EVALUATING PERFORMANCE FOR PROCUREMENT: A STRUCTURED METHOD FOR ASSESSING THE USABILITY OF FUTURE SPEECH INTERFACES

by

Martin Andrew Cruickshank Life

University College London

Submitted for the degree of Doctor of Philosophy in the University of London

August, 1991

ABSTRACT

Procurement is a process by which organizations acquire equipment to enhance the effectiveness of their operations. Equipment will only enhance effectiveness if it is usable for its purpose in the work environment, i.e. if it enables tasks to be performed to the desired quality with acceptable costs to those who operate it. Procurement presents a requirement, then, for evaluations of the performance of human-machine work systems. This thesis is concerned with the provision of information to support procurers in performing such evaluations.

The Ministry of Defence (an equipment procurer) has presented a particular requirement for a means of assessing the usability of speech interfaces in the establishment of the feasibility of computerized battlefield work systems. A structured method was developed to meet this requirement, the scope, notation and process of which sought to be explicit and proceduralized. The scope was specified in terms of a conceptualization of human-computer interaction: the method supported the development of representations of the task, device and user, which could be implemented as simulations and used in empirical evaluations of system performance. Notations for representations were proposed, and procedures enabling the use of the notations.

The specification and implementation of the four sub-methods is described, and subsequent enhancement in the context of evaluations of speech interfaces for battlefield observation tasks. The complete method is presented. An evaluation of the method was finally performed with respect to the quality of the assessment output and costs to the assessor. The results suggested that the method facilitated systematic assessment, although some inadequacies were identified in the expression of diagnostic information which was recruited by the procedures, and in some of the procedures themselves.

The research offers support for the use of structured human factors evaluation methods in procurement. Qualifications relate to the appropriate expression of knowledge of device-user interaction, and to the conflict between requirements for flexibility and low-level proceduralization.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I begin by expressing my appreciation of Professor John Long for his support, not only as a supervisor, but as a friend and colleague. I thank him for offering me the opportunity to study for a PhD at the Ergonomics Unit and for his invaluable advice throughout. However, the thesis could not have been completed without the help of a number of others. I acknowledge the fundamental contribution of the Royal Signals and Radar Establishment, for allowing me to exploit research I performed under Contract 2047/127 (RSRE) as the basis for my thesis[1]. The Science and Engineering Research Council subsequently granted me a studentship, enabling me to complete the research and report it.

Considerable personal debts are owed to Dr. R. Pratt, Mr. S. Roberts, Mr. C. Harding and their staff at RSRE, for their guidance and encouragement while they were involved with the research. At the Ergonomics Unit, Mr. B. Lee provided me with technical assistance over this period, but, perhaps more importantly, his cheerful outlook and friendship helped me overcome the problems that inevitably arise in research work. I must also remember Dr. P. Seymour, without whose ingenuity in software design the experiments would likely never have been implemented.

I am grateful to all those who were directly or indirectly involved in the empirical work: to the staff and students of the Royal School of Artillery at Larkhill, who made possible observations of military tasks; to Birkbeck College (and particularly Dr. J. Laws) for the use of video analysis facilities; to subjects participating in experimental studies; and to Mrs. M. Freedman, for assistance in data analysis. Particular thanks are due to Professor R. King and his staff at the Royal Military College of Science (Cranfield) at Shrivenham, for providing an opportunity to evaluate the assessment method, and to Dr. C. Selby for allowing me to observe her using the method.

Finally, I am indebted to my family and friends, without whom this enterprise could never have been entertained. Particular thanks to my father, Mr. E.A. Life, for his interest, support and encouragement throughout; to colleagues at the Ergonomics Unit (and especially Dr. A. Whitefield), for their many comments and suggestions; and to my friends, at home and abroad for their moral support during the more difficult times. Thank you all.

[1] Copyright for the method described in Chapters 7 to 10 of this thesis resides with RSRE, St. Andrews Road, Great Malvern, Worcestershire.

CHAPTER 1

INTRODUCTION

1.1 Determinants of the performance of human-computer work systems

The British Army is introducing a computer - the Battlefield Artillery Target Engagement System (BATES) - to support the activities of the Royal Regiment of Artillery (RA). The new equipment is intended to improve the effectiveness of the control of artillery resources and the speed of artillery response (Ward and Turner, 1982). It is of concern both to the army and to those who will finance the introduction of BATES that the new system will have the desired effect on the performance of the regiment.

Organisations, such as the RA, may be viewed as systems seeking to achieve goals by performing work. The (ultimate) goal of the RA is the defence of British sovereign interests, and its work relates to the destruction of enemy targets to the benefit of the army or its allies. Members of organisations may use machines to effect their work, so forming human-machine work systems. The RA is such a work system, its members utilizing machines to support their battlefield tasks (e.g. instruments for survey and observation, devices for processing and communicating information, and weapons).

The performance of a system is determined by the behaviour of its components. In the case of human-machine work systems, performance will be determined by the behaviour of its human elements and of its machine elements as the system performs its tasks. A particularly important class of behaviour will be the *interaction* between these various elements.

The effect of BATES on the performance of the RA will be determined, then, by the behaviour of the BATES computer and by the behaviour of artillerymen. It will also depend upon how the computer and soldiers *interact* as they carry out battlefield tasks, such as the resourcing of batteries and the engagement of individual targets. Target engagement performance depends, currently, upon the behaviour of a battlefield observer (locating targets accurately), the behaviour of a radio communication device (transmitting target information to those controlling weapons), and the interaction between device and user (the observer speaking into the radio microphone). In future, system[2] performance will still be determined by the observers locating targets, but then by the behaviour of a *computer* in the transmission of information, and by the interaction between the observer and a computer terminal. The

[2] Unless stated otherwise, the term "system" is used in this thesis to refer to a device and its user interacting together, and not to the device (e.g. computer) alone.

success of the technological change will depend upon whether these behaviours can support the performance required of the artillery system.

This thesis is concerned with means by which those responsible for the introduction of computer technologies to military human-machine systems can evaluate human-computer interaction behaviour. It is assumed that such evaluation can contribute to the development of computerised systems exhibiting enhanced performance.

1.2 Supporting human factors evaluations in military procurement

1.2.1 Military procurement

One of the functions of the U.K. Ministry of Defence (MoD) is to procure equipment for the armed forces. In performing its procurement role, the MoD's objective is "... to acquire for the armed forces the equipment they need to maintain their effectiveness and credibility against the developing threat, to acquire that equipment when it is needed and to do so at a price that can be afforded within the resources available" (Levene, 1987).

The means by which MoD performs its function may be subject to variation due to changes in government policy, but during the 1980s it has been characterized as being primarily executive in nature. The development of equipment has been performed largely by industrial contractors, and the role of the procurer has been to specify requirements for the customer (e.g. the army); to evaluate proposed solutions from competing contractors; to monitor the development process; and to ensure that the delivered equipment meets the requirement (Ministry of Defence, 1987).

Although MoD does not directly design and implement systems, the performance of its procurement function requires the evaluation of systems. Evaluation is required in the identification of inadequacies of an existing system; diagnosis and specification of requirements for a new equipment; in the selection of a viable proposed solution; and in the trialling of the equipment. Because development costs are particularly increased by requirements to rectify design faults after implementation, the military procurement process makes provision for the early evaluation of design options in *feasibility assessments*. Such assessments demand the prediction of the behaviour and performance of systems when they are finally implemented.

1.2.2 The discipline of human factors

Human factors (HF) is a discipline concerned with the development of effective human-machine systems. Other engineering[3] disciplines (e.g. software engineering, mechanical engineering) may frequently share this objective, but HF is distinctive in that its concern is particularly with the behaviour of the human as this relates to the machine. This may be contrasted with the more common situation in engineering, where the focus is on the behaviour of the machine as this relates to people (see Dowell and Long, 1989).

The discipline knowledge supporting HF is embodied in techniques enabling contributions to system development and in knowledge of human behaviour recruited by the techniques. Such knowledge is acquired by HF engineers by training, and it may be refined by experience as knowledge is applied in practice. In addition to the knowledge held by individual HF specialists, academic and industrial research has sought to develop a body of generalizable knowledge relating to machine-user interaction, intended to be directly recruitable to the activities of system development (e.g. Smith and Mosier, 1986, Gardiner and Christie, 1987). Less directly, applied scientific research in anatomy, physiology and psychology may also be used to predict the behaviour and performance of human-machine systems (e.g. Card, Moran and Newell, 1983), and to enhance the design of such systems (e.g. Hammond and Allinson, 1988).

The discipline knowledge of HF is potentially applicable to the procurement of military systems, by supporting the acquisition of systems which exhibit effective human-machine interaction.

1.2.3 The evaluation of system behaviour and performance

Engineers conduct evaluations to determine whether a system will support the performance required of it. The various engineering disciplines may assume differing criteria in effecting evaluations. The criteria employed in HF relate to the adequacy of the system in its support for human behaviour. System development presents requirements for diagnosis and prescription, such that failures to meet criteria are interpreted with respect to a body of engineering discipline knowledge and changes to the system prescribed according to this knowledge.

Because military procurement promotes change within human-machine systems by the introduction of new equipment, one of the procurer's concerns lies with the evaluation of the interaction between equipment and its users, and of its impact on the performance of the system (e.g. U.K. Defence Standard 00-25 Part 12, 1987; Meister, 1986). One way that HF can

[3] The status of the discipline of HF is not at issue at this point: for the purpose of the discussion its contribution to system development is assumed to be equivalent to those of the established engineering disciplines.

potentially contribute to the process of military procurement is by supporting the evaluation function; for example, it may contribute to the diagnosis of failures of human-machine systems to meet the performance required of them and to the prescription of interventions to optimize interaction behaviour.

1.3 This thesis

1.3.1 The problem to be addressed

The effectiveness with which HF discipline knowledge has the potential to contribute to discipline practice is determined largely by the completeness of the knowledge and by its accessibility to those who use it. This thesis addresses part of the problem of expressing HF discipline knowledge to support early system evaluation within military procurement.

The thesis describes the development of a particular embodiment of HF discipline knowledge - a structured evaluation method, the scope, process and notation of which are explicitly conceptualized and proceduralized (Silcock, Lim and Long, 1990). The structured method is offered as a solution to a particular problem in procurement - a need for individuals lacking specialist knowledge of HF to determine the usability of speech interfaces as part of the assessment of system feasibility. This particular case is used to advance the more general argument in favour of structured evaluation methods as means of enhancing the effectiveness of HF evaluations. The contribution of the work is intended to be, then, both specific (i.e. the development of a method to support a particular type of HF evaluation task conducted by a military procurer) and general (i.e. an advancement in the way HF discipline knowledge might be expressed to support procurement effectively).

1.3.2 Thesis structure

The remainder of the thesis may be viewed as comprising three parts. The first part consists of this chapter and Chapters 2, 3 and 4, which identify a general problem of ineffectiveness in procurement and present the rationale for methodological support in order to solve it. Chapter 2 elaborates on the process of military procurement and on the mechanism of its conduct. In particular, it describes the temporal phases of procurement activity, and identifies the requirement for the early assessment of system feasibility. Chapter 3 reviews means by which engineering discipline knowledge may be captured such that it is available to the practitioner. The discipline of HF is distinguished from other disciplines concerned with the design of effective human-computer systems. Chapter 4 considers the evaluation of human-machine interaction performance and approaches to the assessment of systems. Structured HF methods supporting empirical evaluations of feasibility are identified as one means of enhancing the effectiveness of procurement.

The second part describes a particular instance of the general problem - potential ineffectiveness in the procurement of systems involving speech interfaces - and one solution to this problem - a structured method for assessing the usability of speech interfaces. Chapter 5 analyses the requirement, while Chapter 6 presents a rationale for the structured method, contextualized with respect to the background presented in Part 1. Chapters 7, 8 9 and 10 describe the procedures and notation of the method, with illustrations of its application.

The final part of the thesis evaluates the contribution of the research, both in its solution of the specific problem of assessing the usability of speech interfaces and more generally. The method was tested in the context of a small procurement project, and this evaluation is described in Chapter 11. Chapter 12 draws conclusions on the basis of the results of the evaluation, and it considers implications of the research for procurement, for HF evaluation and for the further development of structured evaluation methods to support HF practice.

CHAPTER 2

MILITARY PROCUREMENT

2.1 Introduction

This chapter describes the process which organizations undertake to acquire systems enabling them to operate more effectively: procurement. Its particular concern is with the procurement undertaken by military organizations, such as that comprising the U.K. armed services and their associated government department, MoD. The chapter begins by analyzing the requirement for organizations to acquire systems which will increase their effectiveness. Section 2.3 describes the procurement function and identifies the agents involved in the process (system developers, users and procurers). The process itself is described in Section 2.4; to facilitate administration, MoD divides the process into stages, distinguished by the activities undertaken by the procurer as system development proceeds. Military users impose novel and rigorous demands on machines, with the consequence that military procurers are intimately involved at all stages of the process, and particularly in the identification of the requirements for systems and in evaluating designs.

Section 2.5 presents evidence of ineffectiveness which has led MoD to identify a need for the better assessment of system feasibility at early stages of procurement. The chapter concludes with evidence suggesting that feasibility assessments should be rendered more accurate by the use of experimental methods of evaluation. This evidence is used later to support the argument for a structured method for empirical evaluation.

2.2 Human-machine systems within organizations

2.2.1 Machines to enhance organizational effectiveness

Organizations seek to bring about change to aspects of the world in which they operate, and they do this by performing work. The issue of the nature of work and the criteria for evaluating its consequences will be explored more fully in Chapter 4; for now it will be asserted that the products of work exhibit an attribute of *quality*, and that sustaining a desired level of output quality imposes *costs* on elements comprising the organization (Dowell and Long, 1989). In this discussion, "quality" will be taken to mean those attributes of a product which enable it to fulfil its functions, so it will include attributes of quantity, consistency and timeliness, as well as notions of fitness for purpose. Although "costs" to an organization will be expressible ultimately in financial terms, the concept is used here in the sense of the depletion of immediate resources, both material and human. The quality of the products of work and the costs incurred achieving the required quality, together, constitute the *performance* of the organization.

Organizations which produce physical artefacts (i.e. manufacturers) will make explicit their objectives with respect to output quality. They will seek to fulfil those objectives by the control of production volume and by the operation of quality assurance mechanisms; for example, manufacturers typically employ product testers and inspectors to ensure that products meet their intended specifications. For other types of organization, such as those offering services, quality requirements may be implicit, and actual output quality only open to indirect assessment; for example, the quality of the work of a police force might be reflected in the general levels of reported crime in its operating area. In general, however, the objectives of organizations will refer implicitly or explicitly to operation to some standard of quality.

One definition of organizational effectiveness relates the quality of output to the costs incurred in achieving it (Dowell and Long, 1989). Improved effectiveness might be achieved either by an enhancement of quality without the incurrence of additional costs, or by the maintenance of quality with reduced costs. In the case of a police force, then, improved effectiveness might be achieved by reducing crime in its area without utilizing extra resources, or by maintaining crime at existing levels with fewer resources. One way that an organization is able to improve its effectiveness is by acquiring machines to support the work of people within the organization.

2.2.2 Machines to enhance military effectiveness

In times of conflict, military organizations seek to gain supremacy over equivalent organizations representing an enemy. Supremacy may be achieved by direct action, for example, by an armed force gaining possession of territory held by an enemy through the destruction of the enemy's forces; or, indirectly, by a force exhibiting such *potential* superiority that its enemy surrenders its territory without fighting. In peacetime, military organizations seek to deter aggressors by making known their potential superiority over the aggressor's forces. The (actual or potential) superiority of a force in an armed conflict is, other things being equal, a measure of the (actual or potential) quality of the work of the force.

One way of achieving military superiority is by the recruitment of a larger force than that of the enemy. Such a force can sustain attrition for a longer period and may overwhelm its opposition by weight of numbers. An alternative strategy is to improve the effectiveness of a smaller force by the provision of machines to enhance the quality and/or reduce the costs to its members when fighting. In the case of the Royal Artillery, BATES is intended to improve quality by, for example, increasing the speed of artillery response and improving the efficiency with which resources are deployed; costs might be reduced by eliminating the need for individuals to keep track of resources by means of manual accounting procedures.

In the latter half of the twentieth century, members of the North Atlantic Treaty Organization (NATO) - a military organization - has pursued the second strategy of technological investment in response to a threat posed by a military organization perceived as an enemy - the Warsaw Pact (WP)[1]. The political system of the WP has enabled it to maintain very large military forces relative to NATO. One means by which equity has been maintained between the protagonists is by NATO investing heavily in advanced technology, such as computers, which will offer its end users a fighting advantage over those of the WP.

2.3 The procurement of computer-based worksystems

Although the term is perhaps most commonly used in the context of large organizations, the process by which machines are acquired to meet organizational performance requirements, within available resources and at the right time, is termed "procurement". In a small organization, the people who actually use the procured machines to support their work ("end users") may also be the procurers, whereas in a large organization (such as the military) there will tend to be greater specialization and, hence, partitioning of the activities of the procurers and of end users. In large organizations a procurement sub-system may operate to acquire machines, distinguishable from end user sub-systems. Nevertheless equivalent activities are likely to be undertaken in both small and large organizations, and functional distinctions may be made generally between the end user and the procurer.

Organizations which use computers to support their operations ("user organizations") may be distinguished from organizations which develop and produce computers ("developers"). For the purposes of this discussion, organizations of the first type will be taken to include external consultants who act as advisors for the user organization in the selection and installation of computers. "Developers" will be taken to include retail agents who distribute, market and maintain computers, as well as those who design and manufacture them. A procurer will seek to acquire from a developer machines which will enable the user organization to meet its desired performance criteria within an available budget. The performance criteria will include the required product quality and costs, both organizational (such as maintenance requirements) and costs to be incurred by individual end users. The

[1] At the time of writing the threat of the WP is receding as a consequence of political change in Eastern Europe; however, new threats have emerged in the Middle East. The general argument concerning the military strategy of the Western nations in response to threats continues to be relevant.

procurer, then, acts as the agent of the end user, ensuring that, within other organizational constraints, the equipment acquired meets the requirements of the user.

MoD is the organization responsible for the U.K. contribution to NATO. MoD provides the mechanism of control of the UK armed forces and maintains their effectiveness. One of its functions, then, is to ensure that its forces are adequately equipped. Fighting members of the three armed services (army, navy and air force) constitute end-users of military equipment, and MoD supports a procurement sub-system - the Procurement Executive (PE) - to acquire machines enabling the services to maintain an adequate fighting capability (see, for example, Levene, 1987).

2.4 The process of procurement

2.4.1 Procurer involvement in product development

The establishment of an operational system in a user organization may involve interactions between the processes of development and procurement. The typical sequence of system development activities has been documented elsewhere (e.g. Sommerville, 1985); for the purposes of this discussion it will be assumed to exhibit seven phases:

- user requirements analysis and definition
- preliminary design specification
- detailed design specification
- implementation
- evaluation[2]
- production
- maintenance

The extent and form of the interaction between development and procurement will be determined by the novelty of the user organization's requirement and by the extent to which fulfilment of the requirement demands investment in the development process. Where the user organization's requirement is a common one for which machines have been developed previously - say, a requirement for word processing facilities for secretarial staff - the involvement of the procurer is likely to be small and indirect. For example, as a potential customer within a general market, the procurer might incidentally contribute data to the developer's market research, and hence to the establishment of a general user requirement. Given the availability of a range of potentially suitable products, the procurer's objective

[2] Evaluation is recognized to take place throughout development. The phase explicitly designated *evaluation* includes activity in which the implemented version of the system is tested against the original requirement

will be the evaluation of alternative systems on the market, in order to select the one best suited to its requirement. Modification to suit the needs of end users, if required at all, is likely to be carried out by the user organization, rather than by the developer.

Requirements of user organizations for medium and large scale computerized systems (e.g. company accounting systems), while having characteristics in common with the requirements of other user organizations, will typically have unique features. Under such circumstances, the developer may be required to adapt extant systems to match the performance required by the user organization ("variant design"). The procurer is likely to be involved in the specification of requirements to enable customizing by the developer (e.g. so that the new system is compatible with the user organization's existing manual system); in the selection from alternative detailed design options; and in evaluation prior to acceptance.

The involvement of the procurer in system development becomes greater still where the requirement is novel and where existing products cannot fulfil it. The agreement of the requirement and the selection of an appropriate solution will place an important demand for effective interaction between the developer and the procurer. Furthermore, under such circumstances, the procurer may be required to make a substantial investment of resources (financial and other) in the development process and will, consequently, have a direct financial interest in its outcome.

Military threats evolve continuously. Enemies will identify and exploit weaknesses in their opponent's capabilities, so there is a requirement for military organizations to monitor their own performance and to negotiate the acquisition of new machines when necessary. Because the WP presents a large and sophisticated potential opposition, the maintenance of adequate performance requires MoD to exploit the WP's weaknesses in novel ways, by making technologically advanced devices available to the U.K. armed services. Very often, existing products will not meet service requirements, and MoD finds it necessary to invest substantial financial resources in the development of new products and even new technologies: £9 billion was spent by PE on equipment procurement in 1985/86 (Jordan, Lee and Cawsey, 1988). The procurement process of MoD consequently exhibits intimate links with the process of product development.

2.4.2 The UK military procurement process

MoD (1987) characterizes procurement in terms of seven phases, which may occur following the informal identification by users (i.e. UK armed services) of a requirement for equipment:

- Concept Formulation
- Feasibility
- Project Definition

- Full Development
- Production
- In-service
- Disposal.

These phases are distinguished by their inputs and outputs, and by the activities of the procurer, now briefly summarised.

(a) *Concept Formulation.* Concept Formulation translates the user's needs into a formal demand on the PE to initiate the development of a new system, expressed as a Staff Target. Generation of the ST requires an exchange of views between the user, the PE (which must evaluate the requirement with respect to additional organizational considerations) and technical specialists within MoD and industry.

(b) *Feasibility.* Following acceptance of the ST by a superordinate committee within MoD, an assessment is made "to establish technical feasibility, cost, duration, risk and demand on resources". The assessment may be achieved by analytic or empirical methods, and may be undertaken either by MoD research establishments or by external organizations (sometimes working in competition). The product of the assessment is a feasibility study report which is used by the PE, either as a prompt to seek further information or to create a formal Staff Requirement, which includes plans for subsequent development activities for submission for higher approval.

(c) *Project Definition.* Approval of the Staff Requirement enables a detailed planning phase, in which evaluations of the technical solutions made during Feasibility are verified, performance requirements are set and the outline design specification is formulated. At this stage "realistic assessment" is made of the cost and duration of development, and estimates of the cost of the machines in production. Project Definition is normally performed under the authority of technical and administrative managers within MoD by the contractor who will undertake subsequent development and production.

(d) *Full Development.* Full Development constitutes the detailed design of the machines and implementation to enable evaluation with respect to the Staff Requirement. It is normally undertaken by the contractor who has undertaken Project Definition according to the previously made plans.

(e) *Acceptance.* Acceptance is admitted by the PE when it is satisfied that the developed solution "...meets the Staff Requirement and is suitable for service use".

(f) *Production.* Although the Production phase essentially succeeds Development, in practice some aspects of production may be undertaken when the product is in the final stages of its development in order to meet operational requirements. The procurement function during this phase, therefore, demands evaluation of the risks of advancing the manufacture of elements of the system. More generally, the concern is to ensure a smooth transition from Development into Production, such that the performance of the product is maintained with economy of manufacture.

(g) *In-service.* The procurer is concerned with monitoring the performance of the implemented system and with initiating action to make limited modifications to the design, if necessary. The procurer also ensures the provision of training and maintenance facilities to the user.

(h) *Disposal.* The PE advises the user on the disposal of machines which are no longer useful, establishing that they cannot fulfil requirements presented elsewhere in the military organization.

The procurement process bears similarities to that of development but is super-ordinate to it. The potential for the procurer to influence the system development process is, in the case of UK military procurement, very strong. PE has a technical involvement in concept formulation, feasibility assessment and, under the present scheme, in preliminary design. PE further controls the progress of development by the formal evaluation of the products of each phase. It has been intended that this close involvement between developer and procurer will ensure the fulfilment of MoD's performance requirements within the constraints on resources set by Government.

2.5 Enhancing the effectiveness of UK military procurement

Although PE supports a large and highly skilled administrative and technical workforce, some procurement projects have, in recent years, been subject to criticism of their management performance. Part of this criticism may be attributed to a general political philosophy of the Conservative administration, which has sought to maximize the cost-effectiveness of the operations of government. However, in some notable cases, there has been a failure on the part of the PE to recognize the over-ambitiousness of technical solutions until there has been an unacceptably high expenditure of MoD resources. Such a case is the Nimrod Airborne Early Warning (AEW) system, which failed to meet the user's technical requirements and which was cancelled in 1986, despite very large government expenditure on its development. The Nimrod AEW project must be viewed as an instance of ineffective procurement.

In recognition of such failures, the emphasis of the PE's role has been changing (Levene, 1987). Increased emphasis is being placed on competition in tendering for contracts. It is further assumed that one way of improving the likelihood that design solutions will be feasible is to transfer the responsibility for product specification more completely to the developer. This has been achieved by the introduction of "Cardinal Points Specifications" (CPSs), which are generated by PE at the Concept Formulation stage. Rather than specifying in detail the structure and behaviour of the desired system which is to be implemented by an industrial contractor, the CPS sets performance requirements and broad design constraints. It is then left to the contractor to decide how to fulfil the requirement and to produce the detailed specification.

Although delegation of technical responsibility has the potential to enhance the effectiveness of procurement, there remains a continued requirement for PE to perform technical evaluations throughout the procurement process. If a technical solution to a requirement is weak, it is important that the weakness is recognized as early as possible. Although, in principle, changes may be made to the design of a system throughout the procurement cycle, the cost of implementing modifications increases as development proceeds (Fairley, 1986). The primary reason for the increase in cost is that modification of one system component is likely to have implications for the design of others which interact with it. As a consequence, even a small modification to one component can result in a requirement for widespread redesign of the system as a whole.

In the case of U.K. military procurement, Jordan, Lee and Cawsey (1988) have reported that "£3-4B of each year's equipment budget may be associated with costs which were not foreseen when projects started; about £1-2B may be associated with costs not foreseen when projects entered Full Development". Although there had been earlier recognition that 15-25% of development costs should be spent in Feasibility and Project Definition (Downey, 1966; Rayner, 1971), UK practice has been such that on certain major projects only 8% of development costs have been apparently spent prior to Full Development, and only 1% during Feasibility. The clear implication is that greater resources should, in future, be applied to feasibility assessment.

Jordan et al suggest that the technical difficulty of projects typically becomes manifest only when equipment elements are instantiated and integrated, and they argue that in technologically novel systems the judgement of feasibility on the basis of analysis alone is over-optimistic. They report: "A key characteristic of a successful development programme is that it is based on experience learned from practical work with hardware (or software) and integration". Jordan et al recommend an expanded role for the experimental feasibility study, such that commitment to a Staff Requirement only occurs after a demonstration of the hardware of the system and its integration (see Figure 2.1).

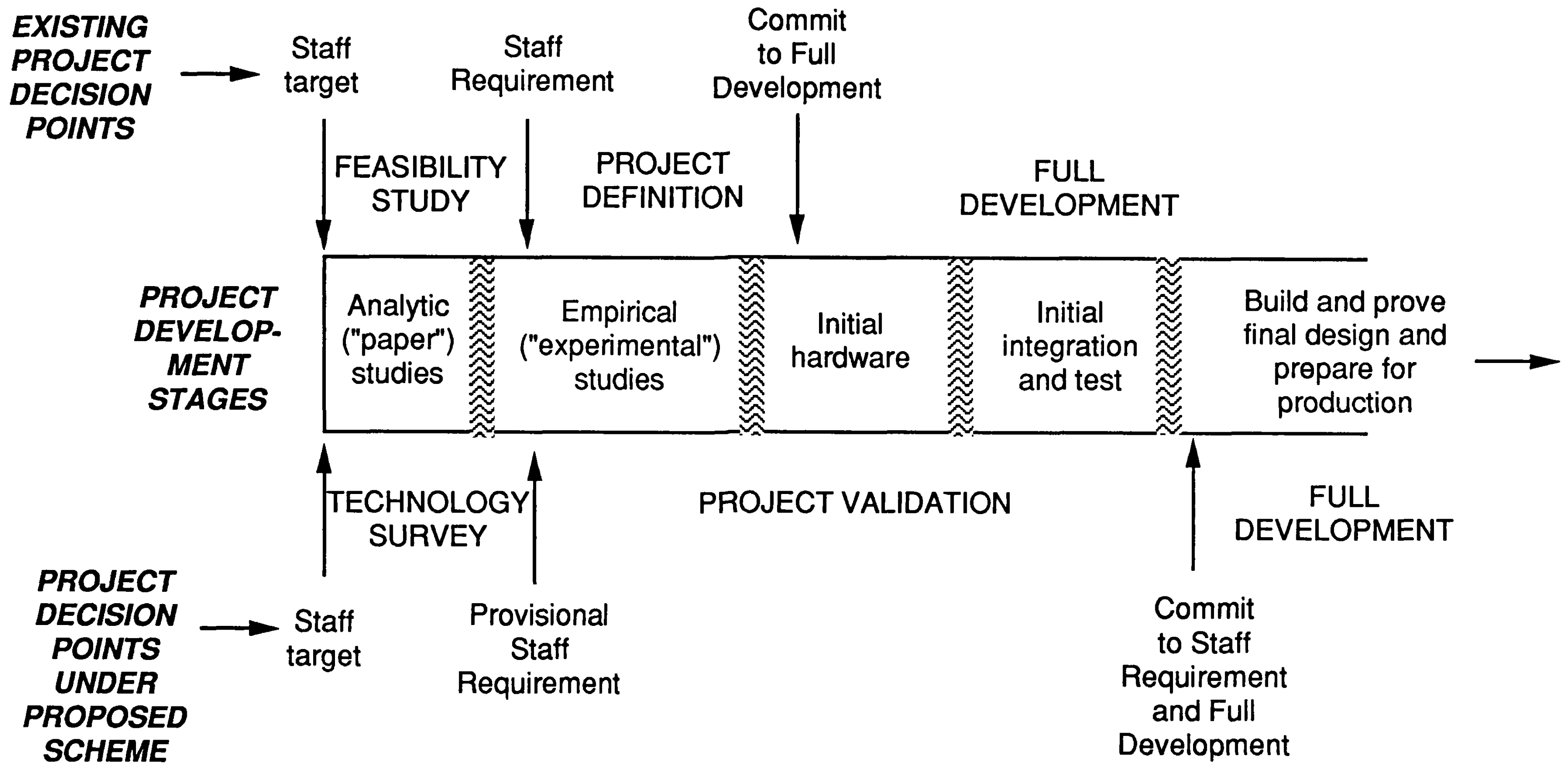

EXISTING PROJECT DECISION POINTS
Staff target
Staff Requirement
Commit to Full Development
FEASIBILITY STUDY
PROJECT DEFINITION
FULL DEVELOPMENT
PROJECT DEVELOP-MENT STAGES
Analytic ("paper") studies
Empirical ("experimental") studies
Initial hardware
Initial integration and test
Build and prove final design and prepare for production
TECHNOLOGY SURVEY
PROJECT VALIDATION
FULL DEVELOPMENT
PROJECT DECISION POINTS UNDER PROPOSED SCHEME
Staff target
Provisional Staff Requirement
Commit to Staff Requirement and Full Development

Figure 2.1: Stages in U.K. military procurement, showing modifications proposed by Jordan et al, 1988

This chapter has reviewed research which identifies, then, a source of ineffectiveness in U.K. military procurement. MoD and its sub-contractors sometimes fail to predict the technical ambitiousness of projects at an early stage, with the result that projects either subsequently fail for technical reasons; or that they result in unforeseen (and unacceptable) additional expenditure. A solution to this problem, identified by MoD, is to conduct empirical system feasibility assessments prior to full development.

Human factors is one of the disciplines which might contribute to the assessment of system feasibility, by predicting whether performance will be compromised by failures in the interaction between machines and their users. Chapter 3 now considers the sources of human factors knowledge which might be recruited to the procurement of military systems.

CHAPTER 3

SUPPORTING THE PRACTICE OF HUMAN FACTORS

3.1. Introduction

Human factors practitioners can contribute to the development and procurement of effective human-machine systems; the present chapter is concerned with the provision of human factors knowledge to support the practitioner. The chapter begins by defining the concept of the "discipline". The discipline of human-computer interaction (HCI) includes the sub-disciplines of human factors (HF) and software engineering (SE). Discipline knowledge is ascribed a status according to the means by which it is derived and expressed; most HF discipline knowledge is ascribed "craft" status, because it is informal and expressed as heuristics. Section 3.3 identifies classes of HF practitioners, and it considers alternative ways in which HF knowledge may be made available to practitioners.

It is observed that the contribution of HF to system development and procurement is not as effective as might be hoped. Part of the ineffectiveness is due to the incompleteness and informality of existing knowledge of human-computer interaction and to its inaccessibility to many practitioners. In system development, one solution to the lack of such knowledge lies in explicit and generalizable methods for solving problems. Structured analysis and design methods (SADMs) are now being extended so that HF concerns are taken into account; Section 3.4 describes such methods and reviews the support they offer to various types of practitioner.

Structured methods have potential for enhancing the effectiveness of HF's contribution to military procurement. However, while extended SADMs may support those with a background in HF, they are not well-suited to the needs of individuals lacking knowledge of HF. This observation forms part of a later argument for structured methods which are supported by knowledge of human-computer interaction. Such methods could enable non-specialist procurers to conduct HF evaluations.

3.2 The discipline of human factors

3.2.1 Human factors as a discipline supporting system development

Long and Dowell (1989) define the term "discipline" as "the use of knowledge to support practices seeking solutions to a general problem having a particular scope". In terms of this definition, the discipline of HCI addresses the general problem of "designing humans and computers which interact to perform work effectively". HF and SE may be considered sub-disciplines contributing to the solution of the general problem of HCI. A developer may primarily seek to design a system element with respect to human behaviour: alternatively, the developer may primarily seek to design with respect to computer behaviour. These contrasting orientations reflect the distinction between the disciplines of HF and SE as they contribute to system development.

To illustrate this distinction, consider the design of a word processor to support secretarial work. An SE approach to ensuring that spelling accuracy is of an acceptable standard might be to incorporate a spelling check program in the computer. The rationale underlying this solution is the assumption that human performance has a tendency always to be "errorful", while computers perform well-specified functions without errors. By incorporating a spelling checker, computer behaviour (largely) makes good the inadequacies in human performance. An HF approach to the same problem might seek to identify the human behavioural causes of spelling errors and design the system with respect to these. Design solutions might include the provision of a spelling checker if typical users were observed to have poor knowledge of the English language; but a modified keyboard or a course in typing might be prescribed if the cause of the errors were observed to be a consequence of miskeying only. The HF approach, then, emphasises design to complement or support particular human behaviour; while SE emphasises design to complement or support computer behaviours.

The design of an effective human-computer system requires contributions from both disciplines. The SE solution to unacceptable spelling errors might overlook the real cause of low output quality. Although the solution might enable the system to achieve the requisite quality in its output, a design fault in the keyboard would likely engender frustration in users. By designing to support human behaviours, HF has the potential to enhance quality *and* minimize costs to the user in the system.

3.2.2 Classes of discipline.

In addition to a general problem, Long and Dowell identify the other defining characteristics of a discipline as being practice and knowledge. Practice constitutes the activities undertaken in order to solve the discipline problem. In the case of HF, the activities might include the observation of the behaviour of a computer user; interpretation of the behaviour in terms of a psychological theory; and specification of the requirements for a computer more compatible with the behaviour of users. Such activities could contribute to the solution of the problem of designing human behaviours interacting effectively with the behaviour of the computer.

The knowledge to support practice is the product of study within a field defined by the scope of the discipline problem. Pursuing the example of HF practice, HF discipline knowledge will include, for example, knowledge of psychological theories and knowledge of techniques for behavioural observation. Long and Dowell observe that such knowledge may exist either privately, in the experience of practitioners; or publicly, in documents such as texts, journals or computer-based representational media.

Three classes of discipline are distinguished which are concerned with the development of artefacts: craft disciplines, applied science disciplines and engineering disciplines. Disciplines of all three types may address the same discipline problem (for example, that of designing effective human-computer systems); however, they may be classified with respect to their differing practices and the knowledge supporting their practices.

Craft disciplines are characterized by practices based upon the processes of (system) implementation followed by evaluation; and the practices are supported by knowledge which is heuristic and informal. In the context of HF, Long and Dowell cite, as an example of craft practice, iterative user interface development in which a prototype user interface is progressively modified according to the results of observational studies. Knowledge supporting such development may take the form of heuristics and craft guidelines (e.g. "simple operations should be simple, and the complex possible"). An important feature of craft knowledge is that it is not formal, so its scope is frequently undefined. It relies on the judgement of the practitioner to decide its applicability to particular cases. It is not testable, so it cannot guarantee to be effective and it cannot be generalized unequivocably.

Long and Dowell define an applied science discipline as "one which recruits scientific knowledge to the practice of solving its general problem". Scientific knowledge is explicit and formal, testable and generalizable. It may exist in the form of scientific theories or prescriptions derived from such theories. For example, in the case of HF, applied science knowledge in the form of psychological theories of attention (e.g. Wickens et al, 1983) might be used to select between alternative options for information display (say, between a visual information presentation and an auditory presentation).

Long and Dowell argue that applied science knowledge, while it may predict behaviour, cannot predict system performance. The practice of applied science disciplines is characterized by specification and implementation followed by evaluation. Relative to the support offered by craft knowledge, scientific knowledge will increase the probability of generating a successful implementation. However, there could still be no guarantee that, when implemented, a system designed according to applied science knowledge would exhibit the desired performance, so subsequent evaluation is necessary.

Engineering discipline practice is characterized by specification in advance of implementation with the objective of design for required (system) performance. Its discipline knowledge is prescriptive and expressed in the form of engineering principles, which "may enable designs to be specified for artefacts which, when implemented, demonstrate a prescribed and assured performance". Long and Dowell cite, as an example of an engineering principle, a principle derived from Kirchoff's Laws for the specification of an electrical network whose behaviour (e.g. distribution of current) would solve a design problem concerning the power supply of an amplifier. However, Long and Dowell are not able to exemplify equivalent HF engineering principles in the context of HCI. HF remains a predominantly craft discipline, gaining some support from applied science.

3.3 Knowledge to support discipline practice

Practitioners are people seeking to solve specific discipline problems. The knowledge they recruit may be personal (in the form of mental structures developed and maintained by each individual) or public (generally accessible published information). Individual practitioners will make differing use of available public information, depending on the quality of their personal knowledge. The following sections consider how three types of practitioner ("specialists", "generalists" and "casual practitioners") are differently served by personal knowledge and by public sources of discipline knowledge as expressed in research findings, prescriptive information and explicit methods.

3.3.1 Personal knowledge held by practitioners

The three classes of practitioner described here may be viewed as occupying relative locations on a continuum defined by personal discipline knowledge.

(a) *"Specialists"*. The knowledge held by specialists may be extensive with respect to the domain of discipline problems but is particularly detailed with respect to a defined part of the domain. For example, specialists in the human factors of speech technology will have a general knowledge of the interaction between people and machines, but they are distinguished by the completeness (relative to other practitioners) of their knowledge of the interaction between people and speech operated computers. Specialist knowledge will include knowledge of the entities relevant to a particular subset of domain problems, such as the structural and behavioural characteristics of the entities. For example, an HF specialist in speech technology will possess mental structures appropriate for the conceptualization of speakers/listeners, of speech I/O devices and of tasks involving speech communication. Specialists will also be familiar with the consequences for performance of the behavioural interaction between domain entities. In the case of speech interaction, for example, a specialist might be able to predict the problems a naive user would have when entering information by means of an isolated-word speech recognizer.

In addition to knowledge of domain entities, specialists possess knowledge of the procedures for conducting discipline activities. Procedures might be relevant to the extension of discipline knowledge (i.e. knowledge of research techniques and methods relevant to their specialism); to the use of discipline knowledge to solve discipline problems (i.e. knowledge of applications within the specialism); or to the maintenance of discipline knowledge (i.e. knowledge relevant to the presentation of the products of their work, for example, in learned journals). Specialists will also establish a body of "meta-knowledge" ("knowledge about knowledge" - see, for example, Barr and Feigenbaum, 1981) relevant to their specialism. Such meta-knowledge would include a perspective on the current state of knowledge within the discipline (e.g. on its coverage and quality); on their own knowledge; and on procedures for gaining access to appropriate sources of knowledge.

(b) *"Generalists"*. By comparison with specialists, generalists possess knowledge which is general with respect to the set of discipline problems. For example, within the discipline of HCI, an HF generalist might possess knowledge of the interaction between people and computers which is not specific to particular classes of person or device. Such a practitioner will possess mental structures relevant to the characterization of humans and computers in general and of their interaction in the performance of tasks. As a result of training and of personal experience generalists may possess some knowledge pertaining to specific classes of device or user, but this knowledge is less complete and detailed than that held by relevant specialists.

Generalists possess knowledge of procedures for conducting domain activities, but this, too, is general, and it is likely to be incomplete with regard to special techniques and methods relevant to small subsets of domain problems. The meta-knowledge of a generalist is also incomplete. Such practitioners would have a partial view of the current state of knowledge in particular areas and of the quality of research supporting domain knowledge. However, their general knowledge of the discipline should enable them to locate relevant knowledge sources for evaluation and recruitment as required.

(c) *"Casual practitioners"*. Casual practitioners attempt to solve discipline problems but do not possess discipline knowledge other than that which has been acquired serendipitously; that which is fortuitously analogous to knowledge of their own discipline; or that which is derivable from notions of "common sense". They are unlikely to have developed mental structures for the characterization of domain entities appropriate to support HF analysis and are unfamiliar with relevant discipline procedures. Casual practitioners possess little or no meta-knowledge of the discipline and rely on weakly-founded criteria for the access and evaluation of information relevant to specific discipline problems. Engineers specializing in other disciplines (such as electrical and software engineering) would typically have little knowledge of HF. On occasions when they attempt to solve HF problems, they might be classed as casual practitioners.

3.3.2 Public discipline knowledge.

The development of a discipline is characterized by the establishment of a body of public knowledge accessible to practitioners. As indicated in the last section, discipline knowledge may pertain to the structure and behaviour of entities with which the discipline is concerned - in the case of HCI, computers and people - or it may pertain to the methods by which discipline problems are addressed. Dowell and Long term the former "substantive" knowledge, as distinct from "methodological" knowledge. In the case of disciplines concerned with the design of artefacts (such as HCI), this knowledge may be expressed either as research findings, primarily intended to constitute a body of knowledge potentially recruitable to discipline activities; or in forms intended to support discipline practice directly, for example, prescriptive design information and methods[1].[2]

(a) *Research findings.* Research has the objective of extending discipline knowledge (substantive and methodological), and its findings are reported in technical journals and at technical meetings. Research studies may take the form of theoretical analyses, controlled experiments or descriptions of the solution of domain problems (application case studies). However, the successful utilization of the research literature requires, firstly, identification of, and access to, relevant published sources; and, secondly, interpretation of general findings with respect to specific instances.

The number of HF research publications is large and disparate, and findings relevant to the development of particular types of systems (such as speech-based systems) are distributed widely. HF specialists in speech technology possess knowledge enabling the identification of relevant sources; and, by virtue of specialists' position and reputation, they are likely to have ready access to such sources. Although the HF generalist may be less familiar with individual publications, such practitioners are potentially able to ameliorate the weaknesses in their private knowledge by effective use of library sources. However, casual practitioners do not possess private discipline knowledge appropriate for the identification of relevant sources and may not have ready access to such sources.

[1] These classes are not mutually exclusive; for example, a document reporting the results of a research study may also advance prescriptive guidelines.

[2] The review presented here is general with respect to human-computer work systems; Chapter 5 presents an equivalent review which is specific to systems involving speech technology.

Unfortunately, the literature is also incomplete and not coherently organised, and non-specialists may find it difficult to determine whether or not it is likely that previous research has been performed which is potentially relevant to their needs.

It was observed in Section 3.3.1 that, although some of the research literature relevant to system development reports findings in the manner of applied science, much of the work described has been performed in the manner of craft (i.e. it is informal and implicit). It is rarely the case that the circumstances of the performance of a research study correspond directly to those presented by a particular design problem. The results have to be interpreted with respect to the features presented by specific cases. HF specialists possess private knowledge to support interpretation and extrapolation from craft and applied science findings; however, generalists may be less successful and casual practitioners may fail, because they are unable to identify critical features of systems which render research findings relevant (or otherwise) to the case of concern to them.

In summary, the HF research literature may be utilized effectively by specialists involved in system development, and it provides less accessible support for HF generalists. It is poorly suited to the needs of casual HF practitioners.

(b) *Prescriptive information.* One way of rendering substantive knowledge derived from research more appropriate for application in system development is to re-express it in a prescriptive form. For example, the practitioner might be told how to design a particular aspect of the user interface, in order to support a particular human behaviour. A potential solution to the poor accessibility of research findings to many practitioners lies in the availability of prescriptive information intended to be applied directly to domain problems. Prescriptions may take the form of guidelines derived from the findings of applied science, of guidelines arising from the experience of practitioners ("craft" guidelines) or of principles derived from applied science or engineering theory. As stated previously, there are, as yet, no HF design principles. However, volumes of general design guidelines and standards have been produced which seek to prescribe features of systems to facilitate device-user interaction (e.g. Smith and Mosier, 1986; Gardiner and Christie, 1987; U.K. Defence Standard 00-25, 1987).

HF specialists possess domain knowledge enabling them to interpret design guidelines with respect to specific cases. The guidelines may serve a check function, helping the practitioner to ensure that important HF issues have been considered. Although such prescriptive information is also intended to be usable by non-specialists, non-specialists typically encounter difficulties in locating relevant guidelines (the set is incomplete); in relating generally expressed guidelines to the specific features of individual cases; and in dealing with conflicting guidelines (de Souza, Long and Bevan, 1990). The latter difficulties may be attributed to the craft nature of HCI guidelines (e.g. expression in the form of heuristics): their successful application demands private discipline knowledge which is likely to be incomplete in the case of HF generalists and absent in casual practitioners.

In summary, although guidelines may offer a "check list" facility to specialists and, less effectively, to HF generalists, they are not necessarily well-suited to the needs of casual practitioners.

(c) *Methods.* Methods express information concerning how to bring about a desired change in the state of objects within the scope of the method, i.e. they contain procedural knowledge to support the performance of discipline tasks. Some of the limitations of completeness and applicability of prescriptive HF information may be resolved by the use of generalizable methods. These support activities such as the diagnosis of inadequate interaction behaviour and the prescription of appropriate intervention for the particular case under investigation (e.g. Meister, 1986)

Discipline knowledge may be expressed more or less formally, and it may be attributed a status according to the guarantee offered by the method that application of the procedures will have the intended outcome. A distinction may be drawn between "engineering" methods, the procedures of which are principled and expressible in formal terms; and "craft" methods, which are more or less heuristic and informally expressed (see Long and Dowell, 1989). Engineering methods will enable the production of an implementable specification of a product of a desired quality. Craft methods, although improving the probability of attaining a successful outcome, can offer no guarantee of quality. The quality of outcome in the latter case will be in part dependent upon the ability of the individual in adapting the procedures to suit the immediate circumstances.

As asserted previously, the discipline knowledge of HF is predominantly heuristic and informal. At present, there are no HF system development methods which can assume engineering status: existing HF techniques, such as task analysis, video analysis and operational evaluation, are exploited in the manner of a craft. Their procedures are expressed in general terms, and skill in applying them during the procurement cycle is developed by learning from other practitioners or by the trial and error of direct personal experience. In general, then, current HF methods are not well-suited to application by non-specialists.

3.4 Enhancing the contribution of HF to system development

3.4.1 Ineffectiveness in HF's contribution to development

It is a common observation that many human-computer systems do not perform tasks as well as might be desired by organizations (i.e. their task quality is inadequate); it is also observed that users frequently experience dissatisfaction due to the effort and frustration of interacting with computers (i.e. costs to users are unacceptable). One interpretation of these observations is that HF does not make a general and effective contribution to system development (see, for example, Long and Dowell, 1989). Long and Dowell suggest that the ineffectiveness of HF may be attributed to its craft characteristics. The incompleteness and informality of the substantive knowledge of the discipline render it difficult or impossible to apply without subsequent testing, particularly where the practitioner lacks specialist knowledge.

Methodological support offers one solution to the weaknesses in substantive knowledge, by providing practitioners with procedures for solving problems. SE shares many of the characteristics of a craft; like HF it also relies heavily on strategies of implementation followed by testing. In the case of SE, methodological support is now being offered in the form of explicitly proceduralized structured analysis and design methods (SADMs). These contrast with the implicit and informal methods available to support the practice of HF, which frequently fail to be exploitable in system development (Bellotti, 1989, 1990).

The following sections briefly summarise the features and benefits of SADMs and describes how they are now being extended to take account of HF concerns in the design process.

3.4.2 Structured analysis and design methods

Commercial pressures have imposed requirements for enhanced software quality. The term "quality assurance" is widely (and loosely) employed to describe the organizational function concerned with ensuring that produced software performs as specified and that it meets user requirements. Structured Analysis and Design Methods (SADMs) have been advanced to support the design of software more likely to meet these requirements. (Walsh, Lim and Long, 1989). They seek to render more systematic the (craft) activities of software development and hence improve design "through a more efficient and complete process of problem resolution" (Lim, Long and Silcock, 1990).

SADMs have been defined by Silcock, Lim and Long (1990) as software development methods, the scope, process and notation of which are explicitly conceptualized and proceduralized; (see also Madison (1983), Carver (1988) and Hartson and Hix (1989) for comparable definitions). SADMs typically decompose the development cycle into phases or stages; notations are provided to represent the design problem at each stage, and procedures are prescribed to effect transformations of the problem leading to its ultimate solution.

Jackson System Development (JSD) is a SADM used relatively widely in the U.K., and it has also been the subject of research to extend it to support HF contributions to system development (see Section 3.4.3). For these reasons, JSD is now used to illustrate some of the characteristics of SADMs. For fuller descriptions of JSD, see Jackson (1982) and Cameron (1986); and for a review of other SADMs, see Madison (1983).

In JSD, the development process is divided into three main phases: a *Model* stage, during which are represented the entities and processes of the world in which the putative computer system is to operate; a *Network* stage, during which the model is elaborated into a specification for the computer; and an *Implementation* stage, in which the specification is expressed as a runnable program operating on software data structures. Each stage involves the progressive transformation of descriptive representations of the domain or the system by the application of explicit procedures. The scope (products) of the procedures are defined in the method, and the representations are expressed using graphical notations supplied by the method. To exemplify the application of the method, Cameron (1986) has described some of the stages in the development of a simplified library system.

Figure 3.1. presents a Model Stage representation of the world in which the system is to operate; the representation is termed a Process Model. The library processes are represented in terms of actions with respect to a single book within the library. The *scope* of the representation is specified by the method (e.g. a process model is based upon observations of the world of the putative system and is complete with respect to the functionality of the system). JSD *notation*, which is used to express the process model, captures the order in which actions occur (from left to right on the page), and it can communicate features of actions such as iteration (indicated by an asterisk) and selection from options (indicated by a superscript "o"). The notation is, then, explicitly conceptualized. The *process* of developing the representation is also conceptualized (e.g. the objective of the development of the process model is that of "scoping the system"); and it is proceduralized (e.g. "Decide what new process types are needed, and how many instances." "Choose and define suitable connections to existing process types"....- from Jackson, 1987)

Cameron describes how the process model is subsequently elaborated to define the data upon which the various modelled processes act. Figure 3.2 illustrates a product of such elaboration. Cameron observes:

> "... the important points are as follows: the original model process is used as a framework for defining the data to be stored for one book; the meaning of the data is formally tied to the meaning of the actions and their attributes; a data item is local to a process instance; the mechanism for updating the data is part of this definition, not something separate; (and) as the model process executes to keep in step with reality, the data is also kept up to date...."

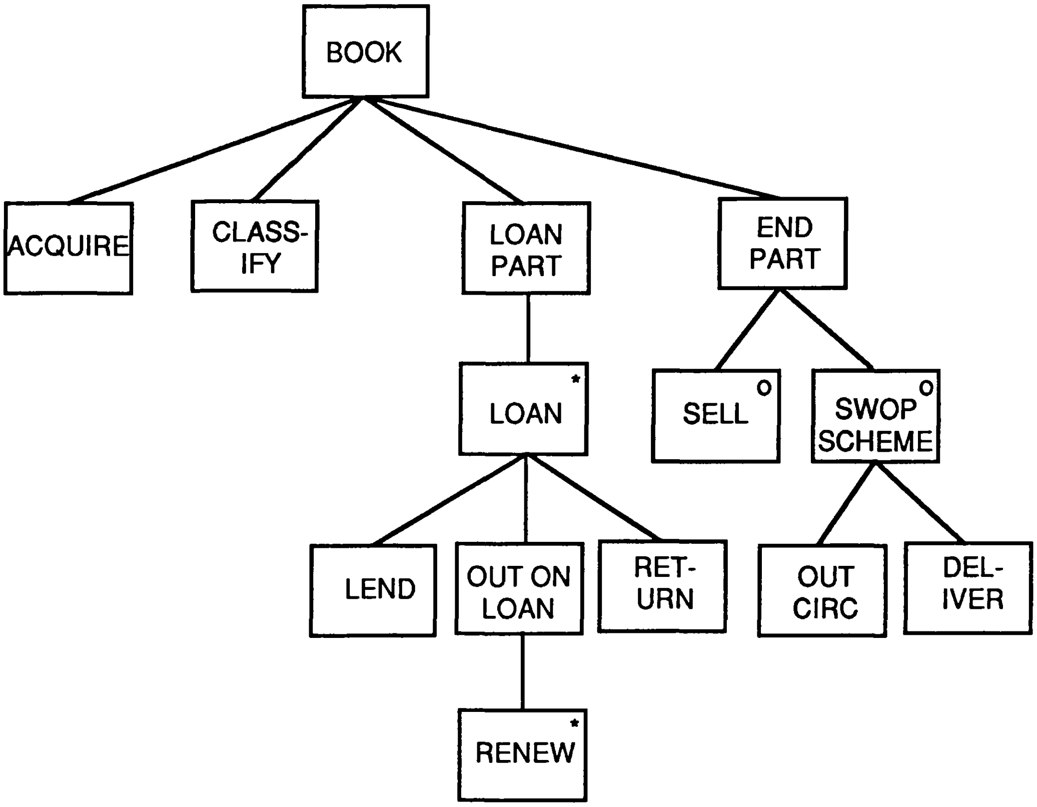


FIGURE 3.1: A process model illustrating JSD notation (from Cameron, 1986)

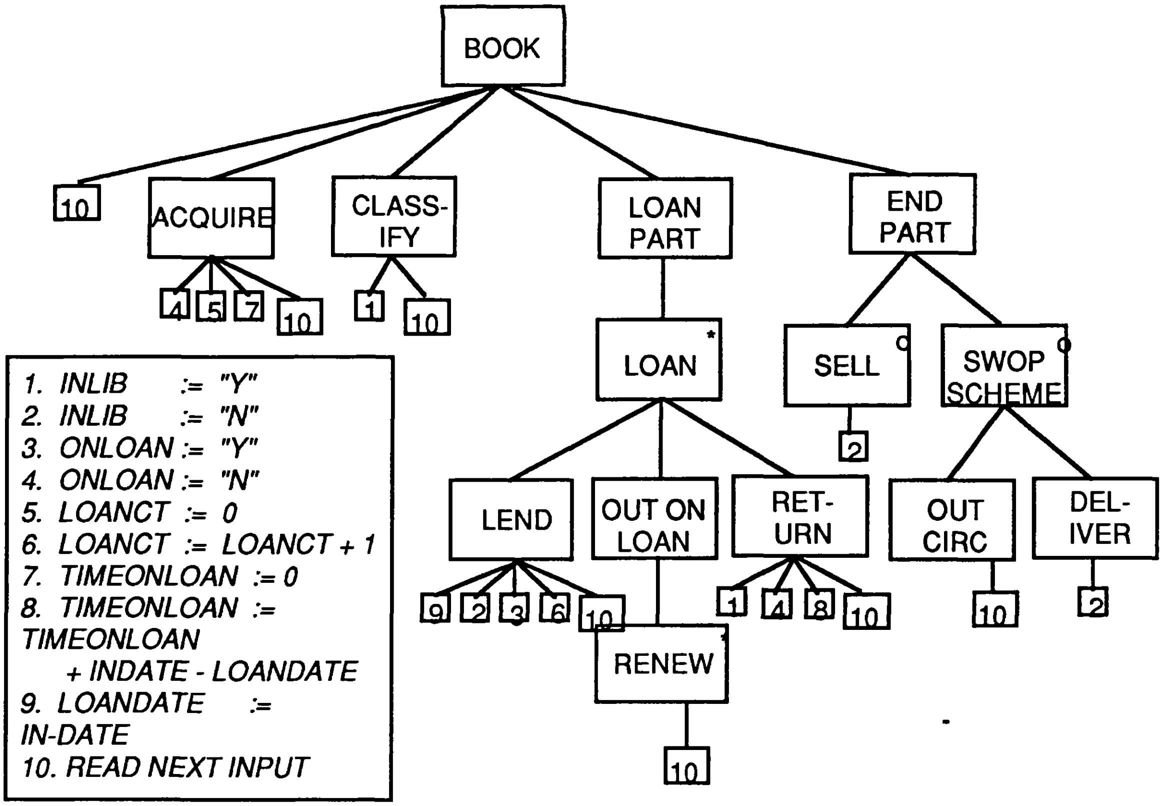


FIGURE 3.2: Elaboration of process model to include data types (from Cameron, 1986)

Again, the scope, notation and process of this step are conceptualized and proceduralized, and the product of the previous step is progressively transformed with the ultimate objective of specifying and implementing an effective computer system.

SADMs (like JSD) cannot claim to generate implementable specifications for systems having a known performance; i.e. they are not engineering methods in the sense of Dowell and Long (1989). They do not recruit engineering principles but, rather, the SE discipline knowledge held by specialist practitioners (Walsh et al, 1989). The performance of the task of systems analysis and design is enhanced by rendering the behaviour of practitioners more systematic. A number of consequent benefits are claimed by Walsh et al:

(1) "*Production of quality:* SADMs support software engineers in making appropriate decisions, although they do not, themselves, make decisions.

(2) *Management of complexity:* SADMs identify the decisions which need to be made, the order in which to make them, and, to some extent, the basis for them. Factors which are independent are separated, and those which are dependent are treated appropriately ("separation of concerns").

(3) *Improvement of communication:* SADMs aid communication between system developers, between developers and managers and between developers and(computer) users.

(4) *Explication of decisions:* SADMs facilitate the review of design decisions and their justification.

(5) *Production of intermediate products:* SADMs aid the verification and validation of the software produced and enhance the efficacy of intermediate user testing (iterative design).

(6) *Improvement of project planning:* SADMs facilitate the setting of project milestones and the estimation of costs. Project management is able to relate project progress to past experience, so that comparison supports present planning."

SADMs offer, then, a means of enhancing the performance of the task of developing systems by the more effective recruitment of craft knowledge. However, in general, SADMs have been concerned with the SE problem of designing with respect to *machine behaviour* and have taken only indirect account the behaviour of the user in the system. Section 3.4.3 now describes an extension of a SADM (JSD) to take fuller account of HF concerns in system development.

3.4.3 Structured Human Factors methods.

Walsh et al (1989) propose that SADMs provide a means by which HF contributions may be made to the task of system development. Specifically, SADMs offer a framework for timing and scoping HF inputs to the design process. Silcock et al (1990) and Lim, Long and Silcock (in press) describe an extension of the existing, SE-orientated JSD method which takes account of HF concerns. Figure 3.3 illustrates JSD*, comprising JSD*(SE) - substantially, the original method - and the parallel activities constituting JSD*(HF). The two sub-methods are described as design streams structured as stage-wise design processes, inter-linked at various points.

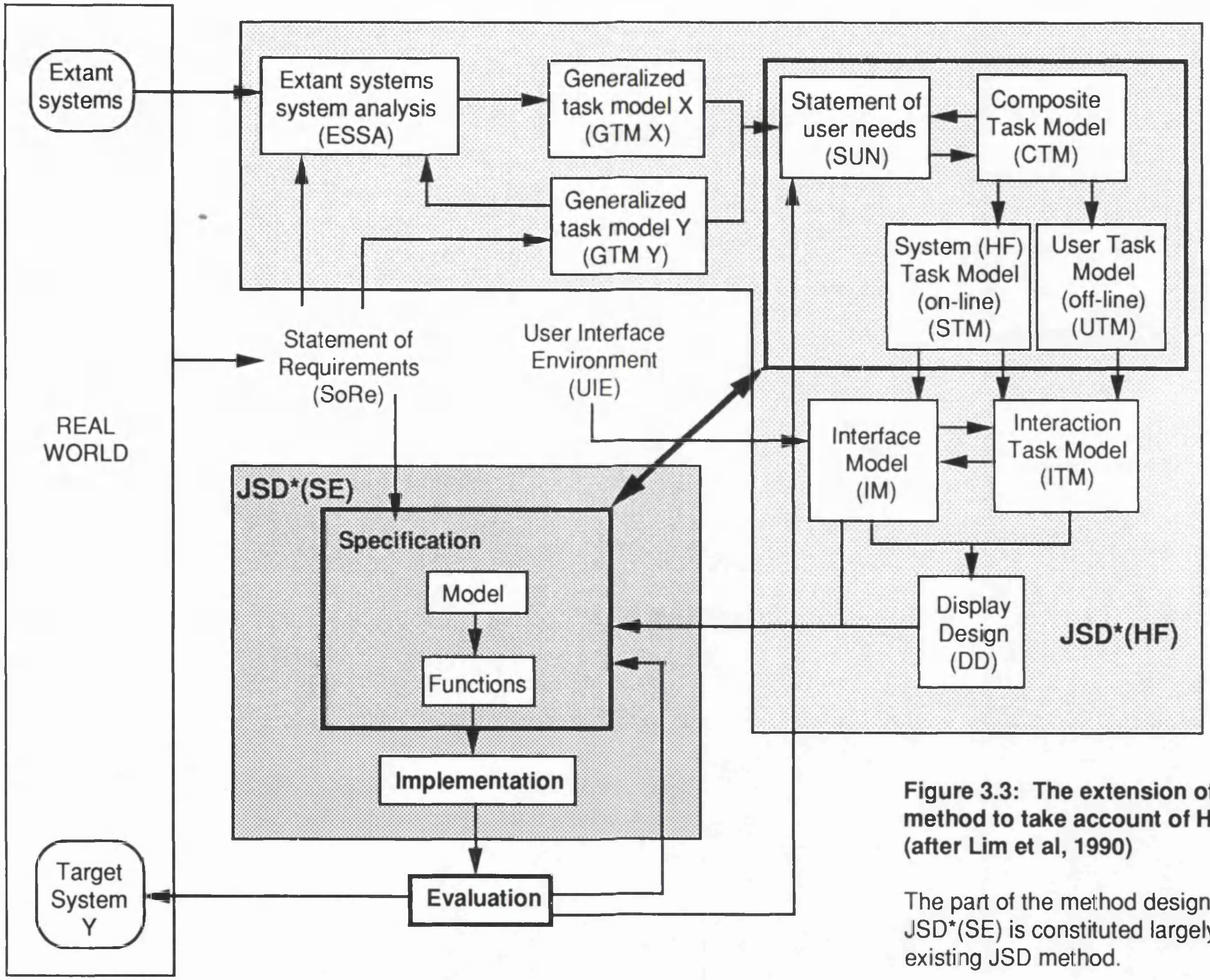

Extant systems
REAL WORLD
Target System Y
Extant systems system analysis (ESSA)
Generalized task model X (GTM X)
Generalized task model Y (GTM Y)
Statement of Requirements (SoRe)
User Interface Environment (UIE)
Statement of user needs (SUN)
Composite Task Model (CTM)
System (HF) Task Model (on-line) (STM)
User Task Model (off-line) (UTM)
Interface Model (IM)
Interaction Task Model (ITM)
Display Design (DD)
JSD*(HF)
JSD*(SE)
Specification
Model
Functions
Implementation
Evaluation

Figure 3.3: The extension of the JSD method to take account of HF concerns (after Lim et al, 1990)

The part of the method designated JSD*(SE) is constituted largely of the existing JSD method.

The stages of JSD*(HF) exhibit the defining characteristics of SADMs (i.e. explicit conceptualization and proceduralization). Silcock et al (1990) provide a number of illustrations of this. For example, the Extant Systems System Analysis stage in Figure 3.3 (termed the stage of Task Description by Silcock et al) involves the collection and analysis of information concerning the (human-computer) system task as it is currently performed. The scope of the method, as it relates to this stage, is defined by the set of extant systems; and the process is decomposition, each system task being expressed as a hierarchical set of actions. JSD notation may be used for the purpose of representing the dynamic aspects of the task (see Section 3.4.2); and this is supplemented with a tabular representation of the static aspects. Silcock et al indicate that the Task Description supports various functions, such as exposure of the allocation of activities between human and machine elements of extant systems; ..."design aspects of the user interfaces; and user problems and needs".

Structured HF methods, such as JSD*(HF), clearly have the potential for enhancing performance of the task of system development by the more-effective recruitment of HF discipline knowledge. Furthermore, they may enhance the performance of tasks addressing the specific discipline problem of HF, by ensuring a systematic approach to problem resolution. In terms of the analysis presented in Section 3.3, structured HF methods potentially support HF specialists by integrating their contribution with those of other specialist practitioners. They could support HF generalists in the same way but, additionally, may provide procedural knowledge enabling the conduct of specialist techniques.

MoD procurers have been instrumental in promoting the use of SADMs. By rendering development systematic, the methods not only enhance product quality but also facilitate project administration. SADMs clearly, then, have potential for enhancing the effectiveness of procurement. MoD has further recognized the value of extending SADMs to the domain of HF (for example, by supporting the extensions of JSD outlined above). However, at present, structured methods are restricted to the processes of systems analysis and design, and do not support other development activities, such as system evaluation. Furthermore, SADMs do not provide prescriptive knowledge: in Walsh et al's terms ..."they do not, themselves, make (design) decisions". HF generalists must, then, rely on other sources of such knowledge. Casual practitioners are likely to be poorly supported even by methods such as JSD*(HF), because they do not possess the necessary discipline knowledge of human-computer interaction.

