actual recording or "tapping" process can be done either in software, with a piece of recording software residing in the user's machine—e.g. "Screen Recorder" (Farallon, 1998) or for the application, under test itself, to be programmed to record interactions of relevance to an evaluation. This approach can provide low-level performance data e.g. task times or more diagnostic data e.g. pathway taken through the software, but it requires access to the source code and programming effort.

The evaluation may be made by examining the keystroke file, for example; however, because this contains low-level data without context, it is difficult to interpret. However the data may be used to provide a replay of the user's interactions in real time by relaying the digital data back to the target system. Thus, the evaluator can rerun the interaction on screen without requiring video.

The advantage of system monitoring, as a source of information for evaluation, is that the recording is non-intrusive and provides a permanent record. This means that it may be reviewed as often as necessary for evaluation.

The main disadvantage is the same as for any other detailed record; there is much data to store and analyse. This means that it can be an expensive technique to implement in terms of the time required to analyse and interpret the raw data.

Clearly, while system monitoring is potentially a useful method it is difficult to implement effectively. Data capture is easy but data reduction and interpretation is extremely difficult.

Capture and analysis
One method of dealing with the lack of context, which is characteristic of the data generated by this method, is to develop a data structure reflecting the internal states of the system (e.g. a state transition diagram) and to apply the inputs to it. This would also support the tracing of user paths through the system. These transition diagrams, reflecting the user interaction, could then be compared with an idealized or optimum performance in order to determine the problematic nodes in the pathway, which in turn may reflect directly on a dialogue problem. However, for all

but simple interactions or systems, the structure would be very complex and difficult to create.

Higher-level software recording of time spent on screens, error messages and functions accessed is a much more accessible level of abstraction.

4. Use of multiple data capture methods

In most evaluations a multiplicity of methods often need to be employed in order to generate appropriate (quantitative and qualitative) data on which to base diagnosis and to develop design recommendations. For instance, by synchronizing a video replay with psychophysiological data and post hoc comments it is possible to replay all three data streams simultaneously, giving a broad picture of a complete interactive session. This provides a much enriched view of the interaction. Similarly subjective data from interviews or rating scales may be cross-referenced with physiological or performance data in order to validate the interpretations from the analysis.

Employing many methods in the evaluation can be said to create a "data explosion". Therefore, unless technology facilitates the process of reducing and interpreting the data, it will remain a less than cost-effective approach.

5. Designing user-based evaluations

5.1. REALISM OF THE EVALUATION SCENARIO

User testing in the laboratory offers a controlled environment for detailed study and data collection, while field studies are more difficult to control but can yield results that are closer to the real situation (Whiteside *et al.*, 1988).

The main strength of laboratory testing is that it allows evaluation to take place in a restricted environment, where variables may be manipulated experimentally in order to determine specific cause and effect relationships. Laboratory testing is appropriate where low-level design details of the system prototype are being tested, as the availability of technical support facilitates detailed analysis. Such evaluations should generally be conducted early in the design cycle to determine elements of the functionality to be provided and the look and feel of the interface. In this instance the variables being tested are predefined and the data recording is selective. Similarly laboratory evaluations are appropriate for final certification testing, where all conditions and control variables need to be specified.

Laboratory evaluations have a tendency to be technology-driven rather than goal-directed. Often, simply having the technology is believed to be reason enough to use it, with the result that more records are made than are necessary and the analysis becomes unduly expensive. It is important that the goals of the evaluation should form the basis for the design of the study, directing the selection of appropriate usability indictors and the related metric.

On a more general level there is a need to consider the organizational context. A product that might seem usable in a quiet laboratory might appear quite different in its real working setting. In an open-plan office where the system users work in a noisy environment, and perhaps work on a number of tasks simultaneously,

different aspects of the product's usability are likely to be highlighted. This emphasizes the need to consider relevant organizational factors or characteristics, so that their effect on the system's usage may be determined as early as possible. Field studies are more appropriate in this context although they are difficult to conduct without intrusion.

5.2. DESIGN OF THE EVALUATION

The design of the evaluation must be considered so that the data yielded is capable of answering the question being asked.