The next chapter considers a particular type of HF contribution to development: that of performance evaluation. Such evaluations can also enhance the effectiveness of military procurement: Chapter 5 will identify a requirement for structured methods to support HF evaluations in procurement.

CHAPTER 4

SYSTEM EVALUATION

4.1 Introduction

Chapter 2 proposed that organizations establish human-machine work systems in order to improve their effectiveness. Improved effectiveness can be achieved by the enhancement of output quality or by the reduction of costs to the organization. Procurement was then defined as a process by which organizations acquire machines in order to achieve desired standards of performance. Evaluations were identified as necessary for system developers and procurers to determine whether systems conform to these standards.

This chapter is concerned with evaluation, particularly as it relates to human-computer systems. A framework is presented which is used to characterize evaluations: evaluations are distinguished by their products, criteria and processes. HF evaluations tend to be conducted after implementation, when opportunities for system modification are limited, and so when the products of HF evaluations may fail to be utilized. A general requirement is identified for HF performance assessments prior to implementation, offering a diagnostic output. Given existing knowledge of HF, an empirical evaluation process is proposed as being appropriate, recruiting simulations to reproduce system behaviour.

The chapter concludes by proposing that the procurement of military systems would be rendered more effective with methodological support for HF evaluation at the stage of Feasibility Assessment. Structured methods for conducting evaluations would offer a potential solution, particularly if they could be adapted to enable use by individuals lacking knowledge of HF.

4.2 Evaluation in the development of human-computer systems

Tasks are distributed as goals (desired changes in the state of objects in the application domain of the human-computer system) and performance requirements. For example, a secretary might be delegated the task of producing a letter to the standards of content and presentation deemed acceptable by the secretary's employer. In this instance, the goal would be completion of the letter and the performance requirement related to the adequacy of content and presentation. As asserted in Section 2.2, the performance of a system is an expression of task quality (i.e. the correspondence between the desired change in the state of objects in the application domain and the change which is actually achieved), and of resource costs (Dowell and Long, 1989).

Dowell and Long draw a distinction between costs incurred by machine elements of a system (e.g. the computer) and those incurred by people (i.e. users). An example of computer costs would be structural wear occurring as a consequence of using the computer to support the performance of a task. User costs include those of developing appropriate mental structures in users to perform tasks (e.g. costs of user training), and costs incurred by users of generating the behaviour which will support the achievement of task goals (expressible, for example, as "workload"). The former class of user costs may, in part, be borne by the organization, while the latter are borne directly by the individual. For example, the secretary producing a letter may experience frustration because the design of the word processor encourages the commission of operator errors: the frustration would constitute one form of user cost.

In Chapter 2, it was observed that evaluation occurs during system development and procurement. Evaluation is necessary to determine the conformity between a system's actual performance and that desired of it (Whitefield, Wilson and Dowell, 1991); for example, evaluations would be necessary to determine whether a particular word processor design could support the requisite task performance. In one sense, "evaluation" constitutes an explicit stage in development, when the performance of the implemented product is checked for conformity with the performance objectives embodied in the original specification of requirements. However, evaluation will also occur at other stages of development, to ensure that potential performance inadequacies are identified and rectified as early as possible. Procurers additionally rely on evaluations at each stage to provide information enabling them to decide whether a project should be allowed to proceed to subsequent stages.

4.3 Approaches to system evaluation

Evaluations may take different forms, depending on their purpose. This section presents a framework for distinguishing different types of evaluation and for relating their form to their purpose in system development. The framework is subsequently used to distinguish appropriate forms of evaluation for application in the assessment of system feasibility.

Whitefield et al distinguish the *products* of evaluations (evaluation statements) from the *process* of achieving them. Whitefield et al's specific concern lies with evaluations of a particular type (human factors - HF - evaluations); however, more generally, evaluations may additionally be distinguished by the *criteria* they employ, which may be HF criteria or others. The following sections consider the products, criteria and processes of evaluation respectively.

4.3.1 Products of evaluation

Whitefield et al propose that the statements generated as the products of evaluations may be divided into two classes. An analogy is drawn between evaluations performed in the context of system development and those performed in medical contexts. Doctors may describe the symptoms of illness identified in patients at presentation; alternatively or in addition, they may offer a diagnosis of the illness which is the underlying cause of symptoms. It is proposed that system evaluations may similarly be divided into those consisting of a non-interpretive report of performance and conformity - a statement of *presentation*; and those offering an interpretation of the causes of the performance - a statement of *diagnosis*.

In terms of the definitions presented in Section 4.2, a statement of presentation would specify actual performance with respect to desired performance: a statement of diagnosis would specify the relationship between actual and desired performance with respect to system behaviour. Pursuing the previous word processing example, it might be determined by observation that a secretary using a particular word processor consistently produces letters with an unacceptably high incidence of spelling errors (i.e. that actual performance does not conform with desired performance). This observation would constitute a statement of presentation. However, it might further be observed that the errors are a result of mis-keying and that the secretary's behaviour would support the generation of error-free text if the keys were re-located (i.e. the discrepancy between actual and desired performance is attributable to incompatibility between the physical structure of the computer and the user's physical behaviour). This would constitute a statement of diagnosis which might be used as a basis for prescribing the use of a different type of keyboard.

4.3.2 Criteria for evaluation

Performance criteria derive ultimately from conceptions of task quality and system costs and are logically distinct from system behaviour. However, developers may seek to isolate the contributions to performance of the behaviour of one or other of the entities comprising the system. Evaluations may be performed, then, against criteria believed to be *correlated* with task-based measures of performance but which isolate specific behavioural contributors to performance.

Because of their different orientations to design, HF and SE are concerned with different aspects of system behaviour (see Section 3.2) and so will assume differing criteria in evaluation. Consider two evaluations of the performance of a system consisting of a user and a word processor with a spelling check facility. A software engineer who is offered alternative spelling check algorithms might make the simplifying assumption of a substantial correlation between overall task performance and the speed with which a spelling checker operates on text files. Alternative spelling check algorithms (exhibiting different behaviours) might be evaluated with respect to the speed with which they enable an expert user to process a text file of known length and containing a specified number of spelling errors. The software engineer may be only secondarily concerned with the behaviour of the user and its impact on performance time: the objective of the evaluation is to select the most efficient check algorithm.

An HF specialist may be concerned with the selection of a spelling check facility which is most compatible with the keyboard operating behaviour of expected users. This specialist may make the simplifying assumption of a correlation between task performance and the speed with which users operate the computer with the spelling check facility. Alternative ways of offering users the option of accepting or rejecting the machine's spelling changes might be evaluated by comparing time taken by representative users as they process a text file of known length and containing a specified number of errors. The HF specialist may be little concerned with the efficiency of the check algorithm and its impact on task performance time: the primary objective of the evaluation is to select the option with features most compatible with user behaviour.

In summary, then, an evaluation will utilize criteria relevant to the use to which the results will be put. SE and HF practitioners will likely perform evaluations against different criteria, because of their differing orientation to the design problem.

4.3.3 Processes of evaluation

The process of system evaluation is a sequence of activities which culminates in the generation of an evaluation statement concerning the system. Value may be demonstrated directly or it may be demonstrated by reasoning: hence, performance evaluations may assume empirical or analytic processes. The empirical process involves the observation of behaviour and the measurement of performance, so it requires behaviour to be instantiated. Behaviour can only be instantiated if the entities of the system are actually present, or if they are represented as "reproductions" which resemble the actual entities with respect to those attributes which determine behaviour. Prototypes, simulations and scale models are examples of such reproductions (Life, 1990). When behaviour is instantiated, performance may be measured against relevant criteria and diagnoses advanced according to available explanations of behavioural phenomena. The illustrations presented in Section 4.3.2 are examples of empirical evaluations utilizing, respectively, SE and HF criteria.

The process of analytic evaluation involves, firstly, the analysis of the system with respect to *a priori* models of behaviour and performance and, secondly, the prediction of its performance according to such models. For example, an HF specialist evaluating alternative design options for a spelling check facility may decide that task completion time is the aspect of task quality most relevant to the design. The specialist might, then, assume a model of human behaviour and performance based upon error-free operation and choice reaction time. Performance might be specified according to Hick's Law (1952) on the basis of the number of alternative keys used in the operation of the spelling checker and their location relative to the user's hands; (for analyses of this type, see Card, Moran and Newell, 1983).

Analytic evaluation assumes, then, the existence of a valid model relating computer attributes (e.g. key locations) to user behaviour (e.g. the selection and implementation of keystrokes) and behaviour to performance (e.g. choice reaction time). It further assumes that the system under evaluation is represented in a way which enables it to be related to the performance model: for example, it must be possible to place values against those of its attributes which, according to the model, are determinants of behaviour; (in the previous example, the frequency of operation of alternative keys must be known, and their relative locations).

Performance evaluations will be performed under constraints which will differentially favour the analytic and empirical processes. Although some of these constraints are pragmatic (e.g. the availability of resources to instantiate and observe behaviour), two classes of constraint are fundamental and are now discussed: knowledge of system behaviour and performance, and the representation of the system to be evaluated.

(a) *Knowledge of system behaviour and performance.* Evaluations impose varying demands for knowledge of the expected behaviour and performance of the system of concern. For empirical evaluation, such knowledge is necessary to scope behavioural observations and to interpret performance for the purpose of diagnosis. For example, an HF specialist concerned with the design of a spelling checker *compatible* with user behaviour may decide to perform an empirical evaluation of alternative spelling check programs. The conduct of the evaluation would require some knowledge of what constitutes "compatible" behaviour, in order that those aspects of system structure critical to behavioural compatibility may be manipulated, and that relevant aspects of system behaviour may be observed. However, even a rudimentary model, perhaps based upon "common-sense" notions of compatibility, could serve this purpose (and, of course, the model could subsequently be refined to take better account of the results of the evaluation).

By comparison, the accuracy and appropriateness of an analytic evaluation is substantially determined by the completeness and appropriateness of the analyst's model of device-user interaction. For example, the analytic HF assessment of alternative spelling checkers on the basis of Hick's Law assumes that the choice reaction model is a complete and appropriate explanation of user behaviour (and predictor of performance) in the context of the selection of keys to operate a spelling checker. Although predictive models of human-computer interaction have been developed and applied (e.g. Card, Moran and Newell, 1983), they are recognized as being restricted in their predictive power, particularly if tasks involve a substantial amount of cognitive processing or involve the commission of errors.

(b) *Representation of the system.* The circumstances of an evaluation will determine the availability of a machine and its users, and of representations of both, for the purposes of observation and analysis. Empirical evaluation requires either the entities themselves or appropriate reproductions to be available for interaction in the context of a representation of the system task. A design specification would not be adequate to support such a representation: it would have to be implemented, at least in part. By contrast, analytic evaluation assumes the description of the task, user and machine with respect to those attributes relevant to the behavioural interaction. Provided the values of relevant attributes are known, or can be inferred, a system specification would be adequate to enable evaluation: it need not be implemented. Other factors being equal, the analytic evaluation described above might be performed prior to the implementation of the keyboard and spelling check program, whereas an empirical evaluation demands their instantiation, either in the form of the actual device or of a reproduction of the device.

In summary, then, the performance of empirical evaluations is favoured when the entities of a work system are implemented (or implementable) and when available models of device-user interaction are incomplete or of unknown validity. The performance of analytic evaluations is favoured when appropriate models of interaction are available and when the system is not available in an implemented form.

4.4 Enhancing the effectiveness of systems

4.4.1 Ineffectiveness in human-computer systems

It was asserted in Chapter 3 that many human-computer systems fail to exhibit the performance expected of them. Timely and appropriate evaluations enable performance to be checked during system development and the system to be modified accordingly. However, Carver (1988) has observed that HF evaluation typically takes place after the system has been implemented, when the system is difficult (and, hence, expensive) to modify. There is a requirement, then, for HF evaluation techniques which are applicable earlier, when the results would be more readily able to influence design.

4.4.2 Approaches to HF evaluation prior to implementation.

In terms of the framework presented in Section 4.3, evaluation techniques are necessary which assume the criteria of *HF*, and which generate evaluation products which contribute to the process of design. During the stage of system specification, designers need to know, not only whether a system will conform to a performance requirement, but also the reasons for failures to conform; i.e. *diagnostic evaluations* are required.

Prior to implementation, an analytic evaluation would be favoured, on the grounds that such evaluations are potentially applicable to specifications; however, Chapter 3 cites evidence to suggest that knowledge of human-computer interaction is generally incomplete and poorly validated. There are few models of interaction sufficient to predict system behaviour and performance. A reliable solution to the HF evaluation problem will tend to utilize, then, *empirical* techniques; and, to be applicable in advance of system implementation, evaluation would demand the utilization of reproductions of system behaviour.

The general problem of poor system usability might be addressed, then, by the conduct of diagnostic HF evaluations utilizing system reproductions (such as prototypes or simulations), supporting an empirical approach in advance of full system implementation. Early HF evaluations which have been conducted and reported in the literature frequently take this form (for examples, see Life, Narborough-Hall and Hamilton, 1990).

4.5 Enhancing the effectiveness of system procurement

Because the establishment of system behaviour supporting desired performance requires the establishment both of machine behaviours and human behaviours, evaluations to support procurement may involve criteria relevant to SE and to HF. In military procurement, statements both of presentation and of diagnosis may be utilized in the evaluation of computers offered to meet users' requirements. Statements of presentation may be adequate as the basis for acceptance (or otherwise) of a machine, according to its conformity with the performance requirement. Statements of diagnosis are necessary for the identification of requirements for modifications to the system that it may better support the required performance.

Feasibility assessments specifically seek to predict whether a system could be developed which would meet a given performance requirement and whether such development would be technically possible within available resources (see Section 2.4.3). The products of feasibility assessments include the identification of preferred design options, specifications of technical risk and statements of cost, duration and demand on resources for product development. Such assessments are required by MoD following the specification of a user requirement but before detailed specification.

Just as system development requires HF evaluations prior to implementation, so does procurement. It, too, requires diagnostic evaluation in order to determine the HF problems to be resolved during Full Development. Jordan et al (1987) have identified a requirement for empirical feasibility evaluations: an observation which is concordant with the view that only such approach could support reliable HF performance assessments.

Empirical HF evaluations using simulations applied at the stage of Feasibility Assessment should contribute to the procurement by MoD of more effective military systems. However, it has already been suggested in Chapter 3 that methodological support for HF is not clearly expressed. It is also not well-matched to the requirements of other types of practitioner involved in development (and, by implication, procurement). The availability of SADMs which take account of HF offer a solution to the problem as it presents itself in systems analysis and design, but SADMs do not extend to evaluation. There remains a requirement for structured evaluation methods which support non-specialists in the conduct of feasibility assessments.

The next chapter describes a specific requirement for an evaluation method for use in the procurement of battlefield computers with speech interfaces. A structured method is proposed to meet the requirement, which is described in detail in subsequent chapters.

CHAPTER 5

A REQUIREMENT FOR KNOWLEDGE TO SUPPORT HUMAN FACTORS EVALUATION IN PROCUREMENT

5.1 Introduction

Chapters 2, 3 and 4 have described, in general terms, a problem arising in military procurement which is attributable in part to the inadequacy of HF discipline knowledge in its support for early system evaluation. A general solution to this problem lies in methodological support for HF practitioners, and a requirement has been identified for explicit HF evaluation methods. This chapter presents a specific example of the general problem, which might be solved by the development of a structured method.

Section 5.2 describes the potential utility of speech-based battlefield computer systems and introduces the Royal Signals and Radar Establishment (RSRE) as an MoD procurer concerned with the acquisition of such systems for the army. In Section 5.3, a specific requirement is identified for a means of determining the suitability of such systems to support battlefield tasks. This requirement arises from a potential for ineffectiveness in procurement, which is analysed with respect to procurement activities, discipline knowledge support, and performance evaluation. A review, in Section 5.4, of previous research addressing the evaluation of speech systems supports the conclusion, in Section 5.5, that RSRE's problem might be solved by the development of a structured evaluation method supported by knowledge of speech interaction. Section 5.6 outlines a strategy for developing a method to meet the requirement.

5.2 The procurement of speech-based battlefield computers

Armies distribute tasks which must necessarily be conducted under unfavourable environmental circumstances, and which may place extreme demands upon both soldiers and military equipment. Computers are increasingly used on the land battlefield to support data processing functions, such as data storage, calculation and communication (for a review, see Ward and Turner, 1982). Conventional, manually-operated computer terminals with visual information displays may be highly effective when a sedentary operator enters alphanumeric information and is able to attend visually to a display. However, such terminals may be less effective when computer operation must occur on the battlefield, where there may be demands for data entry in constrained postures or for concurrent visual attention to other objects in the operator's field of view. Furthermore, keyboards and visual displays tend to be bulky and vulnerable to damage, and they demand a special manual skill (typing) for their rapid operation. For these and other reasons (Blair, 1981), the army has recognized that computers supporting alternative forms of device-user interaction may offer potential task performance benefits on the battlefield. Specifically, speech input and output (I/O) devices may be operated without manual and visual involvement; they are mechanically simple and, potentially, very compact. It is further claimed that they require no special skills for their operation (for reviews, see Knight and Peckham, 1985; Martin, 1989).

In spite of their potential utility, at present there are no examples reported of the use of speech I/O devices in U.K. army applications. An important reason for this is that existing speech I/O devices fail to fulfil the operational demands of army users; for example, poor recognition performance and small vocabulary size reduce the effectiveness of speech data entry, so current recognizers do not support desired task performance. Although existing devices are inadequate, it is expected that speech technology will develop further (Laver, 1987), giving rise, in the foreseeable future, to speech interfaces potentially suitable for army use. However, if such interfaces are to be successfully implemented on the battlefield, they must be designed to meet the particular requirements of battlefield users.

The part of MoD's Procurement Executive concerned with the acquisition of army computer systems is the Royal Signals and Radar Establishment (RSRE). In addition to procurement, RSRE conducts research supporting the development and subsequent exploitation of technologies, including speech technology. RSRE is separately concerned, then, with the establishment of effective army systems and with encouraging the uptake of speech technology research by product developers.

5.3 A problem in the procurement of speech-based battlefield systems

5.3.1 The problem as it relates to procurement

The army and RSRE constitute distinguishable sub-systems within the U.K. defence organization. The army is a sub-system concerned with land-based defence, and it is a user of military equipment. RSRE is a procurement sub-system, concerned with equipping defence sub-systems, including the army. Specifically, RSRE seeks to establish, within government-imposed budgets, human-computer systems in the army. The systems are intended to support performance sufficient for the army to meet its objectives. The procurement task requires RSRE to liaise between computer users and industrial computer developers to ensure that, when implemented, army systems exhibit the required performance.

RSRE has been concerned that procurement may be less than ideally effective in the case of speech-based systems, as there is no means of determining at an early stage whether speech can support requisite task performance. It has been recognized (in line with Jordan et al, 1987) that performance needs to be assessed at the Feasibility phase, if the results of the assessment are to be useful in scoping the work of speech technologists during subsequent system development.

5.3.2 The problem as it relates to discipline knowledge support

In that speech technology is a technology fundamentally concerned with the interaction between people and computers, its development requires support from practitioners both of SE (more specifically, of speech technology - ST), and of HF. RSRE is an established centre of excellence in ST (e.g. Moore and Bridle, 1986); it employs specialists in the discipline and so has been potentially well-placed to perform ST evaluations to support procurement. However, RSRE lacks expertise in HF. Although it might, in principle, sub-contract evaluation work to HF specialists, RSRE would prefer to be able to conduct HF feasibility assessments "in-house", i.e. RSRE requires support to enable their non-specialist engineers to undertake HF practices.

To what extent might such support be offered by existing sources of HF discipline knowledge? Chapter 3 identified HF as a discipline characterized predominantly by its craft practice. As a consequence of its inherently informal nature, practice of this kind tends not to be reported in the public domain. However, there is evidence to suggest that craft is widely applied in the establishment of speech-based systems. Speech dialogues are typically designed according to a strategy of implementation followed by evaluation - see, for example, descriptions of applications such as those of Gill (1990) and Cooke (1990) - and design guidelines have been published bearing craft characteristics (e.g. Smith and Mosier, 1986; Jones, Hapeshi and Frankish, 1989):

- *a special command vocabulary should be designed for voice input*
- *special care should be taken over the design of the command syntax*
- *provide representative template training*
- *multiple criteria should be used to evaluate the efficiency of the speech recognition system*

(examples taken from Jones et al, 1989).

Although some elaboration is normally offered by the authors of such guidelines, the origins of the prescriptions are frequently obscure; they are expressed informally and their scope is implicit.

Applied science knowledge is also recruitable to the design of speech systems. A review of relevant psychological and linguistic research are offered, for example, by Waterworth and Talbot (1989). Psychological theory has been used to develop prescriptive design guidelines; for example, Jones et al (1989) offer guidance on the combination of speech with other means of interaction on the basis of multi-resource models of attention, and guidance on the presentation of feedback on the basis of models of short-term memory.

These applied science guidelines have an explicit rationale and their foundation is open to test and refutation; their generality is explicit, so their applicability may be judged by the practitioner. However, the set of such guidelines is limited by existing psychological theory relevant to the design of speech-based systems. Furthermore, such guidelines fail to specify the relationship between interaction behaviour and system performance and so cannot support specification in advance of implementation. There is no evidence for speech system development according to an engineering conception of HF as espoused by Long and Dowell, and none of the prescriptive knowledge in the HF literature concerned with such systems can assume the status of an engineering principle.

Although HF specialists in speech technology would possess private discipline knowledge, enabling them to utilize the literature with at least some success, this knowledge is likely to be lacking in HF generalists and absent in casual practitioners (see Chapter 3). RSRE engineers attempting to perform HF evaluations would risk performing badly, i.e. there would be a high probability that their evaluations would be inaccurate or that they would incur unacceptable user costs in the process.

A solution to the incompleteness of relevant discipline knowledge was identified in Chapter 3; this lay in the use of generalizable evaluation methods. Bell and Becker (1983) offer some methodological guidance for the design of experiments for evaluating speech recognizers; however, this could not claim to be a complete method. At present there are no explicitly proceduralized methods which would provide the support for casual practitioners such as those employed by RSRE.

5.3.3 The problem as it relates to evaluation

RSRE needs to determine whether a speech interface for a computer would enable battlefield tasks to be performed to the requisite quality with acceptable costs to the system. As RSRE maintains expertise in the evaluation of speech I/O devices against technological criteria, the requirement for performance evaluation is interpreted as a need for a means of determining whether task quality could be sustained with acceptable costs to the user in terms of physical and mental effort, frustration etc. (i.e. for evaluation against HF criteria).

To ensure that speech technologists direct their efforts appropriately, RSRE would need to know, not only whether a system would support desired performance, but also the reasons for performance inadequacies. In terms of the framework advanced in Chapter 4, statements were required, firstly, of the predicted performance of the system at presentation and, secondly, of diagnosis.

Feasibility assessments occur prior to implementation, so the prediction of performance requires either analytic evaluation or empirical studies based upon reproductions (e.g. simulations) of the target system.

The next section considers the research literature relating to the evaluation of speech interfaces. The review focuses on the *types* of research that have been carried out, rather than on the specific findings of studies, the intention being to characterize the research within the framework for evaluation presented in Chapter 4. The review is used subsequently as the justification for proposing a structured method to meet RSRE's support requirements.

5.4 Previous research supporting the evaluation of speech systems

5.4.1 ST evaluations

ST evaluations are concerned with the speech behaviour of computers; i.e. with the behaviour of hardware and software structures, either to support the recognition and understanding of spoken language, or to support the compilation and generation of messages expressed in spoken language. Evaluations assume a co-relationship between task performance and the behaviour of speech recognizers and synthesizers as components of the communication channel between computers and their users. The purpose of the evaluations is to inform the development of I/O devices offering a more effective channel.

(a) *Products of ST evaluations.* In the case of speech recognizers, ST evaluation statements of presentation are based primarily on the frequency with which devices succeed or fail to respond appropriately to input utterances i.e. they are based upon recognition error rate. In the case of speech synthesizers, ST evaluations are based primarily upon the frequency with which people correctly interpret speech output, i.e. intelligibility (e.g. Pratt, 1986); on reaction time to synthesized speech stimuli (Talbot, 1989) or upon subjective evaluations by listeners of speech quality (e.g. Pratt, 1986). For reviews, see, for example, Simpson et al (1985); Knight and Peckham (1986); Jones et al (1988); and Waterworth and Talbot (1989).

Diagnostic evaluation statements are enabled by the analysis of speech behaviour supporting the observed values of recognition accuracy and intelligibility. In the case of recognition, the most common form of diagnosis constitutes an identification of the human speech sounds which the device fails to distinguish due to phonetic similarity. This identification may be with respect to a sample of the population of device users, for example, by means of tests such as the 100 Word Discrimination Test (Simpson and Ruth, 1987); alternatively, more detailed and general diagnosis may be performed by the utilization of speech databases (speech corpora). Speech databases enable devices to be tested against a large number of standardized utterances, covering the range of voice sounds which may be encountered in the context of tasks involving alternative populations of computer users (Doddington and Schalk, 1981; Knight and Peckham, 1984; SAM Partnership, 1988). Recognition accuracy may then be interpreted in terms of the response of the device under test to utterances which have been classified with respect to a range of phonetic features.

Similarly, diagnostic tests of speech intelligibility have been developed, typically based upon the ability of human listeners to discriminate phonetically similar machine-synthesized speech sounds. The Diagnostic Rhyme Test (DRT) requires listeners to discriminate pairs of rhyming, single syllable words distinguishable by differing consonant sounds. The DRT and related Modified Rhyme Test have been used to decompose the performance of speech synthesizers, enabling the identification of the phonetic features of synthesizer output which are inadequate representations of human speech (e.g. Voiers, 1983; Pratt, 1987).

Speech technologists have produced, then, evaluation statements of both presentation and diagnosis.

(b) *Processes of ST evaluations.* ST evaluations have been predominantly empirical, taking the form of tests of the behaviour of implemented input and output devices. Such tests of speech recognizers have been undertaken using standardized speech databases (e.g. Pisoni, 1986) or by observing representative samples of users speaking messages characteristic of specific tasks. Similarly, tests of speech synthesizers have been performed using standardized diagnostic instruments (e.g. Pratt, 1986) or by observing users listening to representative messages (e.g. Marics and Williges, 1988).

In general, such evaluations have used implemented recognizers and synthesizers, rather than reproductions of such devices. However, device simulation has been used where the device has comprised a system of several interacting components. For example, Nakagawa and Ohguro (1988) have evaluated alternative language parsing strategies using a simulated phoneme recognizer which generated errors pseudo-randomly, according to the confusion matrix of a specific phoneme recognizer.

Speech corpora and diagnostic tests of intelligibility enable devices to be modelled within analytic frameworks offered by linguistics. Such models have potential for utilization in explicit ("formal") analytic evaluations of devices. For example, models of recognizers developed under controlled test conditions may (at least, in principle) be used to predict recognition probability with specific vocabularies and with specific populations of users exhibiting known phonetic and voicing characteristics (e.g. SAM Partnership, 1988). However, at present it would appear that most analytic evaluations are qualitative in nature, taking the form of informal assessments by speech technologists on the basis of implicit models of device behaviour.

In summary, documented ST evaluations are predominantly empirical tests of implemented devices. There is, at present, only limited scope for evaluation prior to implementation, although there is increasing potential for the quantitative analytic prediction of device behaviour.

5.4.2 HF Evaluations

HF evaluations are concerned with the behaviour of people as it relates to that of machines. Where the machines are computers supporting speech interaction, a particular concern will lie with the perception, production and comprehension of spoken language. The criteria employed in HF evaluations may ultimately be derivable from notions of task quality and user costs. These criteria may be applied with respect to particular components of tasks of interest to the evaluator; for example, with respect to speech interaction with the computer (as distinct from non-speech interaction with other off-line devices).

(a) *Products of HF evaluation.* Statements of presentation relating to the operation of speech recognizers have reported task quality in terms of time taken to enter data and of user errors (e.g. Visick, Johnson and Long, 1984). Where the required task quality includes requirements for expeditious completion and an acceptably small number of errors in the products of work, the time taken to enter a datum such that it meets the requisite standard constitutes an integrated measure of task quality ("transaction time" - Knight and Peckham, 1985). User costs have only been reported in indirect ways; for example, Dye, Arnott, Newell, Carter and Cruickshank (1990) asked users to rate the acceptability of recognizer behaviour.

Statements of presentation may also relate to the interaction with non-speech devices, where this interaction is concurrent with speech interaction. For example, task quality in the case of air transport is a product not only of the pilot's entering correct information into a navigation computer but also of the pilot maintaining control of the attitude of the aircraft. Evaluation of systems involving speech interfaces have utilized criteria of secondary task performance such as tracking error as indicators of interference imposed by speech data entry (e.g. Mountford, North, Metz and Warner, 1982).

Diagnostic HF evaluations seek to explain system performance variations in terms of models of human behaviour. These models may be informal "common-sense" models held by the evaluator (e.g. an explanation of mistakes in data entry as being due to the device vocabulary being difficult to learn); or they may be more powerful models deriving, for example, from the applied human sciences. For example, the performance of concurrent speech and non-speech tasks has been explained in terms of multi-resource models of attention (e.g. Wickens et al, 1983); and user interaction with devices exhibiting some of the characteristics of natural language understanding have been characterized in terms of the linguistic constructions of users which are exhibited in interpersonal communication (Waterworth, 1982; Morel, 1986).

HF specialists perform evaluations, then, capable of delivering statements both of presentation and of diagnosis with respect to applied science models of human behaviour (as well as informal behavioural models).

(b) *Processes of HF evaluations.* Most HF evaluations of speech-based systems have been empirical studies involving observations of implemented devices. For example, Visick et al (1984) observed user behaviour where representative subjects used implemented recognizers to support tasks involving parcel sorting. Similarly, Simpson et al (1982) observed aircraft pilot behaviour when cockpit auditory warnings were presented by means of synthesized speech.

Where implemented devices have been unavailable - for example, because of a need to evaluate system performance at a stage in development prior to implementation - empirical methods have been applied using simulations of devices. Simulations have been developed in which the behaviour of the device is emulated by a person: Gould, Conti and Hovanyecz (1983); Newell et al (1990) and Dye et al (1990) have evaluated the performance of speech transcription machines in this way; Richards and Underwood (1984) and Morel (1986) have evaluated telephone accessed database enquiry services; and Baber and Stammers (1989) have simulated speech operated computers supporting a process control task. Fraser and Gilbert (in press) offer a review of applications of human simulation, and the means of implementing such simulations is described in more detail in Section 6.6.2 of the present thesis.

The absence of complete and coherent models of speech interaction, and of supporting techniques for the analysis of speech-based tasks, has limited the scope for the analytic evaluation of speech system behaviour. Quantitative analytic evaluations have not been documented; however, HF specialists undertake informal analytic evaluations utilizing implicit models of device-user interaction (Whitefield et al, in press), and such evaluations are likely to have been applied to speech-based systems.

5.5 A structured method as a solution to RSRE's problem

In principle, a strong solution to the requirement for HF knowledge to support non-specialist practitioners might lie in "engineering methods": explicitly proceduralized evaluation methods based upon engineering principles (see Long and Dowell, 1989). However, Long and Dowell acknowledge that HF engineering methods do not yet exist; they also indicate that the development of such methods, and of the principles to support them, would require a large-scale research effort.

RSRE's requirement demands relatively rapid solution within modest resources. The development of an engineering method, if feasible at all, would not be possible within these constraints. More appropriate would be a solution utilizing existing craft and applied science knowledge. One means by which such knowledge has been recruited effectively to computer (software) design has been by the exploitation of SADMs. Similarly structured methods might be developed to support the task of HF evaluation.

A structured method to support the HF evaluation of speech systems might offer explicit procedures enabling the systematic and complete coverage of the evaluation problem. If specified and implemented effectively, such a method should enhance the performance of the evaluation task, offering benefits such as those claimed by Walsh et al for SADMs. However, SADMs are not intended to support "casual practitioners" of computer system design; rather, they assume that the designer possesses knowledge about the behaviour of the system under development and of entities in the domain of application which can be recruited by the method.

A structured method offered as a solution to RSRE's problem would require an expression, not only of knowledge of the procedure for evaluation, but also of speech interaction knowledge, which could be recruited by the procedure. The knowledge would have to be sufficient to support the process of speech interface evaluation, including diagnosis. A structured method supported by knowledge of speech interaction would be novel, then, in two respects: firstly, because it would extend the notion of structured methods to the process of evaluation; and, secondly, because of its incorporated substantive discipline knowledge. The successful solution of RSRE's problem would, therefore, constitute a more general advance of HF discipline knowledge.

Because the method would be applied at an early stage in procurement, it could not be assumed that devices would be available to enable performance evaluation to be performed directly. Assessments of performance might, in principle, be made either analytically, by applying existing HF knowledge to what is known about the system under evaluation in the abstract; or empirically, using simulations of the target system to reproduce its behaviour. However, as the relevant HF knowledge is so poorly specified, analytic assessment would risk being ineffective. For this reason, the structured method offered assumes empirical evaluation, utilizing system simulations such as those employed by Gould et al (1983).

5.6 Strategy for method development.

Although processes for developing methods to support HF discipline practice have recently been described (e.g. Colbert, Green and Long, 1990; Lim et al, 1990), such documentation did not exist at the time of the development of the method described here. As a consequence of this, and of the fact that the notion of a knowledge-supported structured HF evaluation method was itself novel, the development of the method was exploratory in character. The objective, at the outset, was to capture "good practice" as this related to the process of HF evaluation; and to proceduralize it so that the process could be implemented effectively by individuals lacking HF discipline knowledge. The superordinate strategy to achieve this was one of "hypothesize, test and modify".

The strategy was applied at the level of the four sub-methods which were later to constitute the structured method. Each was treated as an entity which supported a particular task contributing to the solution of an evaluation problem; for example, the device simulation method supported the task of generating a simulation of a target device, which contributed to the solution of the problem of evaluating the target system. [Colbert et al (1990) have similarly advocated the view that methods should be treated as products amenable to the same development process as other aids to task performance.] For each sub-method, a process was specified at a high level, to achieve the transformations in the application domain of the method necessary to achieve the task goal. This specification constituted a hypothesis, and it was generated on the basis of the HF literature and the author's private (craft and applied science) knowledge of HF evaluation. The hypothesis would subsequently be subject to confirmation or refutation.

The process was implemented in the context of a limited case study; for example, the process of device simulation was implemented in the context of an assumed requirement to reproduce the behaviour of a speech recognizer that actually existed. If, in the course of the case study, it was found to be necessary to deviate from the previously hypothesized process, the specification of the process was modified accordingly. The revised process was then proceduralized, by rendering explicit the actions taken in the conduct of the case study.

The knowledge of device-user interaction necessary to conduct empirical evaluations of speech interfaces was embodied in "diagnostic tables". The development of the tables required the identification of the classes of interaction knowledge necessary to support the procedures. Existing guidelines for the design of speech interfaces were subsequently analysed with respect to this classification. For example, in order to specify a simulation to evaluate a target system with respect to a specific design guideline, it was necessary to represent certain "critical attributes" of the task, device and user. Such critical attributes were identified (or inferred) for each guideline selected from the literature and expressed in a tabular format.

The final stage in development was the integration of the sub-methods and their application in the context of a case study presenting a requirement for a complete system evaluation. Following this study, requirements for additional modifications of the procedure were identified and implemented.

Chapter 6 now describes the rationale for the design of the method in more detail. The method itself is presented in Chapters 7 to 10.

CHAPTER 6

SIAM - A SPEECH INTERFACE ASSESSMENT METHOD

6.1 Introduction

Chapter 5 has described a problem faced by RSRE in the procurement of speech-based systems. Preceding chapters have given the rationale for structured HF evaluation methods as a potential class of solution to problems of this general type. The present chapter offers a rationale for a particular structured method as a solution to RSRE's problem: a Speech Interface Assessment Method (SIAM).

Section 6.2 presents the general architecture of SIAM and explains the reason for its form. Subsequent sections describe each of the main components of the method and, similarly, present the reasoning behind them. The procedures of SIAM are presented in Chapters 7 to 10.

6.2 General features of SIAM

6.2.1 Scope of SIAM

One of the distinguishing features of a structured method is that its scope is explicit (Silcock et al, 1990). The scope of SIAM is defined by RSRE's problem, which may be decomposed as that presented to *casual practitioners of HF* (Section 3.3.1) in assessing *task performance* (Section 4.2) incurred when *computerized systems with speech interfaces perform battlefield tasks*. Assessments are to be conducted *prior to system specification* (Section 2.4.2) - and are to generate estimates of performance at *presentation* and to offer *diagnosis* of likely failures to meet desired performance (Section 4.3).

6.2.2 Proceduralization of SIAM

As well as their scope being explicit, the processes of structured methods are explicitly proceduralized (Silcock et al, 1990); that is, such methods include instructions for carrying out the processes. To render the process of SIAM tractable for the assessor, it is divided into intermediate stages, each culminating in a representation of part of the problem being addressed. The procedures describe each step that must be taken to generate a representation or to transform an existing representation to some new state. An important concern in specifying the procedures is the population who are expected to be users of the method - casual practitioners of HF. Because such practitioners cannot be assumed to possess any HF discipline knowledge, the procedures are necessarily expressed in detail.

The notations used for expressing representations vary, depending on the precise information being conveyed. These are described with the procedures of each sub-method in later chapters.

6.2.3 Processes supported by SIAM

For the reasons given in the previous chapter, SIAM supports an empirical approach to evaluation, using simulations to represent the system. At a high level, then, the process for solving problems within the scope of the method consists of *simulation development* and the subsequent utilization of simulations in *evaluation*. In an effort to maintain a systematic relationship between the domain of the problem (interactive battlefield computer systems) and the solution of the problem, the development of simulations is partitioned to reflect a generalized structure of such interactive systems. It is assumed that all systems within the scope of SIAM consist of a *device* and one or more *users*, interacting in the performance of a *task*; this decomposition is widely accepted by HF practitioners (e.g. Card, Moran and Newell, 1983; Dowell and Long, 1989; Carroll and Campbell, 1989). Separate sub-methods of SIAM support the assessor in task simulation development, device simulation development, user simulation development and diagnostic evaluation.

Because tasks (abstract entities) are fundamentally different from devices (generally characterizable as determinate concrete entities) and users (indeterminate concrete entities), the details of the processes of developing simulations of each are different. However, at a high level of description, the three simulation methods bear similarities. In each case, the future system is specified, and then features of its components which might determine speech interaction are identified. These features are reproduced in the simulation, when it is subsequently implemented.

The reason for identifying "speech critical" features is that the simulation need only reproduce those attributes of the system necessary for the purposes of a specific evaluation: indeed, it would be wasteful of procurement resources if the simulation were to reproduce attributes of the system which were known to be irrelevant to evaluation. For example, in an assessment of the memorability of a vocabulary for a speech recognizer, unless there were reason to believe that posture influenced peoples' retrieval of information from memory, it would be unnecessary to reproduce postural attributes of a user in a simulation to support the assessment.

The process of identifying critical features, by definition, assumes the existence of criteria. Here, the criteria for simulation derive from an *a priori* model of speech interaction between a computer and a user (Life, 1990). In an assessment of the vocabulary of a speech recognizer, the model might be one representing memory and its support for speech communication. The

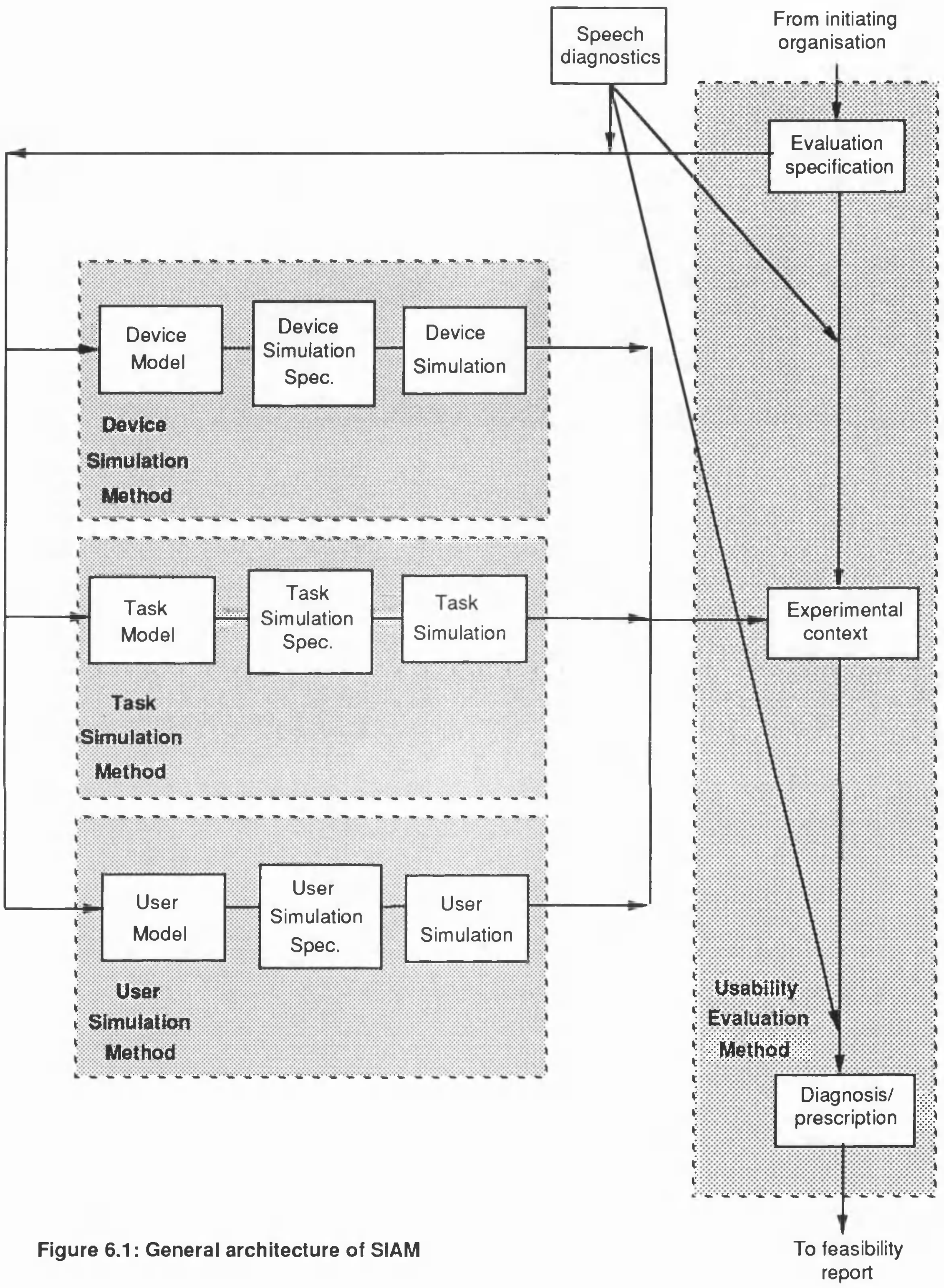


Figure 6.1: General architecture of SIAM

model is necessary, in addition, for interpreting the behaviour of the simulation for the purposes of evaluation (e.g. in suggesting why some vocabulary items are remembered and some are not). It has been asserted in Chapter 3 that casual practitioners of HF are unlikely to possess adequate models of interaction; for this reason, SIAM includes a body of diagnostic information which helps the assessor to develop the necessary models.

Figure 6.1 summarises the structure of SIAM. It consists of three simulation development methods and an evaluation method, these being supported by speech diagnostics. The remaining sections of this chapter give a more detailed rationale for the design of each of these components.

6.3 Knowledge of speech interaction supporting SIAM's process: speech diagnostics

6.3.1 Rationale

Because SIAM is to be applicable by non-specialists, the information to create models of interaction is provided as part of the method. Existing public knowledge of speech interaction between people and computers is expressed in the form of craft and applied science research findings or in the form of guidelines derived from such research (Section 5.3.2). Craft guidelines offer prescriptions without necessarily providing an explicit rationale, and their coverage of issues arising in the design of speech-based systems is recognized as being incomplete. The justification for utilizing such weak knowledge in SIAM is that, although an assessor might not be able to recruit it effectively to predict behaviour and performance analytically, such knowledge would be adequate for the assessor to use in designing an experiment for empirical evaluation; (this argument is elaborated further in Section 6.4). As assessors would be able to observe the behaviour of the simulated system, they would increase their own understanding of interaction, making good some of the inadequacies of completeness and validity in the HF knowledge base. The knowledge would, then, enable particular systems to be evaluated; but on the basis of the results of the experiment, the knowledge itself might be verified and even extended.

In the development of diagnostics, existing craft and applied science knowledge was transformed to render it compatible with the process and proceduralization of SIAM. Specifically, such knowledge had to be *elaborated*, to make explicit its scope, rationale and performance implications; and *interpreted* with respect to the procedures of the method.

Elaboration was necessary because existing sources of HF prescriptive information (e.g. design guidelines) were expressed without explicit statement of their scope, rationale or performance implications. For example, a guideline such as "design the interface dialogue with language compatible with that used for other spoken communication in the workplace", fails to specify the circumstances under which the guideline does or does not hold (necessary for determining its applicability); it fails to explain why adherence to the guideline is beneficial (necessary for diagnosis); and it fails to indicate the likely effect of applying the guideline (necessary for predicting system performance following implementation of the guideline).

Craft guidelines are typically prescriptive in form, and are not immediately suited to supporting the processes of SIAM. For example, information was required to support specification of simulations of tasks, devices and users; to design empirical studies; and to diagnose the causes of sub-optimal system performance. It was necessary, therefore, to interpret the information in the elaborated guidelines with respect to the procedures for carrying out these activities. This was done with the objective of enabling the procedures to refer to specific components of the guideline information for the various purposes described above. For example, support for the specification of a device simulation requires identification of the particular attributes of a device which must be included in a simulation, and these attributes must be partitioned from critical attributes of other system entities, such as the user (see below).

6.3.2 Expression of diagnostic information

The experimental studies which acted as vehicles for the evaluation of the preliminary forms of the usability evaluation method and the device simulation method used a model of device-user interaction in the form of a list of unelaborated and uninterpreted guidelines abstracted from the literature (see Appendices C and D). However, the utilization of such models demanded implicit transformations by the assessor: the requirement upon SIAM was that its process be explicit and applicable by mon-specialists.

In order for the diagnostic information to be accessible to the procedures, it was necessary for the diverse prescriptive guidelines to be represented in a standard form. In the diagnostics supporting SIAM, knowledge of speech interaction between people and computers is expressed as text, structured in a tabular form. The first stage in the elaboration of the guidelines was to express them as expanded production rules, which would make explicit the information lacking in the original prescription. The production rules were of the form:

> IF (condition) THEN (system performance consequence) BECAUSE (interaction model constraint), HENCE (guideline expressed as system design constraint to reduce incompatibility).

For example, in designing speech systems it is regarded as good practice, where possible, to minimize the requirement for the operator to have to learn a vocabulary and syntax for spoken dialogue with a computer which is different from that used for other spoken communication in the workplace (e.g. Jones et al, 1989). The guideline alone is expressed only as a prescription. However, other elements of the production may be inferred on the basis of specialist knowledge of human-computer interaction; i.e.

IF the vocabulary or syntax necessary to operate a speech interface is not the same as that used by the operator when speaking in the working context

THEN there will be an increased probability of lexical and/or syntactical errors in the operation of the computer

BECAUSE there is incompatibility between the knowledge already held by the user and that needed to operate the device, and there is a tendency (particularly under conditions of stress) for more highly-learned behaviour to be elicited than that which is less well-established

HENCE design the interface dialogue with language compatible with that normally used by operators OR give users particular training in the use of the interface.

In most cases, the critical conditions, system performance consequences and rationale for prescriptions were nowhere stated explicitly, and the process of elaboration required some speculative inference. However, it was reasoned that, although interaction models derived in such an unprincipled way would be quite unsuitable for analytic evaluation, they would be adequate to support empirical studies. For example, should the use of the tables result in an assessor proposing ill-founded hypotheses, the inadequacies of the hypotheses would be expected to become evident in the process of testing them experimentally. Repeated use of the tables might, in the longer term, enable the truth of the diagnostics to be better established or the diagnostics to be modified and extended.

Table 6.1 illustrates the form in which the diagnostics are expressed in SIAM. In order to render the IF (critical condition) component of the diagnostics accessible to the procedures of the four sub-methods, interpretation was required with respect to the elements of the human-computer system. The knowledge derived from prescriptive guidelines was decomposed into:

- the circumstances under which the prescription applies (critical system conditions)
- classes of task action bearing on the application of the prescription (critical actions)
- attributes, respectively, of device, user, environmental context, and task, which bear on the application of the prescription (critical system attributes)
- observable consequences of the incidence of critical system conditions, altered by the application of the prescription (behavioural/performance features)
- parameters of behaviour and performance in which the consequences of critical system conditions are observed (critical behaviour/performance parameters)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 1.1 Knowledge required to operate the speech entry device is different from that required to operate the system using alternatively available data entry devices, or to perfrom the task manually.	Device actions	Operating procedure in alternative modes	Skill operating in alternative modes			High probability of operating error (OR evidence that operator is attempting to avoid errors, e.g. by performing slowly) following a change in operating mode (either from speech to the alternative, or vice versa)	Semantic and syntactic error rate. Transaction time following mode change.	*Language incompatibility.* Most recently/frequently used knowledge sources tend to dominate less recently/well established knowledge sources.	Ensure knowledge (e.g. dialogue features) for operating speech interface is compatible with knowledge users already have for doing the task manually or with a non-speech interf ace	See diagnostics 1.2, 1.4
Diagnostic 1.2 Vocabulary and syntax used to enter data by means of speech is different from that normally used in the working environment.	Actions requiring use of language	Device language constraints	Facility with device language and other languages used in the task context.	Speech/language used for off-line task components.		High probability of lexical and/or syntactical errors, characterised by the introduction of language features employed for off-line tasks into computer/user dialogue.	Syntactical and lexical error rate in device operation.	*Language incompatibility.* Tendency for more highly learned responses to be elicited than less well-established ones.	IF device operating language may be modified closer to that used in the working environment without violation of requirements for phonetic dissimilarity, then modify it ELSE give specific training in device language AND/OR encourage change of other language closer to that of device. Use observed error characteristics to guide intervention.	See diagnostics 1.1, 2.1 Supporting documents: Jones et al, 1985
Diagnostic 1.3 Information to be entered is multi-dimensional and continuously scaled.	Data entry actions	Modality and format in which information is accepted by the device			Dimensionality of data. Scaling of data. Time constraints on transactions.	Precision with which continuously-scaled infomation can be transmitted using speech is a function of time (e.g. precision of a geographical co-ordinate is determined by the number of figures in the grid reference). As required precision increases, transmission will take longer and/or will require more effort and/or will run an increased rate of error by comparison with multi-dimensional continuously scaled data entry devices.	Precision of entered data Time to enter data to critical precision.	*Coding incompatibility.* 1. Positive relationship between precision and messagelength when information is transformed from a continuous to a discrete scale. Note also that the message length is multiplied by the number of dimensions of variation. 2. Effort is required to make transformations in 1.	Only use speech for entering continuously scaled and/or multi-dimensional information if a low level of precision is required or if there is low time pressure. OR train users in recodil . (Spatial pointing devices may be more appropriate alternatives.)	Supporting documents: Funk & McDowell, 1982

Table 6.1: Illustration of the expression of diagnostic information in SIAM
- from Diagnostic Table 1.1 (Appendix A): Knowledge incompatibility in speech data input

- rationale for the specified consequences of critical system conditions (interaction model assertion/diagnosis)
- prescription to ameliorate the sub-optimal performance occurring under the critical system conditions (prescriptive options).

Twenty seven diagnostics were derived from the literature. In order to enable them be recruited more easily to specific evaluations, they were grouped according to their rationale. The basis for the classification was to assert that optimal device-user interaction is achieved by ensuring *compatibility* between device, user and operating context (see, for example, Buckley and Long, 1985). Compatibility may be with respect to *knowledge* (i.e. knowledge actually held by the user, relative to the knowledge demanded by the device to achieve the desired level of performance); with respect to *behaviour* (i.e. the actions the user is capable of implementing, relative to those demanded to achieve criterial performance using the device); and/or with respect to the *environment* (i.e. environmental constraints on the user's ability to use their knowledge and to implement actions to achieve the desired level of performance). SIAM includes six diagnostic tables for speech interaction: three based upon guidelines for speech input devices, and three for speech output; in each case, the three reflect underlying rationales attributed to knowledge, behavioural and environmental compatibility.

SIAM is intended to enable the assessment of any speech-based system that might be developed for battlefield use and to be applicable to all battlefield tasks. It was certain that existing prescriptive design knowledge would not be complete with respect to all possible speech-based systems. It was necessary, then, to include within the diagnostic tables "general purpose diagnostics" for the contingency of a failure to identify specific diagnostics corresponding to the problem being addressed by the study. The rationale underlying the proposal of general purpose diagnostics was that they would constitute a generalized form of the specific diagnostics included in each of the six tables. For example, the general purpose diagnostic for behavioural incompatibility was derived from the production rule:

IF users do not have the skills necessary to operate the target device OR users have other skills which interfere with their ability to operate the target device OR the task demands concurrent actions which interfere with one another

THEN system behaviour may fail to support desired performance

BECAUSE the behaviour of which the user is capable does not correspond with the behaviour demanded by the target device to maintain desired performance.

The tables of specific and general diagnostics are reproduced in Appendix A.

6.4 System performance evaluation

6.4.1 Scope of the evaluation method

The scope of SIAM extends to the HF evaluation of proposed speech-based battlefield systems, and, hence, it assumes the performance criteria of task quality and user costs. In assessing feasibility, SIAM enables a procurer to determine whether a proposed system could support desired task quality with acceptable costs to the user, i.e. whether the system would be *usable*. SIAM further supports the identification of requirements for eliminating potential causes of inadequate performance at the subsequent design stage. The scope of the usability evaluation method is, then, the diagnostic assessment of the usability of proposed battlefield systems with speech interfaces.

6.4.2 Processes of system evaluation

The empirical assessment of a system at presentation may be achieved by recording the performance of a simulation of the system under conditions known to be representative of those under which it will ultimately operate. Provided the simulation is known to behave in the same way as the target system, the results may (in principle) be extrapolated to predict operational performance. However, *diagnosis* additionally requires that results are interpreted such that behaviour is explained in a way which is useful subsequently in development: the causes of phenomena must be inferred.

In scientific research, both inductive and deductive inference is employed to extend knowledge. The observation of naturally occurring phenomena may lead a scientist to induce a model to explain the observed behaviour. The model may, subsequently, be used deductively to predict the consequences of manipulating the objects of its concern. However, scientific research demands the validation of models before they are accepted as true, and this may be achieved by systematically testing hypotheses generated on the basis of the model. Induction may be used in system evaluation: the observed behaviour of a system may be explained by a behavioural model, and the model may be used subsequently for prescription. However, such a process cannot be applied when a system is not in a state in which its behaviour may be observed. Prior to implementation, evaluation must rely on some deductive processes.

6.4.3 Rationale for the usability evaluation method

It has been argued previously (Section 5.4.2) that models of speech interaction between people and computers are inadequate for analytic evaluation: they could not be used with reliability for deducing system performance. However, existing models are adequate for *generating hypotheses* of the consequences of design features for system behaviour and performance. The usability evaluation method, therefore, recruits the hypothetico-deductive method of scientific research to system evaluation. The impoverished interaction models embodied in the diagnostics (Section 6.3) are used to design experiments to test hypotheses and simulations capable of supporting the experiments. Such an empirical approach is certainly not novel: it is employed extensively by HF practitioners. What the method seeks to do is to render systematic the evaluation processes which might be used by an HF specialist.

The method assumes three phases in evaluation: experimental design, observation and behavioural interpretation. The experimental design stage, firstly, involves deciding the scope of an experimental assessment; this is supported by the assessor selecting a set of diagnostics having consequences believed to be relevant to the performance required of the target system. The selected diagnostics then support the design of an experiment by identifying dependent and independent variables. Following observation of behaviour in the experiment, the diagnostics are referred to again to assist in the interpretation of the observed behaviours.

The development of the method involved the proposal of a preliminary method and its subsequent implementation in the context of an evaluation of a device to support a battlefield task. The procedure employed in the implementation provided the basis for the final expression of the method.

Preliminary method. A preliminary method was proposed as a flowchart, which constituted a systematic expression of the process of hypothetico-deductive research in the context of system simulation (Figure 6.2). The preliminary method did not support the process of *specifying* observational studies, and it was not explicitly proceduralized.

The preliminary method was applied in the context of a study[1] to assess the usability of variants of an interface to support a computerized version of the task of the artillery forward observation officer (FOO). At the time of the evaluation, the diagnostic tables had not been developed and the "interaction model" was expressed as a set of propositions pertaining to the behaviour and performance of concern in the study. The propositions were derived from HF guidelines. The study identified requirements for elaborating the preliminary usability evaluation method. Specifically, the need was identified for extending it to include the process of specifying the design of the evaluation study; for consideration of the support required for data collection, compression and statistical analysis; and for developing an explicit interface between the method and the interaction model.

[1] The study is described in Appendix D; and the development of the task and device simulations is described in Appendices B and C respectively.

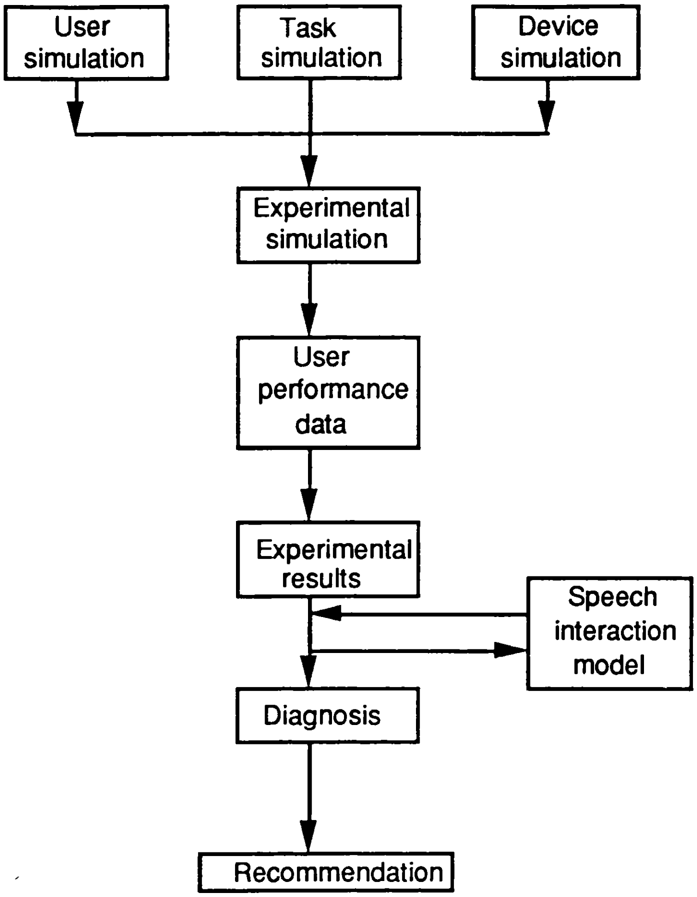


Figure 6.2: A "preliminary method" for usability evaluation (see text)

Enhanced method. The stages of the preliminary method were extended to include specification of the target system and of the problem, and the subsequent development of these specifications in the formulation of a "solution strategy" (including experimental design). The notion of the diagnostic table was advanced as a solution to the requirement for a representation of knowledge of device-user interaction which could be accessed by procedures within the usability evaluation method. It was decided not to elaborate the process of experimentation to include processes such as statistical treatment, on the reasoning that engineers using SIAM would have had training in general research methods and statistics, and that this knowledge could be recruited by them to evaluation.

The elaborated method was proceduralized by explicit identification of the actions performed in the practical application of the preliminary method. Although minor modifications occurred following its application (relating primarily to the interfacing with later versions of the diagnostic tables) the procedures were substantially as presented in Chapter 7.

6.5 Task simulation

6.5.1 Scope of the task simulation method

The central objective of the three simulation methods of SIAM is to reproduce the behaviour of the battlefield system which is the subject of assessment (i.e. the target system). The behaviour of the user and of the computer occurs as a consequence of performing a military task; that is, in the, achievement of a goal which will ultimately derive from the army's objectives. The reproduction of system behaviour requires, then, representation of the task; the function of the task simulation method is to support such representation.

The notion of the task is central to the research and practice of HF, yet, perhaps surprisingly, there is no consensus on its definition. For example, Dowell and Long (1988) define a task in terms of a requirement for a change in the state of the attributes of objects in a domain of work, which is to be achieved with respect to some desired level of performance. Within Dowell and Long's definition, the task (i.e. required attribute state change) is clearly distinguished from the system behaviours which achieve the change. Most other definitions (e.g. Card, Moran and Newell, 1983; Stammers, Carey and Astley, 1990) refer not only to required state changes (goals), but also to the behaviour exhibited by the system in achieving them. The development of SIAM preceded Dowell and Long's conception, and, while acknowledging the merits of their definition of the task, a definition is assumed here in which the task is an expression of the interaction of objects in the domain of work with device and user behaviours in achieving a goal. So the definition includes both goals *and* behaviour. The scope of the task simulation method extends to all army tasks which might occur on the battlefield and which might be supported by a speech-based computerized system. The method reproduces, then, the behaviour of battlefield objects (e.g. enemy targets, friendly forces and environmental features), and their interaction with the target system.

6.5.2 Task analysis

The reproduction of system behaviour necessitates a process of analysis to support the design of a task simulation. Task analysis is widely recognized as a general technique to support HF practice (e.g. Diaper, 1989). The term is used to refer to an extremely diverse set of techniques; its definition will be taken here as *the development of a representation of a task for some purpose.* Purposes of task analysis include the support of activities in system development and in HF research. In system development, task analysis has been proposed to support: the capture of user requirements (e.g. TAKD - Johnson, 1985); system design (e.g. CLG - Moran, 1981; TAG - Payne and Green, 1983); user training (e.g. HTA - Annett, Duncan, Stammers and Gray, 1971); and analytic performance evaluation (e.g. KLM - Card, Moran and Newell, 1983; CCT - Keiras and Poulson, 1985). The methods and techniques offered to support these activities reflect their widely differing objectives, and they have been reviewed, for example, by Wilson, Barnard and MacLean (1987).

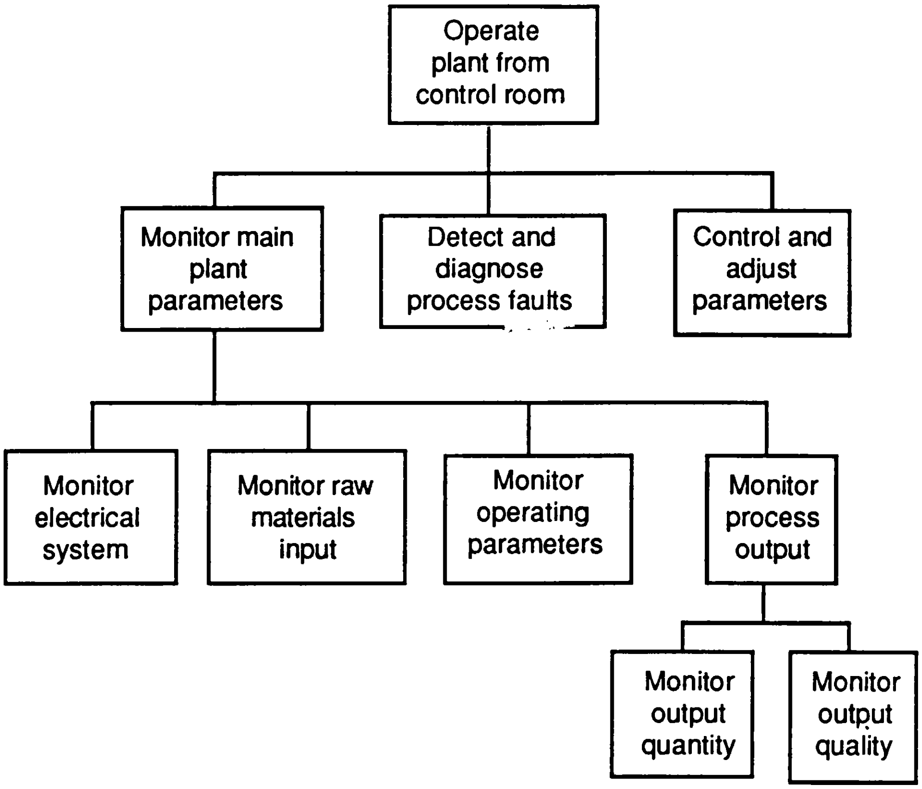

Operate plant from control room
Monitor main plant parameters
Detect and diagnose process faults
Control and adjust parameters
Monitor electrical system
Monitor raw materials input
Monitor operating parameters
Monitor process output
Monitor output quantity
Monitor output quality

Figure 6.3: Illustration of hierarchical task analysis applied to industrial process control activities (after Stammers et al, 1990)

No existing task analysis techniques have been designed specifically for task simulation; however, techniques do exist for reproducing task behaviour in the context of training. In particular, the technique of hierarchical task analysis (HTA) was developed to support the development of operator training programmes (Annett et al, 1970). In HTA, the behaviour of a skilled task performer is decomposed into a hierarchy of actions, so that the desirable behaviour of the skilled person may be segmented into units for teaching to unskilled individuals. The hierarchy may be expressed as a tree diagram (see Figure 6.3).

Van Dijk (1980) has observed that people readily impose structure upon the behaviour of objects, which enable them to describe and to reason about behaviour. The intentional behaviour of other people (i.e. behaviour exhibited in the achievement of a goal) is interpreted as the manifestation of "actions". An action will comprise a coherent sequence of events which, if the action is successful, culminate in the intended goal. Van Dijk further observes that people are able to conceive of actions at different levels of abstraction. Actions may be described at a low level, such as in descriptions of individual (intentional) movements; or at higher levels, where sequences of movements are interpreted as parts of more complex actions. Thus, the low level actions of sitting down, picking up a pen and forming letters on paper to convey ideas may be conceived, together, as constituting the more global action of writing. This may, in turn, be part of a yet higher level action of authoring a book. Van Dijk notes that there is a systematic mapping between people's conceptions of high and low level actions, such that they bear a hierarchical relationship.

Given that people are generally able to interpret the intentional behaviour of others in terms of actions, the concept of action offers a potentially widely acceptable unit for the description of tasks. HTA has exploited this acceptability and, further, renders the action hierarchy in the graphical tree diagram form which exposes both the functional relationships between high and low level actions and features of their temporal relationships (e.g their order of occurrence). Annett et al report that such diagrams are readily assimilable by non-specialists and so are particularly appropriate to support training. As task simulation, in common with training, requires representations of tasks which can (a) support the reproduction of behaviour and (b) be readily comprehended by both non-specialist assessors and target users, HTA has the potential to support a task simulation method.