The *ease of use* of the interface may be tested by a multiple condition study (e.g. by exposing subjects to design variants and then comparing their performance scores or satisfaction ratings). On the other hand the *learnability* of the system might be tested in a longitudinal study (e.g. by plotting the performance measures along a temporal dimension). In both cases, where a number of trials or subjects are being employed, the expense of the evaluation is likely to be high in terms of laboratory time and data analysis, and the cost of the evaluation may be prohibitive. However, such experimental evaluations are valuable sources of valid and reliable data.

5.2. FEEDBACK IN THE DESIGN PROCESS

The evaluation process may be used in order to determine whether a system or product meets certain predefined usability goals. In ideal circumstances the evaluation is conducted early enough to impact the development work within the context of an iterative design cycle. It is important that designers should be aware that the product is expected to enable the user to achieve certain performance goals from the outset or that the system must conform to standards or guide-lines. Indeed, a recent European Community Council Directive (May 1990) on the safety and health requirements at work states that, in designing, selecting, commissioning and modifying software, the principles of software ergonomics must be applied. These requirements came into force in December 1992.

Results of usability assessments should be fed back to designers in a meaningful way, so that they may progress with appropriate design options. It is often stated, as a criticism of human factors work, that the results of evaluations are not presented to designers in a way that enables the recommendations to be readily implemented in terms of improvements to the product. Methods such as video records and verbal records can provide useful edited highlights which may demonstrate the kinds of misconceptions and problems which users experience with the product. This feedback is useful at a gross level, demonstrating where difficulties arise during the interaction. Although some indication of problems with the interface or functionality is valuable, it does not show specifically how the product may be altered or fine-tuned to overcome them. The onus is on the evaluator to bridge the gap from raw evaluation data, to actual interpretation, diagnosis and recommendations which the designer can act upon.

6. Computer-based support for usability evaluation

So far the emphasis has been on gathering a comprehensive range of data streams for cross-referencing during analysis. However, given a rich source of interaction

data, emphasis must also be placed on facilitating the management of evaluations and data reduction processes.

One of the most important functional requirements of a computer-based evaluation system which manages data streams, such as digital or video-recordings, is the capability to review them selectively in order to extract relevant insights which reflect on the evaluation goals being examined. Functions which allow the evaluator to define sections of the data, to be skipped during replay or played at a slow speed, are useful ways of reducing time and effort involved in analysis. The event log could be further analysed to determine the frequency and duration of relevant events.

7. Conclusions

Although the advent of usability evaluation laboratories has meant a greater sophistication in the level of product analysis, the process itself has also become extremely expensive. There is a tendency to invest in the equipment side of the laboratory and to neglect the methodology side. Many evaluations have become technology led, where the kit is used simply because it is available. The evaluation then becomes less goal-directed and inadequate attention is devoted to the formulation of research questions within an appropriate theoretical framework.

The aim of this paper has been to present a framework which describes approaches to evaluation which may be adopted at particular times with questions in mind. It is hoped that the framework facilitates the process of evaluation by mapping the approach which is appropriate onto the given context for any particular evaluation.

The authors would like to express their gratitude to Clive Johnson, Andrew Dillon and Arthur Gardner of The HUSAT Research Institute, for their insightful comments on drafts of this paper.

The stimulus for some concepts developed in this paper came from the Alvey HIMS Project and we acknowledge the value of many discussions with our collaborators in that project.

References

AL-ISA, H. H. (1990). *Experimental activities time recording and analysis system (EATRAS).* MSc Project, Interactive Computer Systems Design, Department of Computer Studies, Loughborough University of Technology, Loughborough, UK.

BELLOTTI, V. (1988). Implications of current design practices for the use of HCI techniques. In D. M. JONES & R. WINDER, Eds. *People and Computers IV. The BCS HCI Special Interest Group.* pp. 13–34. Cambridge: Cambridge University Press.

BOSSER, T. & MELCHIOR, E-M. (1991). The SANE toolkit for cognitive modelling and user-centred design. In M. D. GALER & J. ZIEGLER, Eds. *Methods and Tools for User Centred Design in Information Technology,* pp. 93–125. Amsterdam: Elsevier.

BROWN, C. M. (1988). *Human–Computer Interface Design Guidelines.* NJ: Ablex.