6.5.3 Rationale for a task simulation method

The rationale for the task simulation method is that a simulation suitable for system evaluation may be developed by the progressive transformation of a hierarchical description of an existing version of the target task. It assumes that, although the introduction to an existing task of a novel device (and/or a user with novel attributes) will alter the low level actions exhibited to achieve a task goal, the higher level actions will be unchanged.

A tree notation is utilized to express the hierarchical relationships between actions. Using this notation, a representation of an extant version of the target task (or of a related task) is generated and used as the basis for developing a representation of a "future" version which involves use of the target device. In achieving this transformation, those (low-level) actions which are predicted to be eliminated or changed by the introduction of the new device are deleted from the bottom of the hierarchy. New low-level actions are then generated on the basis of the best available knowledge of the functionality of the target device. The description of the future task is further transformed by the identification of that subset of actions which is critical to speech interaction and so which must be included in the specification of the simulation. This specification is, finally, implemented.

The process of the method was derived by the progressive refinement of HTA, briefly described below and more fully presented in Appendix B. A preliminary method was proposed and trialled in the context of a battlefield observation task. The process of the preliminary method proved unsatisfactory and was consequently modified substantially.

Preliminary method. It was recognized at the outset of the development of SIAM that system evaluation would require the analysis of the system task. The definition assumed for the term "task" included, not only the on-line activities of the operator (i.e. interacting with the computer), but also off-line activities contributing to the achievement of the system's goals. A hierarchical representation was initially chosen because it could capture the functional relationships between on-line and off-line actions in the achievement of task goals.

The differing requirements for analytic and empirical evaluation (Section 4.3.3) were not elucidated until later, and the initial approach taken to task analysis was orientated towards the analytic evaluation of the system. Specifically, it was reasoned that existing knowledge of speech interaction with machines was expressed in ergonomic guidelines for the implementation of speech interfaces. Evaluation required the application of this knowledge to particular cases. Consequently, task analysis demanded characterization of the task with respect to those features which were critical to the application of the guidelines.

Relevant features included dynamic characteristics of the task, such as competition between task actions for the same operator resources (e.g. auditory and visual attention; manual and vocal resources). This assessment was performed by the analysis of a log relating task actions (and the operator resources demanded by them) to time. Also relevant were static characteristics of the task, such as the type and quantity of data manipulated by the system; a static analysis might identify the data to be manipulated as being either spatial or verbal in nature; specify the form of data strings; and indicate the required speed of data entry to the computer. The approach was trialled in the context of the task of the forward artillery observer. Video data were recorded in the context of a trial artillery mission performed by a Forward Observation Officer and a supporting team on a training simulator (see Appendix B). In attempting to perform static and dynamic analyses of the video record, a number of inadequacies of the approach became apparent:

- explicit specification was required of the relationship between observable behaviour, task actions and knowledge held by the FOO supporting his actions
- because of its orientation towards analytic evaluation, the preliminary approach failed to engage with the problem of specifying simulations to support empirical assessments
- the analysis only addressed extant tasks and failed to take account of the changed behaviour following the introduction of the target device
- the relationship between the task representation and the model of human-computer interaction implicit in the guidelines was unclear.

Revised method. Following the failure of the preliminary method, a revised task simulation process was proposed, taking the form presented at the beginning of this section. The revised method was orientated specifically to the development of task simulations (in that the function of the interaction model was to specify attributes to be represented in the simulation, rather than to enable analytic evaluation); and it supported a process of task synthesis to take account of the introduction to the task of the target device. It was, then, clearly intended to support *empirical* evaluation.

Given the new objective of reproducing task behaviour, a clearer rationale was specified for the hierarchical form of description. As intentional behaviour is generally perceived as being attributable to actions (van Dijk, 1980), an action-based task representation was confirmed as likely to be compatible with the cognition of assessors. The method offered heuristics to assist in hierarchical decomposition (such as suggesting that actions should not be decomposed into a very large number of component actions at the layer below); but the intention was to utilize the "natural" abilities of the analyst in decomposing observed task behaviour into coherent actions.

The problem of the relationship between the task representation and the interaction model was addressed later by expressing the diagnostics in the tabular form which was accessible to the process of the method - see Section 6.3.2. This process was proceduralized and successfully applied *post hoc* to the data record of the FOO trial; the results of this application are presented in Appendix B. The procedures are presented in Chapter 8, with examples of task representations developed using the hierarchical notation.

6.6 Device simulation

6.6.1 Scope of the device simulation method

The device simulation method supports the development (i.e. specification, implementation, testing and refinement) of a simulation which reproduces the behaviour of a target device. RSRE's problem derives from a need to assess the feasibility of future devices; the scope of the method extends, then, from computers with speech interfaces of the current generation (e.g. simple isolated word and connected speech recognizers; text-to-speech synthesizers) to future devices exhibiting greater capabilities of linguistic analysis (e.g. devices capable of interpreting and generating speech approximating to natural language).

6.6.2 Human simulation

In designing the method, the implementation of simulations of future computers with speech I/O was recognized as being a central issue, influencing all other stages of simulation development. Speech technology is a member of that class of technologies which derive from a motivation to emulate aspects of human behaviour. Where a device is intended to behave, in certain respects, like a person, reproduction of its "human" behavioural attributes may be achieved by a person acting like the machine. Because people are able to modify their behaviour voluntarily according to previously agreed specifications, they may, in principle, be used to support simulations.

The technique of human simulation has been used, for example, by historians, economists and operational researchers to support the reproduction of complex political, economic and military systems (Hermann, 1967); by the developers of large human-machine systems, such as command and control systems (Parsons, 1972); and, more recently, in HCI research. Chapanis (1975) seems to offer the first description of a variant of the technique being used to study communication within small man-machine systems involving a computer and user; the behaviour of the computer was simulated by a person located in a separate room. The results of Chapanis's observation were used to propose general requirements for communication facilities to support interaction between a person and a machine "co-operating" in the performance of a task.

More recently, expert systems have been simulated to determine requirements to support the communication between such systems and their users (Diaper, 1986); the person simulating the device communicated with the "user" (more precisely, an experimental subject representing the user - see Section 6.7.2) by means of linked computer terminals with conventional keyboards and visual displays. Section 5.4.2 has reviewed the now considerable utilization of human simulation techniques to simulate speech I/O devices. The literature will, here, be referenced only selectively, for the purpose of presenting the rationale for the device simulation method; the reader is referred to Fraser and Gilbert (in press) for further examples of the use of the technique.

To illustrate the technique of human simulation of speech devices, consider a requirement to simulate a system in which a user interacts with a computer by means of speech data entry and by observing visually-presented text output. Such a simulation might be implemented using the arrangement shown in Figure 6.4. The subject representing the user (on the left) is led to believe that his/her microphone is connected to a speech recognizer, while, in fact, it is a link to another person who is located in a separate room (a "system subject", who emulates the target speech interface). Utterances of the user subject are intercepted by the system subject, who enters the data manually into a computer terminal. The data are displayed visually, both at the system subject's terminal and on the user subject's visual display. The device simulation, in this case, comprises the system subject and the information links between the two subjects (the "communication device"). Other types of speech interface might be simulated by modifying the communication device; for example, the computer-computer link might be replaced by a second speech link, enabling the simulation of an interface offering "speech feedback" from the system subject to the user subject.

A simulation such as that illustrated in Figure 6.4 was exploited with some success by Gould, Conti and Hovanycz (1983), who reproduced the behaviour of a speech transcription device (referred to by the authors as a "listening typewriter"). The simulation was implemented

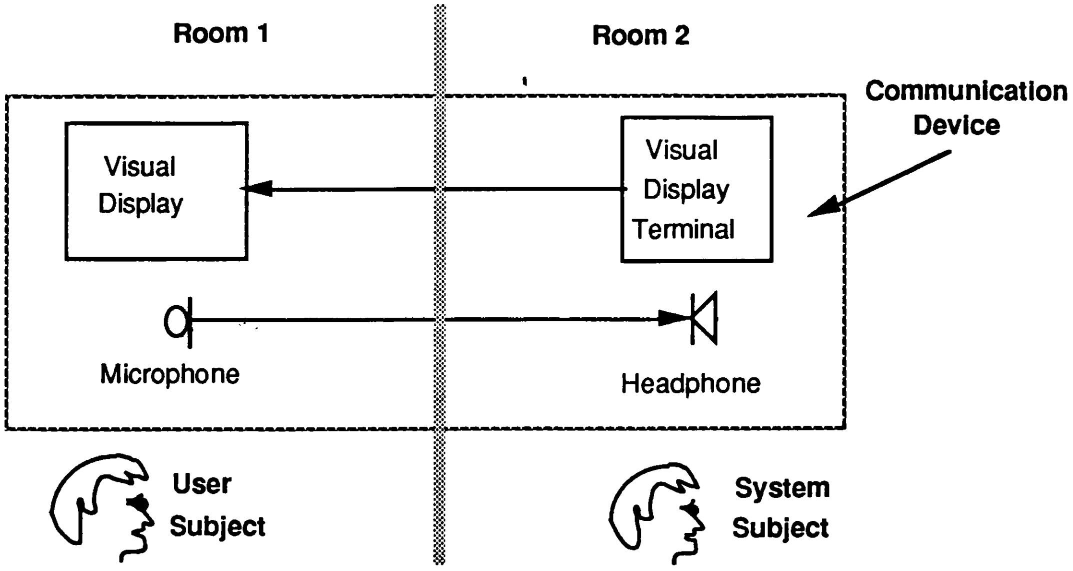


Figure 6.4: A generalized system supporting the human simulation of speech technology.

The system subject emulates the behaviour of the target device (in this case, a speech recognizer writing to a visual display), by transcribing the utterances of the user subject onto the visual display terminal. The typed text is presented back to the user subject who is located in a separate room, unaware of the system subject's presence.

using a high speed audio typist to transcribe the speech of user subjects onto a conventional computer terminal, the text being presented to user subjects on a visual display. The purpose of the simulation was to determine the effect on letter writing performance (particularly speed of writing) of the size of the vocabulary of the speech recognizer. Although Gould et al's report is widely cited, their device simulation has been criticized, for example, by Damper (1988); by Newell et al (1990) and by Dye et al (1990). It has been argued that a typist operating a conventional typewriter layout (QWERTY) keyboard is physically incapable of emulating the transcription speed of a future listening typewriter, because typing rate is slower than normal speaking rate. This has led Newell et al and Dye et al to propose the use of a Palantype (stenographic) system to support the transcription of speech to visually-presented text. Such systems support rapid transcription by the use of a chord keyboard, which enables the entry of entire syllables by the simultaneous depression of multiple keys.

The above criticism of Gould's QWERTY-based implementation are founded on the premise that the simulation does not support fidelity adequate for the purpose of investigating device-user interaction; i.e. the *quality* of Gould's simulation does not meet these investigators' criterion of acceptability. A weakness in Newell's and Dye 's position is that what constitutes "acceptable" performance is implicit and imprecise: indeed, it is probably not possible to determine whether their Palantype simulations constitute adequate representations of their target devices[2].

Nevertheless, the criticism demonstrates the point that human simulations may, themselves, be viewed as human-machine work systems designed to meet particular (if sometimes implicit) performance requirements. In this instance, the system comprises the system subject interacting with the communication device. "Required performance" for a human simulation will include some expression of quality, where this is the degree of correspondence between (relevant) parameters of the behaviour of the target device and those parameters of the behaviour of the simulation (see Life, 1990). A concern of the developer of a human simulation will be to ensure that the communication device offers appropriate support for the system subject in carrying out the simulation task.

6.6.2 Rationale for the device simulation method

The rationale for the device simulation method is that the behaviour of future speech interfaces can be simulated effectively by human subjects (evidence coming from the work reviewed in Section 5.4.2). However, if they are to support assessments of system behaviour and performance, such simulations need to reproduce accurately the relevant features of the behaviour of the target device. As was the case in task simulation, the method assumes a process of simulation development involving specification prior to implementation, with the objective of a device simulation exhibiting the requisite behavioural features. Specification includes that of the behaviours of both the system subject and the communication device.

As with the usability evaluation and task simulation methods, the device simulation method was developed in two phases. During the first phase, a preliminary method was specified in outline and applied; while, in the second, the method was explicitly proceduralized and integrated with other components of SIAM. These two phases are now described in outline; the case study which acted as the vehicle for the development of the method is presented in Appendix C.

Preliminary method. The "preliminary method" was proposed before the architecture for SIAM had been specified. It did not have explicit links with other elements of SIAM, such as

[2] The same authors also draw attention to the potential significance of ensuring that user subjects remain unaware of the existence of the system subject (a potentially important determinant of the psychological fidelity of the simulation). Newell et al (1990) report the findings of a study comparing the behaviour of subjects who were under the impression that they were interacting with a listening typewriter and others who knew that the "device" they were using was a simulation. It was found that the users' reports of the acceptability of the listening typewriter differed between the two experimental conditions: those who believed they were interacting with a machine rated listening typewriters more negatively than those who were aware that they were interacting with a human simulation. The finding, again, illustrates the importance of ensuring simulation fidelity appropriate for the particular evaluation being performed.

with the task simulation method (which was undergoing development in parallel). However, it was recognized that an important element of the putative speech interface assessment method was the utilization of human simulations. Such simulations had, in the past, been developed informally, against implicit performance requirements (see Section 6.5.1). The proposal of a preliminary device simulation method was directed, then, by the need to systematize and to make explicit the process of developing human simulations.

An "ergonomic approach" to simulation development was proposed, on the rationale that the performance of a human simulation is primarily determined by the interaction between the system subject and the communication device. More specifically, performance is determined by the compatibility between the system subject and the communication device: incompatibility will result in sub-optimal behaviour and, hence, performance which fails to meet the requirements for target device assessment.

The ergonomic approach comprised six stages (see Life and Long, 1987):

- specification of simulation elements (i.e. deciding which aspects of the target device are to be represented in the simulation)[3]
- specification of target device performance parameter values (i.e. deciding on the target device performance to be reproduced in the simulation)
- evaluation of the performance of a simulation system comprising a system subject and the simplest possible communication device
- development of system subject/communication device compatibility model (identifying incompatibilities causing inadequate simulation performance)
- specification and implementation of ergonomic intervention
- evaluation of simulation performance following intervention.

It was recognized that, where the purpose of the simulation is to study the behaviour of a device which does not yet exist, specifying the future device (including its expected performance) may be difficult. This, in turn, would make it difficult to evaluate the adequacy of the simulation. Although this issue would have to be addressed at a later stage, it was reasoned that performance setting was peripheral to the determination of whether or not the ergonomic approach was viable. For this reason, the approach was tested by applying it to the simulation of a currently available device, the behaviour and performance of which could be determined empirically.

Appendix C describes three experiments performed in the development of a simulation using the ergonomic approach; (also reported in Life, Long and Lee, 1988). The first experiment resulted in a model of the behaviour of an extant speech recognizer (the target device). The second evaluated the performance, with respect to this model, of a simulation system

[3] The preliminary method assumed the existence of specification of the target device.

comprising a system subject and a simple communication device. In this case, the communication device consisted of two elements: an intercom link, enabling the system subject to hear the user subject's utterances; and a conventional keyboard-based data link, enabling information to be entered by the system subject to appear on a text display visible to the user subject (see Figure 6.4). A third experiment evaluated the performance of the simulation system following the introduction of an enhanced communication device. The performance was found to approximate more closely to that of the target device.

The principle of optimizing the performance of human simulations by means of ergonomic intervention was demonstrated to be viable. However, the experiments resulted in the identification of requirements for modification to the originally proposed "ergonomic approach". The device simulation method required a procedure for specifying the characteristics of the target device prior to specification of the simulation; and simulation specification required specification of the target device functions to be included in the simulation, as well as of "critical parameters" of device behaviour. It was also necessary for the method to support the setting of criteria for the adequacy of simulation performance; and diagnosing inadequacies in the simulation. The stage requiring the implementation of a crude preliminary simulation - solely for the purpose of determining the requirements for the specification of the simulation - was judged to be potentially wasteful of resources and so was eliminated in the enhanced method.

Enhanced method. The enhanced device simulation method was specified within the architecture of SIAM presented in Figure 6.1. It was integrated with the other methods; the structure of intermediate representations was specified; and the processes for transforming representations were proceduralized.

In the light of the requirements for development identified above, procedures were introduced for the generation of a description of the target device; for selecting critical parameters of device behaviour (utilizing the interaction model embodied in the diagnostic tables); and for the specification of ergonomic intervention. One approach to the specification of ergonomic intervention was to develop special diagnostic tables, embodying system subject/communication device compatibility models and analogous to the tables intended for the diagnosis of inadequacies in the performance of the target system. However, the development of such tables was beyond the scope of the present project, requiring empirical research into the behaviour and performance of human simulations. At present, then, ergonomic intervention relies on implicit ("common-sense") knowledge held by the assessor.

In summary, the processes of the device simulation method are those of specifying the device simulation by identifying attributes critical to device-user interaction; implementing a human simulation of the target device (which reproduces the latter's behaviour); and optimizing the performance of the human simulation, by ensuring effective interactions between the system subject and communication device. These processes support the development of device simulations exhibiting the features necessary in assessments of system behaviour and performance. The procedures of the method are presented in Chapter 9.

6.7 User simulation

6.7.1 Scope of the user simulation method

Just as an evaluation of system performance demands representation of target device behaviour, so it demands representation of the behaviour of users of the target device. The user simulation method supports the development of a reproduction of user behaviour appropriate to support empirical evaluation. As RSRE's problem is specifically concerned with army battlefield systems, the scope of the user simulation method covers army users in battlefield contexts.

6.7.2 Processes of user simulation

The human behavioural sciences, such as psychology, use experimental methods in which the behaviour of small samples from specified populations is observed under controlled conditions. The recruitment of subjects demands recognition that there are differences between the behaviours of individuals. Those differences which may interact with the behaviour of concern to the investigator must be controlled, if they are not to mask evidence of experimental effects or invalidate extrapolation of the results beyond the sample. For this reason, experimenters take particular care to ensure that subjects are representative with respect to relevant behaviours, and representativeness is achieved by the selection and training of subjects against specific criteria.

Although the objectives of HF may differ from those of sciences such as psychology, the experimental method is utilized by HF practitioners in an analogous way. The behaviour of a small group of subjects representing a population of device users is observed under controlled conditions, and conclusions are drawn with respect to the population as a whole. The same requirement exists for users to be representative of the target population, and representativeness must, similarly, be ensured by the selection and training of subjects against specific criteria. In HF experiments, where the purpose is to reproduce the interacting behaviours of device and user, the subjects may be viewed as *simulations* of target users.

6.7.3 Rationale for the user simulation method

The rationale underlying the user simulation method is that intentional behaviour is constrained by peoples' physical and mental structures (knowledge); hence, the reproduction of human behaviour requires representation of the physical and mental structure of those whose behaviour is to be reproduced. The reproduction is achieved by selecting people to act as user subjects who possess appropriate physical and mental structures, or who have potential for acquiring those structures by training. It is then assumed that, when the subjects undertake a task having a goal structure and operational constraints equivalent to those presented to target users, they will behave in the same way as target users.

Like the methods for simulating tasks and devices, the task simulation method assumes specification of the target entity (in this case, users), identification of critical attributes to be included in the user simulation, implementation (i.e. the selection and training of subjects) and evaluation to ensure that subjects possess the relevant attributes of target users.

The development of the user simulation method was less complete than that of the other methods comprising SIAM, in that the notation of the method was less formalized and it underwent less extensive trialling. The justification was that SIAM was intended to be applied by engineers employed by RSRE (either directly or under sub-contract). An assumption, at the outset, was that such assessors would have access to populations of soldiers who could be recruited as user subjects. It was reasoned that such subjects would be likely to be representative of target users, so the requirement for a proceduralized method for developing simulations of users was reduced.

However, circumstances could be envisaged under which army subjects would be unavailable (e.g. if an evaluation was required to be performed at short notice). It was further recognized that even subjects from an army population would need to acquire knowledge to perform the simulated task: for example, they would need to be trained in the operation of the simulated device. For these reasons, a user simulation method was necessary, although it could assume a lower priority in the development of SIAM.

The preliminary form of the user simulation method was as a flow diagram. The process was implemented successfully in the context of a speech interface evaluation to be described later. As the outcome of this application was acceptable, the procedure used to implement the process was subsequently made explicit in the form presented in Chapter 10.

6.8 Empirical work in the development of SIAM

6.8.1 Summary of studies supporting development

Section 5.6 characterized the general strategy by which SIAM was developed as one of hypothesizing a process for each sub-method; trialling the process; modifying the process in the light of the results of the trial; and expressing the refined method as a set of explicit procedures. Sections 6.4 to 6.7 have described the implementation of this strategy in the development of each of the four sub-methods of SIAM. The four sub-methods were subsequently integrated, and the procedures of the complete method were trialled together in the context of a project concerned with the development of a novel speech interface for a battlefield computer. The study is used in Chapters 7 to 10 to illustrate the application of SIAM.

In the trial, the method was applied by the author; clearly, then, this test was inadequate as an objective evaluation of the method. Firstly, the author possessed private discipline knowledge, the existence of which would likely result in a failure to expose inadequacies in the expression of the procedures; and, secondly, the author had an interest in the success of the method which would potentially bias the outcome of the test. For this reason, a further trial was conducted, in which SIAM was applied by a commercial consultant in a second study of the same battlefield speech interface. The results of the second evaluation are reported fully in Chapter 11.

Table 6.2 summarises the roles of the various empirical studies undertaken in the development of SIAM.

6.8.2 Presentation of the procedures

The four sub-methods of SIAM are presented in the chapters which follow. Each of the chapters begins with a summary of the sub-method to be subsequently described. The intermediate representations are identified; and the procedure for developing each one is given, with observations on implementing the procedure. The representations and their development are then illustrated. Each chapter concludes with a brief commentary on the sub-method, made in the light of the findings of the evaluation described in Chapter 11.

The present chapter now concludes with an outline of the speech interface development study which is used for illustration purposes in Chapters 7 to 10.

6.8.3 Context for the study illustrating the application of SIAM

The Royal Military College of Science (Cranfield) - RMCS - had developed a novel algorithm for speech signal encoding. The algorithm supported automatic speech recognition with relatively high accuracy and low electrical power demands, and it offered the potential for adaptation to variations in the voice quality of a user over the course of a session of use (Power, Hughes and King, 1986). These features were potentially attractive in devices for the entry of data to battlefield computers, and RSRE had sponsored RMCS in the development of a prototype "voice to data convertor" (VDC) to act as a demonstrator for the

Table 6.2 Empirical studies supporting the development of SIAM

Study title	Functions of study	Reporting
1. The analysis and simulation of battlefield tasks: the case of forward observation	(a) To evaluate the preliminary task analysis method (b) To propose and apply *post hoc* a task simulation method (c) To simulate a task for experimental evaluation of a speech interface to support the task of the artillery observer (see 3 below)	Appendix B
2. System subject assessment study 1: The simulation of a connected word speech recognizer	(a) To trial the preliminary device simulation method (b) To develop a device simulation for the experimental evaluation of a speech interface to support the task of the artillery observer (see 3 below)	Appendix C See also Life and Long (1988)
3. Evaluation of a connected speech recognizer to support the task of the artillery observer	(a) To trial the preliminary usability evaluation method (b) To evaluate the potential utility of speech input in computer support for the artillery observer	Appendix D
4. Evaluation of speech data entry requirements for a computer supporting indirect weapon engagements (study 1)	(a) To test the four sub-methods of SIAM together (b) To advise device developers on the optimal string length for data entry and feedback	Illustrations of the application of SIAM in Chapters 7-10 See also Life and Lee (1990)
5. Evaluation of speech data entry requirements for a computer supporting indirect weapon engagements (study 2)	(a) To evaluate SIAM's support for interface evaluation in the context of a commercial project (b) To evaluate the usability of a speech interface and to suggest enhancements to its developers	Chapter 11

time encoded speech concept. SIAM was applied in the context of a study performed to support the development of the VDC.

The device existed as a prototype which was subsequently to be developed as a demonstrator to illustrate its potential for the support of battlefield observation tasks. The function of the device was to enable the transmission of battlefield information, which was entered by means of speech; feedback was also presented to the user by means of speech, generated by a speech synthesizer. The device developers were concerned with the selection of an appropriate string length for data feedback.

In the prototype, feedback was presented after each word had been entered (i.e. the group size was one item). There were technical advantages associated with this form of dialogue (relating to the facility offered by the VDC for its word templates to adapt to long term changes in the user's voice); however, there were also potential disadvantages for the user, in that he would be forced to enter data in a format probably different from that favoured for the communication of the same information by other means. The study used here to illustrate the application of SIAM intended to support detailed user dialogue design, with a specific evaluation of alternative feedback string lengths (2, 4 or 8 digit strings).

The order in which the elements of SIAM are described in this chapter does not necessarily reflect the order in which development actually occurred, nor the order in which the methods are intended to be applied. In fact, development of the methods and diagnostic tables occurred in parallel, and their use to support evaluations requires the "interleaving" of procedures of the four methods. The order of presentation here is chosen primarily for clarity of explanation.

CHAPTER 7

USABILITY EVALUATION METHOD

7.0 Process of usability evaluation - summary

The process of usability evaluation requires the assessor to develop eight representations which support the specification, implementation and interpretation of experimental studies of system behaviour (see Figure 7.1).

1. Following negotiation with the procurement organization which is initiating the feasibility assessment, a *preliminary problem specification* is agreed, constituting statements of technical issues to be addressed in the assessment and of the resources available to support assessment.

2. The target system is analysed, to identify the task, device and users. A *preliminary system specification* is generated for the purpose of scoping the assessment.

3. A model of device-user interaction is specified by selecting diagnostics which reflect the concerns identified in the preliminary problem specification (i.e. by *configuring the diagnostic tables*).

4. A *solution strategy* is developed, in which the configured diagnostics are used to specify an experiment. The diagnostics operationalize the model of interaction with respect to empirical observation, specifying critical parameters of the system to be simulated (utilized by the simulation methods) and independent and dependent variables of the study.

5. The solution strategy is implemented by integrating the outputs of the task, device and user simulation methods to generate an *experimental context* for the observational study.

6. An experimental assessment is performed, which generates *interaction data* expressed with respect to the dependent variables of the study.

7. The data are analysed for the purposes of diagnosis and prescription (i.e. to generate an *analysis of device-user interaction*).

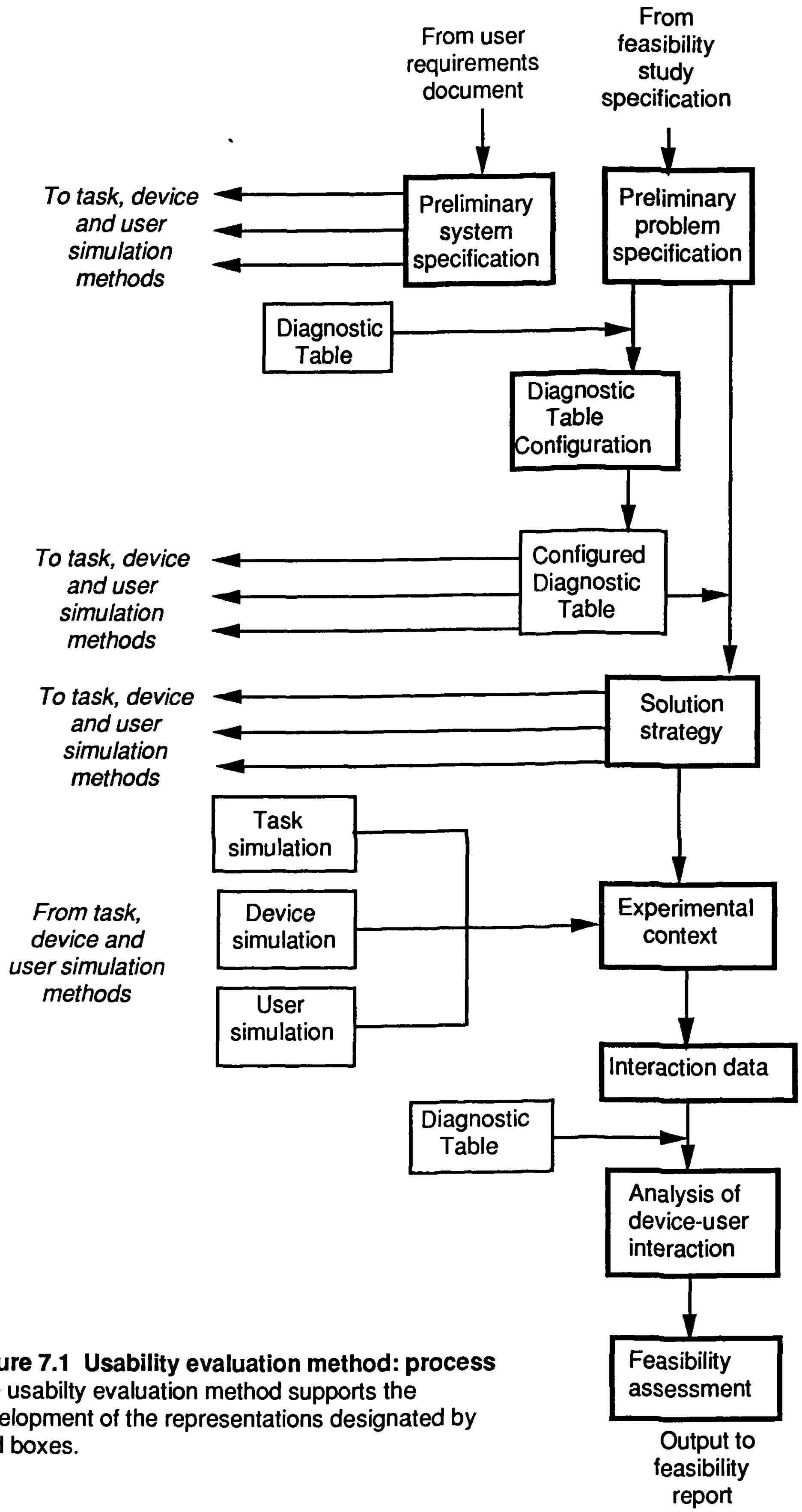

From user requirements document
From feasibility study specification
To task, device and user simulation methods
Preliminary system specification
Preliminary problem specification
Diagnostic Table
Diagnostic Table Configuration
To task, device and user simulation methods
Configured Diagnostic Table
To task, device and user simulation methods
Solution strategy
Task simulation
From task, device and user simulation methods
Device simulation
User simulation
Experimental context
Interaction data
Diagnostic Table
Analysis of device-user interaction
Feasibility assessment
Output to feasibility report

Figure 7.1 Usability evaluation method: process
The usabilty evaluation method supports the development of the representations designated by bold boxes.

8. The analysis is expressed in a format suitable for inclusion in a feasibility report, which constitutes the output of the assessment. The *feasibility assessment* presents an appraisal of the performance predicted for alternative user interface designs, an outline of desirable interface behaviour and a description of the design issues to be addressed during system development.

7.1 Generation of a preliminary specification of the design problem

7.1.1 Preliminary problem specification

Purpose: To enable the planning of the assessment study.

Expression: Two descriptions, coherent with respect to each other:

(a) Technical problem - informal description of the potential set of system design issues pertaining to the implementation of a speech I/O device.

(b) Resource availability - informal description of the finance, manpower, time and physical resources available for the application of SIAM.

7.1.2 Procedure for preliminary problem specification

1. Determine the technical issues to be addressed by the interface assessment. Agreement of the objectives of the assessment will require consultation between the assessors, the project team within MoD and, possibly, users. See Comment i.

2. Determine the resources available for the assessment. See Comment ii.

3. Set out a problem specification under the following headings:

    Technical issue to be addressed
    Time available for assessment
    Number and abilities of investigators available for the assessment
    Availability of field and laboratory facilities for the investigation.

7.1.3 Comments on preliminary problem specification

(i) Potential difficulties for operators of the target device may have been identified by the population of users or by the results of previous analytic or empirical studies of feasibility. However, issues of concern to the procurers are likely to be incompletely specified. The preliminary problem specification expresses the objectives of the study within SIAM's conceptual framework; i.e. SIAM is concerned with determining the implications for system performance of alternative interface options, where performance is defined as the costs to users in achieving a task output of acceptable quality.

(ii) The project manager should be briefed on the relationship between the resources available for the investigation and the status of the assessment, i.e. in general, the probability of an accurate and valid assessment is commensurate with the time and quality and quantity of manpower available for the investigation. See Appendix E.

7.1.4 Example of preliminary problem specification

The developers of the VDC were uncertain as to the optimum number of items of information (target types or numbers) to be entered and fed back to the user in each dialogue transaction. Resources for the study were negotiated, including duration (4 weeks), manning (2 researchers), facilities for evaluation and access to information on the target task, device and user.

7.2 Generation of a preliminary specification of the target system

7.2.1 Preliminary system specification

Purpose: To provide the analyst with information on target task, device and user for planning the development of simulations. See Comment i.

Expression: Three descriptions, coherent with respect to each other:

(a) Task - Informal description of target task objectives, high level actions and contextual constraints.
Specification of relationship with current tasks.

(b) Device - Informal description of target device functionality extracted from Cardinal Points Specification/Staff Target (user requirements document).
Specification of relationship with current devices.

(c) User - Informal description of intended users.
Specification with respect to current army population.

7.2.2 Procedure for preliminary system specification

1. Extract from the Staff Target (ST) or Cardinal Points Specification (CPS) all references to the user interface of the target device. Compose an informal specification of device function and agree it with the procurer, identifying points of difference and similarity with current operational equipment. See Comment ii.

2. Extract from the ST or CPS all references to the intended users of the target device. Compose an informal specification of the intended user, and agree it with the MoD procurement organization. Identify the population of users with respect to the current UK Army population.

3. Extract from the ST or CPS all reference to the task of the user of the target device (the "future task"). In consultation with MoD and/or Army specialists, identify points of similarity and difference between the future task and analogous current task(s). Compose an informal specification of the current task(s) including the objectives, high level actions and contextual constraints on task performance (for example, operation at night; operation while controlling a vehicle). See Comment iii.

4. Check the coherence of the preliminary device, task and user specifications, i.e. make sure that, in principle, the specified users could perform the specified future task, given the specified device. Modify descriptions if necessary, and check with the procurer.

7.2.3 Comments on preliminary system specification

(i) Military systems are complex, comprising interacting sub-systems of which the target system might be one. It is important that the technical scope of the study is explicitly specified with respect to the military system as a whole. SIAM assumes a description of the target system with respect to the *task* that the system performs; the *device* component of the system; and the *user*(s). The scope of the study is expressed within this framework.

(ii) Although SIAM is concerned specifically with the *user interface* of the target device, this is defined in broad terms, to include all aspects of device behaviour which may influence the behaviour of users. It will also include the behaviour of equipment which is not part of the computer but which is used in conjunction with it (e.g. maps, documentation, radio etc.). Equipment is included if it influences the interaction between computer and user(s).

(iii) The task is defined here as the work which will, in future, be *supported by* the target device. Note that it does not refer only to the work of *using* the target device: this is only part of the user's task.

7.2.4 Example of preliminary system specification

A preliminary specification of the device existed in the developers' documentation of the prototype VDC. The intended population of army device users was specified in general terms by referring to the developers (who were, themselves, attached to an army establishment). The task which the VDC was intended to support was similar to that of the artillery Forward Observation Officer (FOO), involving the establishment and maintenance of a battlefield observation post, surveillance and the transmission of target information. Constraints on task performance were that transmission of target information might occur under a wide variety of environmental conditions, and that transmitted data were to be error free.

Figure 7.2(a): Attribution of causes of inadequate system performance

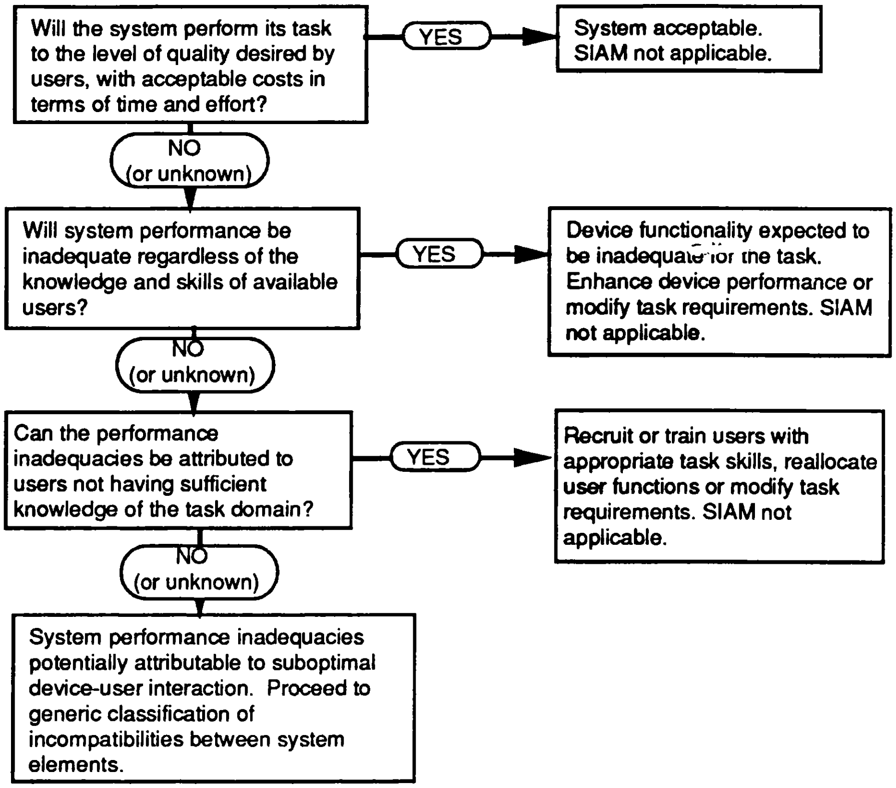
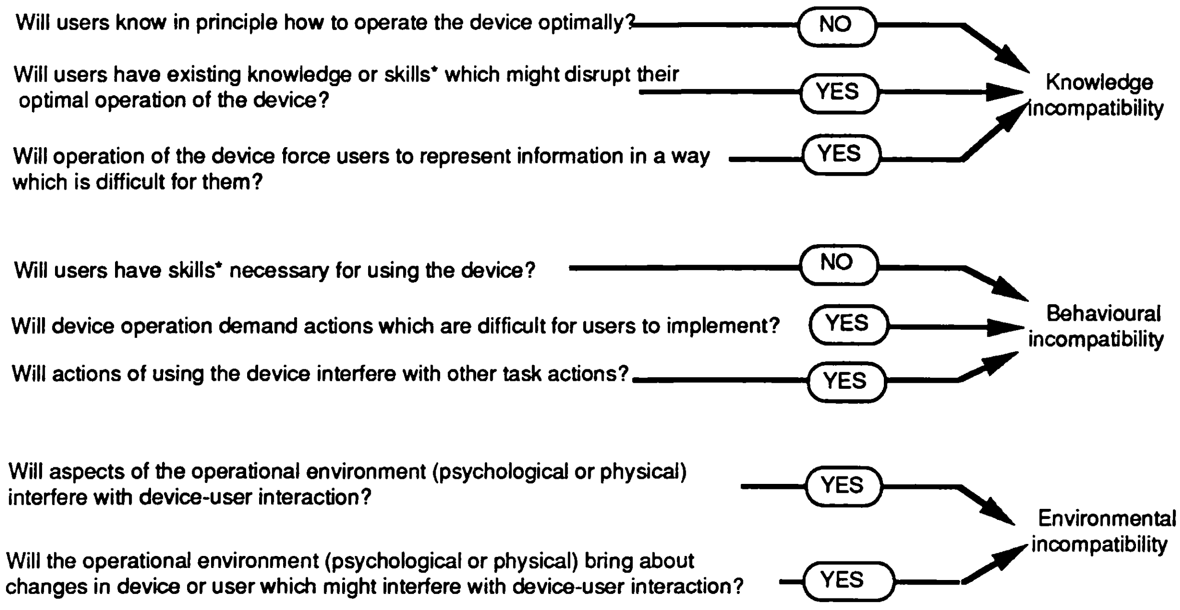

Will the system perform its task to the level of quality desired by users, with acceptable costs in terms of time and effort?
YES
System acceptable. SIAM not applicable.
NO (or unknown)
Will system performance be inadequate regardless of the knowledge and skills of available users?
YES
Device functionality expected to be inadequate for the task. Enhance device performance or modify task requirements. SIAM not applicable.
NO (or unknown)
Can the performance inadequacies be attributed to users not having sufficient knowledge of the task domain?
YES
Recruit or train users with appropriate task skills, reallocate user functions or modify task requirements. SIAM not applicable.
NO (or unknown)
System performance inadequacies potentially attributable to suboptimal device-user interaction. Proceed to generic classification of incompatibilities between system elements.

Figure 7.2(b) Generic classification of incompatibilities between system elements


Will users know in principle how to operate the device optimally?
NO
Will users have existing knowledge or skills* which might disrupt their optimal operation of the device?
YES
Will operation of the device force users to represent information in a way which is difficult for them?
YES
Knowledge incompatibility
Will users have skills* necessary for using the device?
NO
Will device operation demand actions which are difficult for users to implement?
YES
Will actions of using the device interfere with other task actions?
YES
Behavioural incompatibility
Will aspects of the operational environment (psychological or physical) interfere with device-user interaction?
YES
Will the operational environment (psychological or physical) bring about changes in device or user which might interfere with device-user interaction?
YES
Environmental incompatibility

**Note: the term "skill" is used here to refer to the attribute of a facility in the implementation of a process, e.g. typing., speaking.*

7.3 Configuring the diagnostic tables

7.3.1 Diagnostic table configuration

Purpose: Selection of an interaction model appropriate to the technical problem.

Expression: Row entries in diagnostic tables (see Appendix A).

7.3.2 Procedure for configuring the diagnostics

1. Determine the applicability of SIAM to the technical issues under consideration, by applying the decision tree in Figure 7.2(a). See Comment i.

2. Determine classes of potential incompatibilities between system elements by applying the decision tree in Figure 7.2(b). In practice, it is likely that more than one source of incompatibility will be influential. See Comment ii.

3. According to the class of incompatibility; search column 1 of the relevant diagnostic table (Appendix A) for critical system conditions germane to the issue(s) under consideration. Take note of references to other relevant diagnostics in column 11. The selection of relevant rows in the diagnostic tables is termed the "configuration" of the tables.

4. IF critical system conditions are found which address the issues identified in the preliminary problem specification, include these in the diagnostic table configuration. **ELSE** base the diagnostic table configuration on the appropriate general-purpose diagnostic at the end of each table. See Comment iii.

7.3.3 Comments on configuration of the diagnostics

(i) SIAM is specifically concerned with the diagnosis of performance inadequacies caused by failures in device-user interaction. The decision tree in Figure 7.2(a) is intended to help the assessor to decide whether or not expected performance inadequacies are attributable to the user interface or to factors beyond the scope of SIAM (such as limitations of device functionality or users having incomplete knowledge of the task domain)

(ii) SIAM utilizes existing knowledge about the interaction between speech I/O devices and their users. This information is expressed in six diagnostic tables (Appendix A), organized with respect to the three classes of device-user compatibility (knowledge, behaviour and environment), and with respect to speech input to, and speech output from, the device (see Section 6.3). The decision tree in Figure 7.2(b) is intended to help the assessor to identify tables relevant to the problem which was previously specified using Procedure 7.1.

Table 7.1 Example of configured diagnostic table (extracted from Tables 1 and 4, Appendix A).

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 1.4, abstracted from Diagnostic Table 1:Knowledge Incompatibility (speech data input)										
The size of the chunk of information after which feedback is presented is different from that used elsewhere in the task	Actions demanding manipulation of information to be entered into the computer	Device constraints on chunking of data Feedback chunk characteristics Recognition error frequency Availability of aids to memory (e.g. facility for writing down data to be entered)	Familiarity with task data	Factors disrupting memory processes e.g.: - psychological stress - noise - visual distraction	Constraints on structure of task information	Trade-off between: (1) Performance disruption by feedback presented in units smaller than those preferred by users for the representation of domain information; and (2) Performance disruption by cognitive load imposed by having to correct errors at early locations in an entry sequence	Transaction time Subjective assessment of cognitive effort to transform data to preferred format. Mistakes in correction of device errors as function of location of error in the message	(1) Expression of an utterance will be facilitated if the speech representation is compatible with the user's mental representation of the information (2) Error detection depends upon retaining representations of intended and actual data entries. The longer the string, the greater the probability of failing todetect early/ multiple errors (3) Representations in working memory decay over time.[The rate of decay may differ between the auditory and visual modalities.] (4) If speech is used for error correction, the complexity of the command is some function of the distance of the error from the current cursor position. This impacts the information processing load, both by the demand for formulating the correction command and by its interaction with (c) above.	Allow users to enter information in a form familiar to them OR provide extended training to users to familiarize them with the machine information structure.	See diagnostics 1.2, 3.1, 3.3, 4.3, 4.4 Supporting documents: Simpson et al, 1985
Diagnostic 4.3, abstracted from Diagnostic Table 4:Knowledge Incompatibility (speech data output)										
Information is presented in chunks of a size and format different from those used elsewhere in the task.	Actions in which information is acquired from the computer Other task actions in which information is acquired or manipulated	Device constraints on data chunk size Rate of presentation	Familarity with format of task data	Factors disrupting memory process e.g. stress, noise, visual distraction.	Constraints on the structure of task information.	Trade-off between (1) performance disruption by information presented in units smaller than those preferred by users for the representation of domain information and (2) performance disruption by cognitive load imposed by remembering data at early locations in speech output sequence.	Transaction time Frequency of use of "playback"/repeat/ "playslow" facilities Errors: type and frequency Subjective assessment of cognitive effort to transform presented information to preferred format	Information reception will be facilitated if the speech representation is compatible with the user's mental representation of the information; however, representations in working memory decay over time, so early items in a long string may not be recalled accurately	Structure presented information in form familiar to the user. Segment the synthetic speech into grammatically correct elements. Offer "playback" facility Offer redundant visual data presentation. Provide extended training to users to familiarize them with the machine information structure.	See diagnostics 4.4 and 6.1 Supporting documents: Cooper, 1987
Diagnostic 4.4, abstracted from Diagnostic Table 4:Knowledge Incompatibility (speech data output)										
Amount of information (non-redundant chunks of information) in speech output message is large.	Actions in which information is acquired from the computer	Device constraints on data chunk size Rate of presentation		Factors disrupting memory process e.g. stress, noise, visual distraction. Availability of memory support (e.g. facilities for note-taking)		Performance disruption by cognitive load imposed by remembering data at early locations in speech output sequence.	Transaction time Frequency of use of "playback" function Use of mnemonic strategies e.g. note-taking Subjective assessment of effort to remember long strings of data	Representations in working memory decay over time, so early items in a long string may not be recalled accurately	Use shorter chunks of information Provide users with means of recording the information (e.g. pencil and paper) Provide redundant visual feedback Offer "playback" facility	See diagnostics 4.3, 4.7 and 6.1 Supporting documents: Marics & Williges, 1988 RARDE, 1983 Simpson et al, 1985

(iii) General purpose diagnostics enable an assessment on the basis of very limited knowledge of expected device-user interaction. As a consequence, they tend to be over-inclusive in order to avoid missing important impacts of system design on behaviour. If possible, seek the assistance of a human factors specialist in order to render the diagnostic more specific.

7.3.4 Example of diagnostic configuration

Decision tree 7.2(a) was used to determine the relevance of SIAM to the problem identified by the device developers. In all cases, the impact of string length on device-user interaction was unknown, so SIAM was deemed applicable. Within the analysis offered by decision tree 7.2(b), the problem was judged to be an instance of the device requiring users to represent information in a way which was potentially difficult for them: an instance of representational incompatibility.

Diagnostic 1.4 was identified as relevant (see Table 7.1). Column 11 specified other potentially relevant diagnostics, of which two - 4.3 and 4.4 - were also included in the configuration.

7.4 Specification of the solution strategy

7.4.1 Solution strategy

Purpose: Specification of an approach for investigating the technical problem utilizing the available resources. See Comment i.

Expression: Experimental specification comprising:

- list of hypotheses
- list of parameters to be held constant in the experiment
- list of independent variables
- list of dependent variables
- experimental design

7.4.2 Procedure for specifying solution strategy

1. Specify the experimental hypothesis. For each critical system condition, express a null hypothesis [that the behavioural/performance features in column 7 of the diagnostic table, will not be observed in the operation of the target system], and an alternative hypothesis [that the features in column 7 will be observed].

2. Specify the critical independent parameters. For each critical system condition, list the critical system attributes in columns 3, 4 and 5, and critical task attributes in column 6 of

the appropriate diagnostic tables. Differentiate those which will vary in order to test the experimental hypothesis and those which are to remain constant across all conditions. The former are the independent variables of the study.

3. Specify the dependent variables. For each critical system condition, list the critical behavioural/performance parameters in column 8 of the appropriate diagnostic table.

4. Decide the level of description at which behaviour and performance are of concern. See Comment ii.

5. Specify an experimental design enabling the testing of the experimental hypothesis. See Comment i.

PROCEED TO DEVELOPMENT OF:

- TASK SIMULATION: PROCEDURES 8.1 TO 8.6
- DEVICE SIMULATION: PROCEDURES 9.1 TO 9.5
- USER SIMULATION: PROCEDURES 10.1 TO 10.4. (See Comment iii).

7.4.3 Comments on specifying a solution strategy

(i) One of the assumptions of SIAM is that assessors will be familiar with the principles of experimental design. Detailed procedures are not offered to support decisions concerning the sizes of samples, balancing of conditions etc..

(ii) The preliminary problem specification may have been expressed at any level of description (see discussion in Appendix E). The assessor is recommended to distinguish three levels of description, arbitrarily termed the input/output (I/O) level, the communication level and the task level. An evaluation at the input/output level addresses the physical operations of using the target device. The objects of its concern might be data entry devices, such as a keyboard, speech recognizer or joystick, and user actions associated with these might be, respectively, keystrokes, utterances and manual manipulation. Evaluation at the communication level addresses the informational dialogue between the device and the user. The objects of its concern might be the conceptual objects of the system to which the user is exposed. For example, in the case of a database system, they might be files, pages or fields of information, and actions might be entering new information or retrieving items of information. Evaluation at the task level addresses the achievement of work goals. In the case of a Forward Artillery Observer, objects of concern might be enemy forces and friendly forces, and an action might be the provision of indirect fire support for friendly forces.

(iii) Task, device and user simulation development should occur in parallel; however, the early stages of the task simulation method must be performed to enable progress on the others. It is, therefore, recommended that procedures 8.1 to 8.5 be applied first.

7.4.4 Example of solution strategy

An empirical study was specified by relating the diagnostics in Table 7.1 to the target system.

Null hypothesis: that there will be no effect of entry and feedback string length on performance.

Alternative hypothesis: that there will be an effect of string length on system performance and that this will be attributable to feedback being presented in units ("chunks") smaller than those preferred by users, or to the cognitive load imposed when correcting errors at early locations in a string of entered data..

Critical Independent Parameters:

Constant: - *user familiarity with task data*
- *environmental factors disrupting memory (e.g. noise)*
- *constraints on structure of task information (e.g. grid reference structure)*
- *availability of aids to memory*
- *rate of feedback presentation.*

Variable: - *device constraints on chunking of data, and feedback chunking constraints (either 2, 4 or 8 items per string)*
- *recognition error frequency (either 1% or 4%)*

Dependent variables:
- *transaction time*
- *mistakes in correction of device errors, as function of error location*
- *subjective assessment of cognitive effort to ensure correct input*

Because the experimental hypotheses addressed behaviour at the level of communication exchanges and data input/output, the transactions of concern were those which information chunking impacted directly (not with higher level transactions). In the context of the target task, these were transactions involved in the entry of target location information (8 digit grid references).

Experimental design. Time constraints prevented a fully controlled investigation. A within-subjects design was used. A small number of subjects would perform equivalent experimental tasks in which the chunking constraints and the reliability of the device were varied. The *time successfully to enter strings would be measured and the errors analysed. Considerable weight would be attached to subjects' subjective reports.*

7.5 Developing an experimental context

7.5.1 Experimental context

Purpose: Acquisition of system behaviour/performance data

Expression: Operational reproductions of alternative simulated work systems.

7.5.2 Procedure for developing an experimental context

1. Integrate task, device and user simulations, ensuring that these three components remain adequate reproductions when operating in conjunction with one another. Check with the procurer and with army representatives that the system simulation is an accurate reproduction of the target system *for the purposes of the study.* See Comment i.

2. Specify experimental procedure. Abstract from the task simulation the sequence of actions which is expected to be impacted by manipulation of the independent variables of the experiment.

3. Implement experimental control and data collection mechanisms. Automate as much as possible of the experimenter's task in order to standardise the procedure for subjects. This may extend to:
   (i) presentation of instructions to subjects
   (ii) training of subjects
   (iii) initiation of experimental trials under the various experimental conditions (i.e. manipulation of critical system parameters)
   (iv) collection of data (measurement and recording of critical behavioural parameters)
   (v) timing of subject rest periods.

4. Run at least one pilot subject under all conditions of the experiment (see Comment ii).

7.5.3 Comments on development of the experimental context

(i) See Appendix E for a discussion of issues relating to simulation fidelity requirements. It should be noted that the simulation will only seek to reproduce accurately *selected* aspects of the behaviour of the target system. The fact that the simulation is not intended to be a complete reproduction should be explained to army authorities when they evaluate its adequacy.

(ii) Pilot trials are necessary to check the integration of the components of the simulation and to ensure that they may be operationalized in the context of the experiment. The pilot subject should meet the requirements for a user simulation and should be treated as a real subject. He/she should undergo the complete training procedure as well as a full set of experimental trials. Data should be analysed in the same way as is intended for the main experiment to identify likely "floor" and "ceiling" effects, and to ensure that data manipulation procedures are operable, as well as those for running the experiment.

7.5.4 Example of experimental context

Task, device and user simulations were developed, using the respective simulation methods (see Sections 8.6.4, 9.5.4 and 10.4.4). In summary, the simulations took the following form:

Task Simulation: The user subject generated a list of standard messages concerning objects marked on a map. The orders could then be read out verbatim from the list. Each message included an 8 digit grid reference, which was entered in either 2, 4 or 8 digit strings, according to the experimental condition currently in force. All data were subject to the recognition errors of the device which had to be corrected during data entry.

Device Simulation: The device simulation was supported by the system subject (SS) communicating with the user subject (US) via an intercom link. The speech of the SS was distorted to assist the illusion of synthetic speech. The SS was aided in the insertion of "recognition errors" into his output by a computer accessing an error matrix in a look-up table. [The operation of the computer was timelogged, enabling the subsequent calculation of time to enter data.]

User Simulation (user subjects): The two USs were PhD students. Both had used speech recognizers. They were trained in the use of the map and the simulated target device.

The simulations were integrated and operationalized in the experimental design described in Section 7.4.4. Each subject transmitted data concerning eight targets under each condition of the experiment; dependent variables included time to enter grid data and to edit it, etc.

7.6 Generation of interaction data

7.6.1 Interaction data

Purpose: Determination of system behaviour/performance

Expression: List of the observed values of dependent variables

Informal description of observations

7.6.2 Procedure for data collection

1. Run the experiment. See Comment i.

2. Collect, compress and apply statistical tests to the data

3. Record informal observations to assist interpretation of quantitative data.

7.6.3 Comments on data collection

(i) It is assumed that the assessor is experienced in the conduct of experiments.

7.6.4 Example of experimental data

The experiment was run to generate interaction data (time, errors and subjective comments of user subjects). The data collected are now summarized. Because of the small sample size no inferential statistical operations were performed, so the results cannot be considered conclusive.

(a) Transaction time. The transaction time for the entry of the location data was determined by subtracting the time at which the SS pressed the "complete" key from the time that the "grid" key had been pressed (i.e. a measure of total time for the entry of grid data).

Table 7.2 presents the mean time in seconds (and standard deviation) for the device programmed to exhibit 4% recognition errors (3% was the rate actually observed); and Table 7.3 presents equivalent results for the device with 1% errors (where 1.1% was actually observed). As data were collected from only two subjects, the calculation of means for the three conditions was inappropriate; the data for the subjects were analysed individually. Neither subject's data exhibited a consistent effect of chunk size on the time to complete entry of grid data (although, in all cases, the time for 8 digit chunk was the most variable). There was, nevertheless, strongly suggestive evidence that recognition reliability was a determinant of transaction time.

(b) Errors. A total of five mistakes were observed in the output of the two USs. Of these, three were attributable to the SS. Because there were few errors, few conclusions can be drawn from these data on the usability of the target device.

(c) Subjective comments. The comments of the subjects on device usability are summarized in Table 7.4

Table 7.2 *Transaction time (sec) for entry of grid data (observation task) (3% actual recognition errors)*

[n = 8]	*Subject 1*	*Subject 2*
	mean (sd)	*mean (sd)*
2 digit data chunking	*22 (6.95)*	*25 (9.51)*
4 digit data chunking	*26 (8.48)*	*20 (6.61)*
8 digit data chunking	*22 (10.16)*	*31 (30.51)*
	-----	-----
Overall mean	23.33	25.33
	-----	-----

Table 7.3 *Transaction time (sec) for entry of grid data (observation task) (1.1% actual recognition errors)*

[n = 8]	*Subject 1*	*Subject 2*
	mean (sd)	*mean (sd)*
2 digit data chunking	*17 (3.99)*	*18 (3.86)*
4 digit data chunking	*16 (5.07)*	*19 (7.15)*
8 digit data chunking	*21 (20.73)*	*17 (8.19)*
	-----	-----
Overall mean	*18.00*	*18.00*
	-----	-----

7.7 Analysing device-user interaction

7.7.1 Analysis of device-user interaction

Purpose: Interpretation of interaction data

Expression: Statement of results with respect to the experimental hypotheses.

Statement of incidental observations relevant to assessing the usability of the target device.

7.7.2 Procedure for analysing device-user interaction

1. Evaluate the experimental results against the hypotheses set out in the solution strategy. If the dependent variables show the critical features in column 7 of the previously-configured diagnostic tables, reject the null hypothesis, and accept the diagnosis and prescriptions proposed in columns 9 and 10

Table 7.4 Subject comments on device usability (observation task)

Comments common to both subjects:

1. There was a high risk of confusion in the use of "yes" and "no" in the context of error correction, i.e. respond in affirmative to error prompt - "yes" (there is an error in the chunk) - then, when the device asks for confirmation of each digit in the chunk, indicate an error with a negative - "no" (the last digit was incorrect).

2. When using the verification function, it was annoying to have to listen to all the previously-entered data when the concern was with a digit in the last (grid) field.

Subject No.1:

1. 2-digit chunking was optimal for error correction, 8-digit was optimal for rapid entry. If the device was unreliable, 2-digit chunking was preferred; if very reliable, 8-digit chunks were optimal. 4-digit represented the best/worst of both worlds.

2. The subject was not happy with the device prompting for error correction following a time-out: other parts of the interaction were paced by the user. Consistency in this respect would be preferable.

Subject No.2:

1. The subject found it difficult to track errors in 8 digit chunks: sometimes she was not certain that feedback included no errors, especially when tired. She was quite happy with two digit chunks.

2. The subject believed that errors might have a greater likelihood of being overlooked with the more reliable recognition performance, because users might pay less attention to the feedback.

3. The subject naturally grouped the 8 digits into two groups of four.

2. Assign the outcome of step 1 the status of the primary result of the study, i.e. it constitutes an answer to the specified problem.

3. Search the diagnostic tables (Appendix A) for behavioural/ performance features which, incidentally, had been observed as consequences of interaction. If there is a possible match, propose, as a secondary result of the study, tentative evidence in favour of the diagnosis associated with the behavioural/performance feature. See Comment i.

7.7.3 Comments on analysis of device-user interaction

(i) The diagnostics are used in Step 3 to prompt the assessor in the identification of behaviour which may require additional investigation. However, the diagnostics cannot claim to cover all the phenomena potentially observable when people interact with computers. The effectiveness of this stage will be considerably influenced by the inferential skill of the assessor and by the assessor's personal knowledge of human factors (see Appendix E).

7.7.4 Example of analysis of device-user interaction

The data were used to test the experimental hypothesis. Although there was no evidence to suggest that transaction time was determined by data string length, and there were insufficient data to determine an effect of string length on the incidence of errors, there was evidence to suggest that the string length imposed varying cognitive demands on users. The results of the study indicated that, under the conditions simulated, users found that transactions involving the longer string lengths imposed potentially unacceptable cognitive demands. This was diagnosed as being due to a requirement on the user to retain more information in their memory. Small chunk sizes were confirmed as being preferred in the context of the target task and system. An effect on transaction time and errors might have been evident had users not been able to refer to their previously written list of messages to be sent (i.e. if there had been less effective memory support).

Secondary results of the study informed the design of the dialogues for error correction and for verifying previously entered information. Subjects reported difficulties in the use of both facilities (see Table 7.4). It should be emphasised that, because of the small number of subjects tested and the consequent failure to test the statistical significance of the results, any conclusions drawn from the study could only be tentative.

7.8 Presentation of the feasibility assessment

7.8.1 Feasibility assessment

Purpose: Communication of output of SIAM to MoD

Expression : (See Comment i). Text for inclusion in the system feasibility assessment covering:

(1) Appraisal of alternative user interface solutions considered.

(2) Evaluation of preferred solution

(3) Outline description of user interface performance characteristics

(4) Description of issues to be addressed during interface development.

7.8.2 Procedure for presenting a feasibility assessment

1. Set out the alternative options for the user interface of the target system, and state problems identified in the preliminary problem specification.

2. Describe the evaluation of the preferred solution. Specify the assumptions of the study. Specify the primary results with respect to the preliminary problem specification.

3. Where the primary results of the study indicate the need to resolve incompatibilities arising from implementation of the preferred solution, set out the prescriptive options presented in column 10 of the diagnostic table.

4. Where possible, indicate in broad terms the likely performance of the favoured interface option (eg comparative performance relative to alternative options).

5. Propose issues to be addressed either in further feasibility evaluations, or during interface development. These are indicated by unresolved primary findings and any secondary results of the study

7.8.3 Comment on expression of feasibility assessment

(i) The format of the output of SIAM may require adaptation for the purpose of integrating it with other contributions to the feasibility assessment (e.g. pertaining to issues other than the user interface).

7.8.4 Example of feasibility assessment

The evaluation of the VDC used here to illustrate the application of SIAM was not viewed strictly as a feasibility assessment, and the output of the evaluation was not expressed as a feasibility report. However, for the purpose of illustration, a feasibility report based upon the results might be take the form of an elaboration of the following topic headings.

(a) ***Appraisal of alternative user interface options.*** *Description of the different options for entering and receiving feedback on strings of numeric data. Users' behaviour may be disrupted by having to break data down into unfamiliar short strings prior to data entry (i.e. a potential disadvantage of short strings); conversely, disruption may be caused by their having to identify errors at early locations in long strings of feedback (i.e. potential disadvantage of long strings). The feasibility of speech data entry and feedback is dependent on a successful resolution of the trade-off between these factors.*

(b) ***Evaluation.*** *Report of the evaluation summarised in Sections 7.4.4, 7.5.4 and 7.6.4. The modest status of the evaluation would need to be stressed, and its technical assumptions (e.g. there was no attempt to simulate the possible influences of noise and stress on operational performance of the system).*

(c) ***Outline of performance characteristics of the favoured option.*** *The results tended to favour short string lengths, on the grounds that they imposed lower costs on users. The short*

string option offers better performance, provided users have memory support in the form of a list of data to be entered. The results of this limited study would be inadequate to make absolute quantitative predictions of system performance.

(d) ***Issues to be addressed during development.*** *The results of the study indicated that recognition reliability was the primary determinant of system performance. A study to determine what would constitute acceptable recognition performance would need to be determined in a further feasibility study.*

The existing strategies for error correction and for verifying entered data presented difficulties for users. The software supporting these functions require further HF studies at the design stage.

7.9 Critical commentary on the usability evaluation method

The usability evaluation method, as expressed in this chapter, assumes a hypothetico-deductive approach to system assessment. The strength of this approach (indeed, the rationale for its utilization) lies in the fact that empirical data are used as the foundation for evaluation. The acknowledged weakness of HF discipline knowledge of speech interaction is made good by simulating the behaviour of the target system, so testing the assumptions underlying the simulation and enabling the assessor to use his or her implicit knowledge for diagnosis and prescription. However, the hypothetico-deductive procedures of the usability evaluation method are less appropriate when the purpose of the study is not to test hypotheses but, rather, to explore the behaviour of a system in order to identify its strengths and weaknesses. Exploratory studies such as this are common where the usability of an existing device or prototype is to be tested, and they rely on a strategy of induction rather than deduction. The procedures of the usability evaluation method are potentially restrictive, then, in that they are inappropriate for supporting exploratory evaluation.

The usability evaluation method offers procedures for selecting and configuring the information which constitutes a model of device-user interaction to support simulation development (i.e. procedures for configuring the diagnostics). The tabular format of the diagnostics was chosen as a means of decomposing information on device-user interaction in a way which would be accessible to procedures of the method. However, although systematic, the rationale underlying the decomposition may fail to be comprehended by assessors (see Chapter 11), with the consequence that the selection of appropriate diagnostics is rendered difficult or impossible. Comprehension may be further impeded by the use of technical terminology in expressing the diagnostic information or by expression at a high level of description (as in the case of "general purpose" diagnostics).

The usability evaluation method raises, here, an issue of level of description which is relevant to all aspects of the proceduralization of SIAM. There is a conflict between the need to provide detailed information to enable casual practitioners of HF to perform assessments, and yet to offer a method which is sufficiently flexible in its expression to render it applicable in a wide range of investigative contexts. Possible resolutions of this conflict are discussed in Chapter 12.

CHAPTER 8

TASK SIMULATION METHOD

8.0 Process of task simulation - summary

The process of task simulation requires the assessor to generate six intermediate representations of the task (see Figure 8.1).

1. Task information gained from indirect sources (such as through interviews with domain experts, consultation of training manuals etc) is represented within the hierarchical notation, resulting in a *preliminary task description*

2. Existing task(s) are observed which are believed to have a goal structure in common with that of the target task, resulting in a record of *extant task data* (e.g. a video record or a transcript of video recorded data).

3. The task data are analysed to provide information to generate an *expanded task description.* Analysis involves relating the hierarchical expression of the preliminary task description to the data record, with the objective of identifying errors in the description and of increasing its level of detail.

4. A *future task description* is synthesized in two stages, involving the *deletion* of those actions expected to be changed by the introduction of the target device, and the subsequent *generation* of new actions. The process of generation involves relating the elaborated task description to a model of the functionality of the target device.

5. Actions critical to device usability are identified in the future task description by the intersection of the future task description with the model of device-user interaction assumed for the evaluation (see Section 6.6.3). The result is a *future task model*[1]: a hierarchical representation of the task simulation.

[1] In this method (and equivalently in the other simulation methods of SIAM), the term "model" is used to distinguish a *description expressing the task attributes to be included in the simulation* from preceding descriptions which do not seek to take account of the technical concerns of the current evaluation. Although this may seem somewhat arbitrary, the choice of terminology was originally based on practice in operational research, where a "simulation model" refers to a specification of the elements of the target system (and their inter-relationships) which may subsequently be implemented as a simulation (see, for example, Emshoff and Sisson, 1970).

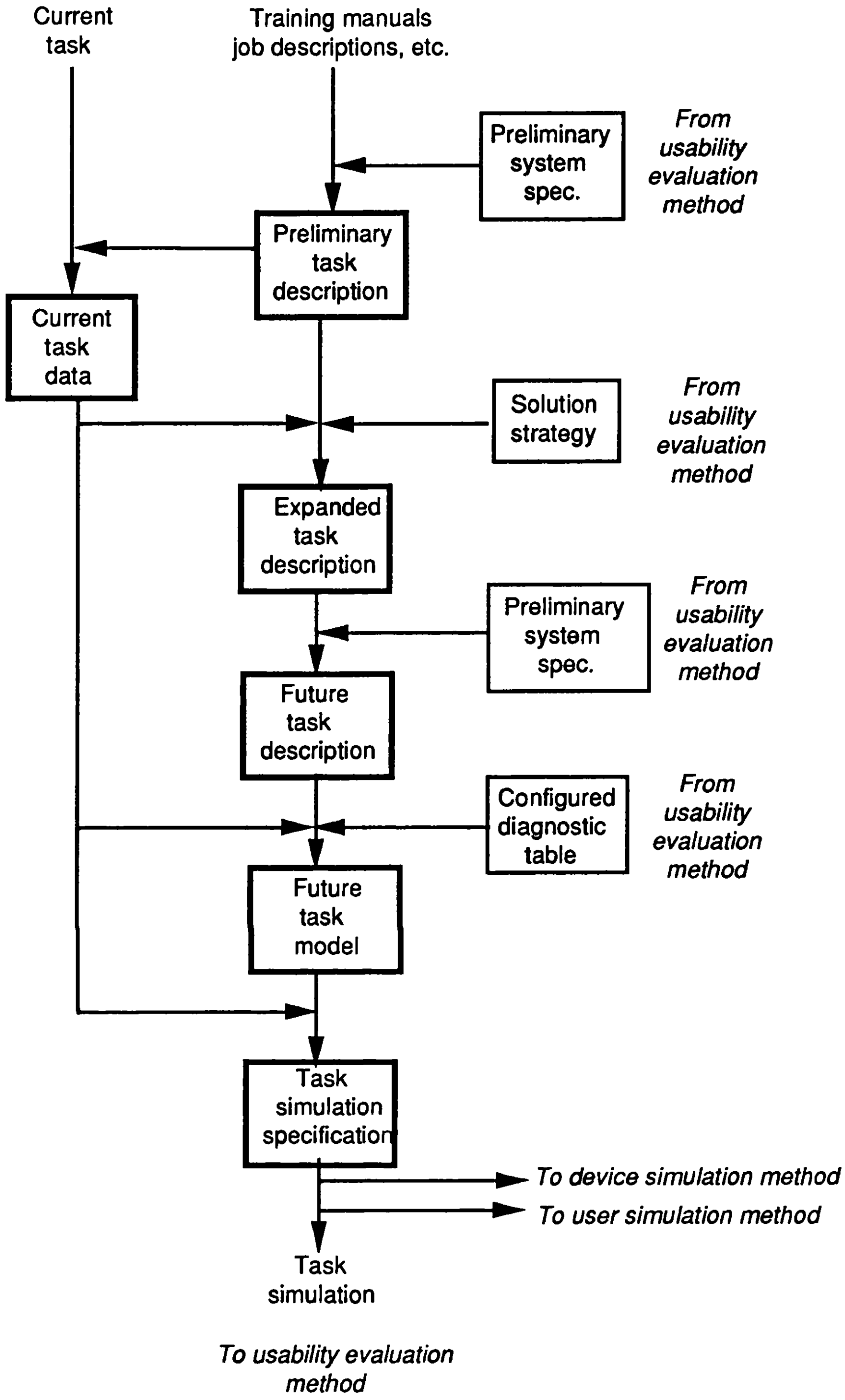

Current task
Training manuals job descriptions, etc.
Preliminary system spec.
From usability evaluation method
Preliminary task description
Current task data
Solution strategy
From usability evaluation method
Expanded task description
Preliminary system spec.
From usability evaluation method
Future task description
Configured diagnostic table
From usability evaluation method
Future task model
Task simulation specificatio
To device simulation method
To user simulation method
Task simulation
To usability evaluation method

Figure 8.1 Task simulation method: process

The task simulation method supports the development of the representations designated by bold boxes; representations developed by means of other sub-methods are designated by unemphasized boxes.

6. Finally, the hierarchical description of the simulation is used to generate an implementable *task simulation specification.*

8.1 Generation of a preliminary task description

8.1.1 Preliminary task description

Purpose: To provide the assessor with information about the task to enable planning of an observational study.

Expression: Hierarchical description of task actions, to reveal how actions within sequences depend upon the results of previous actions, and how actions in a sequence relate to superordinate goals. [In addition to the explicit hierarchical description, the analyst should have sufficient informal knowledge of the language of the task, and of the devices used to perform it, to be able to interpret actions in terms of the hierarchical description.]

8.1.2 Procedure for generating the preliminary task description

1. Collect information on the current task. Having identified the relationship between the future task and current task(s) from the preliminary task specification (part of the preliminary system specification), seek information from as many of the following sources as possible:
   (a) job description of person(s) currently performing the task
   (b) procedures/training manuals describing the current task
   (c) observations of personnel training session(s)

   Check your understanding of the task with an officer who has had experience of it.

2. Develop a hierarchical description of the current task.
   (a) Identify the military function of the person(s) performing the task, e.g. "to support manoeuvre arm in achievement of objectives".
   (b) List the smallest number of activities which completely describe how the purpose is achieved. Within these, identify activities which could potentially impact the target task: these will comprise the top level of the hierarchical description (see Comments i-v).
   (c) For each of the activities in the top level, list the smallest number of actions which completely describe the activity. Repeat the breakdown on each of these actions, and continue the process until further analysis would describe actual movements, or until insufficient detailed information is available to allow further analysis.

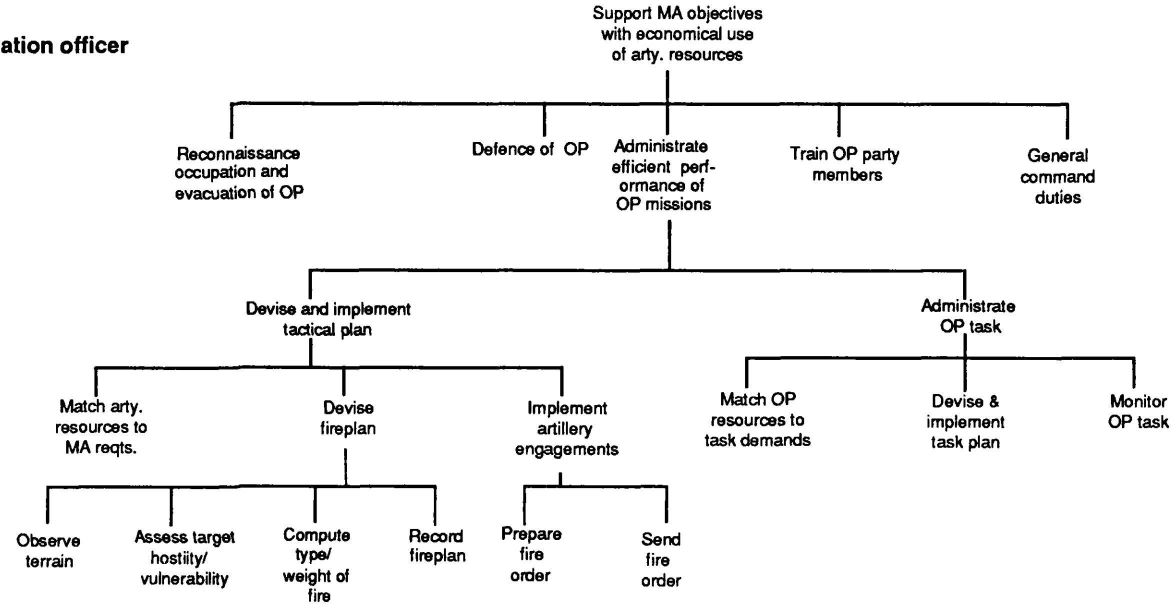
3. Confirm the hierarchical description. Confirmation of the description should be sought from as many army experts as possible and the description modified to conform with the consensus of opinion.

Figure 8.2: Preliminary task descriptions

The description of the forward observation officer was modified by a domain expert to generate a description of the target task.

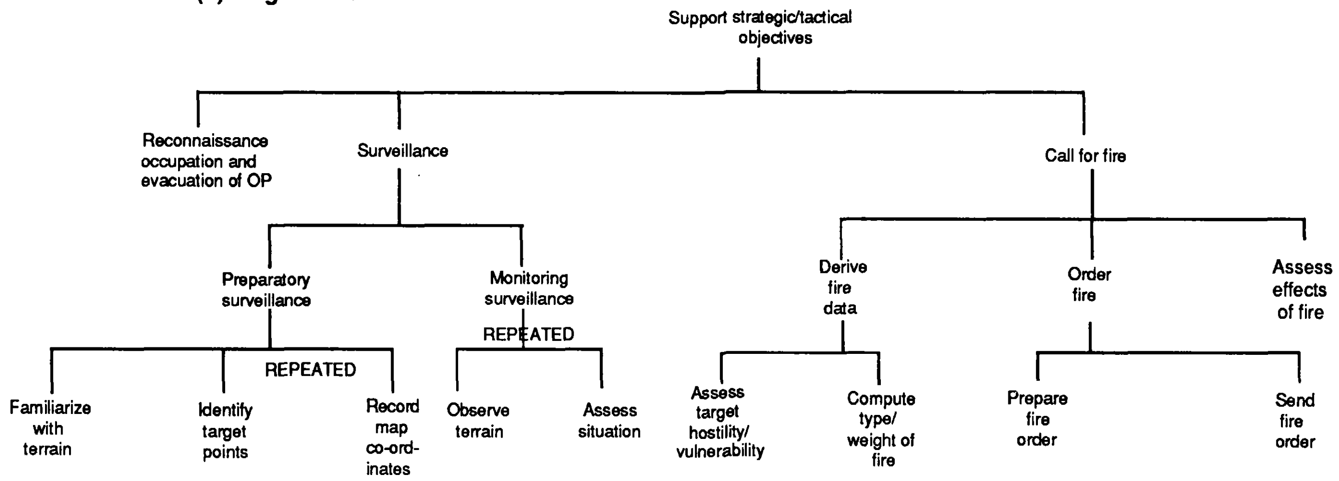
KEY:
OP = observation post
MA = manoeuvre arm

(a) forward observation officer



Support MA objectives with economical use of arty. resources
Reconnaissance occupation and evacuation of OP
Defence of OP
Administrate efficient perf-ormance of OP missions
Train OP party members
General command duties
Devise and implement tactical plan
Administrate OP task
Match arty. resources to MA reqts.
Devise fireplan
Implement artillery engagements
Match OP resources to task demands
Devise & implement task plan
Monitor OP task
Observe terrain
Assess target hostiity/ vulnerability
Compute type/ weight of fire
Record fireplan
Prepare fire order
Send fire order

(b) target task.



8.1.3 Comments on preliminary task description

(i) An action is defined as intentional behaviour. It may be a high level action (e.g. "devise fireplan"), or a low level action (e.g. "pick up binoculars").

(ii) As a rule of thumb, no action should be described in terms of more than four subsidiary actions.

(iii) "Actions" may be mental actions logically required to enable an overt (physical) action (e.g. deciding the content of a data message prior to entering it to the computer).

(iv) Actions may be implemented in alternative ways (e.g. "acquire information" might be implemented as observation of the battlefield or as requesting information from a colleague). All the alternatives should be analysed.

(v) Sequential actions should be represented from left to right in the order in which they occur.

8.1.4 Example of preliminary task description

An ex-army officer familiar with the target observation task was interviewed to develop a preliminary task description. It was agreed that the task of the artillery Forward Observer (see Appendix B) should be used as the basis of the description, and that a description of this task would be modified subsequently with the assistance of the domain expert, to achieve a match with the target task. Figure 8.2(a) shows the Forward Observer task, which was modified by the expert in developing the preliminary description of the target task shown in 8.2(b). The target task was characterized as involving no requirement to manage the activities of colleagues and a reduced need to co-ordinate artillery activities with those of other friendly forces.

8.2 Collection of data on the current task

8.2.1 Current task data

Purpose: Source of information on how the task is currently operationalized.

Expression: (1) Video/audio record of behaviour over time

(2) Interpretation of behaviour with respect to task actions: this might be in the form of a written list of actions or a computer representation of events (e.g. a video data analysis)

(3) Informal description of temporal aspects of the task.

8.2.2 Procedure for data collection

1. Observe/record task behaviour. The usability evaluation of the target device is based upon a characterization of the task as it is performed now (see Comment i). The characterization is developed by observation of behaviour, and, as task behaviour may be complex and transient, it is recommended that a permanent video record be obtained for

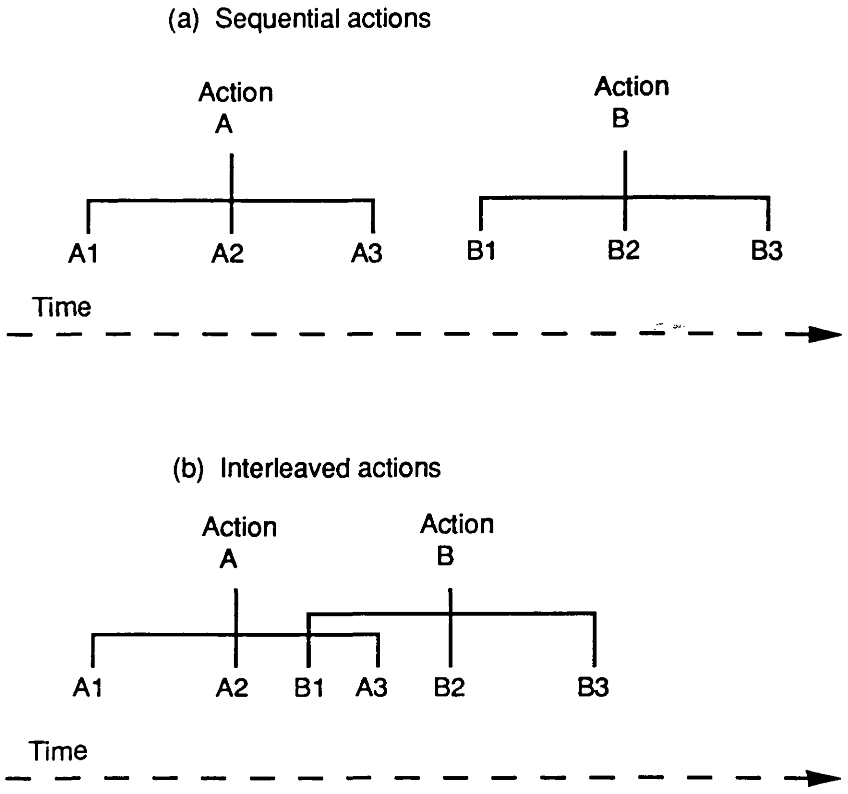


Figure 8.3: Temporal relationships between actions

subsequent analysis. The recording should be retained for reference throughout the evaluation.

2. Analyse task behaviour. The recorded behaviour has to be segmented and interpreted as task actions (see Comment ii). It is recommended that the preliminary task description be used to structure the analysis: the aim should be to account for the behaviour in terms of the actions named in the preliminary task description. The steps are as follows:

   (a) Check the accuracy of the Preliminary Task Description. If observed actions do not correspond to those in the Preliminary Task Description, make a note to modify the description in Procedure 8.3.

   (b) If it is possible to identify consistently actions subsidiary to those at the lowest level of the description, identify them for inclusion in the description in Procedure 8.3. Focus particularly on actions which involve the acquisition, handling and application of task information.

   (c) Note the occurrence of external events which give rise to actions. It may be necessary to simulate events (with appropriate frequency) in the task simulation.

(d) Note actions which occur concurrently (see Comment iii).

8.2.3 Comments on data collection

(i) Where possible, observations should be made on a field trial. Where this is not possible, observations made under (simulated) training conditions may be utilized, but these must be interpreted appropriately: behaviour on a trainer will be substantially different from that exhibited in the field. The subjects of the observation should be representative of users of the target device, with respect to their experience of the task: trainees will not behave in the same way as experienced task performers.

The location of cameras and microphones should be such as to enable observation of task actions and of the inputs and outputs of those actions. For example, if information is obtained by radio or via a computer terminal, it should be possible to determine the information presented to the subject(s) in this way. The resolution of the recording should be adequate to enable subsequent description at the level specified in the solution strategy.

(ii) As a rule of thumb, avoid analysis at very low levels of description unless it is demonstrably necessary for the assessment. It should be remembered that the purpose of the record is to develop and to confirm the hierarchical description, and the video record will be available if low level details of the current task are required during simulation development.

(iii) A superordinate action will be achieved by the completion of a sequence of sub-actions. Under some circumstances, sub-actions of different super-ordinate actions will be performed either simultaneously or in an "interleaved" fashion, where the sub-actions of A are suspended while one or more sub-actions of B are performed (see Figure 8.3). As the requirement to perform actions concurrently is relevant to the suitability of speech interfaces, the requirements will potentially require reproduction in the task simulation.

8.2.4 Example of current task data

It was not possible to observe either target users or Forward Observers performing their task in the field. However, it was possible to observe the task of the Forward Observer in the context of an army training simulator; the performance of the task was recorded on video tape for subsequent analysis (see Procedure 8.3). The collection of video data is also described in Appendix B.

Figure 8.4 (a): Expanded task description: Forward observation officer (Key as in Figure 8.2)

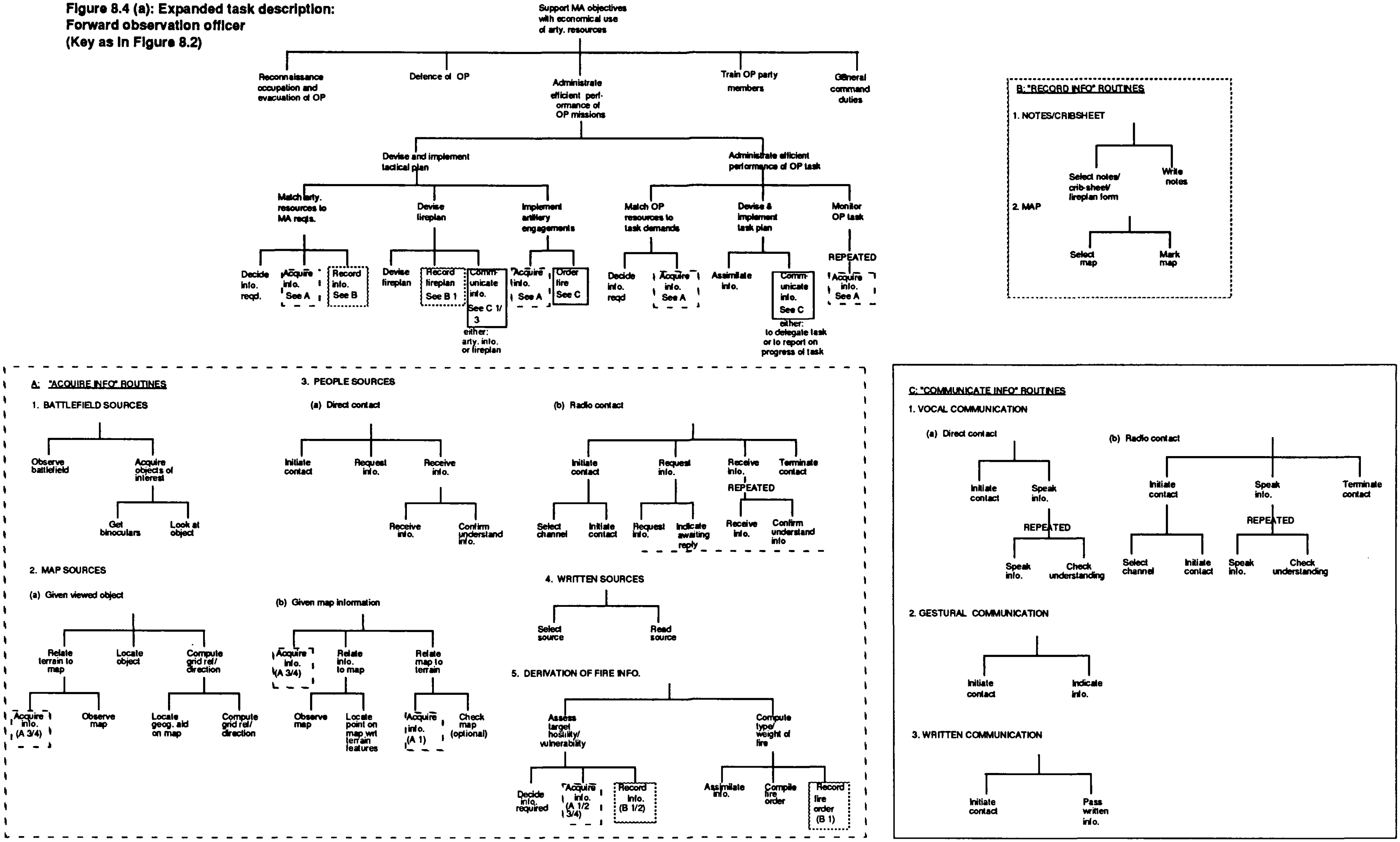
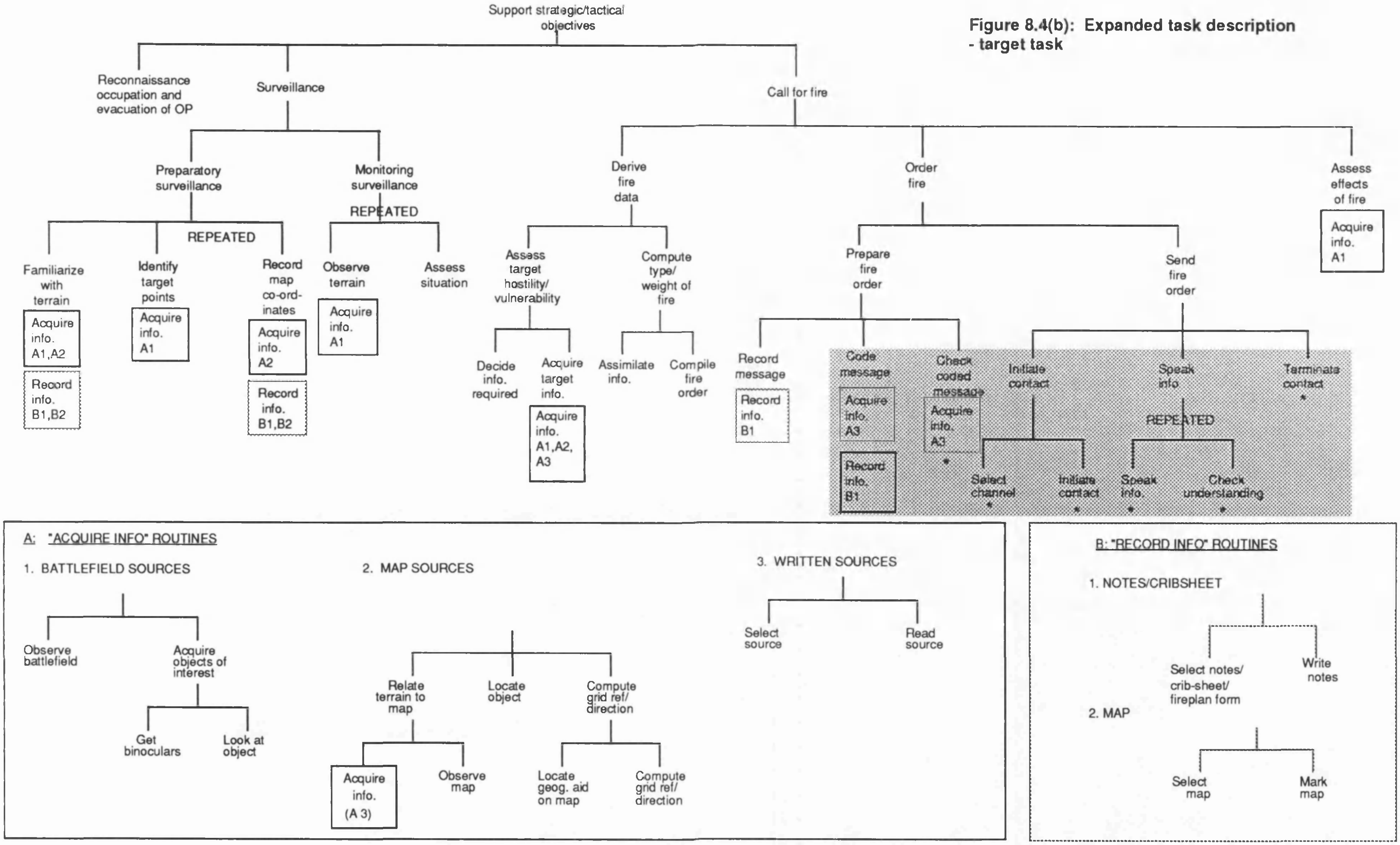


Figure 8.4(b): Expanded task description - target task



8.3 Generation of an expanded task description

8.3.1 Expanded task description

Purpose. Representation of current task.

Expression. Hierarchical description as in 8.1, modified as necessary to include new actions observed in the task and/or sequential relationship not previously recognized.

8.3.2 Procedure for generating an expanded task description

1. Check that the description of the current task is complete at the level specified in the solution strategy.
   - (a) Modify the preliminary task description to account for new actions observed in the video record (Step 2(a) and 2(b) of procedure 8.2.2).
   - (b) Ensure that all subjects' functional utterances (i.e. ignoring irrelevant comments) are attributable to actions at the level in the hierarchy specified in the solution strategy
   - (c) Run through the video tape, stopping at fixed (say, 30 second) intervals to check that all behaviour in the video record is attributable to actions in the description. [The frequency of checking depends on the level of description: lower levels of description demand more frequent stopping.]

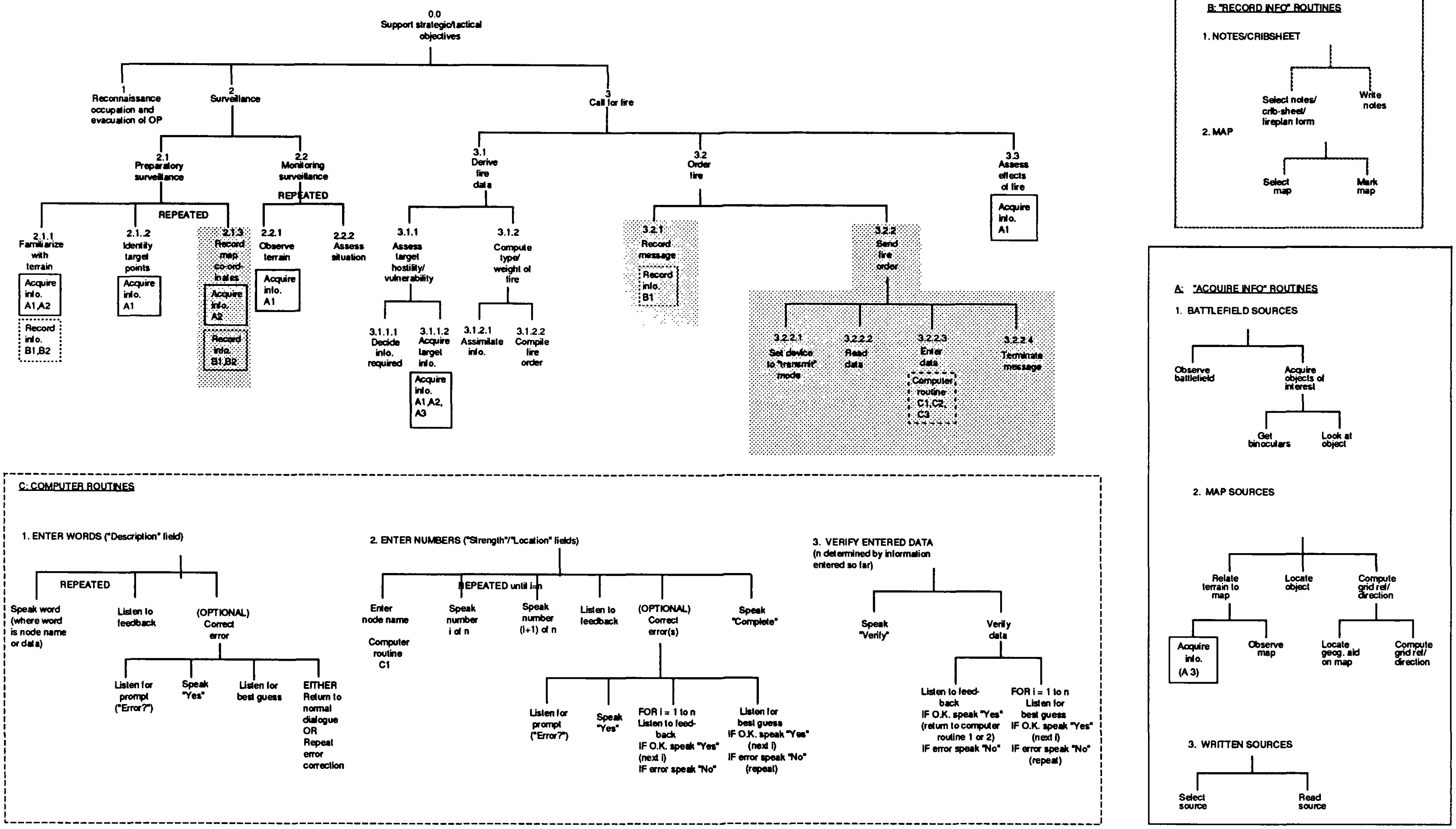
2. Check the sequential relationships between actions.
   - (a) ensure that any action remains completely described by the actions at the level below.
   - (b) ensure that actions are represented in the order in which they occur.

3. Confirm the description with army experts.

8.3.3 Example of expanded task description

The preliminary task descriptions were elaborated by reference to the video record of task performance. The task representation of the artillery observer was extended to a lower level of description (Figure 8.4(a)). Given detailed information on the performance of the Forward Observer's task, the preliminary description of the target task was also elaborated and checked with the domain expert (Figure 8.4(b)). In Figure 8.4, some higher level actions could be implemented in alternative ways; for example the action "Acquire information" could be implemented by observing the battlefield, by reading a map, by reading notes or by consulting colleagues. In such cases, each of the alternative means of implementation was decomposed hierarchically in the boxed sections in the lower half of the diagram; these routines could be appended to the lowest level of the main hierarchy.

Figure 8.5: Future task description (target task)



8.4 Generation of a future task description

8.4.1 Future task description

Purpose: Representation of (hypothesized) future task.

Expression: Hierarchical description as developed by means of Procedure 8.3, modified such that actions rendered unnecessary by the target device are excluded and actions necessary to operate the target device are included.

8.4.2 Procedure for representing the future (target) task

1. Differentiate device and non-device actions. Using the description of the target device in the preliminary system specification, identify in the current task description all actions which will, in future, be achieved by the use of the target device. List these as "device actions". The remainder are "non-device actions".

2. Modify the task description.
   (a) Delete from the current task description all device actions and other actions which will be unnecessary when the target device is implemented. Delete actions lower in the hierarchy which comprise these actions.
   (b) Generate new actions necessary to operate the target device, and locate them in the hierarchy. This is done using what is known about the target device to predict the actions necessary to use it in as much detail as possible.

3. Check the task description:
   (a) ensure that any action remains completely described by the actions at the level below.
   (b) ensure that the preconditions of the new actions are met. For example, the output of preceding action(s) must be appropriate as input to the new action; so if the target device requires, say, the entry of grid data, a preceding action must have generated a grid reference or it must have been obtainable some other way.

8.4.3 Example of future task description

Actions to be achieved in future by means of the target device were identified in the task hierarchy. These are marked with asterisks in Figure 8.4(b). Device actions, and other actions rendered unnecessary by the target device were deleted from the hierarchy. These are shaded in Figure 8.4(b)

New actions were generated to enable task performance using the target device. The documentation of the VDC was the source of device information. It was assumed that the device might be instantiated in one of three possible specifications, differing only with respect to the string length for entry of target location (GRID) data (2, 4, or 8 items). The task *description reflected these alternatives as differences in the value of the variable n in the action routines for computer operation (C2 and C3). The Future Task Description is shown in Figure 8.5.*

8.5 Generation of a future task model

8.5.1 Future task model

Purpose: Representation of actions influencing operation of the target device in the context of the future task.

Expression: Hierarchical task description as in 8.4, modified such that only actions relevant to the study are included.

8.5.2 Procedure for generating a future task model

1. Identify actions which are critical according to column 2 of the configured diagnostic table.

2. For each critical action, specify the objects of the action. In the case of "device actions", the object will obviously be the target device, but hardware and software entities which comprise the device may also be distinguished if the level of description of the task has been low. For example, the object of a database enquiry action might be a page of information in the database. Non-device actions will have as objects other entities in the task domain: tools (such as binoculars, crib-sheets, note pad, radio communication equipment etc.); work colleagues; friendly and enemy forces.

3. For each action and its objects, specify the attributes identified as critical in columns 3, 4, 5 and 6 of the configured diagnostic table. Columns 3, 4 and 5, respectively, identify attributes of the *device, user* and *context* which will be represented in the task simulation. Column 6 identifies dynamic attributes of the task itself which will be represented.

4. Identify the smallest subset of task actions which is representative in terms of entities and critical attributes. The intention is to abstract from the task description a task which may be reproduced in the laboratory as a vehicle to evaluate the usability of the target device (See Comment i).

5. Specify the task model. Delete all actions from the Future Task Description which were not selected in Step 4, and if an action has no critical actions below it in the hierarchy, delete it.

8.5.3 Comments on development of a future task model

(i) If a task comprises a number of actions having entities and critical attributes in common at the level determined in the solution strategy, some of these may be left out of the task model. So, for example, if the task requires map grid references to be calculated and entered for two different purposes, it may only be necessary to include in the task model the actions fulfilling one of these purposes, as it might be assumed that the procedure of calculating and entering the information would be the same in each case. However, ensure that the integrity of linked sequences of actions is maintained, i.e. do not leave out actions which are logically required for the performance of subsequent actions.

8.5.4 Example of future task model

The diagnostic relevant to task simulation development had been specified as part of the solution strategy in the course of applying the usability evaluation method (see Procedure 7.4). The diagnostic specified that actions demanding expression of information to be entered to the computer were potentially critical.

Actions 2.1.3, 3.2.1 and 3.2.2 (shaded in Figure 8.5) all involved the expression of information concerning enemy targets and so were critical according to the criterion of the diagnostic. The objects of these actions were as follows:

Action 2.1.3	*- map*
	- notes/crib/sheet
Action 3.2.1	*- notes concerning target type, strength and location*
Action 3.2.2.1	*- VDC system mode switch (receive/transmit)*
3.2.2.2	*- notes concerning target type, strength and location*
3.2.2.3	*- VDC accepting data concerning target types, strengths and locations*
	- message editor presenting prompts and feedback as synthesized speech (see Section 9.1.4 for fuller details)
3.2.2.4	*- VDC system mode switch (receive/transmit)*

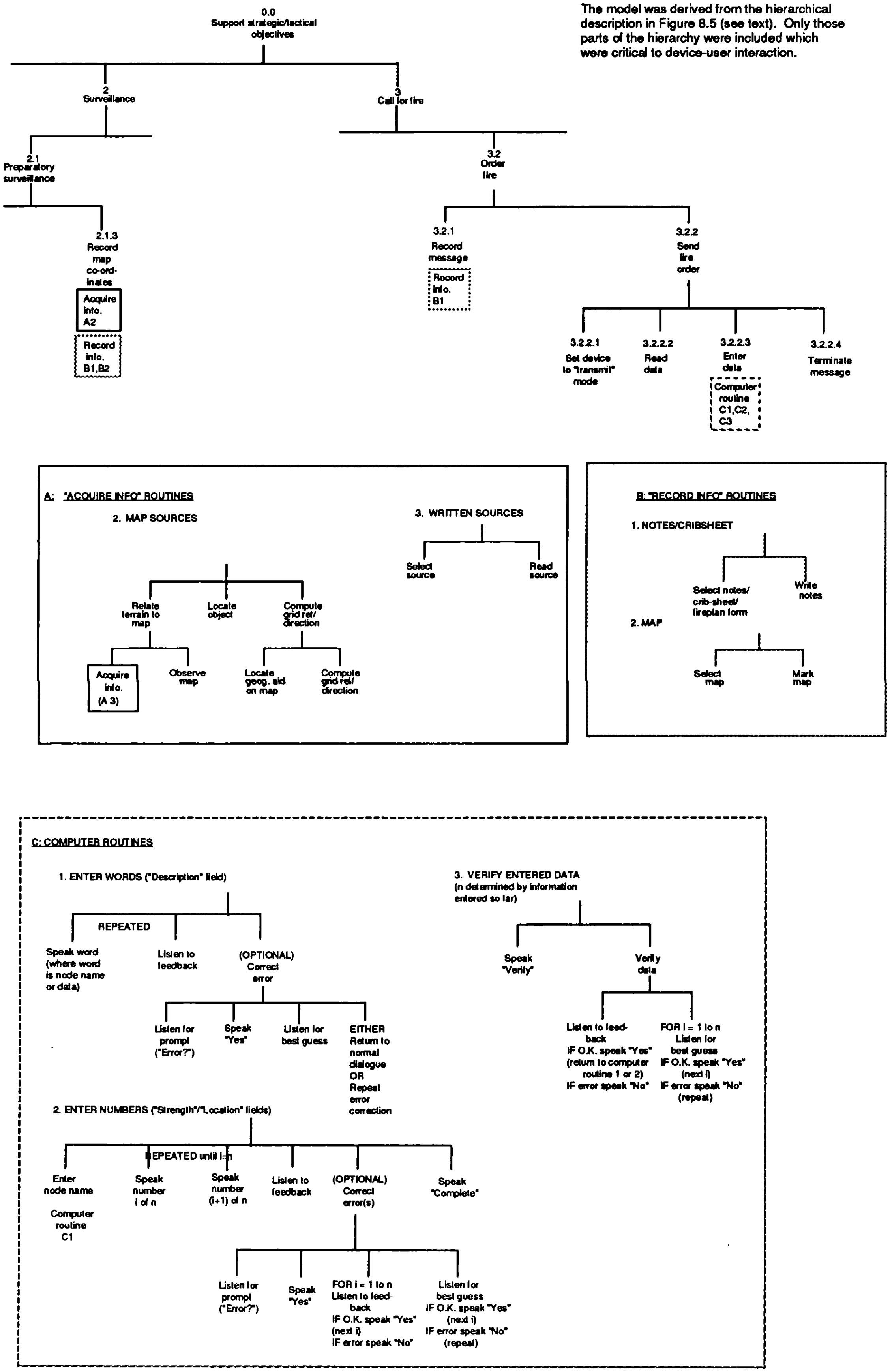
According to the diagnostic, critical attributes of the system were as follows:

Critical device attributes:

- *device constraints on the chunking of data (i.e. 2, 4 or 8 digit strings when entering grid references)*
- *feedback chunk constraints (i.e. 2, 4 or 8 digit strings)*
- *recognition error frequency (i.e. 1% or 4%; see Section 9.1.4)*
- *availability of aids to memory (i.e. record of messages to be sent)*

Figure 8.6: Future task model (target task)

The model was derived from the hierarchical description in Figure 8.5 (see text). Only those parts of the hierarchy were included which were critical to device-user interaction.



0.0 Support strategic/tactical objectives
2 Surveillance
3 Call for fire
2.1 Preparatory surveillance
3.2 Order fire
2.1.3 Record map co-ord-inates
Acquire info. A2
Record info. B1,B2
3.2.1 Record message
Record info. B1
3.2.2 Send fire order
3.2.2.1 Set device to "transmit" mode
3.2.2.2 Read data
3.2.2.3 Enter data
Computer routine C1,C2, C3
3.2.2.4 Terminate message
A: "ACQUIRE INFO" ROUTINES
2. MAP SOURCES
Relate terrain to map
Locate object
Compute grid ref/ direction
Acquire info. (A 3)
Observe map
Locate geog. aid on map
Compute grid ref/ direction
3. WRITTEN SOURCES
Select source
Read source
B: "RECORD INFO" ROUTINES
1. NOTES/CRIBSHEET
Select notes/ crib-sheet/ fireplan form
Write notes
2. MAP
Select map
Mark map
C: COMPUTER ROUTINES
1. ENTER WORDS ("Description" field)
REPEATED
Speak word (where word is node name or data)
Listen to feedback
(OPTIONAL) Correct error
Listen for prompt ("Error?")
Speak "Yes"
Listen for best guess
EITHER Return to normal dialogue OR Repeat error correction
2. ENTER NUMBERS ("Strength"/"Location" fields)
REPEATED until i=n
Enter node name
Computer routine C1
Speak number i of n
Speak number (i+1) of n
Listen to feedback
(OPTIONAL) Correct error(s)
Speak "Complete"
Listen for prompt ("Error?")
Speak "Yes"
FOR i = 1 to n Listen to feed-back IF O.K. speak "Yes" (next i) IF error speak "No"
Listen for best guess IF O.K. speak "Yes" (next i) IF error speak "No" (repeat)
3. VERIFY ENTERED DATA (n determined by information entered so far)
Speak "Verify"
Verify data
Listen to feed-back IF O.K. speak "Yes" (return to computer routine 1 or 2) IF error speak "No"
FOR I = 1 to n Listen for best guess IF O.K. speak "Yes" (next i) IF error speak "No" (repeat)

Critical user attributes:

- *familiarity with task data (i.e. familiarity with format of messages and with grid references)*

Critical context attributes:

- *factors disrupting memory processes (i.e. noise, stress etc.)*

Critical task attributes:

- *constraints on the structuring of task information (i.e., 8 digit grid references with 4 digit eastings and northings)*

The task model comprised all of the critical actions identified above . The future task description was modified to include only these critical actions (see Figure 8.6)

8.6 Generation of a task simulation specification

8.6.1 Task simulation specification

Purpose: Description of entities and their attributes to be included in the simulation, and of the goals of the task which simulated users will be required to perform.

Expression: (1) Narrative description of the task to be performed.

(2) List of entities/attributes impacting task actions.

(3) A description of external events influencing task actions (e.g. events on the battlefield).

8.6.2 Procedure for specifying the task simulation

1. Specify the entities and their attributes to be included in the simulation. These will have been identified in Procedure 8.5. (Step 3) and are now specified for each action in the task model at the level of description determined by the solution strategy. Entities must be included that will enable task model actions to be reproduced accurately at this level of description (see comment i).

2. Specify task attributes to be included in the simulation. These also will have been identified in Procedure 8.5 (Step 3). Refer to the current task data to determine the incidence of battlefield events which initiate/impact actions in the task model e.g. instructions from a senior officer, or fall of shot as a result of sending a fire order.

3. Specify the simulated task.
   (a) Link the actions in the task model within a hypothetical operational scenario: this requires a certain amount of judgement and creativity on the part of the investigator.
   (b) Produce a narrative description of the task scenario.

(c) List the entities and attributes to support the task, including task inputs necessary as a consequence of having deleted actions which would have preceded those in the task model. For example, if the model includes data entry actions, but not the actions of generating the data, then the simulated user must be provided with previously-prepared (simulated) data.

(d) Specify the nature and frequency of external events which pace the current task.

4. Confirm that the narrative description presents a coherent scenario, given the entities and events. Make a subjective evaluation of the relationship between the specified simulated task and the current task. The simulation should represent the task as it is expected to be performed with the future device, at least with respect to the features deemed critical in the configured diagnostic table (see Comment ii).

5. Return to Procedure 7.5 of the usability evaluation method.

8.6.3 Comment on specification of the task simulation

(i) The target device and its attributes are characterized fully in the device simulation method (see Chapter 9). The requirement in the current procedure is to gain only sufficient information about the device to enable the task simulation to be specified.

(ii) The procedure for implementing the simulation is not described. It requires accurate representation of the entities and critical attributes, so that user behaviour is reproduced in the actions of the task model. Precision of representation of critical attributes is the criterion for deciding the means of implementation. Under some circumstances, it may be possible to import either real objects with appropriate attributes (e.g. a real data entry device, real fire plan forms) or simulated entities developed for other purposes (e.g. a training simulator).

8.6.4 Example of task simulation specification

The entities and attributes identified in Procedure 8.5 (Step 3) formed the basis of the task simulation specification. These were linked to create the following operational scenario.

Subjects were required to produce a list of fire orders for entry to the computer from the information given on a marked map. They then entered the information using the VDC, referring to their pre-prepared list. The information consisted of a target description (three individually entered words; e.g. Alpha Artillery Medium); target strength (a single digit between 1 and 9); and target location (eight digit grid reference; e,g, 46669184).

The location data were entered in either 2,4 or 8 digit chunks according to the experimental condition. All entered data were subject to recognition errors of the device and so had to be corrected before new data were entered.

8.7 Critical commentary on the task simulation method

The task simulation method utilizes exclusively a hierarchical tree structure for representing tasks. This kind of representation was selected because of the clarity with which it is able to express the relationship between the actions which are exhibited in the achievement of task goals. However, a hierarchy of actions assumes that behaviour may be decomposed into a distinguishable sequence of discrete elements. Whilst much behaviour is readily decomposable in this way (e.g. spoken dialogues based upon turn-taking), some classes of behaviour may be less so (e.g. continuous manual control behaviour). Furthermore, tree structures which represent the *ordering* of actions may be difficult to generate for tasks in which actions may occur in a variety of alternative sequences (i.e. when a task is not strictly proceduralized). The procedures of the task simulation method may be less appropriate, then, when the task to be represented includes off-line activities of a continuous nature or when sequences of actions are unpredictable.

It is also notable that the method assumes an ability in assessors (either innate or learned) to infer structure in the intentional behaviour of others (i.e. an ability to infer goal-directedness in behaviour and so to interpret behaviour in terms of actions). It further assumes that assessors can acquire sufficient knowledge of the semantics of task behaviour (i.e. the identity of, and relationships between, task goals) to apply their analytic abilities appropriately. It is, therefore, important that the assessor researches the task in adequate depth prior to attempting inferential analysis.

CHAPTER 9

DEVICE SIMULATION METHOD

9.0 Process of device simulation - summary

The process of developing device simulations requires the assessor to generate five representations of the target device (see Figure 9.1).

1. The functions supported by the target device are specified, and its behaviour and performance inferred as they relate to the user's device actions in the future task model (see Procedure 8.5). However, only those aspects of behaviour which are critical to interaction will be included in the device simulation. The specification of the target device is, therefore, reviewed with respect to the attributes identified as critical by the diagnostics, to generate an *elaborated device specification.*

2. The device specification is decomposed to identify those device functions which are already implemented (perhaps in the form of some existing version of the target device), and those which must be simulated. The functions to be simulated are analysed in terms of the allocation of function between the system subject and the communication device. The system subject's task is then derived from the future task model by specifying system subject actions necessary to emulate the target device in the context of each of the actions in the task model. A communication device is specified to support this task. The specification of the system subject's task and of the communication device, together, constitute the *device simulation specification.*

3. The specification is implemented as a *device simulation.*

4. The device simulation undergoes evaluation, during which its behaviour and performance are compared with those in the device specification. If the simulation performance is acceptably close to that in the specification, device simulation development is terminated; otherwise, iterative refinement occurs on the basis of the *device simulation performance results.*

5. The behaviour and performance of the simulation is analysed with respect to the interaction between the system subject and the communication device. In the *analysis of device simulation behaviour,* inadequacies are diagnosed and ergonomic intervention prescribed to make good inadequacies. Intervention takes the form of system subject

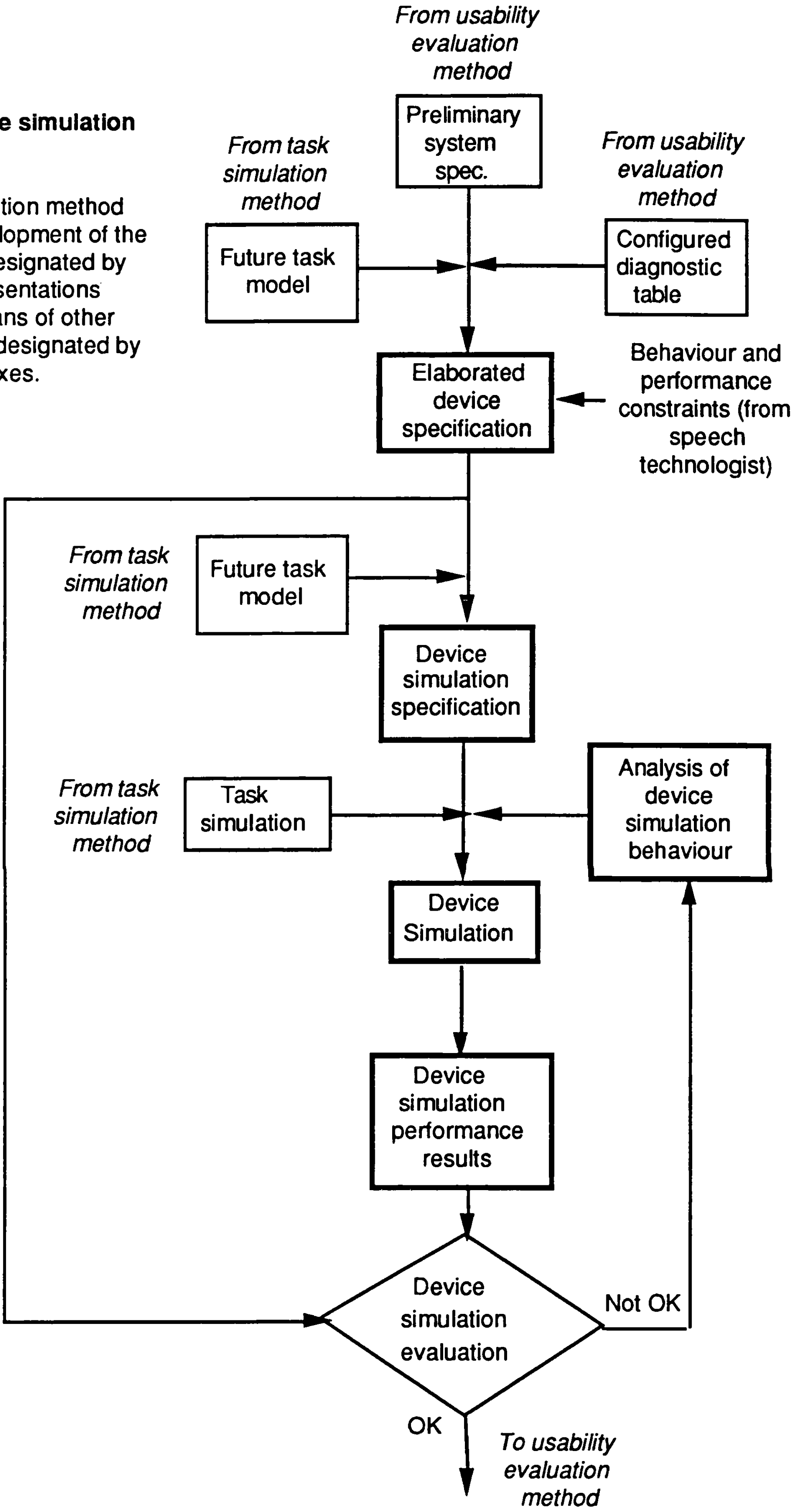

From usability evaluation method
Preliminary system spec.
From task simulation method
Future task model
From usability evaluation method
Configured diagnostic table
Elaborated device specification
Behaviour and performance constraints (from speech technologist)
From task simulation method
Future task model
Device simulation specification
From task simulation method
Task simulation
Analysis of device simulation behaviour
Device Simulation
Device simulation performance results
Device simulation evaluation
Not OK
OK
To usability evaluation method

Figure 9.1 Device simulation method: process

The device simulation method supports the development of the representations designated by bold boxes; representations developed by means of other sub-methods are designated by unemphasized boxes.

selection, training or aiding. Steps 4, and 5 are repeated until simulation performance is deemed acceptable.

9.1 Generation of an elaborated device description

9.1.1 Elaborated device description

Purpose: To enable the specification of the device simulation, and to provide criteria for evaluating it when the simulation has been implemented.

Expression: Two descriptions, consistent with respect to each other:

(a) Device behaviour specification. The form of the description depends upon the complexity of the target device. It includes a description of the user dialogue which supports activities in the future task model. A software engineering representation, such as a state transition diagram, would be appropriate.

(b) Device performance specification. A description of the target device with respect to the critical performance attributes identified in the configured diagnostics (see Procedure 7.3)

9.1.2 Procedure for generating the elaborated device description

1. Identify target device functions to be included in the study (see Comment i). The target device supports users in the performance of the target task by processing information relevant to the task domain (e.g. the battlefield). The processes it performs are termed functions within SIAM; (e.g. data entry, data display, error correction and data processing such as the retrieval of items of information from a database). Identify in the preliminary system specification device functions utilized in the device actions of the future task model, then identify the information which is the input and output of these functions in the case of the target task.

2. Specify the interface dialogue which supports the functions identified in Step 1. Unless the target device has already been implemented (e.g. as a prototype), specification will require the collaboration of the target device project team within the procurement organization. The dialogue is specifiable at different levels, the highest representing the invoking of functions to achieve the super-ordinate goal of the task (e.g. data entry, information retrieval), and the lowest, the physical interactions which enable system operation (e.g. speaking words, displaying items of information). The level appropriate for the study will have been decided as part of the solution strategy (see Procedure 7.4). See Comment ii.

3. Identify critical device attributes. Identification is achieved by referring to column 3 of the configured diagnostic table (see Procedure 7.3). It is also necessary to specify the

level of description at which critical attributes are to be represented accurately in the study (see Procedure 7.4). See Comment iii.

4. Specify the characteristics of the target device to be represented in the device simulation, i.e. characterize the target device with respect to each of the critical attributes identified in step 3. In the case of performance attributes,

   (a) IF the target device is current, determine its operational behaviour and performance with respect to the attributes identified in Step 3. This may be done by experimentation, or, if this is not possible, by less formal observations of the device in use. Take care to ensure that the observations are made under conditions which will elicit representative performance: the vocabulary, device users and operating environment should be as equivalent as possible to the operating context of the target device.

   (b) ELSE, in collaboration with the target device project team, specify the target device in a way appropriate for the prediction of likely behaviour and performance by a speech technologist. The specification might include:

      - necessary system language features (e.g. as specified in the user dialogue)
      - operating context (e.g. physical environment, operator specificity)
      - price constraints
      - implementation constraints (e.g. portability, available technologies).

   See Comment iv.

At the conclusion of this step, there should exist descriptions of:

(a) the interface dialogue, representative with respect to the critical attributes at the level decided in the solution strategy;

(b) any critical performance attributes of the device which must be reproduced accurately.

9.1.3 Comments on the generation of an elaborated device description

(i) It is only necessary to develop a simulation of the device which will behave in a representative way in the context of the *simulated* task, i.e. a complete specification of the device is unnecessary. The elaborated device specification is, then, the target device specification expressed with respect to the actions of the future task model (and, subsequently, also with respect to the critical attributes identified in the diagnostics).

(ii) For the present purpose, it is recommended that the dialogue be specified in a top-down fashion, i.e. first specify device functions to be included (from Step 1, above), specify the sequence of device and of user interchanges to perform the functions, then specify physical actions and the corresponding responses of the device at the level of data I/O. The goal is to specify the user interface in sufficient detail to enable it to be implemented as a simulation. It is important that the simulation is an accurate representation of the target device dialogue at the level(s) of description determined in the solution strategy (Procedure 7.4).

(iii) Just as the device specification may be expressed at different levels of description, so may critical attributes. For example, in the case of the recognition reliability of a speech input device, representation might be accurate with respect to the overall mean frequency of errors (i.e. representative at a high level), or with respect to the occurrence of individual confusions (representative at a low level). In the case of physical attributes, representation might be accurate with respect to the class of display of the target device - say its being a *visual* display device (i.e. representative at a high level) - or it might be accurate with respect to the details of the layout of the display (i.e. representative at a low level). The level of description for interpreting the critical attributes will have been decided in the solution strategy (Procedure 7.4).

(iv) The availability of information about the behaviour of the target device is a primary determinant of the precision of the conclusions of the study (see Appendix E). In the case of a target device which has yet to be specified in detail, the assessor must hypothesize its behaviour and performance on the basis of information elicited from relevant experts. Some of this information may be specified in the Staff Target (user requirements document for the target system); some will have to be proposed by the project team and verified by speech technologists. If speech technologists are unable to commit themselves to prediction, propose approximate performance values and ask speech technologists to assess the likelihood that they will be achieved. Negotiate down to a range of likely performance.

9.1.4 Example of elaborated device specification

A functional specification existed for the VDC in the form of a diagram indicating the device recognition vocabulary pertaining at each node of the dialogue (i.e. it presented vocabulary options for each possible state of the device) - see Figure 9.2. Its expected recognition performance was estimated by the device developers . Relevant diagnostics had been identified as part of the solution strategy in the course of applying the usability evaluation method (see Section 7.3.4). The diagnostics identified critical device attributes as being:

- *device constraints on chunking of data/feedback*
- *chunk characteristics*
- *recognition error frequency*
- *availability of memory aids*

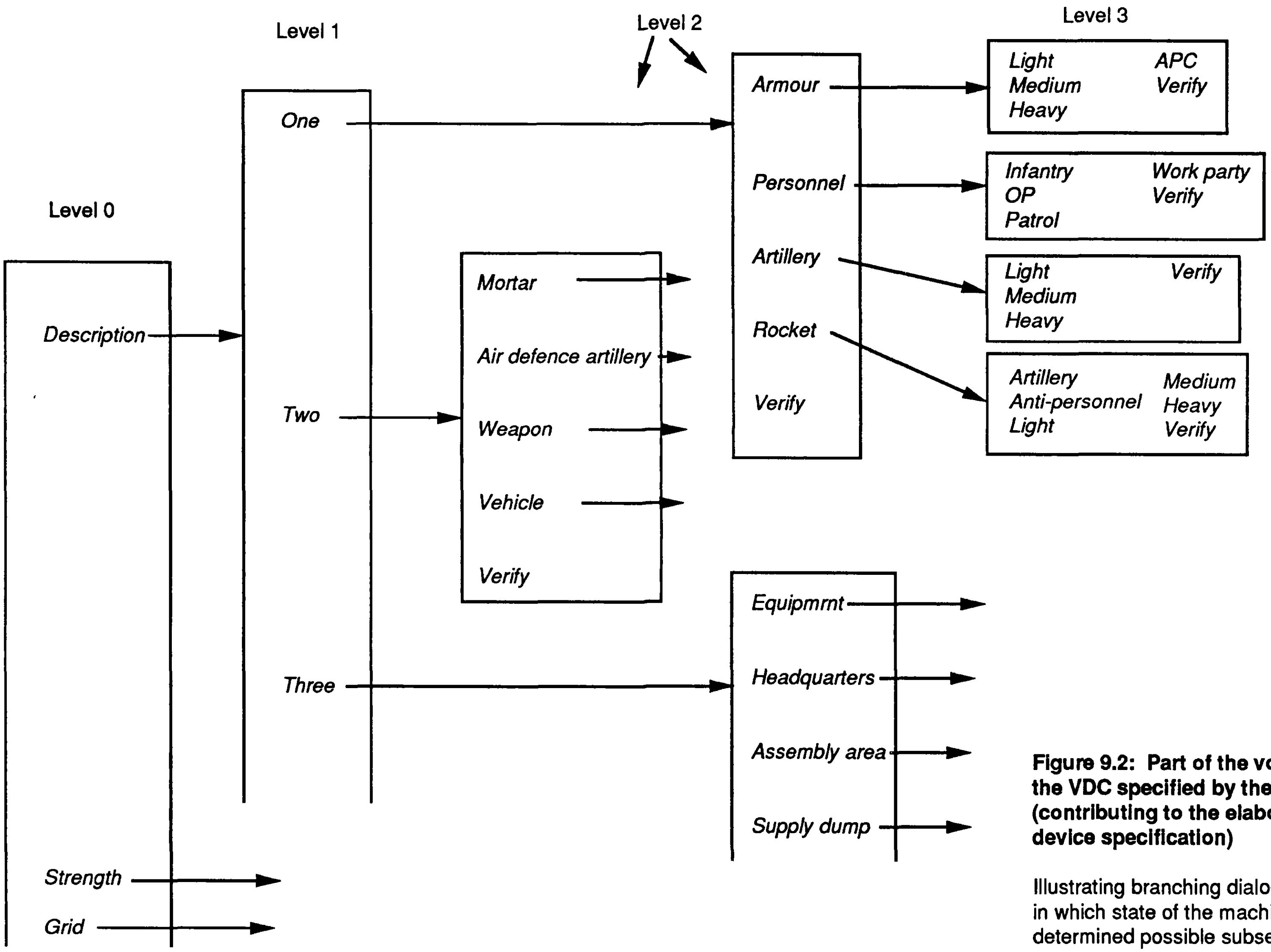


Figure 9.2: Part of the vocabulary of the VDC specified by the developers (contributing to the elaborated device specification)

Illustrating branching dialogue structure in which state of the machine determined possible subsequent states

These attributes were evaluated in the context of the specification of the target device. Device constraints on chunking were that location data were to be expressed in groups of either 2, 4 or 8 items before feedback would be presented; the data could take the form either of words (target descriptions) or numerals (target strength and location); error frequency was either 4% or 1%; and the user had available a memory aid in the form of a list of the data to be entered which had been generated by reference to the map (in Action 3.2.1).

The elaborated device specification comprised, then, the existing dialogue, but expressed with respect to the entry of data in strings of 2, 4, or 8 items; a specification of error frequencies; memory support for users in the form of a list of data to be entered.

9.2 Generation of a device simulation specification

9.2.1 Device simulation specification

Purpose: To enable the simulation to be implemented as a system subject interacting with a communication device.

Expression: (1) Procedural instructions for the system subject

(a) text description of user subject task

(b) system subject action hierarchy

(c) text description of cues for system subject actions

(d) specification of required simulation performance

(2) text description of knowledge required by system subject

(3) specification of the communication device (software engineering representation, e.g. flowchart/state transition diagram)

9.2.2 Specifying the device simulation

1. Identify target device functions which have already been implemented (e.g. in a target device prototype). For example, if the evaluation were for a speech interface to an existing database system, the existing database system would potentially offer the implemented functions. Specify implemented functions which might be used as part of the device simulation (see Comment i).

2. Identify remaining target device functions which must be simulated by means of the communication device and the system subject.

3. In the case of the functions to be simulated, make a preliminary allocation between the system subject and the communication device. Use Table 9.1 as a source of heuristic solutions to allocation. The aim is for the system subject to be able to reproduce all the

Table 9.1 Heuristics for allocating function between the SS and CD

(Key: SS = system subject; TD = target device; CD = communication device)

1. Allocate all TD input/output functions which are not supported by speech to the CD (if they are not supported by a TD prototype).

2. Allocate control of stochastically-determined performance characteristics to the CD (e.g. device error simulation).

3. Allocate storage of unstructured data, and data access functions, to the CD.

4. Allocate low-level, algorithmically-determinable data manipulations to the CD (e.g. arithmetical procedures).

5. Allocate frequently-used repetitive functions to the CD (e.g. generation of regularly-used text strings).

6. Allocate functions involving precisely timed device responses to the CD.

7. Allocate control of complex dialogue sequences to the CD.

8. Allocate speech input/output functions to the SS, unless a speech device with adequate performance is available.

9. Allocate non-deterministic search procedures to the SS.

10. Allocate pattern-matching procedures to the SS.

device functions with the help of the communication device and any functions already implemented.

4. Make a preliminary specification of the system subject's task.
    (a) Identify user subject actions which involve device use in the future task model (see Procedure 8.5). For each user subject device action, specify the system subject actions necessary to the emulate target device by means of the communication device and any available target device implemented functions.
    (b) Check that the allocation of function could realistically enable the system subject to support the required device actions.

5. Specify the communication device.

   (a) For each system subject action in the system subject action hierarchy, determine the communication device function necessary to support it.

   (b) If the communication device cannot support all functions, identify alternative means of supporting the system subject (e.g. using non computer based facilities, such as tables of information on paper).

   (c) Express required functionality in a form suitable for implementation (see Comment ii).

   (d) Review the system subject task specification in the light of the communication device specification.

6. Prepare system subject instructions comprising

   - a description of the user subject task (from the task simulation specification - see Procedure 8.6)
   - the system subject task hierarchy developed in 4(a) above (if necessary, with appended instructions on operation of the communication device and other facilities)
   - a description of potential sources of information which might enhance system subject performance (e.g. visual and auditory cues which provide advance information about future user subject actions). See Comment iii.
   - the required performance specification from Procedure 9.1.

9.2.3 Comments on device simulation specification

(i) If the existing version of the target system may be operated for the purposes of the study, it may be utilized as part of the device simulation. For example, the system subject may be able to use a keyboard-based version of the target device "off-line" to support his/her simulation task.

(ii) SIAM does not extend to the implementation of the specification, so the expression of the specification is left to the discretion of the assessor. Its form is likely to be determined by the engineering skills of the assessor or the preferences of locally-available technician support.

(iii) The simulation task is facilitated for the system subject if continuous information is provided on the behaviour of the system subject. Given knowledge of the current state of the user subject's task (e.g. by means of a television view of the user subject), the system subject can predict likely user subject actions and so reduce his/her own response time.

Potentially predictive cues should be identified and included as part of the system subject's task specification.