CARD, S. K., MORAN, T. P. & NEWELL, A. (1983). *The Psychology of Human–Computer Interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

CHIN, J. P., DIEHL, V. A. & NORMAN, K. L. (1988). Development of an instrument measuring user satisfaction of the human–computer interface *CHI'88 Conference Proceedings, Human Factors in Computing Systems,* pp. 213–288. ACM: Washington, DC.

COUNCIL DIRECTIVE OF MAY 1990. *Official Journal of the European Communities*, (90/270/EEC) No. L156/14.

ERICSSON, K. A. & SIMON, H. A. (1980). Verbal reports as data. *Psychological Review*, **87**, 215–251.

ERICSSON, K. A. & SIMON, H. A. (1983). *Protocol Analysis*. Cambridge MA: MIT Press.

GALE, A. & CHRISTIE, B., Eds. (1987). *Psychophysiology and the Electronic Workplace*. Chichester: Wiley.

HART, S. G. & STAVELAND, L. E. (1988). Development of NASA-TLX (task load index): results of empirical and theoretical research. In P. A. HANCOCK & N. MESHKATI, Eds. *Human Mental Workload*, pp. 139–183. Amsterdam: North Holland.

KAK, A. V. (1981). Stress: an analysis of physiological assessment devices. In G. SALVENDY & M. J. SMITH, Eds. *Machine Pacing and Occupational Stress*, pp. 135–142. London: Taylor and Francis.

KIERAS, D. E. & POLSON, P. G. (1985). An approach to the formal analysis of user complexity. *International Journal of Man–Machine Studies*, **22**, 365–394.

KRIPPENDORFF, K. (1980). *Content Analysis, an Introduction to its Methodology. The Sage Commtext Series*. London: Sage publications.

MAGUIRE, M. & SWEENEY, M. (1989). System monitoring: basis for comprehensive evaluation system? In A. SUTCLIFFE & L. MACAULY, Eds. *HCI'89, People and Computers V, The BCS HCI Special Interest Group*, pp. 375–394. Cambridge: Cambridge University Press.

MORAN, T. P. (1981). The command language grammar: a representation for the user interface of interactive computing systems. *International Journal of Man–Machine Studies*, **15**, 3–50.

MORRIS, D., THEAKER, C. J., PHILLIPS, R. & LOVE, W. (1988). Human–computer interface recording. *The Computer Journal*, **31**, 437–444.

OPPENHEIM, A. N. (1966). *Questionnaire Design and Attitude Measurement*. New York: Basic Books.

PAYNE, S. J. (1984). Task-action grammars. In B. SHACKEL, Ed. *INTERACT'84. 1st IFIP Conference on Human Computer Interaction*, pp. 527–532. Amsterdam: Elsevier Science.

RAVDEN, S. & JOHNSON, G. (1989). *Evaluating Usability of Human–Computer Interfaces*. Chichester: Ellis Horwood.

REISNER, P. (1981). Formal grammar and the design of an interactive system. *IEEE Transaction on Software Engineering*, **7**, 229–240.

SHACKEL, B. (1990). Usability—context, framework, definition, design and evaluation. In B. SHACKEL & S. RICHARDSON, Eds. *Human Factors for Informatics Usability*, pp. 21–37. Cambridge: Cambridge University Press (in press).

SMITH, S. L. & MOSIER, J. N. (1984). *Design Guidelines for User System Interface Software*. ESD-TR-84-190, MITRE Corporation.

THEAKER, C. J., PHILLIPS, R., FROST, T. M. E. & LOVE, W. R. (1989). HIMS: a tool for HCI evaluations. In A. SUTCLIFFE & L. A. MACAULAY, Eds. *People and Computers V. The BCS HCI Special Interest Group*, pp. 427–442. Cambridge: Cambridge University Press.

WALSH, P. A. & LAWS, J. V. (1990). Methods and tools in industry for mobility testing: the STL usability evaluation workbench. *BCS HCI Special Interest Group Meeting*, Camden Town, London, 21 May 1990.

WHITESIDE, J., BENNETT, & HOLTZBLATT, K. (1988). Usability engineering: Our experience and evolution. In M HELANDER, Ed. *Handbook of Human–Computer Interaction*, pp. 791–817. Amsterdam: Elsevier.