9.2.4 Example of device simulation specification

Functions were allocated between the system subject and the communication device as follows, using the heuristics in Table 9.1:

System subject	*Communication Device*
Speech recognition	*Stochastically determined "error" insertion*
Speech synthesis	*Storage of "error matrices" (confusion at word level)*
Error correction function	*Constraints on chunking of numeric data (grid information)*

Actions were identified in the future task model (Figure 8.6) which constituted interactions with the target device (i.e. Actions 3.2.2.1, 3.2.2.2, 3.2.2.3 and 3.2.2.4). For each of these actions, corresponding actions of the system subject were specified, such that the behaviour of the target device would be emulated. Figure 9.3 presents a model of the system subject's task, showing its relationship with the device actions of the user subject.

The communication device consisted of a two-way speech link between the user subject and the system subject; speech transmitted from the system subject was distorted and filtered to simulate current-generation synthesized speech. A television view of the user subject was available to the system subject. The system subject's task was additionally supported by a computer which automated the generation of confusion errors representative of the target device. Individual keys on the computer represented each word in the vocabulary of the target device. Pressing a key initiated a call to a look-up table, which expressed the probability that the entered word would be recognized correctly or confused with another word in the device's vocabulary. The activation of a key, then, resulted in a word being displayed visually, which was either correct or (with an appropriate probability) a "recognition error". Figure 9.4 illustrates the configuration of the simulation.

The system subject's instructions were as follows (abbreviated for the purpose of illustration): "You are required to transcribe all utterances of the user subject onto the computer using the keyboard. Speak aloud what is displayed by the computer, and your utterances will be transmitted back to the user subject. Be careful to read out what appears on the screen, and not just repeat what the user subject said. Although your speech will be distorted to make it sound like synthesized speech, try to minimize the inflections in your voice and to speak as consistently as possible, in order to enhance the "mechanical" effect........"

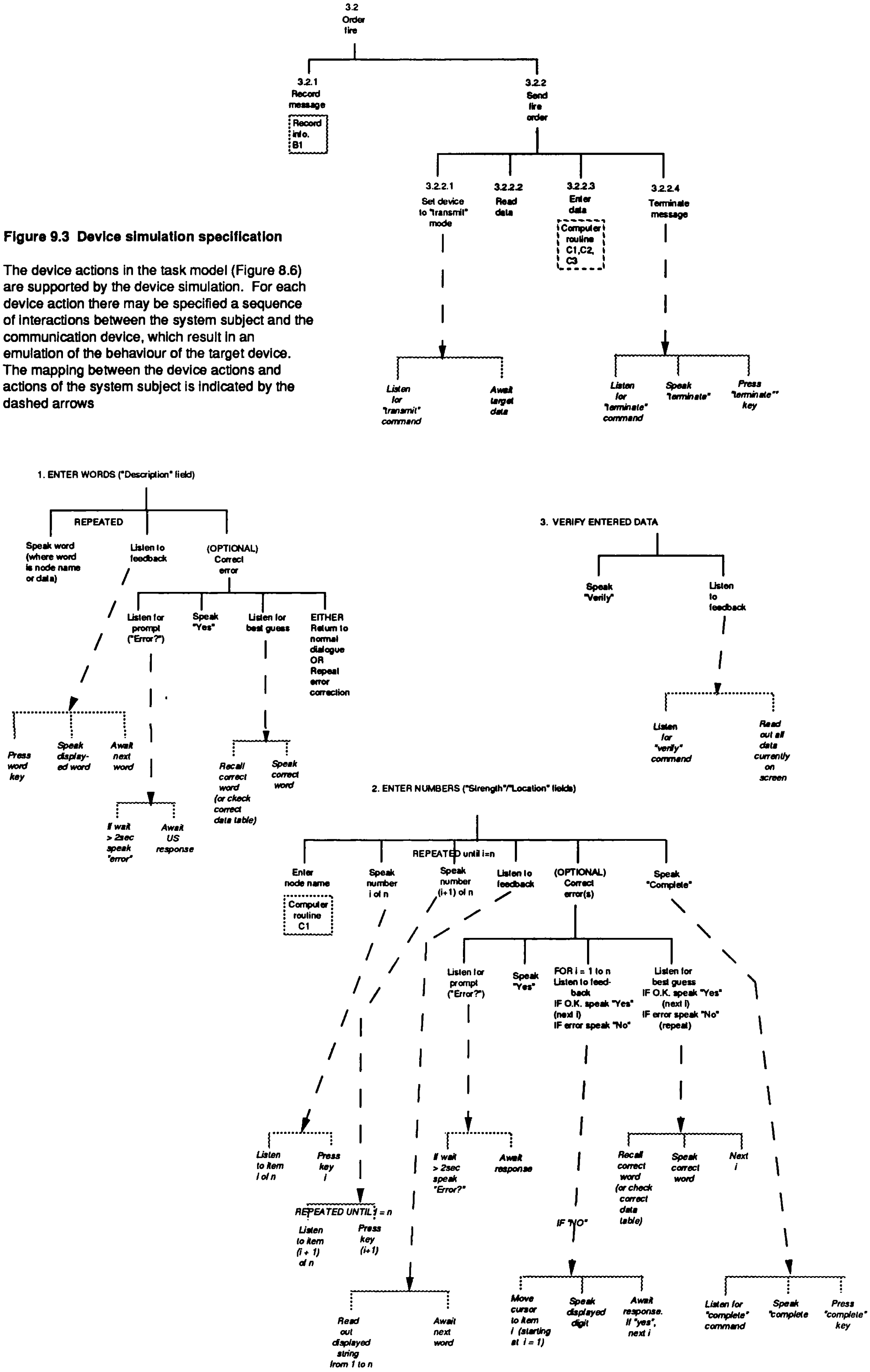

3.2 Order fire
3.2.1 Record message
Record info. B1
3.2.2 Send fire order
3.2.2.1 Set device to "transmit" mode
3.2.2.2 Read data
3.2.2.3 Enter data
Computer routine C1,C2, C3
3.2.2.4 Terminate message
Listen for "transmit" command
Await target data
Listen for "terminate" command
Speak "terminate"
Press "terminate" key
1. ENTER WORDS ("Description" field)
REPEATED
Speak word (where word is node name or data)
Listen to feedback
(OPTIONAL) Correct error
Listen for prompt ("Error?")
Speak "Yes"
Listen for best guess
EITHER Return to normal dialogue OR Repeat error correction
Press word key
Speak display-ed word
Await next word
If wait > 2sec speak "error"
Await US response
Recall correct word (or check correct data table)
Speak correct word
3. VERIFY ENTERED DATA
Speak "Verify"
Listen to feedback
Listen for "verify" command
Read out all data currently on screen
2. ENTER NUMBERS ("Strength"/"Location" fields)
REPEATED until i=n
Enter node name
Computer routine C1
Speak number i of n
Speak number (i+1) of n
Listen to feedback
(OPTIONAL) Correct error(s)
Speak "Complete"
Listen for prompt ("Error?")
Speak "Yes"
FOR i = 1 to n Listen to feed-back IF O.K. speak "Yes" (next i) IF error speak "No"
Listen for best guess IF O.K. speak "Yes" (next i) IF error speak "No" (repeat)
Listen to item i of n
Press key i
REPEATED UNTIL i = n
Listen to item (i + 1) of n
Press key (i+1)
If wait > 2sec speak "Error?"
Await response
Recall correct word (or check correct data table)
Speak correct word
Next i
IF "NO"
Read out displayed string from 1 to n
Await next word
Move cursor to item i (starting at i = 1)
Speak displayed digit
Await response. If "yes", next i
Listen for "complete" command
Speak "complete
Press "complete" key

Figure 9.3 Device simulation specification

The device actions in the task model (Figure 8.6) are supported by the device simulation. For each device action there may be specified a sequence of interactions between the system subject and the communication device, which result in an emulation of the behaviour of the target device. The mapping between the device actions and actions of the system subject is indicated by the dashed arrows

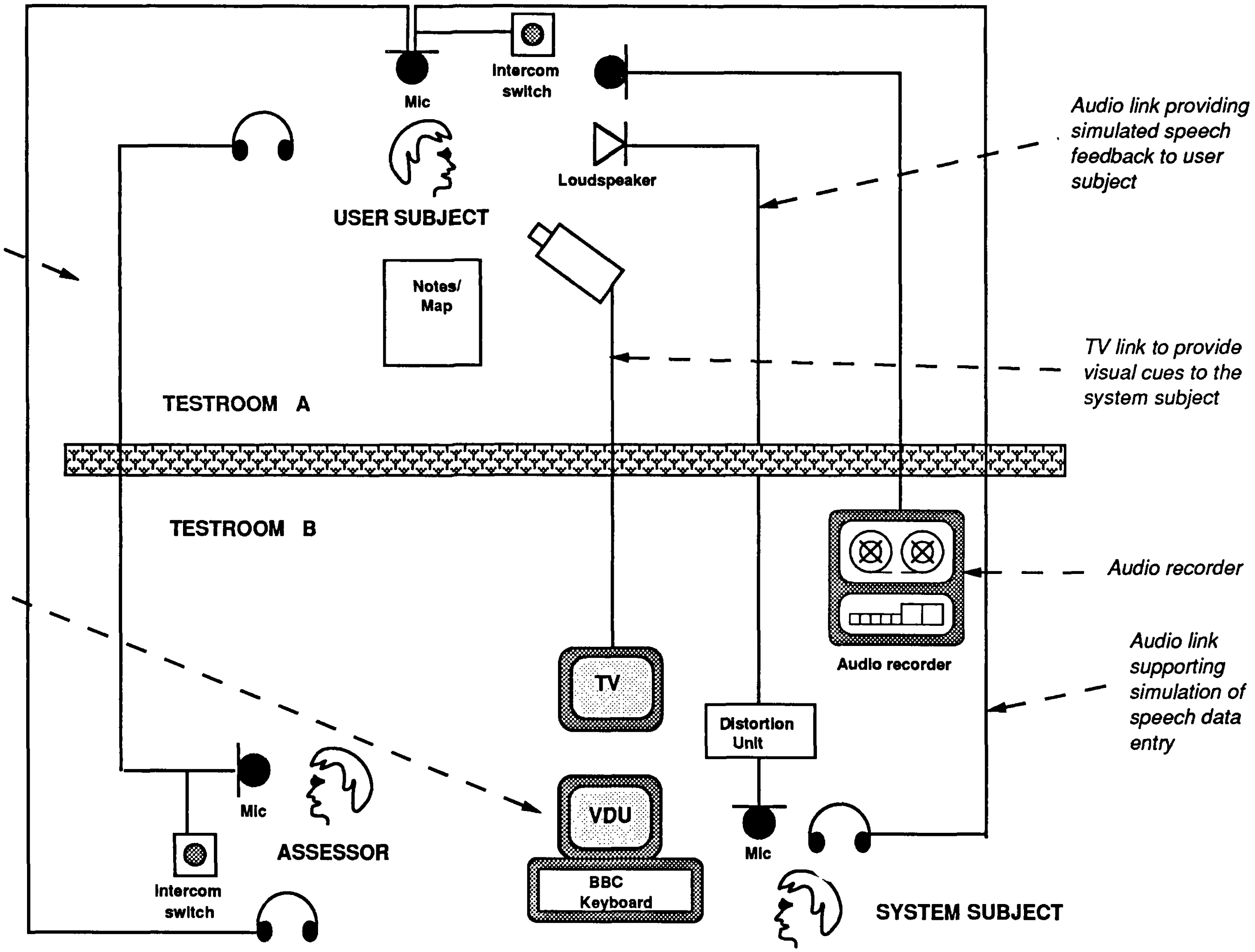

Intercom between assessor and user subject
Computer supporting insertion of recognition errors
Mic
Intercom switch
USER SUBJECT
Loudspeaker
Notes/ Map
TESTROOM A
TESTROOM B
TV
VDU
BBC Keyboard
Mic
ASSESSOR
Intercom switch
Distortion Unit
Audio recorder
Mic
SYSTEM SUBJECT
Audio link providing simulated speech feedback to user subject
TV link to provide visual cues to the system subject
Audio recorder
Audio link supporting simulation of speech data entry

Figure 9.4: Implementation of the simulated VDC

The simulation was implemented on the testbed described in Section 9.3.3.

The instructions included a description of the structure of the messages which the user subject would send, to enable the system subject to predict the type of keystroke that would be required next. For example, following the transmission of "Strength" data, "Grid" data invariably followed.

9.3 Implementation of a device simulation

9.3.1 Device simulation

Purpose: Evaluation of the usability of the target device, within the usability evaluation method.

Expression: Implemented specification in which the system subject emulates the target device using the communication device, implemented device functions and/or non-computer-based information sources.

9.3.2 Implementation of device simulation

1. Implement the communication device specification. (See Comment i).

2. Implement the specification of non-computer support.

3. Check that the device simulation operates in the context of the task simulation by running an informal test of the communication device, in which the assessor acts as a system subject, performing the system subject task with a surrogate user subject

4. Recruit system subjects (see Comment ii). The number of system subjects required will depend upon the size of the usability evaluation. As a general rule of thumb, recruit the smallest number to allow the study to be completed (i.e. ideally, one system subject). The criteria for selection will depend upon the skills required to perform the actions specified in the system subject instructions. In general, recruit for skills which are difficult to automate on the testbed.

5. Train system subjects.
   (a) Enable system subjects to become familiar with their instructions. The assessor should answer any questions the system subject may have regarding the user subject, user subject task, target device and communication device. The system subject should be considered a collaborator in the running of the experiment.
   (b) Enable the system subject to become familiar with the communication device in informal interactions in which the assessor plays the role of a user subject.

(c) Run a sequence of increasingly formal "usability evaluation" trials, during which the assessor plays the role of experimenter and a user subject surrogate performs as in the experiment to be conducted using the usability evaluation method.

(d) Informally evaluate the device simulation against the requirements specified in Procedure 9.1. Refine system subject-communication device configuration until device simulation performance approximates the requirement. Check the adequacy of the device simulation with the target device project team.

9.3.3 Comment on implementation of the device simulation

(i) The form of implementation will be determined by the technical resources available to the assessor. During the development of SIAM, a testbed was developed to support human simulations. The testbed was installed in two testing rooms, thus physically separating the user subject from the assessor and system subject. The two rooms were linked by a computer network (see Lee, 1989), an audio intercom and tape recorder and a one-way video link, enabling the assessor and system subject to observe and record the behaviour of the user subject. The computer network utilized, in this instance, BBC Master microcomputers communicating via Econet. This arrangement could be adapted to support the simulation of simple battlefield tasks and various speech-based devices (see, for example, the simulation described in Section 9.3.4).

(ii) An objective in recruiting system subjects is to minimize variability in the behaviour of the device simulation. Factors likely to determine variability are:

- skill exhibited by the system subject
- fatigue
- individual differences between system subjects.

The first factor may be controlled by recruiting system subjects with favourable potential for skill development, and by training. Individual differences will be eliminated if only one subject is employed; however, this advantage may be reduced if the study is large in scope, resulting in system subject fatigue.

9.3.4 Example of implementation of a device simulation

The specifications of the system subject and of the communication device were implemented on the testbed described above. The generation of representative errors of the target device was supported by a BBC micro-computer. The communication device was tested informally and the system subject practiced the skills necessary to perform the task of device simulation.

9.4 Evaluation of the device simulation

9.4.1 Device Simulation Performance Results

Purpose: Evaluation of device simulation

Expression: Statement of performance inadequacies and their causes.

9.4.2 Procedure for evaluating device simulations

1. Devise an experiment, in which the critical device parameters identified in Procedure 9.1 are monitored over the range of demands on device operation expected to occur in the experiment planned within the usability evaluation method. See Comments i and ii.

2. Compare the obtained performance with the requirement specified in Procedure 9.1.
   IF the difference is judged not to be noticeable to the intended population of user subjects (see Comment iii), proceed to Step 3
   ELSE apply Procedure 9.5.

3. Obtain a subjective evaluation of the device simulation from a speech technologist.
   IF deemed an acceptable representation of the target device, utilize the device simulation in the usability evaluation method (i.e. proceed to Procedure 7.5)
   ELSE apply Procedure 9.5.

9.4.3 Comment on evaluation of device simulation

(i) The evaluation of simulation performance is important, because unless the device simulation reproduces the behaviour of the target device, the results of the usability assessment risk being incorrect. The strongest assurance of an accurate assessment of the performance of the device simulation will be offered by a controlled experimental evaluation against the criterion of the target device performance specification (see Procedure 9.1). An example of such an experimental evaluation is offered in Appendix C. If time and resources are insufficient to allow a fully controlled experiment, it is strongly recommended that *some* evaluation of simulation performance be made, even if it takes the form of an informal test.

(ii) As stated in Chapter 7, it is assumed that users of SIAM will know how to perform and interpret the results of experiments, so the process of running an experimental evaluation is not proceduralized here.

(iii) The correspondence between the performance of the simulation and that predicted for the target device constitutes the fidelity of the simulation. Fidelity should be such that the behaviour of the simulated device elicits equivalent behaviour in the user subject to that which would be elicited by the target device (see Appendix E).

9.4.4 Example of device simulation evaluation

The performance of the VDC simulation was evaluated informally in a small number of test trials. Although the simulation represented the frequency of the target device's recognition errors adequately, some of the feedback responses were unrepresentatively slow, and the feedback to the user subject did not adequately resemble synthesized speech. These failures were attributed to the following causes:

I/O level behaviour:

(a) simulated synthesized speech was modulated like normal human speech

(b) keys were struck with force, producing keying noise

(c) the system subject moved his fingers to the incorrect row of the numeric keypad (which had a "calculator" layout)

Communication-level behaviour:

(d) the system subject forgot to move the cursor to the next line after keying each entry

(e) the system subject expected to find a key corresponding to every possible (legal) utterance of the user subject, and wasted time searching for keys that were not there

(f) the system subject tended to shadow the user subject's spoken utterances, rather than reading from the visually displayed page of information

Task level behaviour:

Fluency in error correction simulation was poor, due to:

(g) the system subject forgetting which was the current target in the list of "correct" responses (particularly when correcting errors on 8-digit chunk trials)

(h) the system subject having difficulty finding the "correct" character/word (the simulated "next most likely" match) to be communicated to the user subject, within the appropriate line in the list of "correct" responses.

(i) the system subject had difficulty ensuring that, following the correction, a further error had not been automatically inserted according to the stochastic rules of the communication device; (and, if it had, re-correcting it while continuing to communicate with the user subject)

9.5 Optimization of device simulation

9.5.1 Analysis of device simulation behaviour

Purpose: Prescription of ergonomic intervention to optimize performance.

Figure 9.5(a): Attribution of causes of inadequate device simulation performance

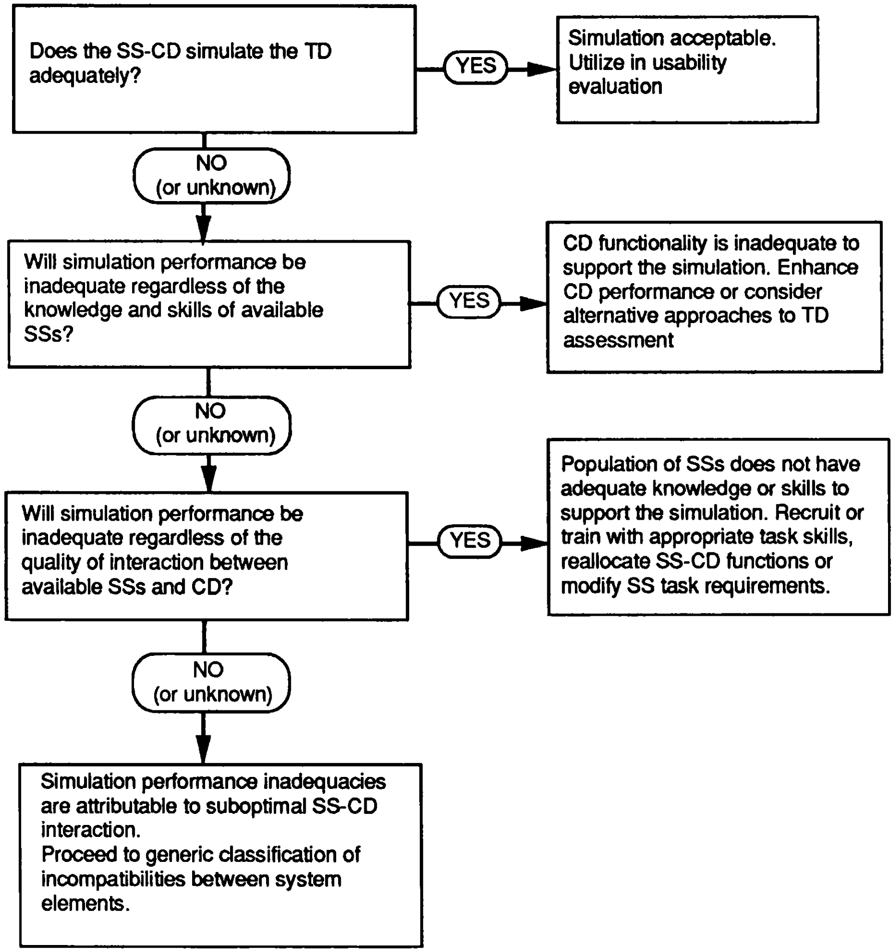
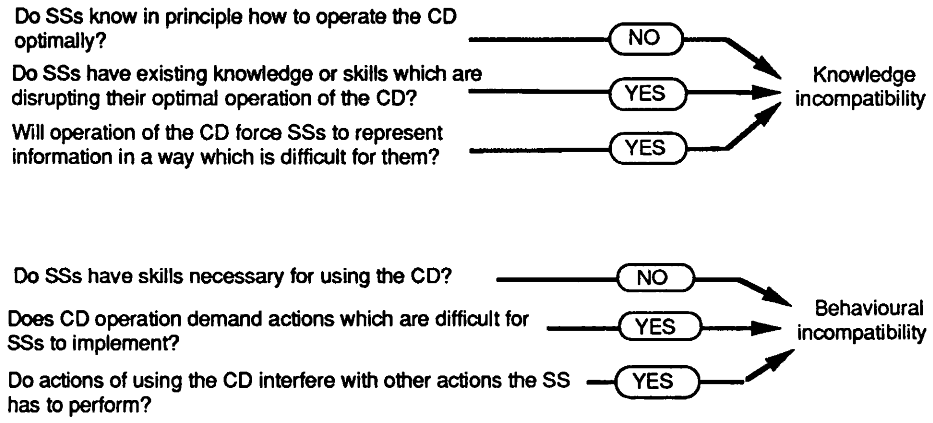

Does the SS-CD simulate the TD adequately?
YES
Simulation acceptable. Utilize in usability evaluation
NO (or unknown)
Will simulation performance be inadequate regardless of the knowledge and skills of available SSs?
YES
CD functionality is inadequate to support the simulation. Enhance CD performance or consider alternative approaches to TD assessment
NO (or unknown)
Will simulation performance be inadequate regardless of the quality of interaction between available SSs and CD?
YES
Population of SSs does not have adequate knowledge or skills to support the simulation. Recruit or train with appropriate task skills, reallocate SS-CD functions or modify SS task requirements.
NO (or unknown)
Simulation performance inadequacies are attributable to suboptimal SS-CD interaction.
Proceed to generic classification of incompatibilities between system elements.

Figure 9.5(b) Generic classification of SS-CD incompatibilities


Do SSs know in principle how to operate the CD optimally?
NO
Do SSs have existing knowledge or skills which are disrupting their optimal operation of the CD?
YES
Will operation of the CD force SSs to represent information in a way which is difficult for them?
YES
Knowledge incompatibility
Do SSs have skills necessary for using the CD?
NO
Does CD operation demand actions which are difficult for SSs to implement?
YES
Do actions of using the CD interfere with other actions the SS has to perform?
YES
Behavioural incompatibility

Expression: Two dimensional matrix:

- class of incompatibility (i.e. knowledge, behaviour or environmental)
- level of incompatibility (i.e. task, communications or I/O level).

9.5.2 Procedure for optimizing device simulation

1. Determine whether inadequate simulation performance is a consequence of system subject-communication device (SS-CD) incompatibility by following the decision tree presented in Figure 9.5(a). IF inadequacies in simulation performance are a consequence of incompatibility proceed to step 2, ELSE consider alternative allocations of function between system subject and communication device by returning to Procedure 9.2.

2. Determine the class(es) of SS-CD incompatibility by following the decision tree presented in Figure 9.5(b). Where the same performance inadequacy is attributable to more than one source of incompatibility (e.g. incompatible representations may be giving rise to incompatible skill demands), then include all potential sources of incompatibility. See Comment i.

3. Prescribe intervention to optimize system performance. Three forms of intervention are possible: *selection* of subjects with more highly developed communication device operating skills; *training* of communication device operating skills; or *aiding* to support SS-CD interaction, e.g. modification of the communication device.

   (a) Knowledge incompatibility:
   - SELECTION:- where the required knowledge is already held by some members of the system subject population, but not all.
   - TRAINING:-where the required knowledge is of a quantity and type readily acquired by the system subject population (e.g. if a learning improvement is evident in the system subject's performance).
   - AIDING:-where the information is of a form which is economically implementable in a usable system subject task aid or communication device modification.

   (b) Behavioural incompatibility:
   - SELECTION:- where appropriate performance dynamics may be exhibited by some members of the system subject population, but not all.
   - TRAINING:-where there is evidence that skills may be developed to achieve the appropriate dynamics within the system subject population.
   - AIDING:-where the dynamic performance may be achieved by an economical implementation of a usable system subject task aid or communication device modification.

Apply the selected prescription to the device simulation specification and return to Procedure 9.3.

9.5.3 Comment on the optimization of the device simulation

(i) The ergonomic intervention appropriate for enhancing simulation performance will vary, depending on the nature of the incompatibility between the system subject and communication device. The same scheme is employed for classifying incompatibility within the simulation system as is assumed in the target system (and, hence, in the organization of the diagnostic tables). Later versions of the method might exploit this equivalence further, by the development of diagnostic tables specifically to support the design of device simulation (see also Section 6.6.2).

9.5.4 Example of device simulation specification

All the behaviours identified in Section 9.4.4 were interaction behaviours, and so could, in principle, be attributed to SS-CD incompatibilities. Diagnosis was inferred by the assessor and prescription made accordingly. Table 9.2 presents sources of incompatibilities and their lowest levels of manifestation. The interaction behaviours to which they relate are indicated by letters corresponding to those used in the list of performance inadequacies in Section 9.4.4.

The following specific prescriptions were made to reduce incompatibility and hence enhance interaction performance.

(a) Task level representational incompatibility was reduced by the assessor producing a list of "correct" entries spoken by the user subject. The assessor and system subject were seated adjacently, so the latter also had access to the list of correct data. The system subject was instructed to tick off each target as it was dealt with, so that a constant tally was maintained of the current state of the task. These interventions should have enabled the system subject quickly to locate the "correct" response corresponding to the machine-generated errors, enabling timely response to the user subject during error-correction routines.

(b) Task-level behavioural incompatibility was reduced by, where possible, automating actions demanded for experimental purposes which were not part of the interaction between system subject and user subject.

It was not possible to override the "error-generating" facility of the communication device, which was counter-productive when the function of the interaction was to correct *errors. This meant that, on occasions, the system subject would have to follow the following sequence:*

- *move cursor to character*

	KNOWLEDGE	BEHAVIOUR
TASK	Correspondence between remembered location in task and the representation of the task presented in the list of "correct" responses (g) (h)	Correspondence between procedure required for interacting with the US and procedures demanded for experimentation (i)
COMMUNICATION	Correspondence between structure of information uttered by the US and device representation of that information (d) (e)	Generation of utterances corresponding to displayed text rather than spoken input from the US (f)
INPUT/ OUTPUT	"Telephone" keypad layout more familiar to the SS than the "calculator" layout (c)	Production of "machine-like" speech sounds (a) Striking keys softly when keing fast (b)

Table 9.2: Analysis of device simulation behaviour

- *speak displayed character*
- *(on hearing user subject confirm the character was an error) consult the table of correct data*
- *read out "correct" character (while simultaneously pressing the appropriate character key)*
- *speak the next displayed character while simultaneously checking that the keypress had successfully corrected the preceding character.*

If the character was still wrong, this would mean having to move the cursor back and try again, while maintaining the communication with the user subject on later characters. Fortunately, the error rate of the target device was low, so this problem did not occur frequently. The only solution in this instance was to ensure that the system subject was highly practiced (trained), so that the system subject's attention could be shared between interaction with the user subject and the requirement to ensure correct data was entered following corrections.

(c) As mentioned above, keys were provided such that, where possible, there was a 1:1 mapping between user subject utterances and system subject keystrokes, e.g. function keys were provided corresponding to node names. This reduced the problem of the system subject having to make a keystroke to most, but not all, user subject responses. However, there was still a requirement for the system subject to use a "down-cursor" command to move between the three fields comprising the target description information and between the fields making up the "chunks" of the grid field in the 2 and 4 item chunk conditions. Such cursor commands did not correspond to user subject utterances. Intensive system subject training was the solution to this.

(d) It was found that there was a natural tendency for the system subject to repeat what he had heard the user subject speak, rather than to read out what was presented on his visual display. This was a phenomenon predicted by Wickens's "Stimulus-Central Processing-Response" hypothesis (e.g. Wickens, Sandry and Vidulich, 1983), which states that, under certain conditions, spoken responses to auditorily presented information are facilitated.

The solution chosen was to train the system subject to suppress the tendency to shadow the user subject's speech and to encourage attention to the visual display, referring to the table of correct data only when an error correction disrupted the system subject's memory for what had been said. This required considerable effort on behalf of the system subject, but the final simulation performance was acceptable.

(e) The tendency for the system subject to expect a "telephone" layout of numeric keys (1,2,3 on the top row), as opposed to the "calculator" layout (7,8,9 on the top row) was predicted by Conrad and Hull (1968). The numeric keypad was, therefore, re-configured in line with the system subject's expectations.

(f) The inadequate fidelity caused by the system subject not speaking in the crudely modulated tones of a synthesizer, and of the system subject striking the keys noisily, were judged to be a result of the system subject not being able to divert attention from other aspects of his task.

The latter problem was eased by moving the system subject's microphone as close to the mouth and as far from the keys as possible. However, the primary solution was to reduce the competing task demands by the interventions already described, and by subjecting the system subject to extended practice. In this way, "speaking synthetically" and pressing the keys lightly became "automated" skills for the system subject. The system subject's final simulation performance was acceptable. Following these ergonomic interventions, the simulation underwent a further informal evaluation and was judged suitable for the purposes of the usability evaluation.

9.6 Critical commentary on the device simulation method

The human simulation technique which is used in the device simulation method offers a powerful means of reproducing many of the attributes of speech interfaces. The method further takes account of the need to develop device simulations exhibiting fidelity appropriate to the needs of evaluation - a weakness in previous reported applications of the technique. However, even given technological support (e.g. by means of aids incorporated in the system subject's communication device), there are limits to the attributes which are reproducible by human simulators and, hence, to the range of device types which may be so simulated. For example, human response times may be lower than those of computers, and system subjects may exhibit behaviours (such as typing errors) which are uncharacteristic of speech I/O devices.

The method sets, as a criterion for adequate fidelity, simulated device behaviour which elicits equivalent user behaviour to that which would be elicited by the target device. Unfortunately, when the target device has yet to be implemented, equivalence in behaviour may be impossible to determine. In such cases, the adequacy of fidelity is left to the judgement of the assessor. In general, however, the method employs experimental evaluation as a means of determining the adequacy of the device simulation. Experimentation at least offers a systematic means of collecting data and of comparing actual simulation performance against that desired. However, it is recognized that, in practice, formal experimentation may be regarded as expensive in resources and that less-rigorous testing is likely to be used (at some cost to the effectiveness of the method).

The device simulation method assumes a hierarchical decomposition of device behaviour, which is complementary to that applied to target tasks by the use of the task simulation method. While such a detailed representation is well-suited to device simulations in which the dialogue is deterministic (i.e. where it is possible to predict unambiguously the state of the device following a new input, given information on its existing state), it may be less appropriate if the device is capable of complex language interpretation. Such devices would have the potential to utilize grammatical context in interpreting entered data; i.e. the current state of the device may be a complex function of previously entered words. In such situations, it may be possible to specify the task of the system subject only at a relatively high level, rather than the low level assumed by the method.

CHAPTER 10

USER SIMULATION METHOD

10.0 Process of user simulation - summary

The process of user simulation development requires the generation of four representations of target users (see Figure 10.1).

1. A *task knowledge description* is developed, documenting user knowledge pertaining to the application domain of the target system and that pertaining to operation of the device. Within the descriptions are identified the subsets of the users' domain and device knowledge which would be necessary to perform the simulated task using the simulated device.

2. Only those attributes of target users which determine device-user interaction need be represented accurately in the simulation. The interaction model embodied in the configured diagnostic tables and the description of user knowledge are intersected to generate a *user subject model*. The model identifies critical attributes of target users for the purpose of their simulation.

3. The user subject model is compared with the available sources of user subjects, to specify requirements for the selection and training of subjects. A specification of this selection and training constitutes the *user subject development programme.*

4. The *user simulation* is implemented by exploiting the user subject development programme in the selection and training of subjects, and in monitoring their behaviour during the evaluation. The objective is to match the behaviour of possible subjects with that specified by the user subject model, in the context of the simulated task.

10.1 Characterizing knowledge of the target task held by users

10.1.1 Description of target task knowledge

Purpose: Specification of the knowledge and skills relevant to the experimental task intended to be held by target users.

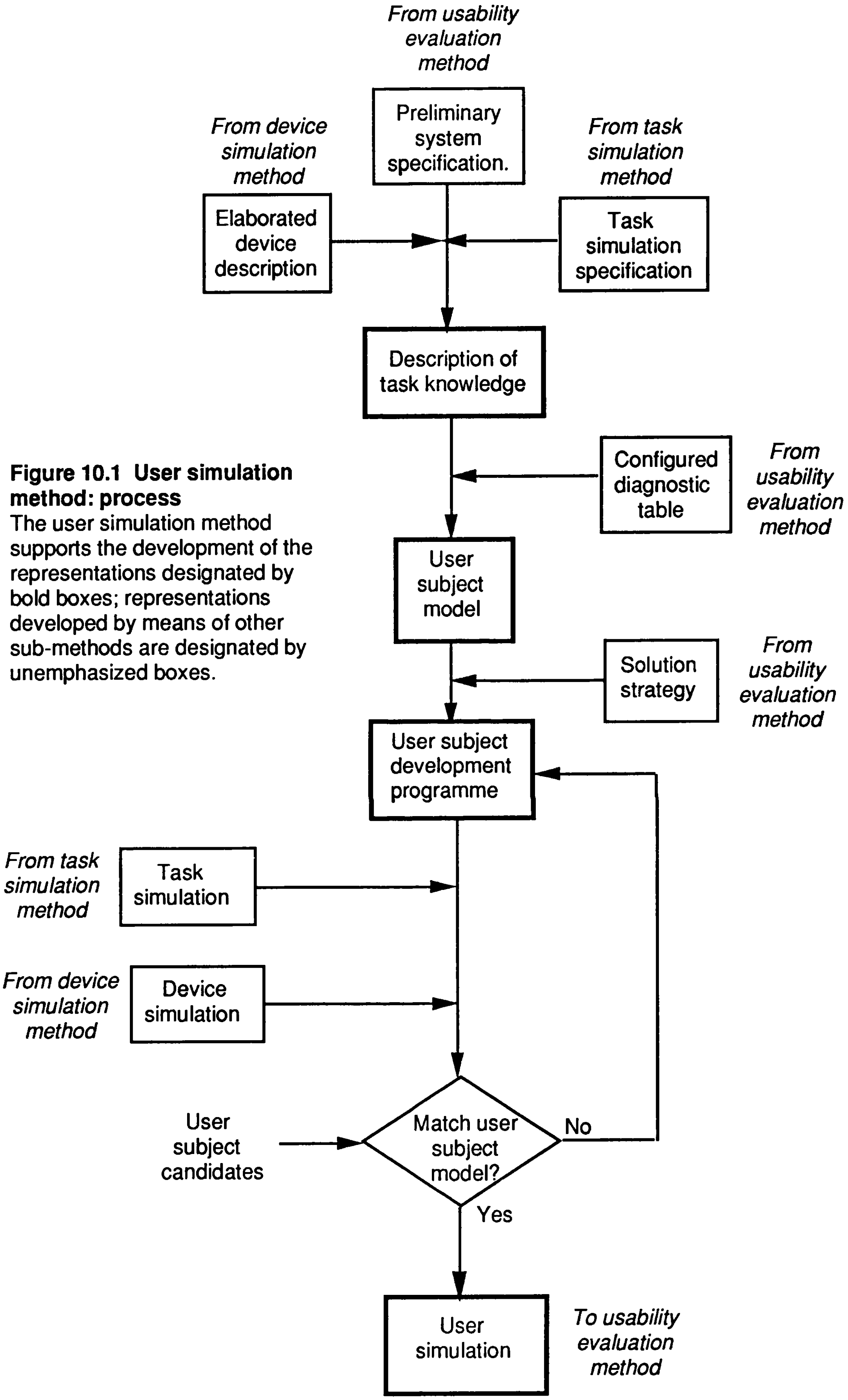


Figure 10.1 User simulation method: process
The user simulation method supports the development of the representations designated by bold boxes; representations developed by means of other sub-methods are designated by unemphasized boxes.

From usability evaluation method
Preliminary system specification.
From device simulation method
Elaborated device description
From task simulation method
Task simulation specification
Description of task knowledge
Configured diagnostic table
From usability evaluation method
User subject model
Solution strategy
From usability evaluation method
User subject development programme
From task simulation method
Task simulation
From device simulation method
Device simulation
User subject candidates
Match user subject model?
No
Yes
User simulation
To usability evaluation method

Expression: Two informal descriptions:

(a) Domain knowledge description, including:

- classification of users with respect to the military hierarchy (i.e. rank and area of technical specialization)
- specification of knowledge and skill expected of target users in the performance of non device actions included in the task model (or their current equivalents)

(b) Device knowledge description, including:

- identification of expected experience of target users with related information systems
- specification of plans for selection and training of users in the operation of the target device.

10.1.2 Procedure for specifying target task knowledge

1. Identify domain knowledge which is expected to be held by the target users (i.e. army knowledge, as opposed to knowledge about the use of the target device). Proceed by elaborating the user description included in the preliminary system specification, as follows.

   (a) In collaboration with the target device project team, identify users within the army organization. This process should identify both rank and area of specialization (e.g. regiment and any specialization within the regiment).

   (b) Use available MoD and army information sources (including specialists from the likely user group) to determine the knowledge and skills necessary to perform the task at present.

   (c) In collaboration with the target device project team, specify the subset of this domain knowledge and any non-device skills necessary to perform the simulated task as specified in Procedure 8.6. See Comment i.

2. Identify device operating knowledge held by target users.

   (a) In collaboration with the target device project team and specialists from the likely user group, identify other information systems (automated or otherwise) to which target users will have been exposed. These may include computer systems, but also command and control systems implemented over unstructured communication media such as radio.

   (b) Review plans for selection and training of users which will modify their existing device operating knowledge. Specify (informally) the knowledge and skills users are expected to possess which will influence how they use the target device as it is described in the elaborated device specification. This specification will be restricted, then, to skills for operating those device functions which are necessary to perform the simulated task.

10.1.3 Comments on the specification of target task knowledge

(i) Note that the objective of the user simulation method is to reproduce the behaviour of users in the context of the *simulated* task. The latter requires only a subset of the knowledge required to perform the task under real battlefield conditions; i.e. it is not necessary to include, in the description, knowledge supporting actions which have been deleted from the task description when generating the future task model (in Procedure 8.5).

10.1.4 Example of specification of target task knowledge

A description of the likely population of users of the VDC was developed with the co-operation of an army authority. This comprised specifications of the knowledge users would possess with respect to the observation task and their expected knowledge of device operation. The intended users were army officers, who could be assumed to be highly knowledgeable about the task of battlefield observation and highly trained in the operation of their equipment. It was assumed that these users would understand the functionality of the communication system of which they and their target device were a part, and would have had extensive practice at using the device in an operational context. They were assumed, then, to exhibit operating competence at the levels of task, communication exchanges and data input/output.

10.2 Specifying a user subject model

10.2.1 User subject model

Purpose: Specification of knowledge and skills to be held by user subjects.

Expression: List of critical user attributes and description of target users with respect to this list.

10.2.2 Procedure for specifying a user subject model

1. Identify critical user attributes in column 4 of the configured diagnostic table (see Procedure 7.3)

2. Specify target user attributes. Review the description of task knowledge with respect to the critical attributes, and so specify which aspects of user knowledge and skill are critical to the assessment. See Comment i.

10.2.3 Comments on the user subject model

(i) The specification should identify the level of description at which subjects must be representative with respect to critical attributes. For example, representativeness with respect to knowledge of device operation might be at the level of the task - say, knowledge of the high level functions of which the device is capable; and/or at the level of communication exchanges - say, knowledge of the content and structure of the dialogue; and/or at the I/O level - say, knowledge of the importance of generating consistent speech tokens to a recognizer.

10.2.4 Example of user subject model

The requirement for user subjects was that they be representative of the target users with respect to those aspects identified as critical by the configuration of the diagnostic manual and the task model. The diagnostic manual indicated that they be representative of users with respect to their familiarity with task data. However, the levels of task of concern to the present study were those of device-user communication exchanges, and data input/output. Furthermore, representativeness only needed to extend to that aspect of the task involving communication of task data.

The requirement for subjects, then, was that they be representative with respect to the operation of the device for entering target data, i.e. highly competent at the levels of communication exchanges and data input/output. There was a potential interaction between domain knowledge and device operation which was not addressed by this experiment. This related to conflict between familiarity with the form of grid references (e.g. an 8-digit reference might be represented familiarly as two 4-digit groups) and the constraints on chunking imposed by the device (e.g. 2, 4 or 8 digit groups). The intensive training envisaged for target users in device operation was such as to make this interaction probably insignificant for performance.

10.3 Specifying a user subject development programme

10.3.1 User subject development programme

Purpose: Specification of requirements for user selection and training.

Expression: Informal descriptions of subject population and selection and training programmes.

10.3.2 Procedure for specifying a user subject development programme

1. Specify the number of user subjects demanded by the solution strategy (from Procedure 7.4).

2. Identify potential sources of user subjects. The requisite number must be available over the duration of the study, and, given training, must meet the requirements of the user subject model.

3. Specify selection criteria. Given the potential subject population and the envisaged subject training programme (see step 4, below), state the minimum requirements for relevant skills and knowledge to be met by user subject candidates. These might be qualitative requirements (e.g. ability to calculate map references) or quantitative requirements (e.g. ability successfully to engage three targets in five minutes).

4. Specify training programme. Specify instructions and practice exercises which will enable user subject candidates with the minimum skills identified in Step 2 to attain the knowledge and skills required by the user subject model. Ideally, training might take the form of preliminary verbal instructions, providing background knowledge of the task and device; demonstrations of individual task elements; practice of individual task elements; and, finally, practice on a fully-implemented task simulation operating in conjunction with a device simulation.

10.3.3 Comments on user subject development programme

(i) The subject group should be as homogeneous as possible with respect to the critical attributes. As a rule of thumb, try first to recruit army subjects from the target user group. If this is not possible, select alternative sources on the principle of minimizing training requirements.

10.3.4 Example of user subject development programme

Subjects were to be selected according to the criterion of their potential for acquiring skills necessary for performing the simulated task with the requisite degree of facility. User subjects were required who possessed, or who had the potential for developing, high levels of competence for data entry using speech interfaces. They would require training and practice in the operation of the three specific speech interface implementations to be utilized in the experimental task. In particular, they would need to develop facility in the use of the data entry, verification and editing functions of the device.

Given that subjects would be familiar with speech input/output devices, it was necessary to provide additional information on the function of the VDC in support of target engagement, and to instruct them in the particular procedures for its use in the experimental task.

Following a 15-minute period of oral instruction, subjects were to practice the task until they were judged to exhibit competence in device operation (i.e. fluent performance). Before each experimental condition, the subjects would send two messages to familiarize them with the particular device characteristics for that trial.

10.4 Implementation of the user simulation

10.4.1 User simulation

Purpose: Development of user subject pool

Expression: Group of competent user subjects

10.4.2 Procedure for implementing a user simulation

1. Implement the selection and training procedures specified in Procedure 10.3. Continue training until each subject consistently reaches criterial performance, or until it is recognized that the subject is incapable of reaching the criteria (in which case the subject should be rejected).

2. In the course of experimentation, check that criterial performance is being maintained, and, if necessary, provide opportunities for additional practice under each of the various conditions of the experiment.

3. Return to Procedure 7.5. of the usability evaluation method.

10.4.3 Comments on user simulation

None

10.4.4 Example of user simulation

Two subjects (students) were selected and trained according to the specification presented above.

10.5 Critical commentary on the user simulation method

The user simulation method is less completely specified than the other sub-methods of SIAM (see Section 6.7.3). Although fuller specification of the method would have been desirable, the method as expressed at least seeks to ensure that the representativeness of subjects (as simulations of target users) is given due consideration by the assessor. SIAM recognizes, then, that the validity of conclusions drawn from studies involving subjects selected from a population other than that of the target users must generally be open to some question.

CHAPTER 11

OPERATIONAL TRIAL OF SIAM

11.1 Overview

This chapter describes an evaluation of SIAM, intended to demonstrate its operation in the context of procurement and to identify potential inadequacies in its design. The chapter begins with a brief review of research addressing the evaluation of structured analysis and design methods (SADMs). It is concluded that previous studies have primarily addressed the suitability of methods for representing certain classes of system, rather than the adequacy of their support for the general practice of systems analysis and design.

The rationale for the trial is expressed in terms proposed by Dowell and Long in their conception of human-computer interaction engineering (see Chapter 4). The trial sets out to determine the impact of SIAM on the *quality* of a user interface assessment and the *costs* to the assessor in evaluating a user interface. In the trial, an assessor uses SIAM to evaluate a prototype interface for a computer to support a battlefield observation and communication task; SIAM is subsequently used to assess the performance of an enhanced version of this device. The results are discussed with respect to the support provided by SIAM to the assessor.

The evaluation serves to test the application of SIAM and so to identify its strengths and weaknesses. The results are used to suggest enhancements to SIAM, but also, more generally, as a basis for discussing the scope of the potential contribution of structured evaluation methods to procurement (Chapter 12).

11.2 Rationale for the evaluation of SIAM

11.2.1 A task based framework for evaluating procurement methods

Parallels may be drawn between work performed by systems in which people are supported by computers and work performed by systems in which people are supported by methods. Both computers and methods are intended to enhance task performance. Just as the concern of the computer designer is to design a system to deliver desired task quality with acceptable cost to the computer user, so a concern of the method developer should be the design of a system (a method and its user) to deliver desired task quality with acceptable costs to the *method user*. In view of this similarity, Dowell and Long's (1989) conception of the human computer work system is adaptable as the basis of a framework for the evaluation of a procurement method such as SIAM.

The class of task which SIAM seeks to support is the evaluation of speech interfaces. Assessment is to be performed to some desired level of quality with acceptable utilization of resources. The achievement of the assessment task goal demands not only behaviours directly to assess the feasibility and usability of target systems, but also behaviours *enabling* assessment, such as activity planning and the communication of information within the procurement system. SIAM improves performance by providing engineers with knowledge to support both types of behaviour.

If SIAM is to be considered effective, impacts should be observable on the behaviour of procurers and on the products of their behaviour. The products would be expected to exhibit enhanced quality, with respect to completeness, precision, accuracy and speed of delivery. The behaviour to achieve the desired quality should also impose fewer costs on the engineer, which might be inferred given observations of more efficient behaviour (systematic and error free) and/or reports of lower demands for mental or physical resources.

11.2.2 Previous work

SIAM is a structured evaluation method. Few, if any, such methods have been developed previously, and there have been no attempts to evaluate methods of this type. However, SADMs share important features with SIAM, most particularly in the explicitness of their scope, notation and proceduralization (see Chapter 4). This similarity supports SIAM's claim to share the benefits of SADMs, such as assurance of product quality and support for project activities, such as communication and planning (see Walsh et al, 1989). In view of these shared features and claims, previous evaluations of SADMs potentially offer a starting point for an evaluation of SIAM.

In fact, as Bubenko (1986) has observed, there have been few systematic evaluations of SADMs. Those that have been performed have not addressed directly the quality of the final products (i.e. systems developed using methods), and the criteria they have used have been somewhat indirectly concerned with costs to users (system designers). An analytic study by Madison et al (1983) compared nine SADMs with respect to a large number of features, related to factors such as the coverage of the stages of the design process, proceduralization, notation, hardware/software support and appropriateness for the representation of system entities. These features determined the general suitability of the methods for application in the development of specified classes of system, but the results of the study could not be used to predict performance of specific system development tasks.

Floyd (1986) compared five SADMs by using them in the design of a (hypothetical) library system. Again, the quality of the final system was not of primary concern and is not reported; (indeed, it is not clear that the specifications were actually implemented). The costs to users are described informally, as reports of particular problems encountered in the use of each method in developing and manipulating representations prescribed for the development of the system.

The studies cited above do not enable the evaluation of SADMs with respect to their delivery of the benefits claimed by Walsh et al. Rather, they set out to provide the reader with a framework to reveal the qualitative differences between methods. In the case of Madison et al this is an analytic framework, and in the case of Floyd it is a common case study. The differences are expressed with respect to the production of system representations rather than to the production of working systems.

The contribution of SADMs to system development at a project level is alluded to in a paper by Jones (1989). Jones considers the work of eight different students in four successive academic years, which contributed to the development of an information system for a small manufacturing company. The study is unusual because it reports on the use of structured methods in a business environment. Although the work described was not intended primarily as an evaluation of methods - its main function was as a teaching exercise - Jones reports favourably on their contribution to the success of the students' work. Specifically, he concludes that the methods enabled the the assimilation of information from previous projects and from the customer company; they enabled students quickly to gain high and low level views of the system; they facilitated the scheduling of work; they contributed to the consistency (and high quality) of the design output; and they facilitated communication between individuals with little or no previous development project experience. There are clearly parallels between the positive outcomes described by Jones and the benefits of SADMs listed by Walsh et al. However, Jones did not attempt to make a formal evaluation of a particular method. The students used a variety of techniques and tools, and there is no evidence to indicate how they might have fared *without* structured methods.

To summarise, evaluations of SADMs have tended to concentrate on the adequacy of their notations for characterizing systems, and they have had classificatory objectives. Such studies are important, in that they are necessary to enable practitioners to select methods appropriate for specific applications. However, there have been few (if any) attempts to evaluate the procedures offered by SADMs, and it would appear that none have assessed with respect to the class of benefit identified by Walsh et al: there has been little or no systematic assessment of the adequacy of the support of SADMs for the general practice of systems analysis and design.

There are likely to be a number of reasons for this. For example, it is difficult to perform evaluations in the context of real system development projects, particularly if they are of a large scale. Such projects are typically of long duration, involve distributed activities and are conducted idiosyncratically. Difficulties also arise in the specification of appropriate metrics for performance measurement and in setting performance criteria as a basis for evaluation. Unfortunately, then, the literature relating to SADMs does not offer appropriate models for evaluating structured methods. An evaluation of SIAM with respect to its support for task performance would constitute a novel enterprise.

11.2.3 Design of the evaluation

Objectives. An evaluation of SIAM might be performed to serve a variety of purposes; for example, it might seek to demonstrate the truth of the claims of the method (validation); to prove that, when applied, the procedures result in the development of the specified products (verification); or, less formally, to demonstrate that the method has utility and to identify requirements for enhancements. It was recognized that the development of SIAM, as reported so far, was not complete. Although the method had been applied by its developers (see the experiment used for illustration in Chapters 7-10), application had not been attempted by an independent practitioner. At the extant stage of development, then, the requirement was primarily for a demonstration of utility and identification of problems in use.

General method. Validation studies require that factors influencing the behaviour of concern to the investigator are fully controlled. While an experimental test of validity may offer powerful evidence for or against a scientific claim, the requirements for experimental control risk rendering the circumstances of the test unrepresentative of conditions outside the laboratory. In view of this, tightly controlled experimentation was not regarded as an appropriate strategy for testing SIAM in its existing state. A preferable strategy was to evaluate the method in a realistic context, which would expose problems likely to be encountered in practice. The context chosen was a system development project involving an HF practitioner.

Measuring performance. The most direct indicator of the performance of the assessment task supported by SIAM would have been the accuracy with which procurers decided the suitability of speech I/O for battlefield computer systems. If SIAM were effective, one would expect to see evidence of more successful systems in use by the army. However, measuring performance in this way was not practical within the timescale of the present project. An alternative approach to measurement was for the assessor to evaluate subjectively the quality of the assessment when the method was used. Similarly, assessors could report on the extent to which they incurred resource costs when running an assessment.

The SIAM evaluation study. Subjective evaluations were collected from a single subject of the performance of the assessment system supported by the method. Performance related to task quality and user costs, including aspects of quality pertaining to the benefits identified by Walsh et al. In addition to the performance measures, the subject's behaviour was compared with that prescribed by the method. Deviations were interpreted either as errors (indicating mistakes by the subject or the inadequacy of the method in prescribing appropriate behaviour); or as disagreement between the subject's opinion of the optimal procedure and that recommended by the method. The behavioural information provided a means of interpreting performance data and, subsequently, for identifying requirements for enhancements to the method.

Because the study involved only one subject, inferential statistics were not performed. It was accepted that the results of the study could only be taken as indicative, and that it would be desirable for controlled experiments to be carried out at a later stage.

11.2.4 Context

The trial was conducted in the context of the VDC demonstrator project used previously in this thesis to illustrate the application of SIAM. The VDC was recognized by both RSRE and RMCS as being somewhat frustrating to use, and this opinion had been confirmed by the preliminary evaluation performed during the development of SIAM (see the results presented in Sections 7.6.4 and 7.7.4). The poor usability was attributed to the design of the user interface dialogue.

Neither RSRE nor RMCS possessed the expertise in human factors necessary for enhancing the device in the time available to the project. RSRE and RMCS approached a human factors consultancy - London HCI Centre (LHC) - to perform a full HF evaluation of the first prototype and to propose enhancements. SIAM was to be used to structure and support the study. The study was to consist of two phases: Phase 1, in which the existing prototype was to be evaluated; and Phase 2, in which interface enhancements proposed by LHC would be evaluated in a system simulation.

The work of this study did not correspond precisely to the class of activity which SIAM had been developed to support. Notably, SIAM assumes a requirement for assessment before implementation, whereas in this instance a prototype device existed. In addition, SIAM is intended to assess feasibility and technical risk, rather than to specify details of an interface. Furthermore, SIAM assumes little or no human factors knowledge in the assessor, whereas this assessment was to be carried out by an organization with human factors expertise (albeit with no previous experience of speech technology).

Nevertheless, the method claims to be adaptable, and many of the features of the proposed study were shared by feasibility assessments; for example, the study demanded system simulation (to evaluate options for interface enhancements), and it would clearly involve performance evaluation and behavioural diagnosis. The project was, therefore, viewed as an appropriate vehicle for a trial, although it was recognized that the results would require appropriate interpretation in the light of these discrepancies.

11.3 Evaluation of SIAM - Method

11.3.1 Techniques for data collection

The study was to be performed in the context of a commercial project, so it was not possible for data collection to interfere with the normal conduct of the work. In view of the complexity of the phenomena under observation, data were collected by means of interviews. The present author ("experimenter") had a subsidiary role[1] in the interface assessment, and so he was in a good position to discuss the behaviour of the assessor (subject) when she deviated from the method. Data relating to the subject's knowledge were collected in pre-planned interviews, but these were conducted in such a way as to enable the subject to express herself freely. These semi-structured interviews were tape recorded, analysed and subsequently summarised.

The evaluation of the subject's task behaviour and performance utilized data collected by means of interviews structured around a questionnaire, completed when the subject had finished each of the stages of SIAM. The questionnaire (see Appendix F.1) comprised five sections. Section A was concerned with whether or not the subject had generated the representation in question, and, if so, with its correspondence to the structure prescribed by SIAM. Section B sought the subject's opinion of the *quality* of the representation, as this related to the primary output of the task (i.e. the assessment of the target device). It specifically investigated the completeness, level of description and accuracy of the representation for its purpose in the assessment: it was intended to address the first of Walsh et al's benefits of structured methods, that is, the production of (product) quality.

Section C was also concerned with the quality of the representation, but here as it related to its contribution to project organization, that is:

[1] Although the experimenter was available as a consultant with respect to the use of SIAM, he only intervened when requested by the subject. In practice, his involvement related to the implementation of simulations specified by the subject on the experimental testbed, and support for the subject in proposing enhancements to the demonstrator. In the first case, he acted as a system subject and as a technician implementing the simulation according to the subject's specification; in the second, he acted as a "sounding board" for her proposals for interface modifications. His role was, therefore, subordinate, and particular care was taken to avoid influencing the subject's behaviour.

- to the subject's understanding of the problem under investigation
- to the subject's subsequent planning of the project
- to the subject's communication within the project
- to support for the subject's subsequent actions.

Section D of the questionnaire addressed the subject's assessment of the mental *costs* of producing the representation and the extent to which SIAM supported the process. The final section (E) was concerned with the subject's task behaviour, identifying non-conformities with the procedure prescribed by SIAM and identifying errors made by the subject (i.e. behaviour which did not have the desired consequences).

11.3.2 Procedure

Three sets of data were collected:

- a preliminary assessment of the subject's knowledge of device assessment
- a selective evaluation of the contribution of the output of the previous study performed using SIAM
- evaluations of behaviour and performance while using the method in the device assessment task

The procedures used for the collection of these data are now described.

(a) *Preliminary assessment of subject knowledge.* The subject's existing knowledge was expected to be a significant factor influencing deviations from the procedures of the method, either due to rational disagreement or due to inadequate ability. Also, the subject's professional experience would determine her criteria for her subjective evaluation of the method. The interpretation of behaviour during the assessment therefore required consideration of the knowledge brought to the task by the subject.

The evaluation of the subject's knowledge was performed by means of semi-structured interviews before and after initial exposure to the method. The intention was to identify differences in the view held by the subject and the view advanced by the method; to determine whether the subject's view was changed following exposure to the method; and to identify where (if at all) the subject actually disagreed with the method.

(b) *Evaluation of previous output of SIAM.* A potential advantage of structured methods lies in the portability of information between separated phases of long-duration projects (e.g. Jones, 1989). An opportunity therefore existed to evaluate SIAM with respect to the carry-over from the previous study of information relevant to the proposed LHC work. Following exposure to the method, the subject read the report of the earlier study (Life and Lee, 1989). Three of the representations were judged to be particularly relevant to the present assessment:

- expanded task description
- SS task model
- user descriptions

The subject evaluated the quality of each with respect to its expected contribution to the work of the present assessment. This was done in the context of an interview structured around Section C of the evaluation questionnaire. Finally, the subject was asked her opinion of the value of structured methods in general and of the value of SIAM in the context of the RMCS work.

(c) *Task behaviour and performance.* The procedure used for the collection of data was intended to interfere minimally with the subject when she was carrying out the work. The experimenter and subject completed an evaluation questionnaire at the time, or shortly after, each stage of the method had been completed. On the two occasions on which the subject deviated substantially from the method, an extended interview was conducted to determine the reason and to determine how the subject intended to surmount the problem. Within the constraint imposed by the sponsor (that the method should be used to conduct the assessment), the subject was free to perform her activities in the order she chose, and to judge when a deviation from the prescribed method would enhance task performance.

11.4 Evaluation of SIAM - Results

11.4.1 Preliminary assessment

The interviews recorded prior to and following exposure to SIAM were analysed, and the outputs compared. The detailed results are presented in Appendix F.2. The subject exhibited more potential competence as an ergonomist than is assumed for users of the method. Her approach to ergonomic evaluation was compatible with that of the method, in that she recognized the potential value of simulation to support an empirical approach. She also recognized the terminology used in the expression of the method, and accepted the conceptualization based upon distinctions between device, user and task.

Later fundamental disagreement with the method was not expected. The subject recognized the potential contribution of the method in ensuring completeness, and she accepted the value of explicit intermediate representations, at least for non-expert users of the method. However, her own competence would subsequently enable the subject to evaluate the contribution of the method's procedures and to deviate in a controlled manner. Her experience was expected to provide relevant criteria for the subjective evaluation of the method.

11.4.2 Evaluation of previous output of SIAM

Table 11.1 presents the subject's responses to questions in Part C of the questionnaire, relating to the support for the present study of four of the representations produced during the previous assessment of the VDC.

Table 11.1 Evaluation of representations developed previously using SIAM
Ratings on a scale of 0 - 5 for questionnaire items C1, C2 and C3 (see Appendix F.1)

	Expanded task description	System subject task model	User description
Question C1: Contribution to understanding of the assessment problem	4	4	3
Question C2: Contribution to project planning	3	4	3
Question C3: Contribution to communication	-	4.5	-

The discussions concerning each representation are now summarised.

(a) *Expanded task description.* The device developer had already given the subject an overview of the target task, but the subject felt that the method output provided information at a much lower level of description; for example, it provided information on the sequences of actions. This had given her a more complete understanding of the problem. She intended to take the representation as a starting point for a revised task description, using it as a basis for discussion with RMCS on her assumptions about the task.

(b) *System subject task model.* The representation increased the subject's understanding of the problem of implementing a device simulation, by providing an insight into the necessity of considering the details of the system subject's activities. She expected to use it as a model for developing an equivalent diagram for the new simulation. She expected that the representation would be important for explaining the task to the system subject.

(c) *User descriptions.* These were derived from the device developer's description of the users, which was the same as those given to the subject. This reduced the effective contribution of the descriptions to the subject's knowledge. However, she reported that they had value in confirming what she had been told. She found it interesting (and probably useful) to distinguish domain and device knowledge, she expected the representation to be useful as a model for the development of future user descriptions (e.g. in demonstrating the level of detail appropriate when using the method).

In general, the subject rated highly the quality of the contribution of the previous output to the organization of the new work. Following her exposure to SIAM and to its previous output, the subject expressed the following opinions regarding the advantages and disadvantages of structured methods.

Advantages:

- they ensure completeness
- they encourage a systematic approach (by emphasizing planning)

Disadvantages:

- they do not ensure adequate implementation of the procedures
- they might result in superficial analysis, because the assessor does not feel it necessary to think for him/herself
- they could encourage wordiness without content
- they might tempt assessors to force problems into an inappropriate mould
- might be more appropriate for some classes of problem rather than others

When asked her opinion of the likely contribution of SIAM to the proposed assessment task, the subject responded as follows:

Advantages:

- previous use of method had given a good appreciation of the usability problems of the device
- SIAM seemed appropriate for some of the classes of problem previously identified (i.e. dialogue based)

Disadvantages:

- some problems identified by RMCS would be difficult to study experimentally
- it was difficult to run human simulations outside the laboratory
- naive users would not be able to distinguish which elements of the method were important in the context of a specific evaluation (could waste a lot of time developing unimportant representations).

Table 11.2 **Questionnaire responses - Usability evaluation method (Phase 1)**
Ratings on a scale of 0 - 5 (see Appendix F.1)
Comments of the subject are reported in Appendix F.3.1 (comments refer to items marked in bold)

	Prel. prob spec.	Prel. syst. spec.	Diag. table conf.	Soln. strat.	Expt. con-text	Data	Anal. of int'n	Feas. rept.
A. Conformity with SIAM Is rep. necessary?	y	y	y	y	y	y	-	-
Was the rep. developed?	y	y	n	y	y	y	-	-
Implicit or explicit	exp	imp	-	exp	exp	exp	-	-
As prescribed?	n	y	-	n	**n**	y	-	-
B. Product quality Does the rep. - include relevant classes of info?	y	y	-	**n**	**y**	y	-	-
- show correct level of description?	-	y	-	**n**	**y**	y	-	-
- characterize what it is supposed to?	y	y	-	**n**	**y**	y	-	-
C. Support for project Helped understanding?	3	n/a	-	**3**	**3**	5	-	-
Helped planning?	4	3	-	4	-	5	-	-
Helped communication?	y3	n	-	n	y4	y4	-	-
Enabled next procedure?	y	-	-	y	y	y	-	-
D. Assessor costs Mental effort to develop rep.?	**2**	-	0	**2**	**2**	**2**	-	-
Support from SIAM?	4	-	**2**	3	**3**	**2**	-	-
E. Assessor behaviour All steps performed?	y	-	n	n	y	n	-	-
Steps performed differently?	n	-	y	n	**y**	y	-	-
Mistakes?	n	-	-	-	n	n	-	-

11.4.3 Assessment behaviour and performance

The results are organized with respect to the two phases of evaluation work: evaluation of the existing prototype and evaluation of the enhanced interface. In each phase, the results of the evaluation of each of the four sub-methods comprising SIAM are presented; firstly, as a table of the subject's questionnaire responses; and, secondly, as a commentary on her responses.

(a) Assessment behaviour and performance: Phase 1 (Device prototype)

(i) Usability evaluation method

The responses given by the subject to the questionnaire are presented in full in Appendix F.3.1, the basic data being summarised in Table 11.2.

Conformity with SIAM's representational structure. The subject generated five of the eight representations prescribed by SIAM. The ones not completed were the analysis of the interaction and the feasibility report (neither of which were appropriate in Phase 1 because it constituted only a preliminary investigation); and the configuration of the diagnostics (see below). The representations which were produced were explicit, except for the preliminary system specification. The form of this differed because the device under investigation actually existed as the prototype, and task and user specifications had already been produced in the earlier investigation. The data generated in the device evaluation experiment also deviated from that specified by SIAM by being less formally expressed.

In summary, deviations from SIAM's representational structure occurred, but most were attributable to the fact that Phase 1 of the study was not a feasibility assessment but an exploratory prototype evaluation. This factor contributed to the failure of the subject to configure the diagnostics; however, the fact that the subject was unable to adapt them for her purposes suggests an inadequacy in the method.

Product quality. The subject's comments on the quality of her representations generally indicated that she believed them to be adequate for the assessment. Ignoring her failure to configure the diagnostics, she only expressed concern over the quality of her solution strategy. She was uncertain whether her strategy had been specified in sufficient detail and, because her study was exploratory, she had felt it inappropriate to specify hypotheses as prescribed by SIAM.

Quality of support for project organization. All of the representations which were developed were judged by the subject to have contributed positively to her understanding of the issues under investigation, to her planning of subsequent activities and (where necessary) to her communication with others involved in the project. They were also judged to provide adequate foundations for the further development of representations.

Where the method adequately supported the development of representations, these contributed positively to the conduct of the project; however, the failure to configure the diagnostic table required the subject to recruit her own knowledge of human factors and to develop her own (implicit) model of device-user interaction. This enabled subsequent successful progress, but it resulted in her subsequently being forced to deviate from SIAM's procedures and to incur substantial assessor costs.

Assessor costs. The subject reported that the development of representations for usability evaluation required substantial effort, and the successful configuration of the diagnostic manual was found to be impossible. SIAM's procedures were judged not to be helpful in this, nor in the collection of experimental data[2]. In the latter case, this was because the procedures were not detailed; however, the experience of the subject enabled her to proceed successfully. SIAM's procedures contributed positively in the preliminary specification of the assessment problem, of the solution strategy and of the experimental context.

Assessor behaviour. The subject reported deviations from SIAM'S procedure in the development of all representations except for the preliminary problem specification. These deviations were attributed to the following reasons:

- the present study was qualitatively different from that assumed by SIAM
- the diagnostics were not configured successfully
- the subject preferred to carry out prescribed steps in a different order.

In view of the importance of the diagnostics in the application of the method, an extended interview was conducted with the subject to identify causes of the problem. A transcript of the interview is presented in Appendix F.3.2. The main problems identified were as follows.

- The process of selecting diagnostics relevant to the problem under investigation was inappropriate when the objective of the study had been to *identify* potential problems. In principle, any or all of the diagnostics were potentially relevant.

- Although "general purpose diagnostics" had potentially offered a solution to the problem above, the potential was not fulfilled because these diagnostics were expressed at too high a level to be readily applicable.

- Where diagnostics had been derived from the research literature, some of the terminology was obscure. As a consequence, the relevance of some diagnostics failed to be appreciated by the subject.

(ii) Task simulation method

The responses given by the subject to the questionnaire are presented in full in Appendix F.3.3, the basic data being summarised in Table 11.3.

Conformity with SIAM's representational structure. The subject generated five of the six representations required by SIAM for the development of a task simulation. She did not collect data on the current task; (a) because it was not possible to gain access to observe the task; and (b) because the previous study had generated a description of the task deemed to be close to that assumed in the present study. In all other cases, representations were explicit

[2] In fact, SIAM assumes an investigator with knowledge of experimental design, so this should not be regarded as a shortcoming of the method.

Table 11.3 Questionnaire responses - Task simulation method (Phase 1)
Ratings on a scale of 0 - 5 (see Appendix F.1)
Comments of the subject are reported in Appendix F.3.3 (comments refer to items marked in bold)

	Prel. task desc	Task data	Exp. task desc.	Fut. task desc.	Fut. task model	Task sim. spec.
A. Conformity with SIAM Is rep. necessary?	y	n	y	y	y	y
Was the rep. developed?	y	-	y	y	y	y
Implicit or explicit	exp	-	exp	exp	exp	imp
As prescribed?	n	-	n	n	**n**	**n**
B. Product quality Does the rep. - include relevant classes of info?	y	-	y	**y**	y	y
- show correct level of description?	y	-	y	**y**	y	y
- characterize what it is supposed to?	y	-	y	**y**	y	y
C. Support for project Helped understanding?	3	-	4	4	**2**	**3**
Helped planning?	4	-	3	4	**2**	**3**
Helped communication?	y4	-	y4	y3	y3	n
Enabled next procedure?	y	-	y	y	**n**	**y**
D. Assessor costs Mental effort to develop rep.?	5	-	5	**2**	**2**	**2**
Support from SIAM?	-	-	-	-	-	-
E. Assessor behaviour All steps performed?	-	-	n	n	n	n
Steps performed differently?	-	-	y	**y**	y	y
Mistakes?	-	-	n	y	-	-

and similar in structure to that required by SIAM. This was not unexpected, because the starting point for the task representations was a previous output of SIAM.

Product quality. In all cases, the subject considered that her representation was adequate for its purpose in developing a task simulation suitable for the evaluation.

Quality of support for project organization. Of the five representations generated, four were viewed as being at least useful in contributing to the subject's understanding of the requirements for simulating the task, to her future planning and to her communication with other people in the project. The future task model was less useful, because the subject had only generated it explicitly *after* she had implemented the simulation: it was thought only likely to have a function in documentation of the project.

Assessor costs. There was a marked contrast between the effort reported by the subject as required for the generation of the early task representations (preliminary task description and expanded task description) - assessed as trivially easy - and the later ones (future task description; future task model and task simulation specification) - which required substantial effort. This was attributable to the fact that the early representations could be generated by simple modifications to existing task descriptions, whereas the later ones demanded analytic and generative skills for the production. The following were reported as presenting difficulties:

- deciding appropriate levels of description
- intersecting the device model and task model
- sustaining demands for accuracy in representation
- designing the simulation

There were substantial deviations from the procedures specified by SIAM (see below). Because of this, the study did not provide data useful in the evaluation of the impact of the procedures of the task simulation method on assessor costs.

Behaviour. In all cases, the behaviour of the subject deviated from that indicated by SIAM. This was attributed to the following reasons:

- availability of existing representation which removed the necessity for some steps
- non-availability of an explicit configuration of the diagnostics
- preference for a bottom-up approach (resulting in the future task model and task simulation specification being developed in parallel).

The subject's existing discipline knowledge enabled the representations to be developed successfully. An extended interview was conducted to determine the criteria she had used for specifying the simulation. This is included as Appendix F.3.4. One error was reported in the specification of the future task description. This was due to the user forgetting to include an action in the description. The mistake was apparently not a consequence of the design of the method.

(iii) Device simulation method

The responses given by the subject to the questionnaire are presented in full in Appendix F.3.5, the basic data being summarised in Table 11.4.

Conformity with SIAM's representational structure. Only the first of SIAM's six representations was developed in Phase 1 (an elaborated device description). This was explicit, but it differed substantially from the structure prescribed by SIAM because the actual device was available for direct evaluation. There was, consequently, no need to

Table 11.4 Questionnaire responses - Device simulation method (Phase 1)
Ratings on a scale of 0 - 5 (see Appendix F.1)
Comments of the subject are reported in Appendix F.3.5 (comments refer to items marked in bold)

	Elab. dev. desc.	Dev. sim. spec.	Dev. sim.	Dev. sim. perf. data	Anal. of dev. sim. behav
A. Conformity with SIAM Is rep. necessary?	y	-	-	-	-
Was the rep. developed?	-	-	-	-	-
Implicit or explicit	exp	-	-	-	-
As prescribed?	y	-	-	-	-
B. Product quality Does the rep. - include relevant classes of info?	-	-	-	-	-
- show correct level of description?	-	-	-	-	-
- characterize what it is supposed to?	-	-	-	-	-
C. Support for project Helped understanding?	-	-	-	-	-
Helped planning?	-	-	-	-	-
Helped communication?	-	-	-	-	-
Enabled next procedure?	-	-	-	-	-
D. Assessor costs Mental effort to develop rep.?	-	-	-	-	-
Support from SIAM?	-	-	-	-	-
E. Assessor behaviour All steps performed?	-	-	-	-	-
Steps performed differently?	-	-	-	-	-
Mistakes?	-	-	-	-	-

develop a device simulation, so other stages of the device simulation method were not applicable.

(iv) User simulation method

The responses given by the subject to the questionnaire are presented in full in Appendix F.3.6, the basic data being summarised in Table 11.5. The responses are discussed below.

Table 11.5 Questionnaire responses - User simulation method (Phase 1)
Ratings on a scale of 0 - 5 (see Appendix F.1)
Comments of the subject are reported in Appendix F.3.6 (comments refer to items marked in bold)

	Desc. of task k.	User subj. model	User subj. dev. prog.	User sim.y
A. Conformity with SIAM Is rep. necessary?	y	y	y	t
Was the rep. developed?	y	y	y	t
Implicit or explicit	exp	exp	exp	exp
As prescribed?	n	n	n	n
B. Product quality Does the rep. - include relevant classes of info?	y	y	y	y
- show correct level of description?	y	y	y	y
- characterize what it is supposed to?	y	y	y	y
C. Support for project Helped understanding?	3	-	-	-
Helped planning?	4	4	3	-
Helped communication?	y3	y3	n	n
Enabled next procedure?	y	y	y	y
D. Assessor costs Mental effort to develop rep.?	3	4	4	2
Support from SIAM?	3	3	**2.5**	2
E. Assessor behaviour All steps performed?	y	**n**	y	y
Steps performed differently?	y	y	y	**n**
Mistakes?	n	n	n	n

Conformity with SIAM's representational structure. All four of SIAM's user representations were developed by the subject. They were all explicit and conformed to the prescribed structure.

Product quality. The quality of the user representations was judged by the subject to be adequate for the development of a suitable user simulation. The subject expressed some concern that the level of detail of the description of the user's task knowledge would be adequate; however, no further information was available to add to the description. The concern did not derive, then, from inadequacies of the method.

Quality of support for project organization. Only the description of the users' task knowledge was judged by the subject to have contributed substantially to her understanding of the issues pertaining to the development of an adequate user simulation. In the case of the others, the subject did not feel that this was an important aspect of their function, and so she preferred not to assess them in this regard. However, all the representations were viewed at least as helpful in planning (or as fundamental in enabling subsequent activities). Where they were required for communication, the subject's representations were adequate for the purpose.

Assessor costs. Only the implementation of the user simulation was reported by the subject as having demanded substantial effort. This was because subjects needed to be recruited, trained and encouraged in performing the task. Unfortunately, the procedures of the method did not alleviate these demands. Otherwise, the procedures were judged to be helpful.

Behaviour. There were some slight deviations from the procedure specified by SIAM, although the procedures generally corresponded with what the subject would have done under her own initiative. The ordering of some of the steps was modified, and some were constrained by the impossibility of direct access to target users. The only deviation arising from the inadequacy of the method was due to the non-availability of configured diagnostics. The subject used her own (implicit) model of interaction to identify critical features of users.

(b) Assessment of behaviour and performance: Phase 2 (enhanced interface)

(i) Usability evaluation method

The responses given by the subject are presented in full in Appendix F.4.1, the basic data being summarized in Table 11.6. The responses are discussed below.

Conformity with SIAM's representational structure. The usability evaluation method prescribes a total of eight representations; however, only six were considered appropriate to Phase 2. The preliminary system specification was unnecessary, because a detailed specification had been produced during Phase 1; and the study was not a feasibility assessment, so the final report would have to differ from that prescribed by SIAM.

The representations which were generated were generally explicit, and they corresponded to the structure required by SIAM. The exceptions were the solution strategy, which was implicit but very similar to that produced during Phase 1; and the analysis of device-user interaction, which was based on the subject's own model of device-user interaction. The latter was caused by the subject encountering further difficulties with the diagnostics (see below).

Table 11.6 Questionnaire responses - Usability evaluation method (Phase 2)
Ratings on a scale of 0 - 5 (see Appendix F.1)
Comments of the subject are reported in Appendix F.4.1 (comments refer to items marked in bold)

	Prel. prob spec.	Prel. syst. spec.	Diag. table conf.	Soln. strat.	Expt. con-text	Data	Anal. of int'n	Feas. rept.
A. Conformity with SIAM Is rep. necessary?	y	n	y	y	y	y	y	-
Was the rep. developed?	y	-	y	y	y	y	y	-
Implicit or explicit	exp	-	exp	**imp**	**exp**	exp	exp	-
As prescribed?	n	-	n	n	n	n	y	-
B. Product quality Does the rep. - include relevant classes of info?	y	-	n	y	y	y	y	-
- show correct level of description?	y	-	y	y	y	y	y	-
- characterize what it is supposed to?	y	-	?	y	y	y	y	-
C. Support for project Helped understanding?	2	-	0	3	-	5	3	-
Helped planning?	3	-	1	4	3	4	3	-
Helped communication?	n	-	n	y	n	y4	y4	-
Enabled next procedure?	y	-	n	y	y	y	y	-
D. Assessor costs Mental effort to develop rep.?	4	-	2	3	2	2	3	-
Support from SIAM?	3	-	1	2.5	2	2	1	-
E. Assessor behaviour All steps performed?	y	-	y	y	y	y	n	-
Steps performed differently?	y	-	y	y	y	y	y	-
Mistakes?	n	-	?	n	y	n	n	-

Product quality. The subject generally reported satisfaction that her representations would enable an adequate system evaluation. However, the main exception was the configuration of the diagnostics: the subject was unable to locate a diagnostic appropriate to the issues of concern to her [3]- the allocation of function between speech and the manual entry of simple (binary) commands - and therefore felt that her selection of diagnostics did not cover all the relevant classes of information. As a consequence, she also felt that the diagnostic configuration was not an accurate representation of the issues under investigation. The subject noted that the accuracy of her analysis of device-user interaction was compromised by the

[3] In fact, a "general purpose diagnostic" in the table was appropriate, but it was expressed in such a way that its relevance was not evident to the subject.

small subject sample; however, this was due to the circumstances of the study and was not attributed to SIAM.

Quality of support for project organization[4]. Although the subject did succeed in configuring the diagnostics (albeit not to her satisfaction), she found that doing this had the effect of confusing her. The subject reported that the remaining representations did contribute (rather obviously) to her understanding of the problem: the experiment was the mechanism by which "the problem" was to be solved.

Concerning the contribution of the representations to the planning of the project, the subject found all except the diagnostics useful. The point was made, however, that the contribution was to subsequent *technical* decision making, rather than at the level of project management. This also is discussed later. Only the solution strategy, experimental data and analysis of device-user interaction were to be used in communication. In all cases, the subject felt her representations would be facilitative.

The problems in configuring the diagnostics and the consequent concern of the subject over the adequacy of her representation was a potential impediment to further progress. For this reason, she utilized her own (implicit) model of interaction in subsequent procedures, rather than relying on the diagnostics.

Assessor costs. The configuration of the diagnostics, the development of the experimental context and data interpretation all demanded substantial effort. In general, SIAM's procedures did not provide good support with respect to these activities (particularly to the use of the diagnostics and to data interpretation). This could be attributed to the subject's resorting to her own procedure as a consequence of the problem with the diagnostics. In other cases, the subject tended to use SIAM's procedures as a checklist, to ensure that she had covered everything.

The following comments were made concerning the subject's difficulties in configuring the diagnostics:

- some diagnostics seemed to be most relevant to specific classes of system, making it difficult to assess their relevance to systems in general
- it was difficult for the subject to map her internal model of the critical features of interaction onto the diagnostics

[4] In her assessment of the contribution of representations to her understanding of the problem, the subject commented that the assessor had to possess a clear understanding of the problem before the procedures could be applied. Making the representation explicit did not contribute further to this understanding. It was evident that this item on the questionnaire was not addressing directly whether the method was supporting the "management of complexity". The issue was discussed further in the concluding review (see Section 11.4.4).

- it was difficult for the subject to assess the implications of selecting one diagnostic in preference to another
- it was difficult to assimilate the information in the diagnostics expressed in the way they were.

Assessor behaviour. With the exception of the analysis of device-user interaction (different due to the non-use of the diagnostics), the subject progressed through the steps in the procedures specified by SIAM. However, there were deviations in the manner in which they were performed. The steps in the preliminary specification of the problem were simplified, because Phase 1 had resulted in a detailed specification of the problem to be addressed.

The problems with the diagnostics resulted in the subject trying her own approaches in an effort to develop a configuration with which she was satisfied. She felt that the decision tree intended to support the subject in selecting an appropriate table might cause the subject to overlook potential sources of device-user incompatibility because incompatibilities might only occur if representations or skills were demanded in combination. The decision tree tends to guide the user in evaluating interaction knowledge and skills in a piecemeal fashion.

She also found that the instructions to select diagnostics on the basis of searching a single column of the table (for critical system conditions) was not useful. She found the information in the designated column inadequate to decide whether a diagnostic was relevant and had to look at all the columns to understand the diagnostic properly. As a consequence of these inadequacies in the procedure, the subject searched the tables exhaustively, rather than selectively (as indicated by the procedures).

Because of her decision to use her implicit model of interaction, the subject specified her solution strategy, developed the experimental context and analysed the data without reference to the diagnostics. No inferential statistical tests were applied to her data, because of the small number of subjects. Concerning errors, the subject felt that she had not performed a sufficient number of pilot trials in the development of the experimental context. The fidelity of the device simulations was slightly compromised, but this was not thought to have seriously affected the results of the assessment.

To summarise, as in Phase 1, the subject encountered problems in the use of the diagnostics. In Phase 1 the problems could be attributed (at least in part) to the fact that the study was of a form different to that assumed by SIAM. However, this was not the case with Phase 2, which sought to evaluate a specified set of problems. The decision of the subject to use her own interaction model rather than compromise the quality of output was appropriate in the context of this study. However, the strategy resulted in her incurring substantial costs to maintain task quality and, thus, clearly indicates an inadequacy in the method.

Table 11.7 Questionnaire responses - Task simulation method (Phase 2)
Ratings on a scale of 0 - 5 (see Appendix F.1)
Comments of the subject are reported in Appendix F.4.2 (comments refer to items marked in bold)

	Prel. task desc	Task data	Exp. task desc.	Fut. task desc.	Fut. task model	Task sim. spec.
A. Conformity with SIAM Is rep. necessary?	n-	n	n	n	y	-
Was the rep. developed?	-	-	-	-	y	-
Implicit or explicit	-	-	-	-	exp	-
As prescribed?	-	-	-	-	n	-
B. Product quality Does the rep. - include relevant classes of info?	-	-	-	-	y	-
- show correct level of description?	-	-	-	-	y	-
- characterize what it is supposed to?	-	-	-	-	y	-
C. Support for project Helped understanding?	-	-	-	-	3	-
Helped planning?	-	-	-	-	**2**	-
Helped communication?	-	-	-	-	y3	-
Enabled next procedure?	-	-	-	-	y	-
D. Assessor costs Mental effort to develop rep.?	-	-	-	-	**3**	-
Support from SIAM?	-	-	-	-	**2.5**	-
E. Assessor behaviour All steps performed?	-	-	-	-	n	-
Steps performed differently?	-	-	-	-	y	-
Mistakes?	-	-	-	-	n	-

(ii) Task simulation method

The responses given by the subject are presented in full in Appendix F.4.2, the basic data being summarized in Table 11.7.

Conformity with SIAM's representational structure. The development of the task simulation in Phase 2 assumed no change in the form of the current task and only small modifications to the future task (related to the changes to interaction behaviour consequent of modifications to the device-user dialogue). For this reason, the first four task representations were equivalent to those produced in Phase 1.

An explicit future task model was produced, corresponding in structure to that prescribed by SIAM, and it was implemented successfully. However, the implementation was the same as that used in Phase 1, apart from changes to device operating procedure. The subject's behaviour in implementation was, therefore, not representative with respect to the use of the method, and it is not reported further. The following sections only relate to the development of the future task model.

Product quality. The future task model was judged to be adequately complete, detailed and accurate for the purpose of developing the task simulation.

Quality of support for project organization. The development of the future task model was judged to be useful in supporting the assessor's understanding of the problem of simulating the task, but it contributed little to planning: the representation of the device was the primary concern at this stage, and the task model was only indirectly relevant to this. The main contribution of the task model was envisaged to be in communication to the others working on the project: the subject intended to include it in the final report of the work.

Assessor costs. The subject reported that the development of the model required little effort, because it involved only modification of the Phase 1 model. Because of this, there was a reduced dependence on the procedures to enable the work. The subject identified a logical problem with the proceduralization, in that the method supports the development of a low level representation of the task and, hence implicitly, a low level representation of the device dialogue. However, at this stage, the assessor only has a relatively crude model of the device developed as part of the preliminary system description. Although not a problem in the present study (as information about the device was available), an assessor would be expected to find this difficult in a feasibility study.

Behaviour. The subject deviated substantially from SIAM's procedure in developing the task model. This could be attributed partly to the requirement to use an implicit interaction model, instead of diagnostics, to identify critical task features; and partly to the fact that it was necessary only to modify an existing model, rather than to specify it completely.

To summarise, Phase 2 made unrepresentatively small demands on SIAM in the development of the task simulation. The data collected were insufficient to evaluate properly the contribution of the task simulation method in Phase 2.

(iii) Device simulation method

The responses given by the subject are presented in full in Appendix F.4.3, the basic data being summarized in Table 11.8.

Conformity with SIAM's representational structure. The subject developed all of the representations required by the device simulation method.

Table 11.8 **Questionnaire responses - Device simulation method (Phase 2)**
Ratings on a scale of 0 - 5 (see Appendix F.1)
Comments of the subject are reported in Appendix F.4.3 (comments refer to items marked in bold)

	Elab. dev. desc.	Dev. sim. spec.	Dev. sim.	Dev. sim. perf. data	Anal. of dev. sim. behav
A. Conformity with SIAM Is rep. necessary?	y	y	y	y	y
Was the rep. developed?	y	y	y	y	y
Implicit or explicit	part exp	part exp	exp	**exp**	imp
As prescribed?	y	y	n	y	y
B. Product quality Does the rep. - include relevant classes of info?	y	y	y	y	y
- show correct level of description?	y	y	y	y	y
- characterize what it is supposed to?	y	y	y	y	y
C. Support for project Helped understanding?	4	4	5	**3**	-
Helped planning?	5	4	4	**4**	-
Helped communication?	**y2**	y3	n	n	-
Enabled next procedure?	y	y	y	y	-
D. Assessor costs Mental effort to develop rep.?	**2**	**2**	**2**	**2**	-
Support from SIAM?	**2**	**2.5**	**3**	**2.5**	-
E. Assessor behaviour All steps performed?	y	y	y	y	-
Steps performed differently?	y	y	y	y	-
Mistakes?	n	n	n	n	-

The five representations generated successfully tended to differ from those proposed by SIAM by frequently being implicit and less formally expressed. The main reason for this was that the subject had to meet a deadline and so was unable to complete the representations as she would have wished. In the case of the device specification, the recognition performance and dynamics of the enhanced ("future") device was assumed to be the same as that of the existing prototype, although the performance of the prototype was not precisely specified and so remained implicit.

In specifying the device simulation, the requirement to state the CD functions was small, so these were not made explicit. The system subject (the experimenter) had had previous experience as a device simulator and was familiarized with the task in a verbal discussion rather then by means of a textual description. However, an explicit system subject action hierarchy was generated. In the evaluation and enhancement of the device simulation, the performance was also represented implicitly, and diagnosis and prescription was performed iteratively until performance of the simulation was judged to be adequate against that of the prototype device.

In general, then, although the various representations were generated, the adherence to SIAM's structure was less rigorous than had been the case with the other sub-methods. This was attributable to the shortage of time available to develop the device simulation.

Product quality. In spite of the informality of some of the device representations, the subject viewed them as sufficiently complete, detailed and accurate for the purpose of developing an adequate device simulation.

Quality of support for project organization. The subject generally rated the quality of the support of the representations for project activities highly. All contributed to her understanding of the problems of simulating the device and also to her view of the usability of the dialogue. They were judged to be of considerable value in planning subsequent activities in simulating the device. There were relatively few requirements to use the representations in communication. The process of developing the device simulation was not of direct concern to project management elements of the procurement system, and communication at a technical level occurred within discussions over the developing simulations.

Assessor costs. All the procedures in the development of the device simulation demanded substantial effort. In the early stages, the design activities to specify the future device and, subsequently, the simulation, incurred cognitive costs. The implementation then placed considerable demands on both the assessor and the system subject in the maintenance of an adequate representation of the performance of the target system, as did the iterative process of evaluating and enhancing the performance of the simulation.

SIAM's procedures tended not to reduce these costs significantly. This was partly because the activities which presented most demands (such as design) could not be proceduralized. However, in many cases the circumstances of the study prevented the procedures being carried out in full (see below).

Assessor behaviour. All of the representations were developed using procedures differing from those specified by SIAM. Although there was one instance in which this was clearly a consequence of an inadequacy of the method (i.e. the failure to configure the diagnostics), most of the deviations could be attributed to external factors. For example, in the

Table 11.9 **Questionnaire responses - User simulation method (Phase 2)**
Ratings on a scale of 0 - 5 (see Appendix F.1)
Comments of the subject are reported in Appendix F.4.4 (comments refer to items marked in bold)

	Desc. of task k.	User subj. model	User subj. dev. prog.	User sim.y
A. Conformity with SIAM Is rep. necessary?	**n**	**n**	y	n
Was the rep. developed?	-	-	y	-
Implicit or explicit	-	-	imp	-
As prescribed?	-	-	y	-
B. Product quality Does the rep. - include relevant classes of info?	-	-	y	-
- show correct level of description?	-	-	y	-
- characterize what it is supposed to?	-	-	y	-
C. Support for project Helped understanding?	-	-	**2.5**	-
Helped planning?	-	-	3	-
Helped communication?	-	-	y	-
Enabled next procedure?	-	-	y	-
D. Assessor costs Mental effort to develop rep.?	-	-	3	-
Support from SIAM?	-	-	**2.5**	-
E. Assessor behaviour All steps performed?	-	-	y	-
Steps performed differently?	-	-	n	-
Mistakes?	-	-	n	-

development of an elaborated device description, SIAM assumes the direct involvement of speech technologists in specifying a future device; however, in this study the subject had to familiarize herself with the technological constraints on the implementation of enhancements and then specify them. Other procedures were performed cursorily because of a shortage of time; this was particularly the case in the evaluation and enhancement of the device simulation.

In summary, task quality was maintained, although user costs were high. The procedures did not apparently reduce them, although the deviations from the prescribed procedure call into question an evaluation of the procedures on the basis of the present data. The deviations were, in the main, a consequence of factors external to the method. The subject did utilize the representational structure of SIAM and found this valuable, particularly in the organization of the work.

(iv) User simulation method

The responses given by the subject are presented in full in Appendix F.4.4, the basic data being summarized in Table 11.9.

As with the task simulation method, representations of the user could be recruited directly from Phase 1. The only representation judged to be necessary in Phase 2 was the user subject development programme, which included different instructions to user subjects on the operation of the device. This modification was small, and the method could not be said to have contributed substantially to its development.

11.4.4 Concluding interview

On completion of the evaluation of the VDC, the subject was debriefed and an assessment elicited from her of the overall performance of the task. The details of the interview are presented in Appendix F.5, but the main points are now summarised.

Conformity with SIAM's representational structure. The subject generally felt that she had conformed to SIAM's representational structure, in that she produced most of the required representations in some form. The subject expressed the view that the notation offered for representing the task was appropriate for discrete, sequential actions, but might be less suitable for other types of behaviour (e.g. continuous actions, such as those used in drawing). She also showed misgivings over the notation used for specifying the target device and, subsequently, the device simulation.

Product quality. In the judgement of the subject and the experimenter, the quality of the assessment was high, particularly given the operating constraints acting on the project. The subject felt that she could probably have produced an adequate assessment without the method[5], and the preliminary assessment of her knowledge of ergonomic evaluation would support this view.

Quality of contribution to project organization. Because the present project was of short duration and involved few staff and resources, it did not impose heavy demands for organization and management. The contribution of SIAM to the "management of complexity" was, then, difficult to assess, because the management element of the project was not judged to

[5] In the opinion of the author, the study would have been performed less systematically and reported less rigourously had SIAM not been used.

be "complex". However, the method supported communication with the device developers. It also facilitated the transfer of information from the previous study of the VDC.

Assessor costs. It was not possible to determine whether use of the method had reduced overall costs, although it probably altered their character. It reduced a certain element of stress by helping the assessor to ensure that everything had been done. However, it added costs by imposing the overheads of generating more documentation, of learning the method and of having to express concepts in terms of the method. The latter two overheads might be expected to reduce with increased experience of the method.

Assessor behaviour. The subject did deviate from SIAM's procedures. The deviations were due to:

- incompatibility between the method and the form of the present assessment;
- defects in the presentation of diagnostic information;
- shortage of time

The subject tended to adhere more closely to the procedures in Phase 1 than in Phase 2. This was due to increasing familiarity with the method, to being able (later on) to generate adequate representations by modification of those produced earlier, and to the shortage of time in Phase 2.

In conclusion, the subject considered that the method had been facilitative at a structural level. While the procedures had assisted in specification activities, they had not contributed so effectively to the final implementation of the simulations or to usability testing. The low level of proceduralization had presented problems where the immediate requirements of the assessment diverged from those assumed by the method, either because of the differing objectives of the study or because of inadequacies in other aspects of the method (i.e. the expression of diagnostic information).

11.5 Summary of results

(a) The evaluation of the VDC was successfully completed using SIAM. The task quality was judged to be high, both by the subject and by the device developers who were to utilize its output. The evaluation was conducted systematically, and the study was fully reported.

(b) User costs were altered, if not substantially reduced, by the utilization of SIAM. The subject reported a reduction of some stress in the conduct of the assessment, by the method providing a better assurance of completeness. However, additional costs were incurred in generating more detailed and explicit documentation and in expressing the problem in terms of the method.

(c) The subject generally adhered to the process of SIAM at a high level and developed most of the prescribed recommendations in some form. With one important exception (the configuration of the diagnostics), the deviations that did occur were attributable to the study differing to that for which SIAM was originally intended.

(d) SIAM's representations facilitated project planning and communication. Their contribution to the "management of complexity" was less clear; in part because the evaluation was circumscribed, but also because the relevant questionnaire item was ambiguous. Discussion with the subject suggested that the process of developing representations probably did enhance her comprehension of the problem.

(e) The diagnostics failed to support the process of SIAM. Their failure was attributed to:

(i) specific diagnostics being expressed in terminology which was not accessible to the subject;

(ii) general purpose diagnostics being expressed at too high a level to be usable;

(iii) the information contained in diagnostics being decomposed in a format which was not compatible with the subject's personal representation of the problem;

(iv) the process of selecting and integrating a set of relevant diagnostics not resulting in a model of device-user interaction which was compatible with the subject's personal representation of the interaction.

(f) The subject deviated from the procedures of SIAM on a number of occasions. Deviations were attributable to:

(i) the assessment being of a form incompatible with the assumptions of SIAM (e.g. evaluation of a prototype being incompatible with the assumption that SIAM will be used in feasibility assessments);

(ii) the procedures of the method being beyond the capability of the subject to implement (e.g. configuration of the diagnostics);

(iii) the procedures of the method being judged by the assessor to be less effective than her own preferred procedures (e.g. development of an implicit model of interaction, instead of configuration of the diagnostics);

(iv) procedures having to be modified to accommodate the products of deviations from earlier procedures (e.g. in the process of developing a simulation, an implicit model of interaction was intersected with representations of the work system, instead of there being proceduralized reference to the configured diagnostics); and

(v) lack of time, preventing complete implementation of procedures (e.g. failure to implement controlled experimental evaluation of the device simulation in Phase 2).

Some of these causes - for example, shortage of time - did not, in the study, indicate inadequacies of the method. However, others may be attributed to limitations of the conceptualization and proceduralization of the scope, notation and process of the method.

(g) The study constituted only a partial test of SIAM, as some parts of the method were not fully utilized (especially, parts of the device simulation method); and because, at early stages of the task simulation method, the subject was able to utilize representations developed in a previous study. These weaknesses were an unavoidable consequence of the test being conducted in the context of a commercial system development project.

11.6 Conclusions

The test of SIAM was modest in scope and limited in power; however, it served to demonstrate a number of the strengths and weaknesses of the method. There is evidence that the quality of speech interface evaluation was enhanced by the support of SIAM. User costs were apparently not reduced, but SIAM does not claim to eliminate the assessor's work; rather, to render it systematic, and so more effective. It was expected that assessment would demand effort, even when supported by SIAM. Within this perspective, costs incurred by the subject were probably acceptable. Overall, then, the results suggest that SIAM *does* improve the performance of evaluations of speech interfaces.

Although the assessment task described here was completed successfully, it must be acknowledged that the personal knowledge of HF possessed by the subject - within the classification of Chapter 3, she was an HF generalist - enabled her to complete the task in spite of the failure of the diagnostics. The same performance could not have been achieved had the assessment been conducted by a casual practitioner of HF. If SIAM is to be utilized by the population of assessors for which it was originally intended, procedures need to be modified to take account of the difficulties encountered in this study. In particular, further development is required with respect to the expression of interaction knowledge and its recruitment by the procedures. Furthermore, the necessity to deviate from the procedures as a consequence of the specific characteristics of the evaluation problem suggests inflexibility. Flexibility would need to be improved if SIAM were to be used by non-specialist procurers.

SIAM shows considerable promise, then, but it exhibits some important weaknesses. Chapter 12 considers the implications of the findings for the further development of SIAM and for structured HF evaluation methods in general.

CHAPTER 12

CONCLUSIONS AND IMPLICATIONS OF THE RESEARCH

12.1 Introduction

The work described in this thesis has addressed a particular problem presented in the procurement of speech-based battlefield computer systems. The solution offered - a structured evaluation method - constitutes a contribution particularly relevant to this problem. However, the research can claim also to have contributed to the solution of the more general problem of providing support for the conduct of HF evaluations in procurement.

Chapter 12 evaluates the contribution of the research on the basis of the results of the trial of the method presented in the previous chapter. The contribution is considered with respect both to the specific and to the general problems. Implications of the work are identified for the planning of military procurement; for the conduct of usability evaluations; and for the further development of methods to support the practice of HF.

12.2 Contribution to the solution of RSRE's specific problem

12.2.1 Conclusions of the trial of SIAM

Chapter 5 has described RSRE as an organization concerned with the procurement of battlefield computer systems, which is also involved in the development of speech technology. RSRE recognized potential ineffectiveness in the procurement of battlefield computers as this related to speech technology. Part of the problem was attributable to there being no means to determine, prior to specification, whether or not speech interaction would support the performance desired of work systems. RSRE did not employ HF specialists in the domain of speech technology, so the organization required a means to enable non-specialist procurers to conduct appropriate performance assessments.

The solution offered to RSRE's problem was SIAM: a structured method for the empirical assessment of the performance of speech-based computer systems. The results of the operational trial of SIAM indicate that the method facilitates the effective evaluation of existing speech interfaces (e.g. the prototype VDC), and the assessment of the performance of future speech interfaces (e.g. the enhanced version of the VDC). The trial further demonstrated that SIAM can support assessors lacking *specialist* knowledge of the human factors of speech technology.

SIAM has been shown, then, to meet a number of RSRE's requirements; however, the trial did identify weaknesses in the method, and the trial itself had limitations. Specifically, testing was conducted in the context of a project in which the predominant activity was design, rather than performance assessment; it involved the evaluation of only one type of speech interface; and it studied the behaviour of just a single assessor. The trial of SIAM cannot, then, be regarded as definitive; however, its results are instructive as regards RSRE's procurement problem. The following sections consider the implications of the results for RSRE's activities and for the further development of SIAM.

12.2.2 Implications for the procurement of military speech systems

The evaluation of SIAM was conducted in the context of a small user interface development project - the development of a demonstrator for a novel voice to data converter (VDC). At one level of description, the entire project was an assessment of the feasibility of a speech-based computer to support certain battlefield observation tasks; it was, by definition, a feasibility study. However, at a lower level, the development of the VDC demonstrator may be viewed as a complete procurement project in microcosm, in which RSRE (the procurer) performed many equivalent functions to those which would be performed in a large scale procurement project. For example, RSRE was involved with the developers (RMCS) in performance setting, evaluation and iterative design, and with processes to support the decision of MoD to proceed further with the product.

The study described in Chapter 11 supported the later stages of design of the VDC: the device had been implemented and was in the process of being refined. It was notable that the procurers (and developers) had recognized shortcomings in the implemented user interface which might have compromised the performance of the system for which it was intended. The study subsequently performed sought to identify these shortcomings and to prescribe user interface enhancements.

As a result of the study, changes were made to the implementation of the VDC software; however, the requirement for these late modifications would have been reduced, or even eliminated, had HF performance assessments taken place *in advance* of implementation. SIAM could have been as readily applied at an earlier stage as it was during design. Although the costs of failing to conduct early performance assessments were small in this instance, in larger scale projects the consequences of such failure might well have been unacceptable. The implication is that the use of SIAM prior to detailed specification would result in an improvement in the effectiveness of the procurement of battlefield computers with speech interfaces.

12.2.3 Implications with respect to HF evaluation

Although the trial was primarily concerned with the effectiveness of SIAM's support for the assessor, the results have indirect implications for HF evaluation techniques recruited by the method: for the hypothetico-deductive approach to evaluation; and for the human simulation of speech-based systems.

(1) *Hypothetico-deductive evaluation.* The hypothetico-deductive approach to evaluation assumed by the method was appropriate where the assessor could specify in advance a problem to be addressed (and, hence, a hypothesis to be tested); however, the approach failed where the concern was with the *identification* of problems. This failure was evident in the evaluation of the VDC prototype and is attributable to the requirements for *inductive* rather than deductive investigative techniques under such circumstances.

Induction requires the inference of the general from specific instances. In the trial of SIAM, the instance was the inadequacy of the behaviour of the system to support desired performance, and the inference was that of relating system behaviour to a generalizable model of the interaction between people and computers (i.e. diagnosis). Logically, then, behaviour had to be instantiated in order that inadequacies might be identified. This presents a problem for simulation-based methods, because simulations are designed to reproduce selected aspects of the behaviour of system entities and their interaction. Inadequacies will not be identified unless the simulation happens to reproduce the behaviours which are the source of inadequacies.

A simulation to support inductive reasoning might be developed iteratively, by exploring the consequences of behaviours *potentially* critical to performance. In the first instance, such behaviours might be reproduced with low fidelity; the reproduction of behaviours showing some evidence of an adverse effect on performance might subsequently be refined in order to ascertain the significance of the effect. However, the identification of problems in this way would be speculative. Reliability would be dependent, firstly, upon the completeness of information concerning the target system (e.g. the availability of a prototype); and, secondly, upon the skill of the investigator. It would be difficult (or impossible) to offer detailed procedures for a casual practitioner to conduct such evaluations.

(2) *Human simulation.* Human simulation techniques have been exploited previously for implementing simulations of advanced speech I/O devices, and the research described here confirms the viability of the techniques for reproducing the behaviour of simple recognizers and synthesizers. However, the second experiment reported in Appendix C

demonstrates limitations of human performance in accurately reproducing the behaviour of a speech interface.

Provided the behaviour of the simulation is such that the user subject behaves in the same way as if he or she were interacting with the target device, these limitations are unlikely to be critical. However, human simulation is inappropriate when this condition cannot be met. The paradigm would seem to be weak, for example, when the concern lies with the low-level dynamics of interaction, where it is critical to reproduce the temporal aspects of device behaviour precisely. Human simulation may also fail where the constraints on the simulated dialogue are so complex that the system subject cannot compute them and behave appropriately in real time - such constraints might occur in reproducing the behaviour of devices supporting "natural language" dialogues (see, for example, Morel, 1986). Although, in principle, the capabilities of the simulation system may be extended by offering sophisticated aids to the system subject, there will be limits to the cost effectiveness of interventions of this sort.

12.2.4 Implications for the further development of SIAM.

Those parts of the assessment process for which SIAM was found to offer only incomplete support in the present study would be expected to be performed less adequately in the context of a more complex assessment task and/or a less knowledgable assessor. Further research and development is necessary, therefore, if SIAM is to represent a complete solution to RSRE's requirement. A full solution would require, firstly, that the currently-recognized defects in SIAM be rectified; and, secondly, that the method be validated as a solution, by the application of more complete and rigorous evaluations.

(1) *Improvements to SIAM.* The quality of the assessment observed in the operational trial was adequate (or better than adequate). It may be supposed, then, that deviations from the procedure prescribed by SIAM potentially are manifestations of behaviour which is better able to support desired performance than that intended by SIAM's procedure. Here, the major deviations reported in Chapter 11 are interpreted as indicating requirements to modify the notation and process of SIAM, in order that it may support evaluations which should be within its scope.

Modification of notation. One reason why the subject of the study might have deviated from SIAM's procedure would be if SIAM's notations had been poorly suited to the representation of relevant information. All the descriptive representations developed using the method are based upon textual notations. In some of these (e.g. preliminary system description) text is expressed in natural language; however, such representations are subsequently transformed to a more structured format to support the application of the procedures of the method. These structured representations are based upon text organized in systematic ways; for example, task descriptions are expressed as tree diagrams; descriptions of the target user and the specifications of simulations take the form of lists of attributes etc.. Such structured representations facilitate reference to particular components of the textual expressions.

The tree-diagram utilizes spatial configuration, as well as text, to convey meaning; for example, the temporal order of actions is represented by their relative horizontal position on the page. The notation generally proved successful for representing task structure in the context of the trial; however, the subject expressed reservations concerning its effectiveness for conveying interactions which were not readily decomposable into discrete features. Although speech interaction comprises discrete actions (utterances and the reception of information uttered by correspondents), non-speech aspects of tasks may be continuous in nature (e.g. continuous manual control actions). Non-speech action may, then, be incompletely described within the notation of SIAM.

However, in spite of the fact that it may not be possible to decompose continuous (non-speech) task elements to the same low level as is possible for speech task elements, there is no reason to believe that they cannot be analysed at a level *adequate* for the purpose of developing system simulations. The hierarchical notation would appear adequate for expression of task structure within the scope currently specified for SIAM[1], but it might be less adequate if the scope of the method were extended to include tasks involving predominantly continuous elements.

Text structured as tables is used to express the substantive knowledge of human-computer interaction embodied in the diagnostics which support SIAM. Part of the problem associated with the use of the diagnostics (Section 11.4.3) may be attributed to the conceptualization of their tabular format. Specific inadequacies identified by the subject were:

- failure to convey meaning clearly for purposes of identifying the relevance of diagnostics; and

- ineffectiveness of general purpose diagnotics due to the abstract nature of their expression.

The rationale underlying the organization of the tables was that guidelines are expressible as productions of the form: IF....THEN.. ..BECAUSE....HENCE.... However, in the course of the development of SIAM, a decomposition of prescriptive information

[1] The hierarchical notation was noticeably laborious in the development and modification of task representations - particularly of the task of the system subject. A solution to this would be the development of a tool for devloping hierarchical representations. Edmondson and Johnson (1990) have developed such a tool support hierarchical task analysis.

based upon these four "headings" was recognized to be inadequate for interfacing with the procedures of SIAM. Additional decomposition was subsequently made of the attributes of separate components of the target system assumed by the architecture of SIAM (e.g. task, device, user and context). The format for the expression of evaluation knowledge was specified, then, primarily on pragmatic criteria. One explanation for the failure of the subject to make use of the diagnostics might be that the conceptualization of the decomposition was inappropriate for the expression of diagnostic information concerning speech-based systems. The production rule representation was somewhat inflexible and was apparently incompatible with the assessor's mental representation of the problem and its solution.

The further development of SIAM demands a review of the rationale for decomposing diagnostic information to support the method. Ontological, rather than pragmatic, criteria should offer a more coherent decomposition of interaction knowledge; but the diagnostics should be expressed in a form compatible with the mental representations of domain knowledge held by practitioners. A possible starting point for further research to this end would be a study of the behaviour of HF specialists performing evaluations. The form of the specialists' representations would be inferred, as would the process by which the representations were recruited for the purpose of prescription (e.g. deSouza et al, 1990).

Modification of process. The general adherance of the subject to the process of SIAM, coupled with the successful outcome of the study, is compatible with the view that the process is effective at a high level, i.e. its conceptualization is appropriate. However, the appropriateness of its low level expression (its proceduralization) has been brought into question, particularly as relates to the recruitment of evaluation knowledge in the diagnostic tables. Deviations from the prescribed procedures were a consequence either of the procedures being flawed, or by their being inappropriate under the particular circumstances of the study.

A *flawed* procedure would not offer an optimal solution to the assessment problem, so an assessor possessing relevant knowledge (such as the subject in the study) would choose to deviate from the procedure. In the study, the procedure supporting the specification of a model of human-computer interaction by selecting a set of "relevant" diagnostics was apparently not effective. An alternative strategy (say, one of progressive refinement of a general diagnostic) might be more effective, if appropriately proceduralized.

An *inappropriate* procedure would offer an optimal solution to the assessment problem, had the problem been in a different state. In the test of SIAM, deviations from the procedures could be attributed either to the assessment being of a form for which the method was not intended (e.g. the development of an interaction model to support the inductive processes of Phase 1 of the study); or to the assessor having previously deviated from procedures, with the consequence that products of earlier procedures violated the input assumptions of subsequent procedures. In the study, such deviations occurred when the assessor made good the method's shortcomings (e.g. in the configuration of the diagnostics). Although there was no evidence of this in the present study, such deviations might also have been a consequence of *errors* committed earlier in the process, or by the *absence of information assumed* by the method to be available.

Two approaches might be taken to repair the flawed and inappropriate procedures of SIAM. The first approach assumes that SIAM's scope is redefined, such that the method would, in future, be applied by individuals possessing at least general knowledge of HF; the second assumes adherence to the currently intended scope of SIAM, i.e. applicability by casual practitioners.

Approach (a). Because SIAM is intended to be used by casual practitioners, its level of description is set comparatively low. For example, the specification of a solution strategy (Procedure 7.4) instructs the assessor to specify the experimental hypothesis by reference to the seventh column of the diagnostic table; the independent variables by reference to columns two, three, four and six; the dependent variables by reference to column eight etc. By definition, the level of description of a procedure will exhibit a direct relationship with the extent to which the procedure is generally applicable. It is unsurprising, then, that, given such a low level of description, some of the procedures are of limited generality and cannot be applied when their input assumptions are violated.

Generality might be increased by raising the level of description of the procedures; for example, in the case of specifying a solution strategy, the assessor might be instructed to design an experiment on the basis of less-systematically decomposed evaluation knowledge, by inferring appropriate experimental parameters. However, such modification would reduce the accessibility of the method to assessors lacking personal discipline knowledge. Although the procedures would support the specification of a wider range of types of study and might be implementable by HF specialists or generalists, they would not be sufficiently detailed for application by casual practitioners (at least, not without extensive exemplification of the inference processes required).

Approach (b). An alternative approach to the enhancement of SIAM would be to re-specify ineffective low level procedures and then extend the existing set to include procedures appropriate for different classes of study. For example, a set of procedures might be added to be used where an assessment is exploratory in nature and where, as a consequence, no usability problems are identifiable to orientate the design of the study. Such an extension of the procedure set would demand the inclusion of procedures appropriate to the context of the assessment; e.g. IF the evaluation is to determine the usability of an *existing* device THEN use procedure set A, ELSE use procedure set B. Increasing the generality of SIAM by extending the procedures would, in principle, allow its application by users lacking private discipline knowledge; however, the set of possible types of assessment study is large and it would be a substantial task to ensure complete coverage of the set. This second approach to enhancement would demand, then, large research resources for its successful implementation.

(2) *Validating SIAM.* The informal test presented in Chapters 10 and 11 was justified on the grounds that, at the time of the evaluation, the development of SIAM was incomplete. The primary objective was to identify requirements for improving the method, rather than to offer a validation of the method in its finished form. However, following the implementation of enhancements such as those proposed above, a requirement remains for a fuller evaluation of SIAM; for without formal testing, the truth of the knowledge embodied in SIAM remains unknown.

A number of the limitations of the existing study might be rectified relatively easily. Clearly, any formal evaluation would require a larger sample of subjects, which should be more fully representative of the population of method users (i.e. procurers lacking HF knowledge). Furthermore, the task performed by subjects should be representative of that for which the method was intended (i.e. evaluation prior to device implementation); and evaluations should be conducted across the range of target devices within SIAM's scope.

However, other weaknesses of the study would be less easy to make good. For example, the implicitness of the evaluation criteria used in the test (i.e. subjective comparison of task performance using SIAM against the situation had SIAM not been available) was unsatisfactory. One reason for choosing this criterion was the absence of other yardsticks; yet, in many ways, the comparison was trivial, and would have been more so had the subject possessed less personal discipline knowledge. As long as there are no alternative methods to support speech interface evaluation, comparative studies present difficulties in the specification of criteria for testing SIAM's claim that it enhances performance. It might be possible to compare performance using *later* (improved) versions of SIAM against that using the first version (hence, evaluating the improvements, rather than the method as a whole). However, such variant evaluation could only constitute a partial solution to the problem.

Formal evaluation of structured methods is also rendered difficult by the "craft" nature of the knowledge they embody. Because many of the procedures recruit personal knowledge held by the assessor, speech system assessments will invariably exhibit considerable variation. Such variation would tend to mask differences in performance contingent upon the use of a particular set of procedures. Variations in performance will tend to increase with the variability in assessors and with the complexity of the assessment they undertake. Such sources of masking variability are controllable, for example, by selecting and training "assessor subjects" carefully, to maximise homogeneity in the sample; and by comparing the performance of only simple assessment tasks. However, these interventions tend to reduce the "ecological validity" of the test.

In the light of these observations, a complete validation of a structured method, based upon comparitive experimentation, would be difficult to operationalize and, in practice, may not be feasible. The further evaluation of SIAM might usefully follow a strategy of smaller scale testing, the results of which, when viewed together, would more fully reveal the method's support for procurement; e.g.

(a) further informal testing with more representative subjects, tasks and target devices, leading to further procedural enhancement;

(b) experimental studies addressing critical parts of the method (e.g. hierarchical task representation; recruitment of diagnostics for simulation design) carried out under controlled conditions, comparing performance using later and early versions of the procedures;

(c) documented case studies of the method in use in procurement projects, which might be evaluated by independent HF experts against their criteria of "good practice".

12.3 General contribution of the research

12.3.1 General conclusions of the trial of SIAM

RSRE's specific problem is an instance of the more general problem of the ineffectiveness of the procurement of technologically advanced systems (Jordan et al, 1988). The problem is attributable in part to the inadequate provision of HF evaluation techniques applicable prior to system specification, and to the inadequacy of substantive knowledge to support the practice of HF by non-specialists. The trial of SIAM demonstrated that structured evaluation methods can support individuals who lack specialist knowledge in the conduct of early HF evaluations in domains with which they are not directly familiar. Hence, such methods potentially increase the effectiveness of procurement. The research has, then, demonstrated the potential for extending the notion of the structured method beyond the task of systems analysis and design, complementing the development, elsewhere, of SADMs which take account of HF concerns (e.g. Lim et al, 1990).

Interestingly, Walsh et al (1989) explicitly state that the benefits of SADMs do not include the making of design decisions, i.e. SADMs do not make available substantive design information to practitioners. This preclusion implies that SADMs are not claimed to support practitioners who do not possess personal design knowledge (i.e. "casual practitioners" of design). SIAM sought to extend the application of structured methods to users without domain knowledge, by the provision of substantive knowledge intended to be compatible with the procedures of the method. Although not demonstrated conclusively, it would appear that casual practitioners would not be supported as intended. The findings of the the trial suggest, then, that, while the research succeeded in extending the benefits of SADMs to the novel domain of HF evaluation, it did not extend the benefits to include support for casual practitioners.

12.3.2 Implications for the procurement of computer systems

Procurement was conceived in Chapter 2 as a process to establish work systems meeting organizational requirements. Distinction was drawn between procurement to meet novel and/or large scale requirements (where there is likely to be an intimate relationship between the procurer and the product developer), and procurement to meet a common requirement (which might be met by the acquisition of devices "off the shelf"). Military procurement falls into the former class, being conducted as a phased process, the stages of which bear a systematic (superordinate) relationship with the stages of product development. Evidence cited by Jordan et al suggest that some military procurement projects have failed as a consequence of an inadequate allocation of resources at the stage of feasibility assessment. Jordan et al particularly emphasise the importance of empirical feasibility assessments to establish subsequent technical risk in procurement projects.

The research described in this thesis cannot claim to have contributed substantial evidence to prove or disprove Jordan et al's contentions. To achieve such proof would, in principle, demand long term studies in which the methods used at the feasibility stage were related to the quality of procured systems. However, the results of the operational trial of SIAM tend to be concordant with the views of Jordan et al, who argue in favour of empirical feasibility studies, with the following observation:

> "..We found that projects typically reveal their technical difficulty only when hardware is built and tested and when integration of sub-systems is attempted. Earlier judgement of the technically feasible based on paper studies, modelling or extrapolation appears over-optimistic".

Where behavioural models are incomplete, provided it is possible to instantiate system elements and so reproduce system behaviour, empirical studies will likely offer more accurate assessments of system performance. The thesis is able, then, to offer theoretical support for Jordan et al's argument in favour of experimental studies of system feasibility. Such experimental studies of feasibility would likely be effective, not only in military procurement, but also in the procurement of behaviourally complex systems in civilian domains.

12.3.3 Implications for the evaluation of human-computer systems

In Chapter 4, evaluations were distinguished by their products (i.e. whether they are statements of presentation or of diagnosis); by their criteria; and by their processes. The evaluations supported by SIAM are intended to be diagnostic, and they assume the criteria of task quality and user costs incurred in achieving desired quality. Contrasts were drawn between analytic and empirical processes: the absence of adequate models of speech-based human-computer interaction precluded the former, so SIAM assumed an empirical approach to evaluation. Simulations were demanded if the method was to be applicable in advance of system implementation. Experience gained in the development of SIAM might, then, extend to other contexts demanding diagnostic usability assessment.

The test of SIAM demonstrated a role for induction in system evaluation which had not been recognized at the outset of the development of SIAM. The strategy of specifying simulations before implementing them may be difficult, if not impossible, to apply where the simulation is to support inductive processes (see also Section 12.2.3). The present work suggests, then, that pre-specification of simulations might be exploited further to support hypothetico-deductive methods but not inductive methods.

The applicability of the human simulation (Wizard of Oz) technique recruited to the method has been shown to be limited in applicability to the simulation of target devices whose behaviour it is within a system subject's ability to reproduce. In the case of speech interfaces, device behaviour is, itself, an emulation of human behaviour (speech communication); thus, the likelihood of a person being able to achieve the requisite behaviour to support a simulation is relatively high. The Wizard of Oz technique is likely to be, then, particularly appropriate as a means of implementing simulations of devices having the underlying rationale of reproducing human behavioural characteristics. Other researchers (e.g. Diaper and Warren) have utilized the technique for the simulation of expert systems (which seek to emulate the behaviour of the human expert); and the technique has been proposed by Life and Long (1987) as a means of simulating user interfaces for the human supervision of semi-autonomous robots (which seek to emulate human manipulatory skills).

While potentially suitable for simulating "intelligent" devices, one might speculate that the technique would likely fail to provide adequate reproduction of machines which have been designed to support tasks whose performance is limited by *"inadequacies"* of human behaviour. Examples of such tasks might be large scale rapid data processing and tasks involving highly repetitive activities. Other implementation techniques would be expected to be more suitable for the simulation of devices in these cases.

12.3.4 General implications for supporting the practice of HF

Chapter 3 characterized the discipline knowledge of HF as predominantly craft knowledge. As such the practice of HF is conducted with some success by experienced specialists, but less successfully by individuals lacking experience. Existing sources of substantive discipline knowledge (e.g. HF guidelines) are incomplete and difficult to apply, but generalizable methods were identified as a means of addressing HF problems which overcome some of these limitations.

Structured methods are one means by which practitioners might be supported in the conduct of HF evaluations. These methods enhance the effectiveness of evaluation by rendering the process systematic and complete; but they offer no performance guarantees. As suggested in Chapter 5, potentially more powerful support would lie in engineering methods, recruiting principles based upon prescriptive theories (Dowell and Long, 1989). However, given the present incompleteness and incoherence of existing knowledge of human-computer interaction, the HF engineering method remains a hypothetical notion.

Structured HF evaluation methods might be viewed as intermediate steps in progress towards HF engineering methods for use in system development. A structured method for empirical evaluation, unsupported by substantive domain knowledge, might constitute the first such step. The benefits derived from the use of a method of this kind would primarily be those of systematic and complete coverage of the problem. SIAM sought to progress one step further, by the inclusion of substantive knowledge recruitable by its procedures. Being based upon incomplete and poorly validated knowledge, such "knowledge-based" empirical methods could still not guarantee evaluation performance, but they potentially offer the additional benefit of applicability by individuals lacking specialist knowledge.

A further step might be the development of structured methods for *analytic* evaluation. Even at the conceptual level of Figure 6.1, SIAM is specifically orientated toward the development of simulations and their utilization within empirical context. However, given adequate interaction models and small modifications, the basic architecture might be rendered appropriate for analytic evaluation. Figure 12.1 illustrates such an analytic variant, in which, rather than the development of simulations, analytic (descriptive) models of task, device and user are generated, and the models are convolved to predict

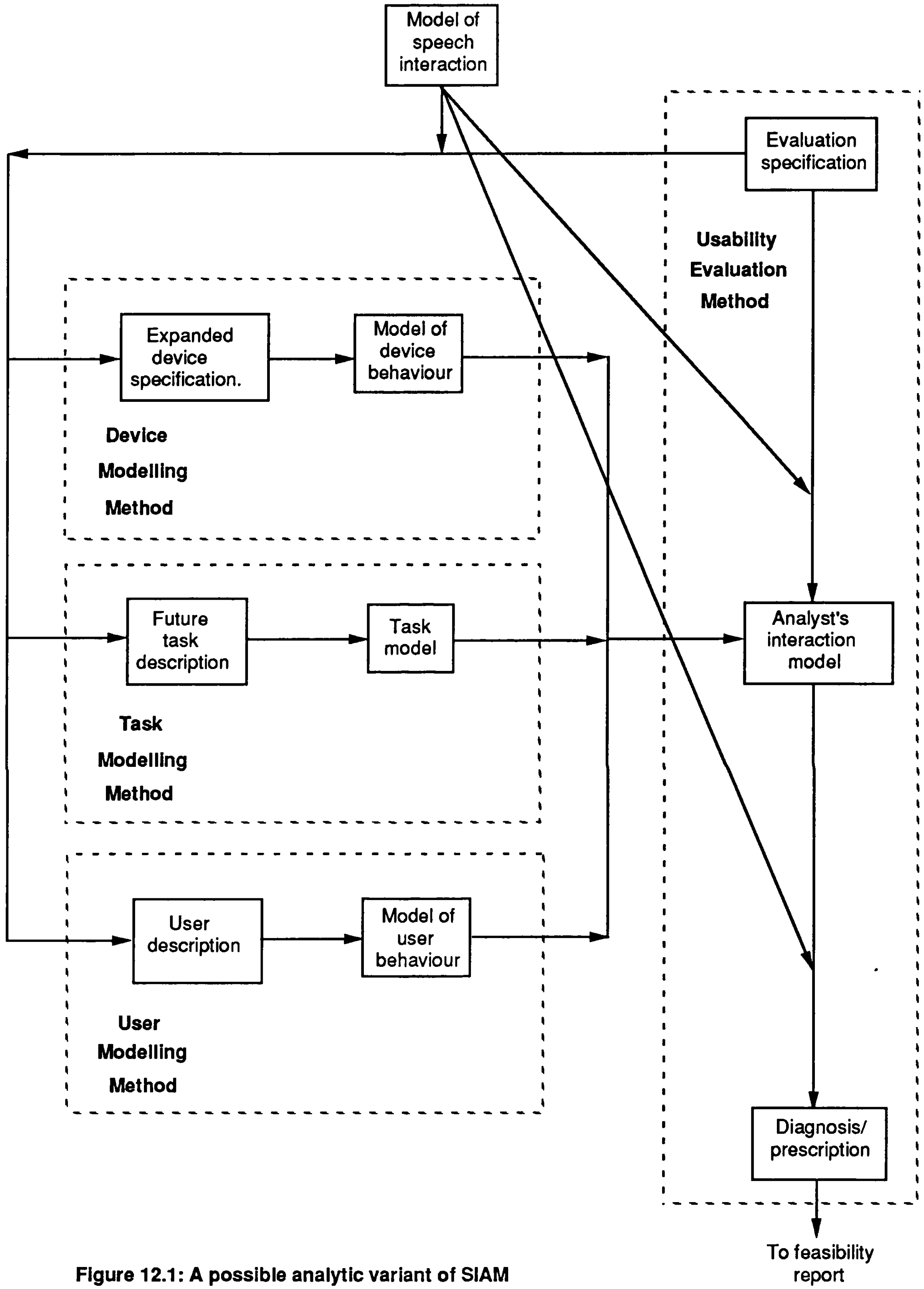


Figure 12.1: A possible analytic variant of SIAM

performance (instead of an experiment being conducted using simulations of the respective elements).

As the substantive discipline knowledge of HF is extended through research and practice, the provision of a knowledge-base adequate to support analysis will become increasingly feasible. Analytic structured methods would, potentially, offer the advantage of evaluation without the (substantial) cost of developing simulations. However, unless supported by a principled knowledge base, there would be, again, no guarantee attached to assessments. An engineering method would constitute the final step in this evolution, offering the assessor an assurance of task performance. At this stage, however, the feasibility of such a method remains unknown.

The attempt to render substantive domain knowledge compatible with a structured method was novel and ambitious. The failure of SIAM in this regard should not be taken to indicate that the development of structured methods for casual practitioners is, in principle, impossible. Rather, it indicates a requirement for further research addressing the decomposition of such knowledge, and its utilization in evaluation and prescription, as proposed earlier in this chapter. Such research could contribute to progress in the ultimate development of HF engineering principles.

The success of SIAM encourages the view that structured HF methods might be developed to support practice in non-speech domains. SIAM is specifically orientated to the evaluation of that class of human-computer system supported by speech communication. However, at the conceptual level expressed in Figure 6.1, SIAM may be viewed as an instance of a family of structured evaluation methods. Such methods would be potentially adaptable to the evaluation of other types of computerized system by the provision of bases of different evaluation knowledge; for example, a method might be developed for the evaluation of visual display formats, recruiting knowledge of visual perceptual organization.

At this level of description, the generality of the family of empirical evaluation methods would be defined by the set of all human-computer systems and by the set of (even informal and rudimentary) evaluation knowledge pertaining to the behavioural interactions between the entities of the work systems. At this level, then, the success of SIAM is promising for the viability of other, similarly structured evaluation methods. However, the declarative (evaluation) knowledge and the procedural knowledge of the various sub-methods of SIAM are not independent, and *details* of the process and notation of SIAM would be less generalizable. For example, the comments of the subject of the operational trial suggest that the representation of some tasks may be difficult within the notation prescribed by the task simulation method (a tree structure). Such notations would appear less appropriate where tasks are conducted through the expression of behaviour less easily identifiable as discrete actions, e.g. continuous control tasks. SIAM in its proceduralized form, then, is likely to be restricted in the extent to which it may be exploited in other contexts.

12.4 Evaluation of the contribution of the research

The thesis began by identifying a requirement for HF discipline knowledge which is applicable in the development of human-computer work systems. Discipline knowledge may be substantive (e.g. knowledge of human-machine interactions), or methodological (e.g. knowledge of processes by which dicipline problems may be solved). The work described here has sought to make a methodological contribution. However, if it is to be applicable, knowledge of either sort must be accessible to those engaged in discipline practice. The particular contribution has been in the attempt to provide methodological knowledge in a form accessible to a specific class of practitioner (i.e. practitioners lacking general discipline knowledge) engaged in a specific class of task (HF evaluation in procurement).

To attempt to develop a structured HF evaluation method for speech interfaces was ambitious in at least three ways: firstly, there had previously been no generally recognized complete method for evaluating speech interfaces; secondly, although structured methods existed for computer systems analysis and design, there had been no previous attempt to embody substantive discipline knowledge, and no structured methods extended to the process of HF evaluation; and, thirdly, the *development* of methods was itself poorly conceptualized.

The work described here was, then, pioneering, and its contribution should be evaluated in this light. No claim is made that the processes underlying the various sub-methods of SIAM are novel: experimentation is widely established in ergonomics practice, as are techniques such as hierarchical task analysis and human simulation. Rather, the novelty lies in the way that the processes have been explicitly conceptualized and proceduralized, and in the enterprise of specifying, implementing and evaluating a novel type of method.

The method offered is not claimed to be perfect, but it has been demonstrated to have utility and to offer potential for further exploitation. Structured evaluation methods should complement SADMs in encouraging the wider consideration of HF concerns in system development and procurement. The most important weakness of SIAM - its failure to provide a successful operationalization of substantive HF knowledge for uptake by the procedures - is itself instructive. The failure exposes a problem (and one likely to be faced more widely) of expressing substantive HF discipline knowledge in a form readily usable by practitioners. It is hoped that the inadequacies exposed here might be instructive to others engaged in HF engineering research.

There would appear to have been no previous attempts to document the development of a practical method from specification of requirements through to post-implementation evaluation; (at least, not in the domain of computer system development). In attempting to cover all stages of the process, the research has been variable in its quality. The conceptualization of the method with respect to its requirement is relatively strong. However, the evaluation has been recognized as being incomplete, both in its breadth - some parts of the method were not fully evaluated - and in its depth - the evaluation could not claim the status of a validation. In defence of the strategy chosen, the field of enquiry was novel, and it was appropriate that an attempt be made to address the problem in total.

In summary, then, the research described here is offered as a pioneering effort, demonstrating the potential of structured HF evaluation methods and identifying requirements for their further development.

REFERENCES

Annett, J., Duncan, K.D., Stammers, R.B. and Gray, M.J. (1971) Task Analysis. HMSO Training Information Paper

Baber, C. and Stammers, R.B. (1989) Is it natural to talk to computers? An experiment using the "Wizard of Oz" technique. In E. Megaw (ed.) "Contemporary Ergonomics 1989", London: Taylor and Francis, 234-239, (1989)

Barr, A. and Feigenbaum, E.A. (1981) The Handbook of Artificial Intelligence Volume 1. London: Pitman (1981).

Bell, D.W. and Becker, R.W. (1983) Designing experiments to evaluate speech I/O devices and aplications. Speech Technology, Jan./Feb. 1983, 70-79.

Bellotti, V. (1988) Implications of current design practice for the use of HCI techniques. In D.M. Jones and R. Winder (eds) "People and Computers IV", Proceedings of HCI'88 Cambridge: Cambridge University Press.

Bellotti, V.M.E. (1990) A framework for assessing applicability of HCI techniques. In D. Diaper et al (eds.) "Human-Computer Interaction - INTERACT '90" Noth-Holland: Elsevier Science Publishers, 213-218, (1990).

Blair, W.D. (1981) Opportunities for voice input and output in military applications. Unpublished report; Royal Signals and Radar Establishment, U.K.. Reference RSRE ZLA/180/02, (1981).

Bubenko, J.A. jr. (1986) Information system methodologies - a research view. In T.W. Olle, H.G. Sol and A.A. Verrijn-Stuart (eds) "Information System Design Methodologies: Improving the practice". Amsterdam: Elsevier Science Publishers. 289-318.

Buckley, P. & Long, J. (1985) Effects of system and knowledge variables on a taskcomponent of "Teleshopping". In P. Johnson and S. Cook (eds) "People and Computers: Designing the Interface". Cambridge: Cambridge University Press.

Cameron, J.R. (1986) An overview of JSD. IEEE Transactions on Software Engineering, SE-12, 2, 222-240.

Card, S.K., Moran, T.P. and Newell, A. (1983) The Psychology of Human-Computer Interaction. Hillsdale, N.J.: Lawrence Erlbaum Associates (1983)

Carroll, J.M. and Campbell, R.L.. (1989) Artifacts as psychological theories: The case of human-computer interaction. Behaviour and Information Technology, 8, 247-256.

Carver, M.K. (1988) Practical experience of specifying the human-computer interface using JSD. In E. Megaw (ed.) "Contemporary Ergonomics 1988", London: Taylor and Francis, 177-182, (1988).

Chapanis, A. (1975) Interactive human communication. Scientific American, 232(3), 36-42, (1975).

Clark, P.E.J. (1986) Implementation of voice input to computer applications. Journal of Information Technology, 1, 39-56

Colbert, M., Green, D.W. and Long, J.B. (1990) The development of methods for the design of computer systems: A case history and some heuristics. In E.J. Lovesey (ed.) "Contemporary Ergonomics 1990", London: Taylor and Francis, 44-49, (1990)

Conrad, R. and Hull, A.J. (1968) The preferred layout for numeral data entry keysets. Ergonomics, 11, 2, 165-173.

Cooke, N. (1990) RAE Bedford's experience of using direct voice input in the cockpit. In Proceedings of Voice Systems Worldwide 1990. New York: Media Dimensions, 135-144.

Cooper, M.B., 1987 Voice input/output - the human factors issues. British Telecom Technology Journal, 5, (3) July 1987.

Damper, R.I. (1988) Practical experiences with speech data entry. In E. Megaw (ed.) "Contemporary Ergonomics 1988", London: Taylor and Francis, 92-97, (1988).

Dew, A.M., Hardwick, B.A., Roach, P.J., Shirt, M.A. and Kirby, H.R. (1986) Voice degradation problems in using automatic speech recognisers. IEE International Conference on Speech Input/Output; Techniques and Applications. 24-26 March,1986. 319-323

de Souza, F.L., Long, J.B. and Bevan, N. (1990) Types of error and difficulty in using human factors guidelines: the case of interface menu design. In E.J. Lovesey (ed) "Contemporary Ergonomics 1990", London: Taylor and Francis, 340-346.

Diaper, D. (1986) Identifying the knowledge requirements for an expert system's natural language processing interface. In M. Harrison and A. Monk (eds.) "People and Computers: Designing for Usability", Cambridge: University Press, pp263-280.

Diaper, D. (1989) Task Analysis for Human-Computer Interaction. Chichester: Ellis Horwood (1989).

van Dijk, T.A. (1980) Macrostructures. An interdisciplinary study of global structures in discourse, interaction and cognition. New Jersey: Lawrence Erlbaum Associates. (1980).

Doddington, G.R. and Schalk, T.D. (1981) Speech recognition: Turning theory into practice. IEEE Spectrum, 18.9, 26-32, (1981).

Dowell, J. & Long, J. (1988) Towards a paradigm for human computer interaction engineering. In E. Megaw (ed) "Contemporary Ergonomics 1988", London: Taylor & Francis. 45-50.

Dowell, J. and Long, J. (1989) Towards a conception for an engineering discipline of human factors. Ergonomics, 32, 11, 1513-1535

Dye, R., Arnott, J.L., Newell, A.F., Carter, K.E. and Cruickshank, G. (1990) Simulating the speech-operated user interfaces of the future: The case of listening typewriters. In M.A. Life, C.S Narborough-Hall and W.I Hamilton (eds.) "Simulation and the User Interface", London: Taylor and Francis, 159-168, (1990).

Edmondson, D. and Johnson, P. DETAIL: An approach to task analysis. In M.A. Life, C.S Narborough-Hall and W.I Hamilton (eds.) "Simulation and the User Interface", London: Taylor and Francis, 147-158, (1990).

Emshoff, J.R. and Sisson, R.L. (1970) "Design and Use of Computer Simulation Models". London: Macmillan (1970)

Fairley, R. (1986) "Software Engineering Concepts". McGraw Hill (1986).

Floyd, C. (1986) A comparative evaluation of system development methods. In T.W. Olle, H.G. Sol and A.A. Verrijn-Stuart (eds) "Information System Design Methodologies: Improving the practice". Amsterdam: Elsevier Science Publishers. 19-54.

Fraser, N.M. and Gilbert G.N. (in press) Simulating speech systems. Accepted for publication in *Computer Speech and Language.*

Funk, K. & McDowell, E. (1982) Voice input/ouput in perspective. Proceedings of the Human Factors Society 26th. Annual Meeting. 218-222.

Gardiner, M. and Christie, B. (1987) Applying Cognitive Psychology to User Interface Design, Chichester: John Wiley (1987)

Gill, K.D. (1990) The role of speech systems within the automotive quality assurance function. In Proceedings of Voice Systems Worldwide 1990. New York: Media Dimensions, 32-45.

Gould, J.D., Conti, J. & Hovanyecz, T. (1983) Composing letters with a simulated listening typewriter. Communications of the ACM, 26, 295-308

Hammond, N.V. and Allinson, L. (1989) Development and evaluation of a CAL system for non-formal domains: The hitch-hiker's guide to cognition. Computers and Education, 12, 215-220, (1988).

Hartson, R.H. and Hix, D. (1989) Towards empirically derived methodologies and tools for human-computer interface development. International Journal of Man-Machine Studies, 31, 477-494, (1989).

Hermann, C.F. (1967) Validation problems in games and simulations with special reference to models of international politics. Behavioural Science, 12, 216-231.

Hick, W.E. (1952) On the rate of gain of information. Quarterly Journal of Experimental Psychology, 4, 11-26, (1952)

Jackson, M.A. (1982) System Development. Englewood Cliffs, NJ: Prentice Hall (1982).

Jackson, M.A. (1987) Course notes: JSD Realtime and Embedded Systems Course, Michael Jackson Systems Limited, September, 1987.

Johnson, P. (1985) Towards a task model of messaging: An example of the application of TAKD to user interface design. In P. Johnson and S. Cook (eds.), "People and Computers: Designing the Interface". Cambridge: Cambridge University Press, 46-62, (1985).

Jones, A.H. (1989) The use of structured design methods and tools for information systems development. Design Studies,, 10, 135-142, (July, 1989).

Jones, D., Hapeshi, K. and Frankish, C. (1989) Design guidelines for speech recognition interfaces. Applied Ergonomics, 20, 1, 47-52.

Jordan, G., Lee, I. and Cawsey, G. (1988) Learning from experience: A report on the arrangements for managing major projects in the Procurement Executive. Report to the Minister of State for Defence Procurement, Ministry of Defence, U.K. (1988).

Kieras, D.E. and Polson, P.G. (1985) An approach to the formal analysis of user complexity. International Journal of Man-Machine Studies, 22, 365-394, (1985).

Knight, J.A. & Peckham, J.B. (1984) A generic model for the assessment of speech input applications. Report: Logica Space and Defence Systems Ltd. Reference 93.4628.

Laver, J. (1987) New horizons in European speech technology. Report of the ESPRIT Workshop on Speech Technology, Aarhus, Denmark. 21-22 May, 1987.

Levene, P. (1987) Competition and collaboration: U.K. defence procurement policy. Journal of the Royal United Services Institute for Defence Studies, 132, 2, 3-6.

Life, M.A. (1987) RSRE/UCL Speech Technology Assessment Project: the analysis of battlefield data communication tasks (Issue 2). Ergonomics Unit, University College London report reference MACL/R/31-34/86/2. February, 1987.

Life, M.A. (1990)
A conceptualization of simulation for user interface development. In M.A. Life, C.S Narborough-Hall and W.I Hamilton (eds.) "Simulation and the User Interface", London: Taylor and Francis, 31-42, (1990).

Life, M.A. and Lee, B.P. (1989)
Evaluation of speech data entry and feedback requirements for a computer supporting indirect weapon engagements. Unpublished Ergonomics Unit report reference MACL/R/18/89/1, (1989).

Life, M.A., Lee, B.P. and Long, J.B. (1988)
Assessing the usability of future speech technology: towards a method. In W.A. Ainsworth and J.N. Holmes (eds) "Proceedings of SPEECH '88: Seventh FASE Symposium." Edinburgh: Institute of Acoustics (1988), 1297-1304.

Life, M.A., Long, J.B. and Lee, B.P. (1988)
Human simulation of speech technology: an illustration of the ergonomic approach. In E. Megaw (ed) Contemporary Ergonomics 1988, Proceedings of Ergonomics Society's Annual Conference, Manchester, 1988. London: Taylor and Francis.

Life, M.A. & Long, J.B. (1987)
The use of people to simulate machines: an ergonomic approach. in E. Megaw (ed) Contemporary Ergonomics 1987 "Ergonomics Working for Society", Proceedings of the Ergonomics Society's Annual Conference, Swansea, 1987. London: Taylor and Francis.

Lim, K.Y., Long, J.B. and Silcock, N. (in press)
Integrating human factors with structured analysis and design: From conception to an Extended Jackson System Development Method. Accepted for publication in Ergonomics.

Lim, K.Y., Long, J.B. and Silcock, N. (1990) Requirements, research and strategy for integrating human factors with structured analysis and design methods: The case of the Jackson System Development method. In E.J. Lovesey (ed.) "Contemporary Ergonomics 1990", London: Taylor and Francis, 38-43, (1990)

Long, J. (1986) Problems in the control of motor output during multi-modal dialogues involving voice. NATO Conference: "Structure de Dialogues Multimodaux a Composante Orale", Venaco, France. 1-5 September, 1986.

Long, J. and Dowell, J. (1989) Conceptions of the discipline of HCI: craft, applied science and engineering. In Sutcliffe, A. and Macaulay, L. (eds) "People and Computers V". Cambridge: University Press. 9-32

Madison, R.N. (1983) Information Systems Methodologies. London: Wiley Heydon Ltd. (on behalf of the British Computer Society). (1983)

Marics, M.A. and Williges, B.H. (1988) The intelligibility of synthesized speech in data enquiry systems. Human Factors, 30, 719-732, (1988).

Martin, G.L. (1989) The utility of speech input in user-computer interfaces. International Journal of Man-Machine Studies, 30, 355-375, (1989).

Martin, J. (1980) The division of attention between primary and secondary tasks in a simulated ships-gunfire control task. Unpublished MSc dissertation. University of London.

Meister, D. (1986) Human Factors Testing and Evaluation. Elsevier Science Publishers (1986).

Ministry of Defence (1987) Compendium of guidelines for project management. Amendment no. 15. DPP(PM), Procurement Executive, U.K. Ministry of Defence (1987).

Moore, R.K. and Bridle, J.S. (1986) Speech research at RSRE. Proceedings of the Institute of Acoustics, 8, 257-263, (1986)

Moran, T.P. (1981) The command language grammar: A representation for the user interface of interactive computer systems. International Journal of Man-Machine Studies, 15, 3-50, (1981).

Morel, M-A. (1986) Computer-human communication. Paper presented at NATO Conference: "Structure de Dialogues Multimodaux a Composante Orale", Venaco, France. 1-5 September, 1986.

Mountford, S.J., North, R.A., Metz, S.V. and Warner, N. (1982) Methodology for exploring voice-interactive avionics tasks: optimizing interactive dialogues. In Proceedings of the Human Factors Society 26th. Annual Meeting, 207-211, (1982).

Nakagawa, S. and Ohguro, Y. (1988) Comparison of parsing methods on continuous speech recognition. In W.A. Ainsworth and J.N. Holmes (eds.) "Proceedings of Speech '88", Edinburgh: Institute of Acoustics, 369-376, (1988).

Newell, A.F., Arnott, J.L., Carter, K. and Cruickshank, G. (1990) Listening typewriter simulation studies. International Journal of Man-Machine Studies, 33, 1-19 (1990)

Parsons, H.McI. (1972) Man-Machine System Experiments. Baltimore: The Johns Hopkins Press (1972).

Payne, S.J. and Green, T.R.G. (1983) The user's perception of the interaction language: A two-level model. Proceedings of the CHI'83 Conference on Human Factors in Computer Systems, New York: ACM, 202-206, (1983).

Pisoni, D.B. (1986) Automatic measurement of speech recognition performance: A comparison of six speaker dependent recognition devices. Unpublished report: Contract nos. 435114 and 562010, IBM Corpooration, T.J. Watson Research Centre, Yorktown Heights, New York, (1986).

Power, R.C., Hughes, R.D. and King, R.A. (1986) Verification archetype updating and token set selection as a means of improving the performance of menu-driven isolated-word recognition systems using Time Encoded Speech descriptions in high acoustic noise backgrounds. In IEE International Conference on Speech Input/Output; Techniques and Applications. London, March, 1986; pp 144-151.

Pratt, R.L. (1986) On the intelligibility of synthetic speech. In Proceedings of the Institute of Acoustics Volume 8, (1986).

Pratt, R.L. (1987) Quantifying the performance of text-to-speech synthesizers. Speech Technology, March/April 1987, 54-64.

RARDE (1983) Code of practice for the application of human factors to the design of military ADP systems. Report attributed to RARDE MA3 Branch.

Richards, M.A. & Underwood, K. (1984) Talking to machines: how are people naturally inclined to speak? In E.D. Megaw (ed) "Contemporary Ergonomics 1984". London: Taylor and Francis.

S.A.M. Partnership (1988) Multilingual speech assessment methods. In W.A. Ainsworth and J.N. Holmes (eds.) "Proceedings of Speech '88", Edinburgh: Institute of Acoustics, 137-143 (1988).

Silcock, N., Lim, K.Y. and Long, J.B. (1990) Requirements and suggestions for structured analysis and design (human factors) method to support the integration of human factors with system development. In E.J. Lovesey (ed.) "Contemporary Ergonomics 1990". London: Taylor and Francis. 425-430.

Simpson, C.A., Coler, C.R. and Huff, E.M. (1982) Human factors of voice I/O for aircraft cockpit controls and displays. Workshop on Standardization for speech I/O technology. National Bureau of Standards, Gaithersburg, Maryland, USA. 18-19 March, 1982. 159-166.

Simpson, C.A., McCauley, M.E., Roland, E.F., Ruth, J.C. and Williges, B.H. (1985) System design for speech recognition, Human Factors, 27, 115-141, (1985).

Simpson, C.A. and Ruth, J.C. (1987) The phonetic discrimination test for speech recognizers - Part 1. Speech Technology, March/April 1987, 48-53. (Part 2 in subsequent issue).

Smith, S.L. and Mosier, J.N. (1986) Guidelines for Designing User Interface Software. Bedford Massachusetts: Mitre.

Stammers, R.B., Carey, M.S. and Astley, J.A. (1990) Task analysis. In J.R. Wilson and E.N. Corlett (eds.) "Evaluation of Human Work: A Practical Ergonomics Methodology". London: Taylor and Francis.

Sommerville, I. (1985) Software Engineering (2nd. Edition). Wokingham, U.K.: Addison-Wesley (1985)

Talbot, M. (1985) Speech technology: Is it working? In P. Johnson and S. Cook (eds.) "People and Computers: Designing the Interface", Cambridge, U.K.: Cambridge University Press, 345-358, (1985).

Talbot, M. (1987) Reaction time as a metric for the intelligibility of synthetic speech. In Waterworth, J.A. and Talbot,M. (eds.) "Speech and Language-Based Interaction with Machines: Towards the Conversational Computer". Chichester, U.K.: Ellis Horwood, (1987), pp 66-88.

Tsang, P.S., Hart, S.G. & Vidulich, M.A. (1986) The effects of display-control I/O compatibility, and integrality on dual-task performance and subjective workload. AGARD Conference "Information Management and Decision-making in Advanced Airborne Weapon Systems". Toronto, Canada, 14-18 April,1986.

Visick, D., Johnson, P. and Long, J. (1984) A comparative analysis of keyboards and voice recognition in a parcel sorting task. In E.D. Megaw (ed) "Contemporary Ergonomics 1984". Taylor and Francis: London. 56-61.

Voiers, G. (1983) Evaluating processed speech using the Diagnostic Rhyme Test, Speech Technology, 1, 4, 30-39 (1983).

Walsh, P.A., Lim, K.Y. and JSD and the design of user interface software. Ergonomics, 32, 11, 1483-1498.

Ward, J.W.D. and Turner, G.N. (1982) Military Data Processing and Microcomputers. Brassey's Publishers Ltd.: Oxford.

Warren, C.P. (1985) The Wizard of Oz Technique: a comparison between natural and command languages for communicating with expert systems. Unpublished MSc dissertation, University of London.

Waterworth, J.A. (1982) Man-machine speech "dialogue acts". Applied Ergonomics, 13, 203-207.

Waterworth, J.A. and Talbot, M. (1987) Speech and Language-Based Interaction with Machines: Towards the Conversational Computer. Chichester, U.K.: Ellis Horwood, (1987).

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

Wilpon, J.G. & Roberts, L.A. (1986) The effects of instructions and feedback on speaker consistency for automatic speech recognition. IEE International Conference on Speech Input/Output; Techniques and Applications. 24-26 March, 1986.

Whitefield, A., Wilson, F. and Dowell, J. (1991) A framework for human factors evaluation. Behaviour and Information Technology, 10, 65-79 (1991)

Williges, B.H. et al.(1986) Using speech in the human-computer interface. In R.W. Ehrich & R.C. Williges (Eds.), Advances in human factors/ergonomics. Elsevier Science Publishers B.V., 1986.

Wilson, M.D., Barnard, P.J. & MacLean, A. (1988) Task analysis in human-computer interaction. In T.R.G. Green, J.M. Hoc, D. Murray and G. Van Der Veer (eds) "Working with computers: theory versus outcome". London: Academic Press, pp47-87.

APPENDIX A

DIAGNOSTIC TABLES

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 1.1 Knowledge required to operate the speech entry device is different from that required to operate the system using alternatively available data entry devices, or to perfrom the task manually.	Device actions	Operating procedure in alternative modes	Skill operating in alternative modes			High probability of operating error (OR evidence that operator is attempting to avoid errors, e.g. by performing slowly) following a change in operating mode (either from speech to the alternative, or vice versa)	Semantic and syntactic error rate. Transaction time following mode change.	*Language incompatibility.* Most recently/frequently used knowledge sources tend to dominate less recently/well established knowledge sources.	Ensure knowledge (e.g. dialogue features) for operating speech interface is compatible with knowledge users already have for doing the task manually or with a non-speech interface	See diagnostics 1.2, 1.4
Diagnostic 1.2 Vocabulary and syntax used to enter data by means of speech is different from that normally used in the working environment.	On-line and off-line actions requiring use of language	Device language constraints	Users' facility with device language and other languages used in the task context.	Speech/language used for off-line task components.		High probability of lexical and/or syntactical errors, characterised by the introduction of language features employed for off-line tasks into computer/user dialogue.	Syntactical and lexical error rate in device operation.	*Language incompatibility.* Tendency for more highly learned responses to be elicited than less well-established ones.	IF device operating language may be modified closer to that used in the working environment without violation of requirements for phonetic dissimilarity, then modify it ELSE give specific training in device language AND/OR encourage change of other language closer to that of device. Use observed error characteristics to guide intervention.	See diagnostics 1.1, 2.1 Supporting documents: Jones et al, 1985
Diagnostic 1.3 Information to be entered is multi-dimensional (e.g. spatial information) and expressed on a continuous scale (e.g. spatial position)	Data entry actions	Modality and format in which information is accepted by the device			Dimensionality of data. Scaling of data. Time constraints on transactions.	Precision with which continuously-scaled infomation can be transmitted using speech is a function of time (e.g. precision of a geographical co-ordinate is determined by the number of figures in the grid reference). As required precision increases, transmission will take longer and/or will require more effort and/or will run an increased rate of error by comparison with multi-dimensional continuously scaled data entry devices (e.g. pointing devices).	Precision of entered data. Time to enter data to critical precision.	*Coding incompatibility.* 1. Positive relationship between precision and message length when information is transformed from a continuous to a discrete scale. Note also that the message length is multiplied by the number of dimensions of variation. 2. Effort is required to make transformations in 1.	Only use speech for entering continuously scaled and/or multi-dimensional information if a low level of precision is required or if there is low time pressure. OR train users in recoding. (Spatial pointing devices may be more appropriate alternatives.)	Supporting documents: Funk & McDowell, 1982

TABLE A.1: DIAGNOSIS OF KNOWLEDGE INCOMPATIBILITY
(SPEECH DATA INPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 1.4 The size of the chunk of information after which feedback is presented is different from that used elsewhere in the task	Actions demanding manipulation of information to be entered into the computer	Device constraints on chunking of data Feedback chunk characteristics Recognition error frequency Availability of aids to memory (e.g. facility for writing down data to be entered)	Familiarity with task data	Factors disrupting memory processes e.g.: - psychological stress - noise - visual distraction	Constraints on structure of task information	Trade-off between: (1) Performance disruption by feedback presented in units smaller than those preferred by users for the representation of domain information; and (2) Performance disruption by cognitive load imposed by having to correct errors at early locations in a long entry sequence	Transaction time Subjective assessment of cognitive effort to transform data to preferred format. Mistakes in correction of device errors as function of location of error in the message	(1) Expression of an utterance will be facilitated if the speech representation is compatible with the user's mental representation of the information (2) Error detection depends upon retaining representations of intended and actual data entries. The longer the string, the greater the probability of failing to detect early/ multiple errors (3) Representations in working memory decay over time.[The rate of decay may differ between the auditory and visual modalities.] (4) If speech is used for error correction, the complexity of the command is some function of the distance of the error from the current cursor position. This impacts the information processing load, both by the demand for formulating the correction command and by its interaction with (c) above.	Allow users to enter information in a form familiar to them OR provide extended training to users to familiarize them with the machine information structure.	See diagnostics 1.2, 3.1, 3.3, 4.3, 4.4 Supporting documents: Simpson et al, 1985
GENERAL PURPOSE DIAGNOSTIC FOR KNOWLEDGE INCOMPATIBILITY (where TD = target device) The knowledge held by users is inadequate or inappropriate to operate the TD OR knowledge users already hold interferes with TD operating knowledge OR users are unable to represent information in ways demanded by the TD	Actions during which potentially incompatible knowledge is used Actions involving creation or manipulation of information to enable above actions	Content/ structure/ medium of information presented by the TD Content/ structure/ medium demanded to enter information to TD Knowledge demanded for TD operation	TD operating knowledge Knowledge of the operation of similar devices Familiarity with the structure and content of information handled in the task	Content/structure /medium of information from non-device sources Content/structure /medium demanded for recording information other than to TD Knowledge demanded for operation of other devices in task context Environmental factors disrupting memory process	Temporal demands of task Order in which task actions are performed	Increased incidence of operating errors AND/OR increased transaction time AND/OR subjective reports of excessive cognitive load	Errors in operation of device Errors in performance of off-line activities Transaction time	Knowledge incompatibility between the TD and users, i.e. knowledge held by users is not compatible with that demanded by the TD to maintain desired performance		

TABLE A.1 (contd.): DIAGNOSIS OF KNOWLEDGE INCOMPATIBILITY
(SPEECH DATA INPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 2.1 Vocabulary demanded for data entry includes words which are phonetically similar to each other.	Data entry actions	Device vocabulary Tolerance of variance in form of input tokens.	User speech characteristics	Factors influencing speech variability, e.g. ambient noise level	Relative frequencies with which speech tokens have to be used	High incidence of device confusion errors, or evidence that user is exerting particular articulatory effort. Size of effect may be influenced by contextual factors and/or by differences between users	Confusion error rate Transaction time	Variability in operator utterances exceeds threshold for mapping unambiguously to computer representation of vocabulary (word templates).	Disambiguate by providing device with contextual knowledge where possible (i.e. use sub-vocabularies) AND/OR select phonetically dissimilar homonyms AND/OR train users to articulate more distinctively. Provide means of editing messages before transmission Utilize phonetic alphabet (Alpha, Bravo, Charlie etc.) if alphabetic tokens must be presented.	See diagnostics 1.2, 2.4, 4.2 and all diagnostics in Table 3 Supporting documents: Clark, 1986 Jones et al, 1985 Knight & Peckham, 1984
Diagnostic 2.2 Users required to perform computer and non-computer operations concurrently, where non-computer operations demand hands and eyes.	Device actions Actions interrupting, interrupted by or simultaneous with device actions	Location of device in workspace Operating procedure	Facility in use of device and in performing non-device actions	Constraints on actions imposed by context, e.g. workspace configuration; need to wear protective clothing; need to free hands to maintain balance etc.	Temporal demands of on- and off-line tasks Hands/eyes/speech requirements of task elements.	Mutual interference between on- and off-line operations. Operator may attempt to reduce by serial operations with rapid alternations between task elements. Increased probability of 'slips' Evidence of heavy demands on operators (fatigue, etc) Evidence that users have to physically re-orientate themselves when switching between task elements.	Transaction time Off-line task performance Operating errors of the "slip" variety	Limited operator resources (hands/eyes) Possible under-utilization of alternative resources (speech/ears)	Consider use of speech interface if data suitable.	See diagnostics 1.3, 2.4, 2.5 and all diagnostics in Table 3 Supporting documents: Simpson et al, 1982 Tsang et al, 1986
Diagnostic 2.3 Users do not have specialized device (e.g. keyboard) operating skills, or are physically unable to utilize such devices so as to meet criterial performance.	Device actions	Location of device in workspace Operating procedure	Facility in use of data entry device(s)	Constraints on actions imposed by context, e.g. workspace configuration; need to wear protective clothing; need to free hands to maintain balance etc.	Temporal demands of task (pacing).	Operator unable to utilize computer at required speed and accuracy.	Transaction time relative to alternative devices Incidence of slips in operation relative to alternative devices	Level of device operating skill inappropriate, inadequate or difficult to implement. Alternative skills might be utilized.	Consider use of speech interface if data suitable OR provide training in the use of alternative devices OR select operators with skill in the operation of alternative devices	See diagnostics 1.3, 2.4, 2.5 and all diagnostics in Table 3 Supporting documents: Jones et al, 1985

TABLE A.2: DIAGNOSIS OF BEHAVIOURAL INCOMPATIBILITY
(SPEECH DATA INPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 2.4 Users required to perform non-computer tasks that demand significant physical effort, such as lifting or handling heavy objects.	Device actions Actions requiring physical exertion	Device hardware configuration e.g. type of microphone and means by which it is mounted	Physical fitness in performing tasks such as lifting.	Workspace configuration .	Temporal demands of task (pacing).	High incidence of device recognition errors.	Transaction time Device recognition errors	Physical exertion modifies acoustic characteristics of speech signal, so that there is no longer a match with machine representation of vocabulary	Carry out the acquisition of training utterances under similar conditions that require physical effort Re-training of templates at intervals. Use those recognisers that can adapt their templates automatically.	Supporting documents: Knight & Peckham, 1984
Diagnostic 2.5 Users are required to make physical movements functionally related to, but temporally decoupled from, a data entry utterance	Data entry actions Functionally-related but temporally decoupled actions	Temporal characteristics of device operating procedure			Temporal constraints imposed by task Temporal relations between actions	Disruption of data entry or of functionally related action	Low level slips in data entry or in functionally related actions	Motor actions having a common mental representation tend to be performed simultaneously	Design the device so that speech input and associated actions can be performed simultaneously	Supporting documents: Long, 1986
Diagnostic 2.6 Information must be entered in conditions of low ambient illumination	Data entry actions	Device operating procedure	Skill in device operation	Ambient light levels	Requirements to enter data Temporal constraints imposed by task	Increased probability of errors in manual data entry	Data entry errors.	Operators require visual feedback of own actions with manual data entry devices	Consider use of speech interface if data suitable	See diagnostics 1.3, 2.4, 2.5 and all diagnostics in Table 3

TABLE A.2: DIAGNOSIS OF BEHAVIOURAL INCOMPATIBILITY (SPEECH DATA INPUT) - continued

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
GENERAL PURPOSE DIAGNOSTIC FOR BEHAVIOURAL INCOMPATIBILITY (where TD = target device)										
Users do not have the skills necessary to operate the TD OR users have other skills which interfere with their ability to operate the TD OR the task demands concurrent actions which interfere with one another	Device actions Actions concurrent with, interrupting or interrupted by device actions	Operating procedure Content/structure/medium of information on the TD Content/structure/medium demanded for entry of information to TD	TD operating skill Familiarity with the structure & content of information handled in the task Familiarity with the operation of competing devices	Workspace layout Contextual factors disrupting user's ability to access or operate display and controls of TD	Domain events initiating concurrent actions Temporal demands of the task	Transaction time is slow AND/OR increased errors of the "slip" variety AND/OR subjective reports of excessive physical or cognitive loads	Transaction time Low level errors device operation Performance of off-line task activities Subjective reports of workload	Behavioural incompatibility between the TD and users, i.e. the behaviour ("skill") of which the user is capable does not correspond with the behaviour demanded by the TD to maintain desired performance		

TABLE A.2: DIAGNOSIS OF BEHAVIOURAL INCOMPATIBILITY (SPEECH DATA INPUT) - continued

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 3.1 Information must be entered in a noisy environment.	Device data entry actions	Tolerance of device to spurious non-speech noises. Capability of elimination by microphone selection. Tolerance of device to varaibility in speech tokens .	Ability to speak consistently regardless variable noise.	Ambient noise level		Variable incidence of errors (false acceptances, incorrect rejection and substitution errors etc.) depending upon ambient noise conditions	Transaction time. Device recognition error frequency.	Speech waveform distorted by users who subconsciously alter speech output to compensate for changes in acoustic environment AND/OR spurious noise from environment triggering device.	Reduce acoustic noise in environment AND/OR provide users with hearing protection AND/OR utilize contact (throat) microphones AND/OR train users to maintain consistency in vocal output. Carry out the acquisition of training utterances under similar noisy conditions to those that will apply during use. Use those recognisers that can adapt their templates automatically.	See diagnostic 3.2 Supporting documents: Clark, 1986 Cooper, 1987 Knight & Peckham, 1984
Diagnostic 3.2 User subject to variable vibration.	Device data entry actions	Tolerance of device to variance in speech tokens.		Ambient vibration characteristics.		Variable incidence of errors (false acceptances, incorrect rejection and substitution errors etc.)	Transaction time. Device recognition error frequency.	Speech waveform distorted by ambient conditions due to tissue elasticity in vocal tract	Minimize operator vibration (e.g. by means of damped seating) Carry out the acquisition of training utterances under similar conditions to those that will apply during use. Use those recognisers that can adapt their templates automatically.	See diagnostic 3.1 Supporting documents: Knight & Peckham, 1984
Diagnostic 3.3 Users subject to extreme emotional stress.	Device data entry actions	Tolerance of device to variance in speech tokens.	Users' resistance to psychological stress	Factors eliciting stress	Time constraints on transactions.	Variable incidence of errors (false acceptances, incorrect rejection and substitution errors etc.) depending upon stress experienced by user	Transaction time. Device recogntion error frequency.	Speech waveform distorted by physiological changes in vocal tract caused by increased arousal.	Select individuals who remain calm. Use adaptive interface.	
Diagnostic 3.4 Data are to be entered for extended periods OR in dry/fume laden atmosphere.	Device data entry actions	Tolerance of device to variance in speech tokens.		Atmospheric conditions.	Temporal features of task.	Variable incidence of errors (false acceptances, incorrect rejection and substitution errors etc.) tending to increase with time on task AND/OR with exposure to the task environment	Transaction time. Device recogntion error frequency.	Speech wavefrom distorted due to inflamation/drying of tissues of vocal tract.	Use adaptive interface. Consider provision of respiratory protection Provide refreshments/breaks in operation.	Supporting documents: Dew et al, 1986 Jones et al, 1985 Knight & Peckham, 1984

TABLE A.3: DIAGNOSIS OF ENVIRONMENTAL INCOMPATIBILITY
(SPEECH DATA INPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
GENERAL PURPOSE DIAGNOSTIC FOR ENVIRONMENTAL COMPATIBILITY (where TD = target device)										
Environmental factors prevent or disrupt the application of knowledge or skills held by users in the operation of the TD	Device actions	Content/structure/medium of information on the TD Content/structure/medium demanded for entry of information to TD TD operating procedure	TD operating skill Protective clothing	Workspace layout Environmental factors disrupting TD-user interaction	Demands for exposure to environmental interference Temporal demands of the task	Transaction time is slow AND/OR increased incidence of TD operating errors AND/OR subjective reports that the TD is difficult to use in the operating context	Transaction time TD operating errors Subjective reports of usability of device in operating context	Incompatibility between the environment and the interaction of the TD and the user		

TABLE A.3 (contd.): DIAGNOSIS OF ENVIRONMENTAL INCOMPATIBILITY
(SPEECH DATA INPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 4.1 Vocabulary and/or syntax generated by the speech synthesizer is different from those normally used in the work environment.	Actions in which information is acquired from the computer	Intelligibility Vocabulary and syntax	Familiarity with device language and competing language	Ambient noise level		High probability of operating mistakes due to misrecognition of output	Reaction time to presented data Errors in interpretation of presented information	*Language incompatibility* Most frequently-used knowledge of language tends to dominate interpretation of presented information. Probability of correct understanding is reduced if information is presented in a form contrary to expectation	Use standard vocabulary/syntax to maximize the accuracy of information reception Output should be free of regional accents and intonation.	See diagnostics 4.2, 4.3 and 6.1 Supporting documents: RARDE, 1983 Simpson et al, 1987
Diagnostic 4.2 Vocabulary used in machine generated speech includes words which are phonetically similar to each other.	Actions in which information is acquired from the computer	Device vocabulary Intelligibility		Factors influencing speech intelligibility, e.g. ambient noise level		High incidence of users' acoustic confusion errors, or evidence that users are exerting particular effort to listen.	Transaction time Frequency of use of "playback"/repeat facilities Subjective listening effort	Phonetically similar words are easily confused by human listener unless supported by contextual knowledge	Use distinctive synonyms to avoid confusing words. Where possible, use polysyllabic words in preference to monosyllabic words so that words are more easily distinguished. Offer "playback" facilities Consider redundant visual presentation of information Present information in familiarly structured sentences to provide a linguistic context to reduce ambiguity. Train users to utilize contextual information to reduce ambiguity.	See diagnostics 4.1, 4.7 and 6.1 Supporting documents: RARDE, 1983 Simpson et al, 1987
Diagnostic 4.3 Information is presented in chunks of a size and format different from those used elsewhere in the task.	Actions in which information is acquired from the computer Other task actions in which information is acquired or manipulated	Device constraints on data chunk size Rate of presentation	Familiarity with format of task data	Factors disrupting memory process e.g. stress, noise, visual distraction.	Constraints on the structure of task information.	Trade-off between (1) performance disruption by information presented in units smaller than those preferred by users for the representation of domain information and (2) performance disruption by cognitive load imposed by remembering data at early locations in speech output sequence.	Transaction time Frequency of use of "playback"/repeat/ "playslow" facilities Errors: type and frequency Subjective assessment of cognitive effort to transform presented information to preferred format	Information reception will be facilitated if the speech representation is compatible with the user's mental representation of the information; however, representations in working memory decay over time, so early items in a long string may not be recalled accurately	Structure presented information in form familiar to the user. Segment the synthetic speech into grammatically correct elements. Offer "playback" facility Offer redundant visual data presentation. Provide extended training to users to familiarize them with the machine information structure.	See diagnostics 4.4 and 6.1 Supporting documents: Cooper, 1987

TABLE A.4: DIAGNOSIS OF KNOWLEDGE INCOMPATIBILITY
(SPEECH DATA OUTPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related consideration
Diagnostic 4.4 Amount of information (non-redundant chunks of information) in speech output message is large.	Actions in which information is acquired from the computer	Device constraints on data chunk size Rate of presentation		Factors disrupting memory process e.g. stress, noise, visual distraction Availability of memory support (e.g. facilities for note-taking).		Performance disruption by cognitive load imposed by remembering data at early locations in speech output sequence.	Transaction time Frequency of use of "playback" function Use of mnemonic strategies e.g. note-taking Subjective assessment of effort to remember long strings of data	Representations in working memory decay over time, so early items in a long string may not be recalled accurately	Use shorter chunks of information Provide users with means of recording the information (e.g. pencil and paper) Provide redundant visual feedback Offer "playback" facility	See diagnostics 4.3, 4.7 and 6.1 Supporting documents: Marics & Williges, 1988 RARDE, 1983 Simpson et al, 1985
Diagnostic 4.5 Multi-layer menus are presented by means of speech output	Actions demanding menu selection	Menu structure	Familiarity with menu structure	Factors disrupting memory process e.g. stress, noise, visual distraction.	Requirements to access menus	Users will be unable to locate desired options in the menu, or will take a long time to do so; and/or they will fail to remember their location in the menu structure.	Transaction time Frequency and type of selection errors Subjective cognitive effort of remembering menu items and/or place in menu structure.	Users have to assemble and re-order information held in their transient auditory memory. This memory load interferes with the ability of the users to carry out the problem solving associated with the menu selection.	Improve the menu by reducing the number of menu items in each menu layer so that it contains no more than 5 selections. Provide user control over the flow of speech output, by means of commands to stop, repeat or move forward and backward through the menu hierarchy. Provide visual representation of menu as a back-up information source.	See diagnostics 4.3, 4.7 and 6.1 Supporting documents: Cooper, 1987
Diagnostic 4.6 Information presented is multi-dimensional and continuously scaled.	Actions in which information is acquired from the computer	Modality and format in which information is presented by the device			Dimensionality of data. Scaling of data. Time constraints on transactions.	Precision with which continuously-scaled infomation can be transmitted using speech is a function of time (e.g. precision of a geographical co-ordinate is determined by the number of figures in the grid reference). As required precision increases, transmission will take longer and/or will require more effort and/or will run an increased rate of error by comparison with presentation by display devices able to represent information expressed in multiple dimensions on continuous scales.	Time to receive data of required precision.	*Coding incompatibility.* 1. Positive relationship between precision and message length when information is transformed from a continuous to a discrete scale. Note also that the message length is multiplied by the number of dimensions of variation. 2. Effort is required to make transformations in 1.	Only use speech for presenting continuously scaled and/or multi-dimensional information if a low level of precision is required or if there is low time pressure. Visual display devices may be more appropriate alternatives.	See diagnostics 4.3, 4.4 and 6.1 Supporting documents: Funk & McDowell, 1982

TABLE A.4 (contd.): DIAGNOSIS OF KNOWLEDGE INCOMPATIBILITY
(SPEECH DATA OUTPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related consideration
Diagnostic 4.7 Intelligibility of speech output is poor by comparison with that of normal human speech	Actions in which information is acquired from the computer	Intelligibility Vocabulary	Familiarity with device characteristics	Factors disrupting memory process e.g. stress, noise, visual distraction.	Rate at which information must be acquired from computer	Poor recall of presented information Unacceptable effort required to understand device output	Misrecognitions of device output Recall of presented data Subjective effort to understand device output	Inadequate specification of acoustic cues to phonetic segments results in increased cognitive demands for phonetic coding. This disrupts the mental rehearsal of presented information, hence recall is impaired	Improve intelligibility (e.g by use of noise insulated headphones) Reduce rate at which information is presented Offer facilities for playing back presented information	See diagnostics 4.2, 4.4, 4.5 and 6.1 Supporting documents: Waterworth & Talbot, 1987
GENERAL PURPOSE DIAGNOSTIC FOR KNOWLEDGE INCOMPATIBILITY (where TD = target device)										
The knowledge held by users is inadequate or inappropriate to operate the TD OR knowledge users already hold interferes with TD operating knowledge OR users are unable to represent information in ways demanded by the TD	Actions during which potentially incompatible knowledge is used Actions involving creation or manipulation of information to enable above actions	Content/structure /medium of information presented by the TD Content/structure /medium demanded to enter information to TD Knowledge demanded for TD operation	TD operating knowledge Knowledge of the operation of similar devices Familiarity with the structure and content of information handled in the task	Content/structure /medium of information from non-device sources Content/structure /medium demanded for recording information other than to TD Knowledge demanded for operation of other devices in task context Environmental factors disrupting memory process	Temporal demands of task Order in which task actions are performed	Increased incidence of operating errors AND/OR increased transaction time AND/OR subjective reports of excessive cognitive load	Errors in operation of device Errors in performance of off-line activities Transaction time	Knowledge incompatibility between the TD and users, i.e. knowledge held by users is not compatible with that demanded by the TD to maintain desired performance		

TABLE A.4 (contd.): DIAGNOSIS OF KNOWLEDGE INCOMPATIBILITY
(SPEECH DATA OUTPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 5.1 Users required to perform concurrent computer and non-computer operations which demand visual attention	Actions in which information is acquired from the computer Concurrent non-computer actions	Information presented to operator by computer -type -rate of output	Facility in scheduling device and non-device actions	Workspace configuration for computer and non-computer actions.	Temporal demands of on-line and off-line tasks.	Mutual interference between on- and off-line operations. Operator may attempt to reduce by serial operations with rapid alternations between on- and off-line task elements. Increased probability of 'slips' Evidence of heavy demands on operators (fatigue, etc) Re-orientation when transferring between task elements.	Transaction time. Errors in acquisition of information from the computer Off-line task performance. Subjective assessment of cognitive effort to maintain desired performance.	Limited operator resources (eyes). Possible under-utilization of alternative resources (ears). Suggestive evidence that eyes resources partially independent of speech understanding.	Consider use of speech data output interface if data suitable.	See diagnostics 4.4, 4.6 and 6.1 Supporting documents: Williges et al, 1986
Diagnostic 5.2 Information must be presented independently of head movement or body position.	Actions in which information is acquired from the computer	Information presented to operator by computer -type -rate of output Location of device in workspace		Constraints on operator's position Constraints on operator's orientation	Temporal demands of task	User forced to divert visual attention to obtain information from computer, with consequent disruption of off-line actions.	Transaction time Errors of omission Subjective reports of fatigue due to attentional switches	The visual field attained by eye movement without co-ordinating head movement is restricted. Auditory signals are omni-directional.	Consider use of speech interface if data suitable	See diagnostics 4.4, 4.6 and 6.1 Supporting documents: Deathridge, 1972
Diagnostic 5.3 Users need to monitor multiple information displays concurrently.	Device actions	Information presented to operator by computer -type -rate of output Location of displays in workspace	Ability to handle information from multiple sources		Temporal constraints imposed by task Requirement to obtain information from multiple sources	User has to alternate visual attention between multiple sources of information, with consequent failures to acquire relevant information	Transaction time High level operating errors such as misinterpretation of information received. Subjective reports of fatigue due to attentional switches	Limited operator resources (eyes). Possible under-utilization of alternative resources (ears). Suggestive evidence that eyes resources partially independent of speech understanding.	Consider use of speech interface for suitable sources of data	See diagnostics 4.4, 4.6 and 6.1 Supporting documents: Deathridge, 1972 Simpson et al, 1985

TABLE A.5: DIAGNOSIS OF BEHAVIOURAL INCOMPATIBILITY
(SPEECH DATA OUTPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related consideration
GENERAL PURPOSE DIAGNOSTIC FOR BEHAVIOURAL INCOMPATIBILITY (where TD = target device)										
Users do not have the skills necessary to operate the TD OR users have other skills which interfere with their ability to operate the TD OR the task demands concurrent actions which interfere with one another	Device actions Actions concurrent with, interrupting or interrupted by device actions	Operating procedure Content/structure/medium of information on the TD Content/structure/medium demanded for entry of information to TD	TD operating skill Familiarity with the structure & content of information handled in the task Familiarity with the operation of competing devices	Workspace layout Contextual factors disrupting user's ability to access or operate display and controls of TD	Domain events initiating concurrent actions Temporal demands of the task	Transaction time is slow AND/OR increased errors of the "slip" variety AND/OR subjective reports of excessive physical or cognitive loads	Transaction time Low level errors device operation Performance of off-line task activities Subjective reports of workload	Behavioural incompatibility between the TD and users, i.e. the behaviour ("skill") of which the user is capable does not correspond with the behaviour demanded by the TD to maintain desired performance		

TABLE A.5 (contd.): DIAGNOSIS OF BEHAVIOURAL INCOMPATIBILITY
(SPEECH DATA OUTPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
Diagnostic 6.1 Speech must be presented in a noisy environment.	Actions in which information is acquired from the computer	Rate of output Speech intelligibility	Familiarity with device output (acoustics, lexicon and syntax)	Ambient noise level		Evidence (such as diminishing secondary task performance) suggests that the user is making particular effort to listen. Increased incidence of errors due to acoustic confusion. Evidence that speech output is overlooked or only partially received.	Transaction time Frequency of use of "playback" function Subjective reports of effort demanded to hear information	Speech output is masked by acoustic noise Auditory memory processes are disrupted by high levels of competing noise	Use distinctive synonyms to avoid confusing words. Where possible, use polysyllabic words in preference to monosyllabic words so that words are more easily distinguished. Offer "playback" facilities Consider redundant visual presentation of information Present information in familiarly structured sentences to provide a linguistic context to reduce ambiguity. Train users to utilize contextual information to reduce ambiguity.	See diagnostics 4.1, 4.2 Supporting documents: Miller & Isard, 1963
Diagnostic 6.2 Emergency situations result in operator anoxia	Device actions involving acquisition of information from device	Information utilized during emergencies Medium of information output Rate of information output			Temporal constraints of task in context of emergency	Impaired acquisition of visual information resulting in reading errors, or long transaction time or reports of effort required to acquire information.	Mistakes arising from misinterpretation of presented information	Greater resistance of auditory sensitivity to anoxia as compared with visual sensitivity.	Consider use of speech interface for emergency data presentation if data is of an appropriate sort	See diagnostics 4.4, 4.6 and 6.1 Supporting documents: Deathridge, 1972
Diagnostic 6.3 Information must be presented without light being emitted.	Device actions involving acquisition of information from device	Medium of information presentation Content and format of information Rate of presentation of information			Constraints on light output Requirements to acquire information	Light output from device unacceptable (e.g. for reasons of operator disclosure) or operator has difficulty extracting information from a visual display due to its low luminance	Display luminance Errors in acquisition of information from visual display Subjective reports of difficulty acquiring information from visual display	Auditory information may be presented without light emission Visual acuity is reduced at low levels of luminance	Consider use of speech for data presentation if data is of an appropriate sort	See diagnostics 4.4, 4.6 and 6.1 Supporting documents: Deathridge, 1972

TABLE A.6: DIAGNOSIS OF ENVIRONMENTAL INCOMPATIBILITY
(SPEECH DATA OUTPUT)

Critical System Conditions	Critical actions	Critical device attributes	Critical user attributes	Critical context attributes	Critical task attributes	Behavioural/ Performance features	Critical behav/perf. parameters	Interaction model assertion/diagnosis	Prescriptive options	Related considerations
GENERAL PURPOSE DIAGNOSTIC FOR ENVIRONMENTAL COMPATIBILITY (where TD = target device)										
Environmental factors prevent or disrupt the application of knowledge or skills held by users in the operation of the TD	Device actions	Content/structure/medium of information on the TD Content/structure/medium demanded for entry of information to TD TD operating procedure	TD operating skill Protective clothing	Workspace layout Environmental factors disrupting TD-user interaction	Demands for exposure to environmental interference Temporal demands of the task	Transaction time is slow AND/OR increased incidence of TD operating errors AND/OR subjective reports that the TD is difficult to use in the operating context	Transaction time TD operating errors Subjective reports of usability of device in operating context	Incompatibility between the environment and the interaction of the TD and the user		

TABLE A.6 (contd.): DIAGNOSIS OF ENVIRONMENTAL INCOMPATIBILITY
(SPEECH DATA OUTPUT)

APPENDIX B

DEVELOPMENT OF THE TASK SIMULATION METHOD

ADAPTED FROM "THE ANALYSIS AND SIMULATION OF BATTLEFIELD TASKS: THE CASE OF FORWARD OBSERVATION"

(Ergonomics Unit report dated November, 1988, reference: MACL/R/41/88/1)

B1 INTRODUCTION

The work described in this appendix supported the development of the method for task simulation (see Chapter 8). The task simulation method (TSM) evolved in the course of a study of the forward artillery observer (Forward Observation Officer - FOO), the aim of which was to develop a laboratory simulation for assessing the impact on task performance of an artillery targeting computer with a connected speech recognizer. The report on which this appendix is based was intended to perform three functions:

(1) to document the development of a simulation of the FOO task
(2) to explain the evolution of the TSM
(3) to demonstrate and to evaluate the application of the TSM.

It therefore begins with a brief description of the approach proposed by Life (1987) as starting point for task analysis. The application of the approach to the FOO task is described and critically reviewed. Section B3 presents the revised and extended TSM approach. Section B4 describes the application of this new method, post hoc, to the FOO task data, and Section B5 considers refinements to the method indicated as necessary following its application to the FOO data.

B2 THE INITIAL APPROACH TO ANALYSIS

B2.1 Summary of the approach

Analytic and empirical assessments impose differing requirements for the representation of battlefield tasks. The former demand a task model amenable to the application of existing knowledge of the usability of speech input/output (I/O) devices, and the latter a task model suitable as the specification of a simulation. The approach to task analysis which was originally proposed actually concentrated on *analytic* assessment (and was subsequently recognized to be poorly suited for empirical assessment). In the first version of the method, it was reasoned that existing knowledge was expressed conveniently in ergonomic guidelines for the implementation of speech interfaces. Task analysis demanded, then, characterization with respect to those task parameters critical to the application of the guidelines.

The analysis of the task was structured according to a two factor classification of the guidelines. Guidelines addressed, on the one hand, either the *compatibility* of the device with the type of data being mediated, or the *interactions* that occur with the execution of non-device actions in the task; and, on the other hand, either the *representation* (and processing) of information, or the *physical actions* taken by operators to bring about change in the world. Table B1 illustrates the classification with four of the speech guidelines.

The approach to analysis involved describing the task in terms of a hierarchy of actions, then applying what were termed "static" and "dynamic" analyses. The static analysis intended to expose those aspects relevant to assessment with respect to *data compatibility.* It could be done on the basis of a statistical description of task and data characteristics (e.g. the frequency and length of data messages to be entered; the way in which the data was coded). Dynamic analysis determined the incidence of *procedural interaction* and hence the occurrence of competition between task actions for the same operator resources. This required observation of the task being performed and assessment of concurrence in the performance of task actions. The dynamic analysis was expressed as a log relating task actions (and the operator resources demanded by them) to time. This log was subsequently searched for "critical patterns", such as the incidence of demands for the same resource (e.g. use of hands or eyes) within a predefined time frame. The task representations and their use were exemplified in the context of an imaginary intelligence/ reconnaissance task.

Table B1: A preliminary classification of ergonomic guidelines for the implementation of speech interfaces

	Representational	Physical
Data compatibility	e.g. Use vocabulary/syntax compatible with normal speech when implementing speech I/O	e.g. Do not use speech input when task conditions cause the speech waveform to vary
Procedural interaction	e.g. use speech when data I/O must be performed concurrently with task features using non-verbal resources	e.g. Use speech I/O when omni-directional or hands-free operation is demanded

A rudimentary procedure for analysis was proposed, then. This involved:

(1) hierarchical task description, on the basis of indirect methods (such as reading task training manuals and interviewing domain experts), and direct methods (such as observation of the task being performed).

(2) static analysis, in which observational data would be assessed with respect to the compatibility with speech coding of information transmitted by operators.

(3) dynamic analysis, in which data recorded in real time (e.g. video data or that recorded by on-line monitoring techniques) would be searched for concurrent actions.

This procedure was followed in the development of a simulation of the task of the FOO.

B2.2 Applying the approach to Forward Observation

UCL required a task to act as a vehicle for developing their method. The reasons for selecting the FOO task to trial the task analysis approach were (1) that it was representative of an important class of military tasks (observation), and (2) that it was due to be computerized, and (3) that it was an example of a task in which speech I/O was potentially useful. The analysis of the task was to support the development of a simulation suitable for the assessment of the suitability of a connected word speech recognizer for the FOO to enter indirect fire orders to an artillery targeting computer. It was carried out during 1987 with the support of RACISG, Royal Artillery Regiment, Larkhill.

B2.2.1 Hierarchical task description

Four stages were proposed for the development of a hierarchical description: familiarisation, field observation 1 (e.g. observation of trainee FOOs operating an engagement simulator), field observation 2 (e.g. observation of qualified FOOs on a field training exercise) and a follow-up interview.

(a) **Familiarisation** UCL and RSRE visited RACISG to obtain a preliminary overview of the task from domain experts. This, and information extracted from other MoD sources, formed the basis of a preliminary description of the task. The hierarchy was based on the concept of the "task feature" - originally conceived as being "task elements that, individually, may occur in several different tasks, but the particular combination of which defines each task uniquely" (Life, 1987a). The preliminary description is presented in Figure B1.

Generating the hierarchy prompted a number of questions which were addressed on a second visit to RACISG. In addition, UCL observed two artillery training sessions based upon the Invertron simulator (a classroom training system intended for the instruction of trainee observers in Fire Discipline). These led to a refinement of the model.

Figure B1: Familiarization with FOO task

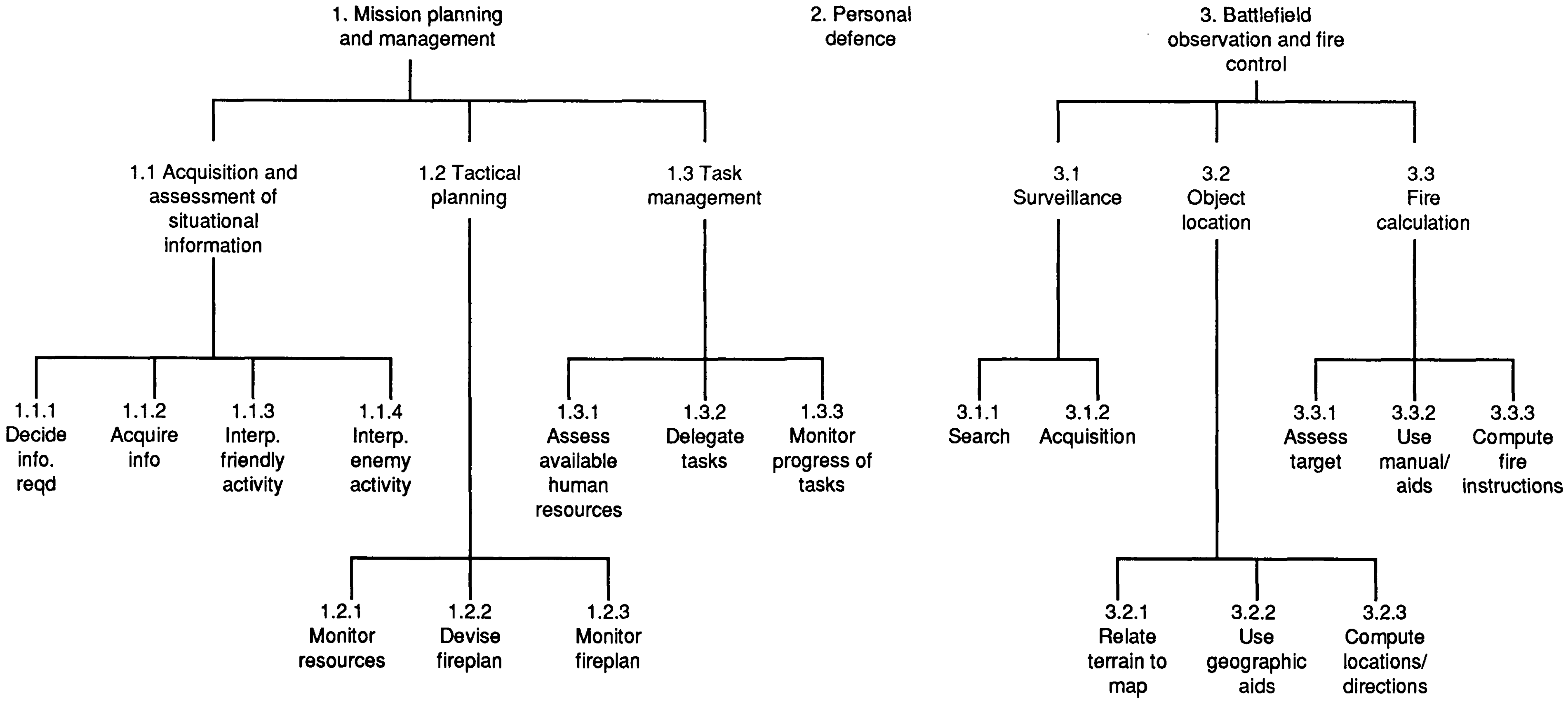
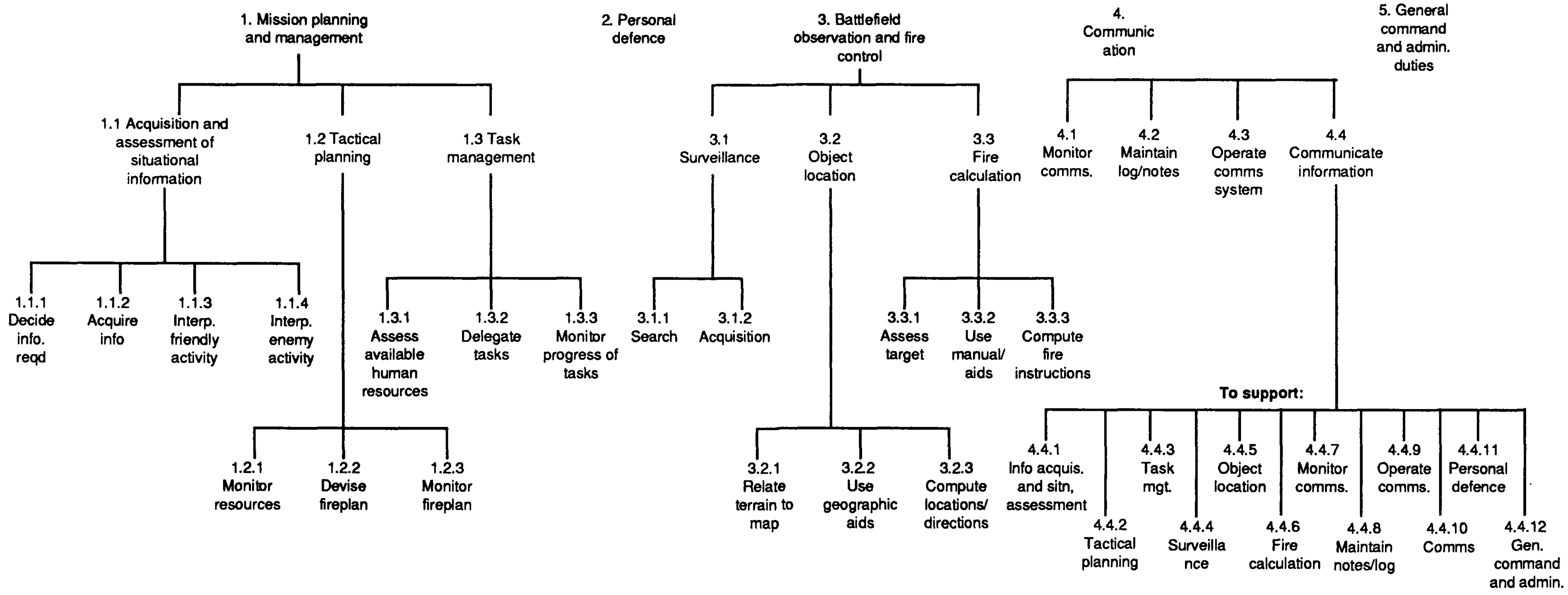

1. Mission planning and management
2. Personal defence
3. Battlefield observation and fire control
1.1 Acquisition and assessment of situational information
1.2 Tactical planning
1.3 Task management
3.1 Surveillance
3.2 Object location
3.3 Fire calculation
1.1.1 Decide info. reqd
1.1.2 Acquire info
1.1.3 Interp. friendly activity
1.1.4 Interp. enemy activity
1.3.1 Assess available human resources
1.3.2 Delegate tasks
1.3.3 Monitor progress of tasks
3.1.1 Search
3.1.2 Acquisition
3.3.1 Assess target
3.3.2 Use manual/ aids
3.3.3 Compute fire instructions
1.2.1 Monitor resources
1.2.2 Devise fireplan
1.2.3 Monitor fireplan
3.2.1 Relate terrain to map
3.2.2 Use geographic aids
3.2.3 Compute locations/ directions

Figure B2: Elaborated description of FOO task (following field observation)



(b) **Field observation.** The familiarisation phase gave the investigators a working understanding of the task of the FOO which enabled them to plan data collection. At this stage it became evident that the collection of video data on an outdoor training exercise would be difficult to arrange within the timescale of the project. It was decided instead to collect video data on simulated artillery missions in the context of the Invertron simulator. It was arranged that an observation post (OP) party comprising FOO, assistant and signaller, supported by a mortar fire commander (MFC), would operate in conjunction with simulated manoeuvre arm commander, fire direction centre and battery command post in the completion of two simulated attack missions.

Two video cameras were used to record the actions of the OP party - one located directly in front of the men, and another placed high and to the side, oriented downwards to give a view of the workspaces of the FOO and signaller. Two video recorders were used to record the output of each camera and also the audio interchanges occurring (i) over the simulated artillery radio network and (ii) directly between members of the OP party. In addition, audio-only recordings were made of interchanges over the simulated mortar network, but they were not used subsequently.

Video data were supplemented with written notes, to assist in the interpretation of recorded events. Unfortunately, it only proved possible to record one mission, and this lasted approximately one hour. It was a timed attack demanding the generation of an artillery fire plan, co-ordination with mortar support and the ordering of fire against on-call targets. The recording included briefing of the FOO by the manoeuvre arm commander and illustration of the "management" role of the FOO in co-ordinating the activities of the members of the OP party.

The exercise resulted in immediate refinement and elaboration of the task description. An example of the elaboration is presented in Figure B2. The observation session generated two sets of video taped data which were to be used subsequently for the static and dynamic task analyses.

(c) **Follow-up** Unfortunately, it was not possible to interview the subjects of the simulated mission directly - it would have been instructive to view the recordings with them and to have obtained their comments. However, members of RACISG offered their services in the clarification of recorded events if necessary.

The task description was discussed with RACISG. Aspects of the data which the investigators had been unable to interpret were particularly addressed. It was concluded that the description was an adequate representation of the simulated task for the purpose of analysing the video data. [It was recognized that this task representation would not necessarily characterize forward observation on the battlefield. It was intended that an analytic modification of the representation might be performed later to enhance it.]

B2.2.2 Static analysis

Static analysis demanded a characterization to enable assessment of the video data with respect to compatibility of fire information with speech coding. This required classification of knowledge sources supporting the generation of fire orders, and of the characteristics of fire order messages.

(a) **Analysis of task knowledge representations.** Each task component at the bottom of the hierarchy was subjected to an analytic elaboration, hypothesizing its goals, inputs and outputs. The inputs could be representations (mental or physical objects) or they could be sources of knowledge enabling the construction of new representations. The outputs could be representations or actions to gain access to new input knowledge. It was assumed that, additionally, the construction of new representations demanded domain knowledge, and these sources were also identified.

Table B2: Illustration of static analysis of the forward observation task

TASK COMPONENT: 1.1.1 Decide information required

GOALS: Identification of battlefield information required to support mission planning

INPUTS:

Representations:

Model of battlefield
Model of manoeuvre arm mission
Model of available artillery resources

Overt inputs:

CONTENT	MODALITY	CODE
a. Briefing from MA and FDC	Aud	Ver
b. Reports on mission/resources from OP party	Aud	Ver
c. Observation of battlefield	See 3.1.2	
d. Orders and requests to provide information	Aud/Vis	Ver

OUTPUTS:

Representations:

Model of missing resource, battlefield and mission information

Overt outputs

CONTENT	MODALITY	CODE
a. Requests for resource/mission information	Spe	Ver
b. Orders to OP party to provide information	Spe	Ver
c. Observation of battlefield	(See 3.1)	

ADDITIONAL KNOWLEDGE SOURCES:

a. SOPs
b. Military procedural knowledge acquired by previous experience
c. Knowledge of enemy tactics
d. Knowledge of friendly tactics
e. Knowledge of other people's knowledge

COMMENTS:

- Overt inputs and outputs in this component would be difficult to distinguish from those occurring under 1.1.2
- Interpretation of battlefield situation (1.1.3) is implicit in this component

Table B3: Fire command characteristics

Total number of fire commands	30
Mean number of words per command	13.2 (s.d. = 12.8)
Mean rate of speaking (each command)	2.86 words per sec. (s.d. = 0.83)

For each of the overt inputs and outputs, the form of the information was specified with respect to the modality of input (auditory, visual) or output (speech, manual action, orientation of auditory attention, orientation of visual attention). This analysis was performed utilizing information gained in the interviews with RACISG, from MoD sources and inferred from the video data. The results of the analysis are exemplified in Table B2.

(b) **Analysis of message characteristics.** Fire commands sent by the FOO or by the signaller to the battery command post constitute the information which would be entered into a future target engagement computer. The messages therefore represent "data entries" in the current system, and so were subjected to investigation. Each such message was timed, and the number of words counted; the results are summarized in Table B3. These data were used to specify the minimum acceptable rate of entry of speech data to the simulated device.

B2.2.3 Dynamic analysis

"Dynamic analysis" was intended to characterize the task for assessment with respect to interface of speech with other task elements. Life (1987) proposed that a suitable representation would be a time-sequential tabulation of task actions with respect to the user resources they demand. This would be searched for evidence of speech communication of fire order information interacting with other task elements.

It was reasoned that an appropriate representation could be conveniently developed using a computer-based video analysis system. Such a system - VITAS - was accessible at Birkbeck College (described in Laws et al, 1986). This system, based on the Apple IIE microcomputer, enabled video data to be analysed within a framework of actions specified by the investigator. In practice this meant, firstly, specifying classes of actions (e.g. FOO looking at the battlefield representation, FOO speaking to the signaller, FOO utilizing the map); and secondly, progressing through the taped data recording the onset, offset and identity of each action. The actions were stored in files which could be subsequently searched and integrated for phenomena of interest to the investigator.

Table B4 presents the frameworks of actions used to analyse the FOO task, and examples of data corresponding to the "Speech Activity" of the FOO's group. Such data sets were produced for the FOO, the assistant, the signaller with respect to eye activity, hand activity and speech activity.

Unfortunately, when the files of action data had been generated, problems were encountered with the analysis. It became evident that there were flaws in the logic of the approach to analysis.

B2.3 Problems with the original approach to task analysis

Section B2.2 describes the development of three forms of task representation intended to support the assessment of the suitability of speech I/O for the FOO: hierarchical description, static analysis and dynamic analysis. However, during the dynamic analysis it became evident that these representations were not supporting their function adequately. The weaknesses were of 4 classes:

Table B4: Example output of automated video analysis

KEY TO ACTION CODING CATEGORIES

Callsign (CS):

1	Yes
2	No

Speaker (SP):

1	FOO
2	SIG
.	.
.	.
9	Unknown

Intended listener (LN):

1	FOO
2	SIG
3	ASST
4	Control post
.	.
.	.
9	Unknown

Class of speech act (SA):

1	Declarative
2	Interrogative
3	Directive - Means oriented - OP party
.	
.	
6	Other

Utility (UT):

1	Assertion
2	Confirm/repetition
3	Denial
.	.
.	.
6	Other

Propositional content (P1):

1	Last message
2	Situation
3	Requirements
.	.
.	.
9	Other

Propositional content (P2):

1	Plan activities
2	Adjust fire
3	Order fire
.	.
.	.
6	Other

Propositional content (P3):

1	Warning order
2	Location/direction
3	Target
.	.
.	.
10	Other

```
File name: TALK S4 T92A2

CP SPEECH TWO

Starting visual time code:    00:25:54:08
Starting electronic code:     23081

Finishing visual time code:   00:69:35:48
Finishing electronic code:    88624

Number of actions:            44

No.   Frame   CS   LN   SP   SA   UT   P1   P2   P3   End no.
1     23081   1    4    2    1    2    1              23120
2     23787   1    4    2    1    2    1              23868
3     24050   1    4    2    1    2    1              24102
4     28582   1    4    2    1    2    1              28615
5     31149   1    4    2    1    2    1              31267
6     33306   1    4    2    1    2    1              33359
7     37197   1    4    2    1    2    2              37239
8     37611   1    4    2    1    2    1              37880
9     39136   1    4    2    1    1    2              39334
10    39541   1    4    2    1    4    1              39645
11    39862   1    4    2    1    1    2              39881
.     .       .    .    .    .    .    .    .    .    .
.     .       .    .    .    .    .    .    .    .    .
44                                                     88624
```

(1) the lack of specification of the relationship between behaviour, its interpretation as a hierarchy of task actions and knowledge held by the FOO to support actions.
(2) the orientation of the approach to extant tasks rather than future tasks.
(3) the inappropriateness of the format of the representation for empirical assessment (i.e. simulation development), as opposed to analytic assessment.
(4) the lack of specification of the relationship between the task representation and the model of human-computer interaction that might support analytic assessment.

Furthermore, the static and dynamic analyses had generated large amounts of information, much of which had not proved useful to the analysis.

B2.3.1 The hierarchical representation

The hierarchy was based on the concept of the "task feature". These could be decomposed into lower order features, and it was implicitly assumed that, at the limit, the bottom of the hierarchy could characterize behaviour such as movements. Life (1987) seemed to hold two ill-formed views:

(1) that there was a direct mapping between "features" which describe the task in generic terms and "actions" which represent the operational structure of the task
(2) that there was a direct mapping between low level actions and movements (or other observable behaviour).

One reason for the difficulties with the dynamic analysis was the failure to structure the analysis (and, hence, the VITAS data files) so that it corresponded to the hierarchy. Furthermore, although there was a loose mapping between the data classification and the overt inputs and outputs tabulated in the static analysis , the relationship was not explicit. The static analysis also suffered from the confusion between concepts. "Inputs" and "outputs" and "transformations" *imply* action, but the relationship between a "sub-sub-feature" of a task and task actions was not explicitly specified; (implicitly, "actions" were thought to "support" features presented in the hierarchy).

Finally, the relationship between task features and the FOO's knowledge was unclear. Implicitly, again, a task feature at any level was assumed to be "supported" by knowledge, and that at lower levels, this knowledge could be decomposed such its structure would reflect the hierarchical structure of the task. Upon reflection, this was unlikely to be true. For example, the task feature "Acquisition and assessment of situational information" includes the lower level feature "Interpret enemy activity". The former might be supported by a general model of the battlefield situation and the latter by a specific model of intelligence relating to enemy movements and intentions. There need not be any hierarchical relationship between these models.

B2.3.2 Orientation to current task

Although the speech technology assessment method is primarily concerned with the evaluation of future devices, the initial approach to analysis did not provide a means of characterizing the task in a form modified by the introduction of different technology. The representations described in B2.3.1 are of the *current* FOO task, not the computerized task. Fortuitously, the non-automated, current task involves the transmission of information by voice radio, so it may be assumed to have many features of the computerized task involving speech interface. However, it cannot be a perfect analogy for the future task, and, in any case, the task analysis method must enable the assessment of tasks in which present and future tasks are different at a low level (e.g. when the current task involves the use of a keyboard). In summary, the task representations described in B2.2 required modification to that of the assumed future task before assessment could take place. The existing method offered no mechanism for this.

B2.3.3 Inappropriateness for empirical assessment

Although one of the requirements for the analysis was stated as being to support the development of task simulations for empirical suitability assessment, the mechanism for this was not stated. Implicitly, it was assumed that the task simulation should be comparable to the target task with respect to those descriptive parameters relevant to the application of the ergonomic guidelines for speech interfaces. For example, the guideline "Use vocabulary/syntax compatible with normal speech when implementing speech I/O" dictates that an assessment of the suitability of an interface dialogue structure requires a simulation in which the operator has to use both the computer language and the language "normally" used in the workplace. The representations described in B2.2 did not necessarily carry this information, nor was there a mechanism for specifying a laboratory task on the basis of a description of a "real world" task.

B2.3.4 Unspecified task model/interaction model relationship

Even if the representations of the FOO task had been of some "future" version, they would not have supported analytic assessment because they did not carry all the necessary information. This is illustrated by the previous example: analytic assessment of the compatibility of computer operating language and other spoken language associated with the task would not be possibie, because the actual vocabulary and syntax were not represented in the analysis. (It was "lost" in the act of classifying utterances into generic types during dynamic analysis - see Table B4)

This problem is symptomatic of the fact that some guidelines are appropriate for evaluating specific speech interface implementations (such as that described above), while others address the suitability of speech as a means of interaction in principle (e.g. "use speech when hands-free operation is demanded"). The task representations produced were suitable only for assessment with respect to some of the latter class of guidelines. The problem suggests that these different classes of evaluation demand different representations, and that the class of evaluation is determined by the interests and motivations of the investigator (i.e. which "guidelines" are important to the evaluation he has in mind).

In the FOO analysis, the nature of the evaluation was not explicitly specified in advance: implicitly, in the first instance, the concern was with the general suitability of speech interfaces for the FOO. The importance of scoping the analysis was underestimated, and the form of the guidelines was not directly compatible with the specification of suitable task representations to apply them in analytic evaluation.

B2.4 Modifications required to the approach

The failure of the analysis approach when applied to the FOO task resulted in the identification of requirements for modification. These may be summarized as follows:

1. Clear definition of concepts, such as those of "task", "actions", "behaviour" and "movement".
2. Explicit specification of the relationship between actions, goals, knowledge representations and domain objects.
3. Explicit specification of the relationships between actions in a hierarchy.
4. Procedures for synthesizing a future task from a representation of an extant task.
5. Procedures for specifying a task simulation suitable for empirical assessment from a representation of an extant or future task in the "real world".
6. Compatible formats for representation of a task and for the representation of information relating to the interaction between people and computers using speech (i.e. compatible formats for the task and interaction models).

B3 A MODIFIED METHOD

B3.1 Work, tasks and actions

The overt, intentional behaviour of people may be attributed to the performance of *actions* in the achievement of *goals*, that is, states of the world which they desire. The overt behaviour by which an action is manifested is the final stage of a sequence of events, the earliest stages of which are cognitive. It is assumed that people hold cognitive *representations* of goals and of the present state of the world and that these are the bases for action (see van Dijk, 1980).

A *task* produces an intentional change in the state of an entity by the application of *work*. The desired state(s) of the entities is the *goal* of the task, and it may be expressed as the ideal outcome of the task (Dowell and Long, 1988). The intentional behaviour exhibited when a task is performed may be interpreted, then, as actions to achieve a task goal. People engaged in tasks hold cognitive representations of current and desired states of the entities within the domain in which work is to be applied.

Tasks typically involve several actions; these, and the relationship between them, characterize the functional structure of a task. People describe (and think about) sequences of actions in terms of superordinate actions, i.e. actions are seen as parts of more global actions, (and, hence, goals of actions may be seen as being attained by the achievement of sub-goals). This provides a rationale for describing action in hierarchical terms, and hence, in terms of "macrostructure". A coherent sequence of actions $a_i...a_n$ may be re-described in terms of a single action (A) at a higher level in a hierarchy.

The mental representations $R(a_i)$ or $R(A_i)$ underlying actions a_i or A_i will include the goal of the action, the current state of the world with respect to the action and knowledge of how to transform the current state to the desired (goal) state. This knowledge will be procedural (i.e. the actions necessary to effect the transformation) and declarative (i.e.information about objects and entities which may affect the transformation procedure).

B3.2 Simulating tasks

B3.2.1. Critical task components

A task simulation is utilized to examine the behaviour of users working at a computerized task and seeking to achieve a criterial level of performance. It is important that user behaviour exhibited in the simulation be equivalent to that which would be exhibited in the real (target) system. Given that this target behaviour will be a manifestation of task actions, equivalent behaviour has a high probability of being elicited if an equivalent user is required to achieve equivalent goals within equivalent constraints; that is, if achievement of simulated task goals depends upon the performance of actions constrained as they would be in the target task.

The behaviour of interest is that which arises through the mutual influence of the user and the device in the context of the application domain. Consequently, it is only necessary to simulate those components of the task which determine this mutual influence. If a task T is analyzed into its goals and their associated actions $A(T) =(a_i... a_n)$, the simulated task clearly must elicit the actions A(d) of using the device, but it must also elicit actions A(e) which interact with A(d) (where A(d) and A(e) are subsets of A(T), and A(o) is the complementary subset which contains all actions of A(T) not included in A(d) and A(e)). Interactions may be divided into two classes:

(a) representational interaction: where a_i, a member of A(e), shares a common representation with a_x, a member of A(d), (that is, $R(a_i)$ interacts with $R(a_x)$).

(b) procedural interaction: where actions comprising A(d) interrupt, or are performed simultaneously with, a sequence of actions comprising A(e), or vice versa.

Exemplifying this notation with the FOO task, A(T) is the set of actions involved in "Supporting Manoeuvre Arm Objectives with Economical Use of Artillery Resources". As the task currently involves the compilation of fireplans by entering engagement information on a paper form in manuscript, A(d) would include "Select Fireplan Form" and "Write Notes". In this case, A(e) would include those actions which involve acquiring written information, e.g. "Compute Grid References", because these share the requirement for a common representation (written symbols). If it were observed that the FOO had to stop writing information on a fireplan form in order to look at the battlefield through his binoculars, "Acquire Information From Battlefield Sources" would also be included in A(e), because of its procedural interaction with A(d).

Actions A(d) and A(e) involved in these types of interaction are termed "critical actions". The task simulation will be instantiated as a scenario in which a representative user is required to achieve goals which demand critical actions, within the constraints which, in the real task, act upon those critical actions. Non-critical actions (that is, those members of A(T) which are not members of A(d) and A(e), i.e. members of A(o)) need not necessarily be reproduced in the simulation.

B3.2.2 Critical task parameters

Task analysis is the process of identifying critical actions which will comprise the model of T (and which will underlie the simulation). Actions are critical if they are significant with respect to a model of device-user interaction specified by the analyst according to the purpose of the simulation. Life et al (1988) propose that such a model might be derived from existing human factors guidelines on the design of speech interfaces. A propositional format for ergonomic guidelines enables an interaction model to be related to the antecedent conditions for its functioning, viz: IF (condition) THEN (system performance consequence), BECAUSE (interaction model constraint), HENCE (guideline expressed as system design constraint).

For example:

IF the vocabulary or syntax necessary to operate a speech interface is NOT EQUAL to the vocabulary and syntax used by the operator when speaking in the working environment

THEN here will be an increased probability of lexical and/or syntactical errors in the operation of the computer

BECAUSE there is a tendency (particularly under conditions of work stress) for more highly-learned spoken responses to be elicited than less well-established ones

HENCE design the interface dialogue with language compatible with that normally used by operators OR give particular training in the use of the interface.

More specifically, then, *actions are critical if they create the antecedent conditions for the functioning of that model of device-user interaction assumed for the purpose of the simulation.* These conditions are expressed as "constellations" of system states of the form <parameter>. <state>. The parameters of the conditions are termed"critical parameters", and a "constellation" is a sequence of one or more critical parameters and their states. The relationships between them are expressed using the logical operators (e.g. <,>,=, AND,OR, etc.). In the example presented above, critical parameters would be the vocabularies used by the operator in normal speech and in operating the device, and states would be the words used to express a particular meaning in the two vocabularies. The critical parameters and states are included in the specification of the task simulation (see later).

B3.3 Synthesis of a future task

The discussion so far has assumed that full information exists on the nature of task actions, that is, that the task is extant and observable. However, the method must allow the simulation of tasks with future devices, and such tasks cannot be observed. This section suggests a means by which the structure of a future task may be synthesized from a description of a current task. The current task might be a manual equivalent of the future task (target task), or it might be a different task believed to have characteristics in common with the target task.

The rationale makes the following assumptions:

(1) that the goals of the future task (F) are the same as those of the current task (T) at some level of description

(2) that the constraints on F not related to the operation of the device are the same as those acting within T (e.g. user population, environmental factors constraining the user, domain information, social interactions etc.).

The F will involve the use of a future device, f, which differs from current device d. Just as there are actions A(d), A(e) and A(o) involved in performing the current task T, so there will be equivalent classes of action, A(f), A(e') and A(o'), to support the future task F, hence

$$A(F) = \{A(f),A(e'),A(o')\}.$$

The simulated F must elicit A(f) and A(e'): those task actions which interact (representationally or procedurally) with components of A(f).

The current and future task descriptions will be equivalent insofar as the functionality of d and f are equivalent with respect to the achievement of task goals. In this case, the actions will be the same at a relatively low level of description. An appropriate heuristic for the design of the task simulation in which d and f are functionally equivalent is to assume that the future task description is the same as the current task description. It may further be assumed that critical task components, and hence the task model, will be the same. The simulation of the future task will be composed, then, of the critical task components observed in the current task.

For example, in the case of the FOO task, if the change to the system were the introduction of a re-styled (but functionally equivalent) form for the transcription of the fireplan, or if the radio were to be rendered more compact by the use of smaller electronic components, the description of the task would not need to be modified, i.e. F => T

However, if d and f are not equivalent in functionality the current and future task descriptions will differ with respect to the user's actions. In principle, it should be possible to modify the action hierarchy to take account of f by deleting actions not demanded by f and generating actions demanded by f which were not required by d. An appropriate heuristic in this instance would be to assume the simplest actions for operation given the functional specification of f. For example, if it is assumed (speculatively) that the introduction of a keyboard interface will eliminate the need for a FOO to record information in manuscript on paper, then the action "Record Information on Notes/Cribsheet/Fireplan Form" will be deleted from the description, and a new action "Enter Information Using Computer Keyboard" will be generated to replace it.

Given that the hierarchy may be modified to take account of f, the differing functionality will also result in a change to the nature of the interaction between the user, device and environment in the achievement of task goals. There will not only be a different set of actions between which interactions might occur, but the representations and procedures underlying the actions will likely have changed. It is necessary, therefore, to specify a new set of critical actions, based upon the different functional hierarchy, and modified criteria derived from a model of the user-device interaction which takes account of the different class of device.

Assume that the current FOO fire-planning task is to be computerized such that the new system involves the entry of target information to the system by means of a speech interface, instead of the manual transcription of the information. A(f) would include the action of speaking a grid reference to the device. A(e') might then include, firstly, communicating with the Manoeuvre Arm commander to obtain target information (representational interaction because of the common representation of information by speech), and, secondly, observing the battlefield through binoculars (*predicted* procedural interaction as the FOO can use binoculars while speaking).

This latter example of interaction illustrates that, although it should be possible to identify probable representational interactions in the synthesized task model, procedural interaction cannot be identified unequivocally. The heuristic solution for this problem is to assume that the <u>nearest</u> counterparts in the future task of the critical components in A(T) will be critical too, as will be, of course, actions involved in using the future device, A(f). In the case of a task which exhibits procedural interaction primarily as a consequence of external events (e.g.FOO), this assumption is likely to be valid, as the cause of the interaction (i.e. external events) will be reproduced in the simulation. However, if the interactions observed in the current task are, rather, a consequence of a strategy initiated by the user (i.e. internally generated), the assumption may well be incorrect. In this instance, the analyst must either include all task components in the simulation or, if being selective, must recognize the assumptions being made.

B3.4 Specifying a simulation of a future task

A task simulation must present the system states identified as critical to the usability of speech in the context of a scenario which has a known relationship to the task as it is (or as it will be) performed in the real world. Simulation specification is a process of representing critical parameters such that they elicit critical actions in the form that they would occur in the real world.

Actions are performed to achieve goals within constraints, and the constraints are defined by the desired system performance, factors acting upon the system (such as external events, operating procedures etc.) and by the limitations of the behaviour of the device and the user. The task simulation is specified, then, as a description of the goals of the actions identified as critical, and of the entities (and their attributes)which impact the achievement of these goals. The specification enables the entities to be instantiated and simulated users (i.e. subjects) to be selected and trained so that they may be observed in an operationalization of the task model.

B3.5 A revised task simulation method

The method proposed in response to the shortcomings described in Section B2.4 assumes the rationale presented above. It comprises a sequence of steps, the culmination of each being a transformed representation of the task. Thus the method is expressed as a series of task representations and transformation procedures. The representations are as follows:

(1) Preliminary task description (an action hierarchy based upon indirect sources of task information, e.g. interviews with domain experts, training/procedural manuals, etc)
(2) Extant task data (observational data, e.g. video record of task performance or video-derived data)
(3) Expanded task description (hierarchical action description derived from observational data)
(4) Future task description (description in (3) modified to account for target device functionality)
(5) Future task model (description in (4) modified to include only actions critical to target device usability)
(6) Task simulation specification (description of task goals; list of entities/attributes impacting task actions; description of temporal relations between task events).

Section B4 describes the application of the method to the FOO task data.

B4 SPECIFICATION OF A SIMULATED FORWARD OBSERVATION TASK

It is necessary to assess the suitability in principle of a connected word recognizer as a means of entering data to a hypothetical target engagement computer. For the purpose of illustration of the revised "task simulation method", it is assumed that an additional and more specific question is whether the messages to be transmitted by the FOO are of a length that will impose an unacceptable load on the memory of the operator. This section describes, then, the use of the task simulation method for specifying in the context of the FOO task:

(1) a simulation to assess the general suitability of a connected word recognizer
(2) a simulation to assess the working memory load imposed by a connected word recognizer.

The process of specifying these two simulations is identical to the point of identifying critical task actions, at which stage the interaction models assumed for the two assessments lay down different criteria for criticality.

B4.1 Generation of the preliminary task description

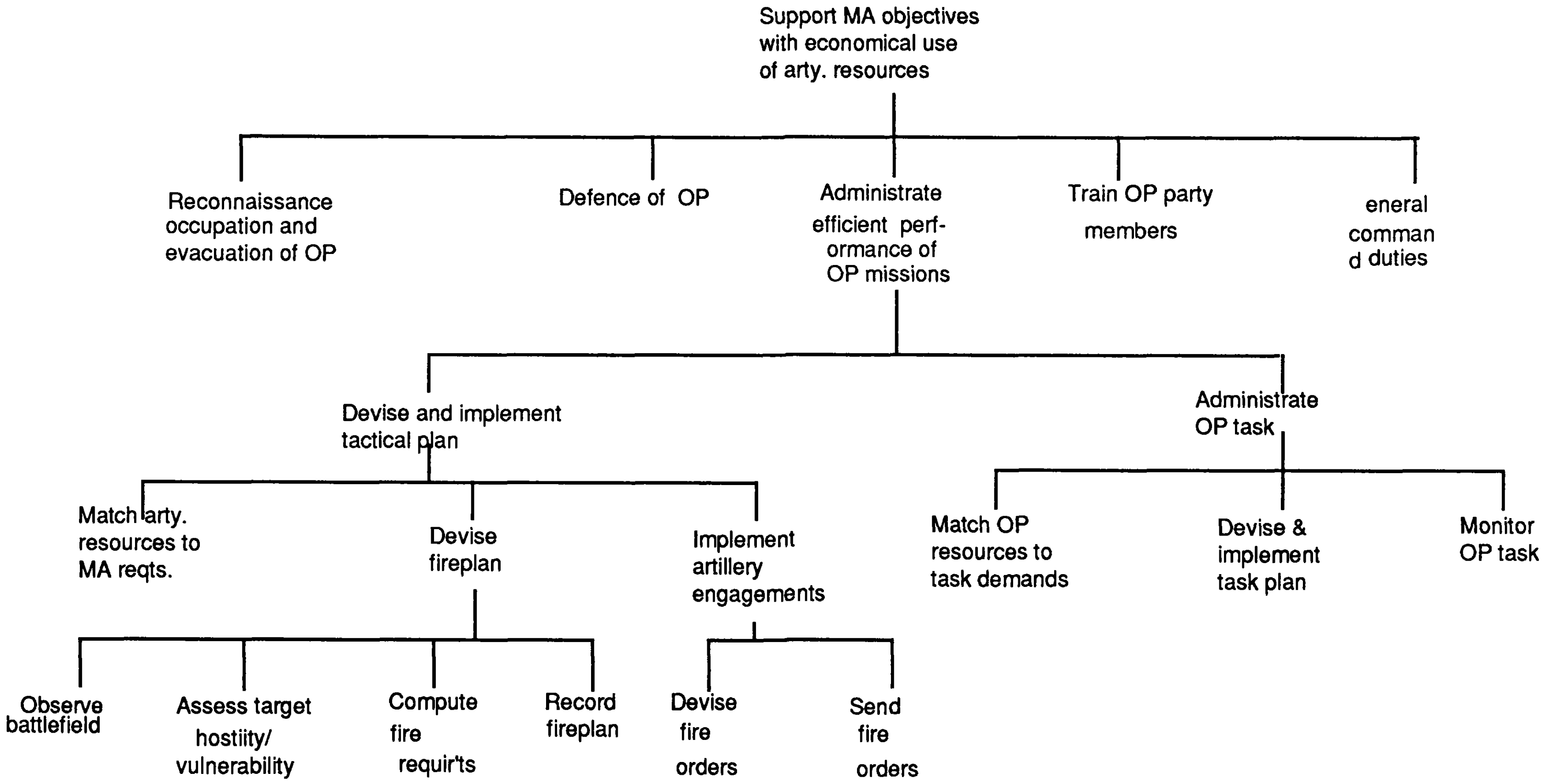
B4.1.1 Collection of information on the extant task

The preliminary collection of information on the FOO task has already been described in Section B2.2.1(a).

B4.1.2 Development of a hierarchical description of the task

(a) The military function of the FOO is to support the manoeuvre arm in the achievement of its objectives with economic use of resources.
(b) The activities which enable this function are:
    (i) Reconnaissance/occupation/evacuation of an observation post (OP)

Figure B3: Forward observation - manual task preliminary task description



(ii) Administration of efficient performance of the OP party in support of MA objectives
(iii) Training of the OP party
(iv) Defence of the OP
(v) Non-specific army command duties

Of these, discussion with domain experts suggested that all but "Administration of efficient performance of the OP party in support of MA objectives" occurred "offline", i.e. they would be unlikely directly to interact with the transmission of artillery information to the battery command post. There was thus only one action at the top level of the hierarchy: the others were deleted by analytic reasoning.

(c) Recursive application of the analytic principle of specifying tне smallest number of lower level actions that completely describe each action generates the hierarchy shown in Figure B3. This might represent a preliminary task description within the TSM; (however, note that it was constructed post hoc, with the benefit of knowledge obtained by detailed task observation).

B4.1.3 Confirmation of preliminary description

This was not done, because it would have repeated the confirmation described in Section B2.2.1(a)

B4.2 Generation of an expanded task description

The video recordings described in B2.2.1(b) were reviewed, and an expanded description generated (see Figure B4). The main hierarchy appears at the top of the figure. Because actions in the bottom layer of the hierarchy were observed to be instantiated in a number of ways, these are expressed as alternative "action sequence modules" which may be located in the bottom layer. These were of three different classes ("Acquire information", "Record information" and "Communicate information"), and they are shown in the lower part of Figure B4.

B4.2.1 Checking for completeness

The description in Figure B4 was developed by an iterative process of checking through the video recording and extending the hierarchy when behaviour was encountered which could not be accounted for within the existing description. It was considered to be complete when it covered all observed actions relevant to the task.

B4.2.2 Checking sequential integrity

A logical analysis of the actions at the lowest level in the hierarchy suggested that their left to right ordering was appropriate. Note, however, that in the video data this integrity was not always maintained, because of procedural interaction (see Section B4.4.2(b)): e.g. on occasion, action sequences were interrupted by a pressing requirement to perform a different kind of action.

B4.2.3 Confirmation of the description

This was not done, for the reason given in B4.1.3.

B4.3 Generation of a future task description

B4.3.1 Assumed target devices

A full, explicit model of the future target engagement computer in its various possible forms was not available to the project. It was assumed that the device would operate by the completion of pages of information which would be transmitted by the user. The fields of information on each transmitted page would correspond to the classes of information currently conveyed when a fire order is communicated by speech via radio networks: these are precisely specified within existing fire discipline.

The comparative assessment of a speech interface against the "default option" of a keyboard and visual display demands specification of both interfaces. The keyboard is assumed to exhibit a conventional QWERTY layout, with special command keys, e.g. for transmitting a

Figure B4: Forward observation - manual task elaborated task description

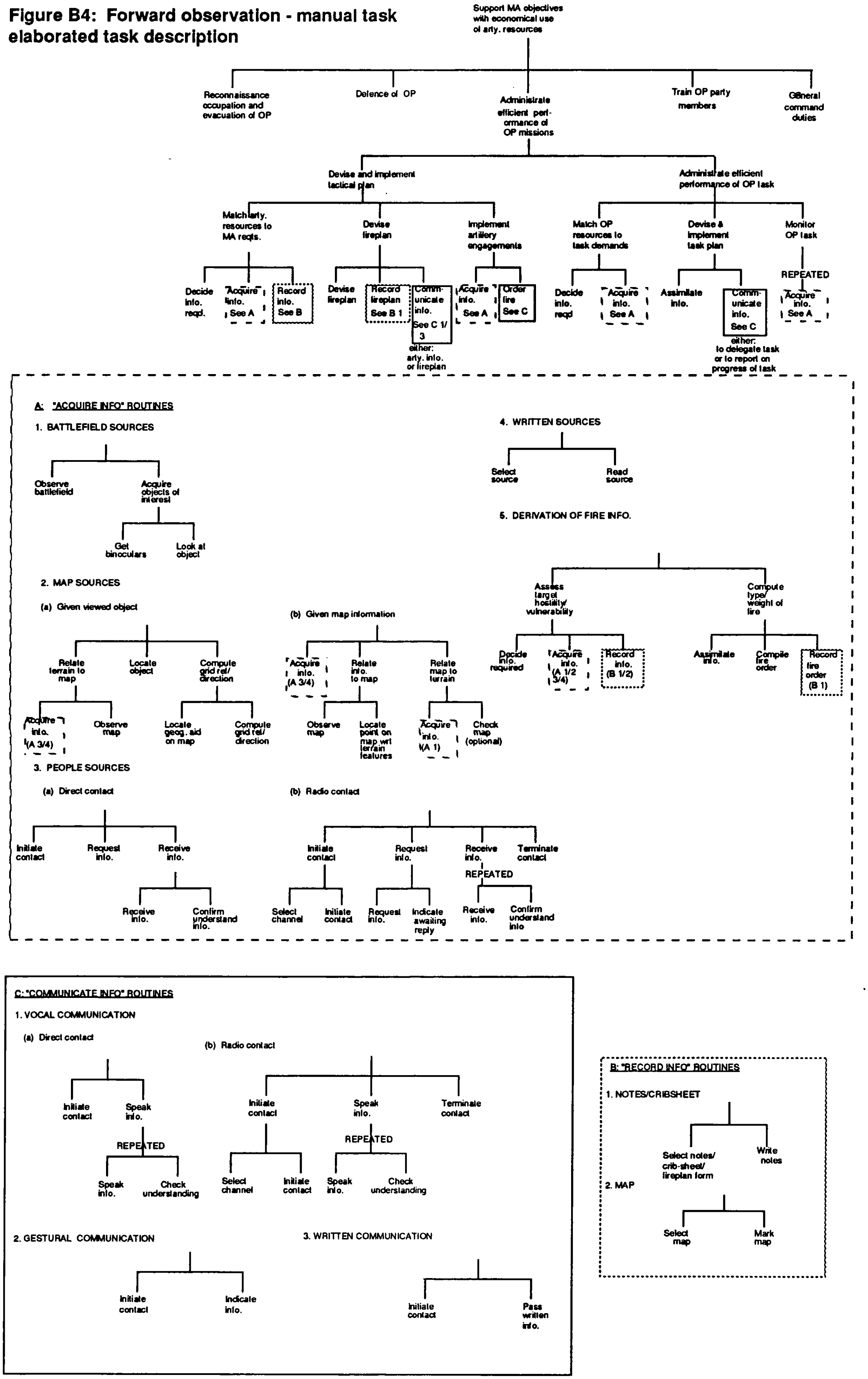


**Figure B5: Future task description: keyboard interaction.
(continued overleaf)**

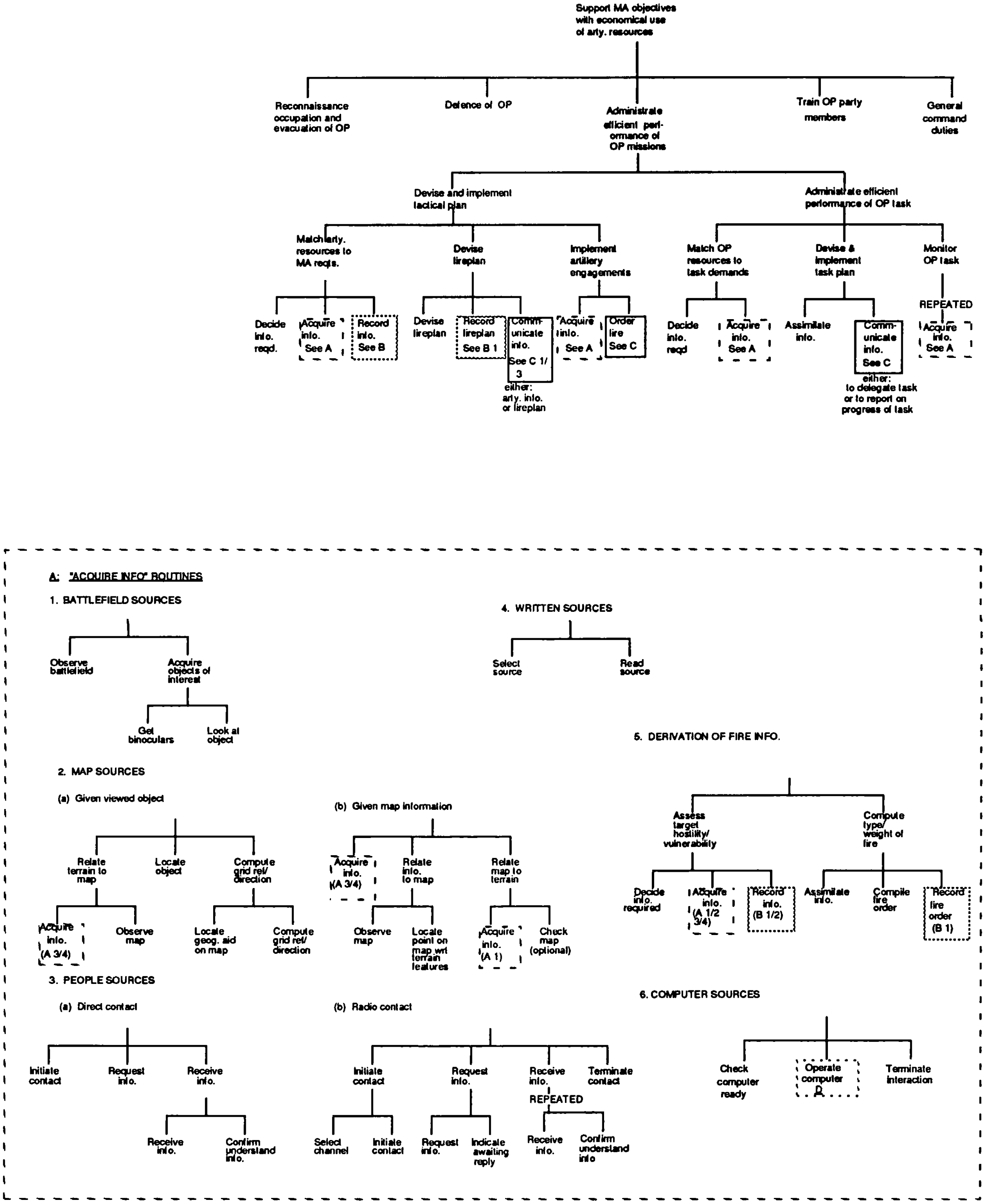
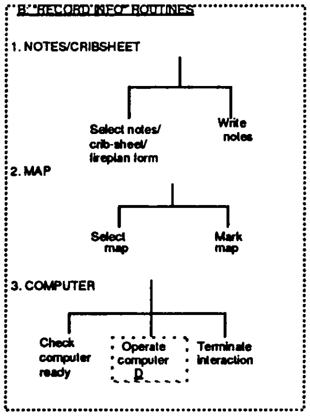
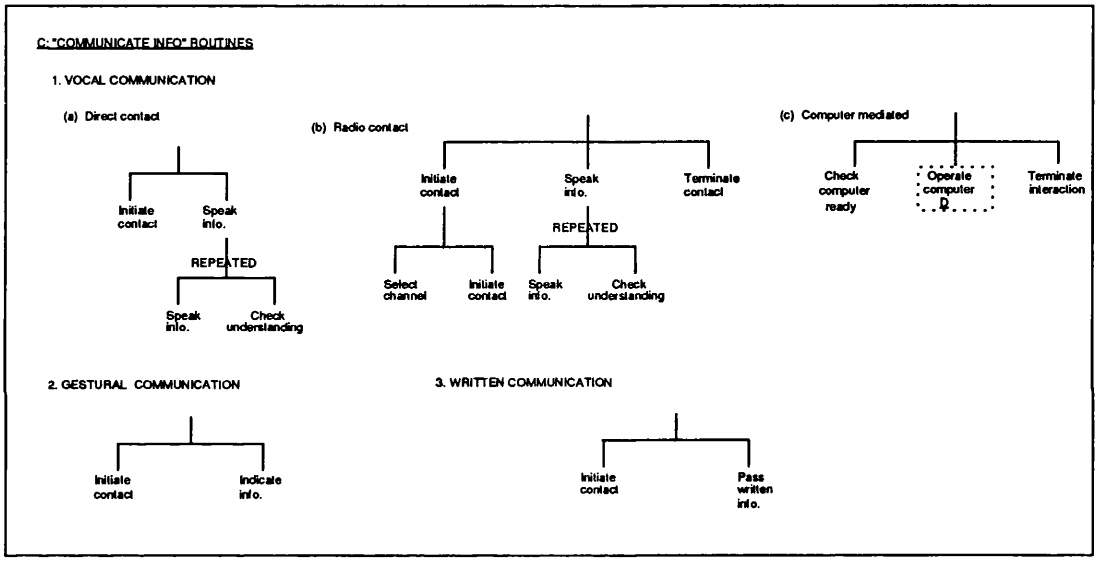


Figure B5: Future task description: keyboard interaction. (continued from previous page)





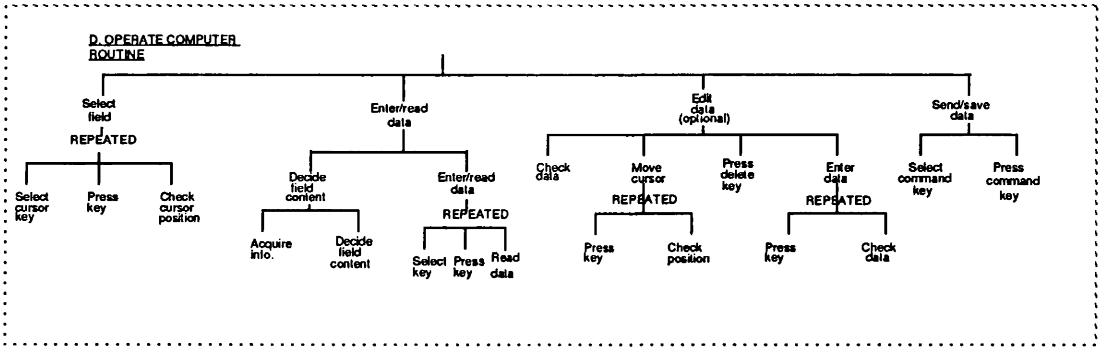
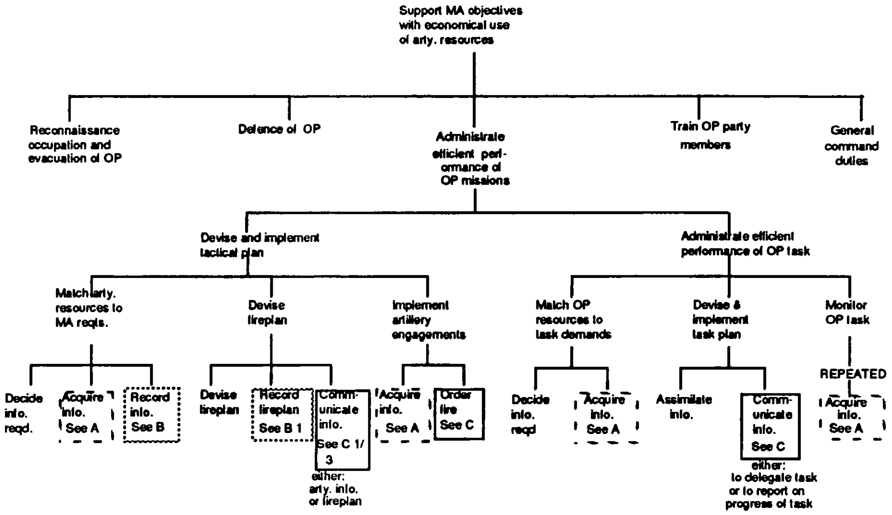


Figure B6: Future task description: speech interaction (Continued overleaf)



Support MA objectives with economical use of arty. resources
Reconnaissance occupation and evacuation of OP
Defence of OP
Administrate efficient perf- ormance of OP missions
Train OP party members
General command duties
Devise and implement tactical plan
Administrate efficient performance of OP task
Match arty. resources to MA reqts.
Devise fireplan
Implement artillery engagements
Match OP resources to task demands
Devise & implement task plan
Monitor OP task
Decide info. reqd.
Acquire info. See A
Record info. See B
Devise fireplan
Record fireplan See B 1
Comm- unicate info. See C 1/3
either: arty. info. or fireplan
Acquire info. See A
Order fire See C
Decide info. reqd
Acquire info. See A
Assimilate info.
Comm- unicate info. See C
either: to delegate task or to report on progress of task
REPEATED
Acquire info. See A

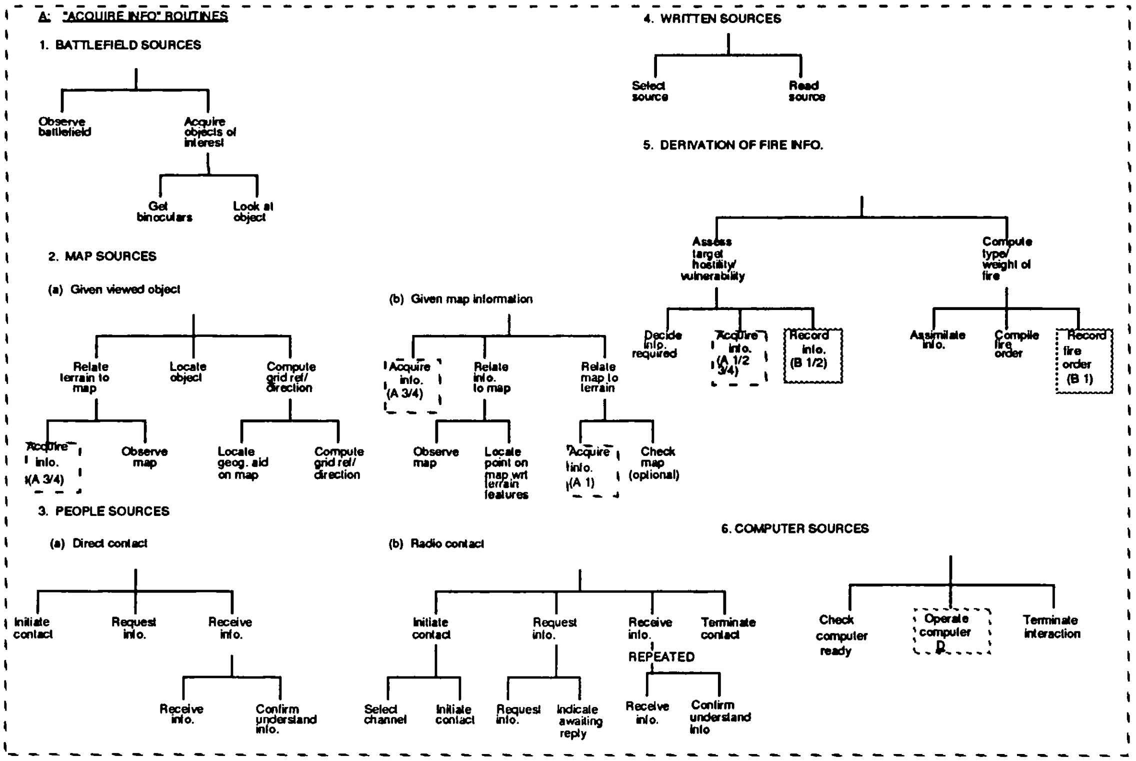
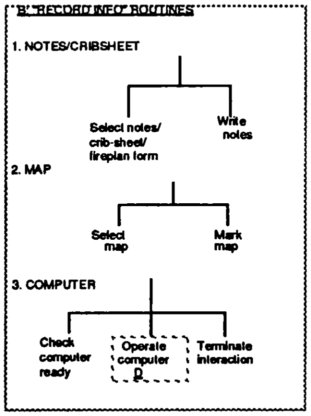
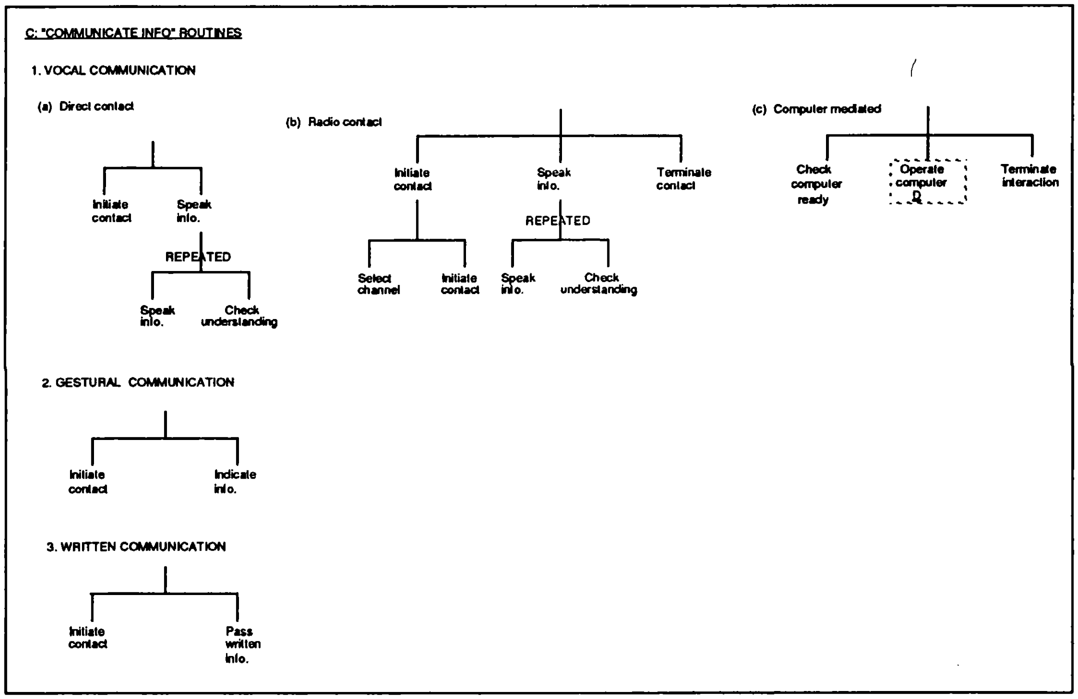
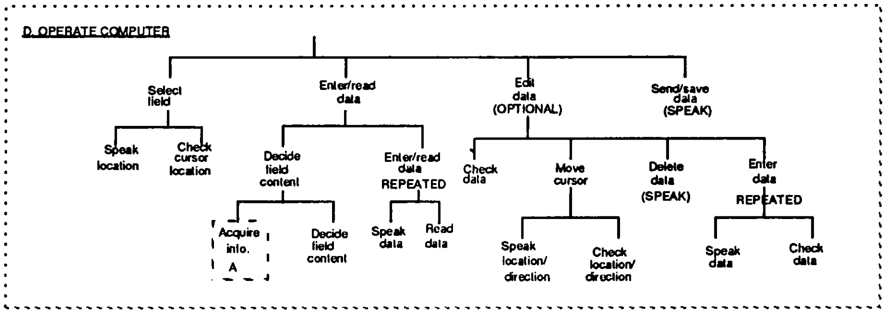


Figure B6: Future task description: speech interaction
(Continued from previous page)







message. This is assumed to write to a visual display presenting a number of information fields to a page. Pages may be edited on the screen before transmission.

The speech interface is assumed to be identical to that described above, except that the keyboard is replaced with a connected word recognizer modelled on Marconi Macrospeak. The vocabulary includes all the legal data (e.g. numerals, target names), field names and system commands necessary to operate the engagement computer.

B4.3.2 Differentiation of device/non-device actions

It was assumed that the computer would replace *radio communication of fire orders* (Action C1(D)) and, possibly, *manuscript notes for the recording of information* (Action B1). These were device actions: all others were non-device actions.

B4.3.3 Modification of the hierarchy

(a) At this point, the procedure specified by the method is deletion of device actions. This was not followed, as radio communication and manuscript notes were to be retained for other purposes (e.g. communication of information other than fire orders) following implementation of the engagement computer. NO ACTIONS WERE DELETED.

(b) New actions were generated to account for operation of the computer. The modified hierarchies for the keyboard and speech versions are presented in Figures B5 and B6 respectively.

The modifications in both cases are manifested as additional "record info" and "communication info" modules which recruit lower level "operate computer" modules.

B4.3.4 Checking the description

The integrity of the hierarchy was checked, both with respect to the completeness of the description of each action and with respect to the input-output relations between original and newly-generated actions.

B4.4 Generation of laboratory task models

B4.4.1 A task model for assessment of speech "in principle"

B4.4.1.1 *The general interaction model*

The interaction model assumed for the assessment of the general suitability of speech is carried in the following propositions[1]:

(1) IF speech/language representation held by the user (e.g. to perform non-device actions) are incompatible with the speech and language representations necessary to operate the device THEN there will be an increased probability of data entry errors BECAUSE there is interference between knowledge representations (representational interaction)

(2) IF data entry actions interrupt, or are interrupted by, other task actions which do not depend upon speech THEN speech data entry may, potentially, enhance system performance BECAUSE speech data entry may occur concurrently with non-speech actions (procedural interaction)

(3) IF the task context inconsistently influences the generation of speech THEN speech data entry will be disrupted BECAUSE the user will not produce consistent speech tokens (contextual interaction)

Critical actions with respect to this model are those which fulfil the antecedent conditions of the propositions, viz:

- all actions which utilize a speech/language representation (actions with potential for representational interaction with speech data entry)
- all actions which, in the video data, were observed to interact procedurally with the transmission of fire orders (i.e. interrupt, be interrupted by or be performed concurrently with the transmission of fire orders).

[1] Note that, at this stage of the research, the diagnostic tables had not been proposed as a means of supporting the development of models of device-user interaction

In addition, those aspects of the task context which could have disrupted speech data entry would have been critical. In the event, data was obtained only in the benign context of an indoor static training simulator, so the sources of potential (contextual) interaction occurring on the battlefield could not be assessed.

B4.4.1.2 *Generation of the general laboratory task model*

(a) **Representational interactions.** The criterion for criticality in an action at the bottom of the hierarchy was its utilization of the auditory and speech modalities in its execution, or its recruitment of word-based or number-based knowledge representations. <u>All</u> actions were critical in this regard, except for

A1: "Acquire information from battlefield sources"
B2: "Record information on map"
C2: "Communicate information gesturally."

(b) **Procedural interactions.** The video record was examined for instances of procedural interaction with the transmission of fire information. All such transmissions are tabulated in Table B5 with identification of preceding, succeeding and concurrent actions.

Table B5 indicates that the following actions interacted procedurally with the transmission of fire information:

A2: "Acquire information: map sources"
A1: "Acquire information: battlefield sources"
A4: "Acquire information: written sources"
B1: "Record information: notes/cribsheet"
C1(a): "Communicate information: vocal communication (direct contact)"
C1(b): "Communicate information: vocal communication (radio contact)"

Adding these actions to those identified in (a), all actions in Figures B5 and B6 were critical except for

B2: "Record information on map"
C2: "Communicate information gesturally".

The laboratory task models for assessment of suitability in principle comprise, then, the descriptions shown in Figures B5 and B6 with the two actions above deleted from them.

B4.4.2 A task model for assessment of potential memory load imposed as a consequence of speech message length

The assessment might be of concern if it were decided to offer a device option in which there were no concurrent visual feedback, but only (say) a synthesized speech repetition of the field when the data had been entered. Such an option, if it could be successfully implemented, would offer the potential for complete hands/eyes free operation and operation by foot-mobile observers.

B4.4.2.1 *The "message length" interaction model*

The interaction model is carried in the following proposition:
IF the number of non-redundant chunks of speech-mediated information is greater than 5 per message AND the operator has no means of recording information for subsequent review THEN there is a high probability that the user will exhibit errors in speech data entry BECAUSE as a rule of thumb, working memory has a capacity of 7+/-2 items.

B4.4.2.2 *Generation of the laboratory task model*

Critical actions with respect to this model of interaction are those enabling the generation of fire information, the communication of fire information by means of the computer, and those associated with the recording of fire information, viz

A1: "acquire info: battlefield sources"
A2: "acquire info: map sources"
A4: "acquire info: written sources"
A5: "acquire info: derivation of fire information"

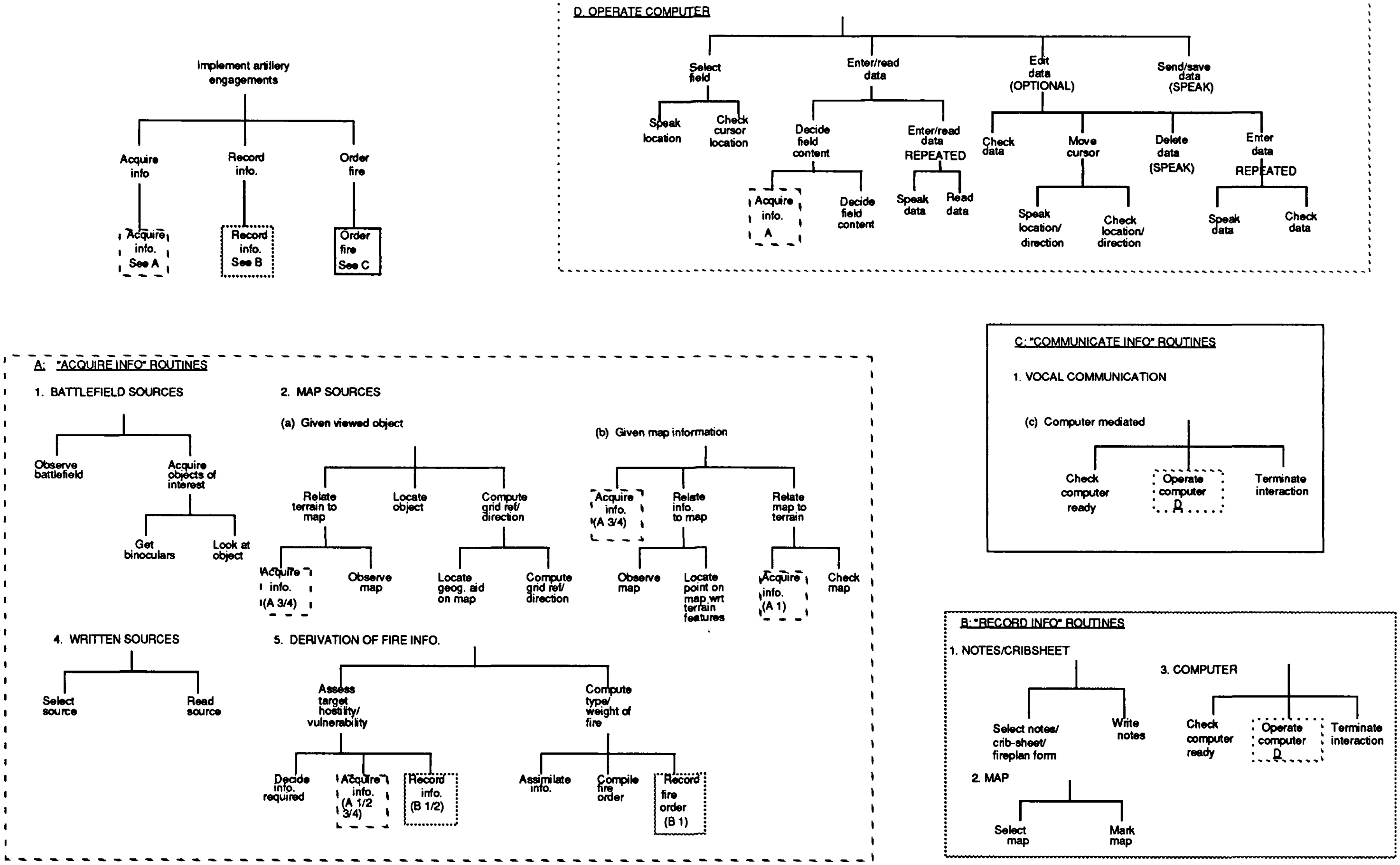
TABLE B5: Log of transmissions of fire information

Time on	Time off	Speaker/ listener	Preceding action	Succeeding action	Concurrent action
16.17	16.34	FOO/Sig	A2	A1	-
18.28	18.31	FOO/Sig	A1	A1	-
19.08	19.09	FOO/Sig	A1	A1	A1
19.40	19.42	FOO/Sig	A1	A2	-
19.34	19.35	FOO/Sig	A2	A1	-
20.47	20.49	FOO/Sig	A2	A2	-
22.55	22.56	AC/Sig	A2	A1	-
23.53	23.54	AC/Sig	A1	A1	A1
24.20	24.21	AC/Sig	A1	A1	A1
24.50	24.51	AC/Sig	A2	B1	-
25.45	25.46	AC/Sig	A1	-	-
25.52	25.53	AC/Sig	A2	A2	-
25.59	26.11	AC/Sig	B1	B1	A4
29.00	29.02	AC/Sig	A2	A4	-
30.38	30.52	AC/Sig	A2	C1/A3(FOO)	A2
31.51	32.05	AC/Sig	A4	B1	-
32.35	32.36	AC/Sig	B1	B1	-
34.08	34.17	FOO/Sig	B1	C1/A3(AC)	B1/A4
34.23	34.41	FOO/Sig	C1/A3(AC)	B1	B1/A4
40.51	40.52	AC/Sig	B1	B1	B1
41.48	42.54	FOO/Sig	C1/A3(MFC)	A2 (AC)/A4 /A2	C1/A3
46.18	46.19	FOO/Sig	C1/A3(MAC)	A2	-
46.30	46.52	FOO/Sig	A4	C1/A3(AC)	A4
50.39	50.42	FOO/Sig	C1/A3(MAC)	A1	A4
55.09	55.17	FOO/CP	A2	-	A4
55.45	55.49	FOO/CP	C1/A3(MAC)	A1	A4
57.20	57.37	AC/Sig	A1	-	-
58.46	58.51	FOO/Sig	C1/A3(MAC)	A1	A1
62.25	62.28	AC/Sig	A1	A1	-
62.44	63.04	AC/Sig	A2	A2	A2
66.22	66.23	AC/Sig	A1	C1/A3(MFC)	-
66.59	67.04	AC/Sig	A1	-	-

FOO - Forward Observation Officer
AC - Assistant Cdr. Sig - Signaller
MFC - Mortar Fire Commander
MAC = Manoeuvre Arm Cdr.

For key to actions, see Figure B2.

Figure B7: Future task model for memory study



B3: "record info: computer"(and, hence also D: "Operate Computer")
C1(c): "communicate info: computer mediated"(hence also D)
B1: "record info: notes/cribsheet"
B2: "record info: map"

Figure B7 illustrates the laboratory task model to assess the suitability of a speech interface with respect to the demands imposed on the users' working memory as a consequence of message length.

B4.5 Generation of a simulation specification

B4.5.1 A simulation for the assessment of a speech interface in principle

(a) **Overview of the laboratory task** The task involves a subject who represents the FOO (user subject - US) being presented with a task which requires the engagement of a number of targets presented on an electronic 2-D graphical display (i.e. a map, but without grid markings) - *acquire information: battlefield sources.* Each target on the display has associated with it a target number, which may be used to obtain further information about the target from an alternative text display. This text display is accessed by continuously pressing an inconveniently located key, thus requiring manual involvement in order to obtain detailed information about a target. This component of the task is intended to be analogous to the use of binoculars to acquire information about targets. The text display informs the subject of the nature of the target (e.g. tank) and the current hostility value of the target: a variable which increases continuously over time at a rate determined by the nature of the target. It is in the interest of the US to deal quickly with particularly hostile targets - *devise fireplan*

When the US has selected a target for engagement, the geographical location is calculated by identifying its position on a paper map which is marked with a grid - *acquire information: map sources and battlefield sources.* The grid reference is used as the basis of a communication to one of four gun batteries calling for fire against the target - *order fire (computer mediated).*

The process is complicated by the fact that each battery has a limited number of ammunition rounds, and that each battery is biased in its aim in a different way. Thus it is necessary for the US to monitor the number of rounds demanded from each battery - *record information; acquire (resource) information from notes* - and, for each battery, to ascertain the adjustment to the grid reference necessary for rounds to hit the target - *derivation of fire information.* From time to time, throughout the engagement, the US will be required to report the current situation to the experimenter - *administrate effective performance of OP task.*

Directly hitting the target "freezes" the hostility value and a near miss reduces its rate of increase. Performance of the subject is assessed by summing the accrued hostility values of all of the targets at the end of the mission, and/or by measuring the time taken to complete the mission, and/or by measuring the resources used to complete the mission.

(b) **Entities and attributes impacting simulated task.**

(i) Representation of the user. The FOO is represented in the laboratory task by a single entity: the US. The US sample must be representative of the FOO population with respect to the following attributes:

- visual search characteristics
- auditory processing characteristics
- characteristics of allocation of attentional resources
- ability to translate between visual spatial representations and numeric representations
- possession of skills in the use of relevant devices
- motivation to perform tasks optimally
- ability to develop strategies to optimize task performance.

(ii) Representation of the application domain. The application domain is represented in the laboratory task by a number of entities:

The battlefield The battlefield is a visual space within which targets may be located, and within which the consequences of the US's own actions may be observed. Targets may be differentiated by their location relative to each other and relative to other objects on the battlefield. The battlefield thus has the attributes of

- spatial representation (2 dimensional)
- accessibility by visual observation
- representation of the consequences of US actions.

Targets A target is an object which is attributed hostility in the context of the US task. Targets are uniquely identified by a number and by their spatial location on the representation of the battlefield. They have associated with them a generic identifier (e.g. "tank", "infantry platoon","missile launcher", "self-propelled gun") and a current hostility value which increases at a rate determined by

- spatial location (i.e. their closeness to friendly forces)
- generic identity (i.e. their potential destructive power)
- proximity of an ammunition round ordered by US.(i.e. the extent to which damage has been inflicted).

Although the spatial location of the target is accessible by direct visual observation of the representation of the battlefield, other attributes of the target may only be discovered by the manual intervention of the US (analogous to the interrogation of an object by the use of binoculars). The other attributes are presented in a table showing, for each target number, its generic type and its current hostility value.

Artillery batteries An artillery battery is the class of entity through which changes may be brought about on the representation of the battlefield. A battery does not appear on the battlefield representation itself, but the consequences of its activation (i.e. "fall of shot") does appear as a permanent marker. A battery is activated by a communication containing the following information:

- grid reference
- target number
- class of target
- number of ammunition rounds
- battery number

(The target number is actually redundant information).

Each battery has associated with it a limit to the number of rounds that it can fire and a fixed error in its aim, i.e. its activation will result in a consequence at a point having a fixed spatial relationship to the grid position actually communicated by the US. If the consequence is brought about at the same spatial location as a target, the latter's rate of increase in hostility value becomes zero; if it occurs adjacent to a target its rate of increase in hostility value is halved. The size of the area influenced by the activation of a battery depends upon the number of ammunition rounds specified.

In summary, batteries have the following attributes

- aiming error
- resource limit (with respect of ammunition rounds)
- activation by a communication containing a grid reference and a number of ammunition rounds
- activation such that the location of the consequence is a direct function of the specified grid reference
- activation such that the size of the area influenced is a direct function of the specified number of ammunition rounds
- activation such that the rate of change of hostility of a target becomes zero when it lies within the area of influence.

(iii) Representations of devices. Devices are entities with which a person interacts to facilitate the initiation of changes in the world. The devices of particular interest in the present context are the instantiations of the user interface of the target computer; however, other devices are involved in the task, and these will be termed "task aids".

Target computer interfaces. The target computer interface is the medium through which communication passes from the US to the battery. The interfaces are of two types: keyboard plus visual display; and speech entry plus visual display. Three physical entities thus are involved in representation of the the target computer interface - keyboard, speech recognizer, and visual display - but only one interface will be represented in a particular performance of the task.

In addition to physical entities, a representation of the interface will include conceptual entities: the classes of information which are themselves represented within the target computer. These conceptual entities are common to all instantiations of the interface and are as follows:

Domain Entities:

- target location, expressed as a six-digit grid reference - target number, expressed as a four digit identifier
- class of target, expressed as a word string of up to 3 words (25 characters)
- ammunition quantity, expressed as a single digit
- battery number, expressed as a single numeric identifier

Device entities:

- page, expressed as 5 fields of information corresponding to the above classes
- cursor, expressed as the current state of the device

The device entities have additional attributes relating to the changes in their state which occur as a consequence of actions which may be performed upon them by the user.

The page is "complete" when all five fields have been addressed by the cursor. A field is "complete" when it has been addressed by the cursor and a "return" command given. A page is "sent" when it is complete and when the "send" command is given. "Sending" the page brings about a consequence on the representation of the battlefield.

The cursor may be located within any field on the page or at the bottom of the page when the page is complete. The cursor moves to the next field on the page when a field is complete. Sending the page clears the fields and returns the cursor to the first field of the display.

Task aids:

Map - A map is a two dimensional representation of a space defined by a system of co-ordinates. In the present case, a map is used by the US to translate spatial locations on the battlefield representation into a numeric grid reference for communication to a battery. The map representation has the following attributes:

- it is a physical object which may be manipulated by the US (e.g. a paper representation)
- it is identical in content to the battlefield representation but with the following exceptions: targets are not represented upon it; the consequences of actions are not represented upon it; it has a grid system imposed upon it, such that the first two digits of an ordinate are specified directly and the third may be calculated by interpolation.

Geographical aid - A geographical aid is a device which facilitates object location by the use of a map. The aid to be represented in the laboratory task is a graduated overlay, to assist in the calculation of the third digit of the ordinate of a location.

Radio communication device - The laboratory task will include a representation of a communication system which is secondary to the task of observation. The communication system will have the following attributes:

- US will be open to the receipt of spoken messages throughout the task
- US may send spoken messages by activating the communication device.

Note-taking device - The laboratory task will include a facility to enable the operator to make free-format notes as an aid to memory and as an aid to calculation. This will have the attributes of free access and of being physically manipulable (e.g. a pencil and paper).

Observation aid - An observation aid is a device which enables visually acquired objects to be examined for the purpose of more detailed analysis. The laboratory task will include a representation of an observation aid with the following attributes:

- activation by continuous manual manipulation
- activation such that it results in the provision of information on objects acquired visually on the representation of the battlefield

(c) Parameter values.

(i) Device representation. Representation of the domain entities within the target computer demands a vocabulary composed of the following classes of word:

- field identifiers
- data identifiers
- command identifiers

The vocabulary needed to perform the task as specified is as follows:

Field identifiers:	**Data identifiers:**	**Command identifiers:**
Grid	Numerals: 0-10	Next
Number	Platoon dug in	Last
Rounds	Platoon advancing	Send
Battery	Tank	
	Rocket launcher	
	Missile launcher	
	Self-propelled gun	
	APC	
	Communications centre	
	Stationary gun	
	Headquarters	

B4.5.2 A simulation for the assessment of a speech interface with respect to message length (abbreviated specification).

(a) Specification of the simulated task The actions in Figure B7 must be elicited by representation of their goals in the simulated context. This is achieved by the specification of a hypothetical operational scenario, which might be as follows.

The (simulated) observer is presented with a number of targets on a battlefield. The task goal is to engage these by generating fire information messages and by entering them to the computer. The fire information must be generated by calculating target grid reference using map and geographical aid; by using a cribsheet to decide the required amount of ammunition according to target identity; and by selecting a battery according to the resources available at each. The observer would be able to record information either in manuscript or by means of the computer (or such facilities might be controlled experimentally). The observer enters the information and continues to engage targets until they have all been eliminated.

The steps to be followed by the observer might be as follows:

- look at battlefield display and find target to engage
- calculate the grid reference using a map
- (record it)
- look up the amount of ammunition required for the target type (crib sheet)
- look up which batteries have resources available (own notes)

Table B6 Task simulation specification

		Entity	Attributes
A	ACQUIRE INFO		
1	Battlefield sources	Target	- Identity - location on battlefield - current state
		Shell effect	- location
2	Map sources	Map	- battlefield topography - grid structure
		Geog aid	- size - function
4	Written sources	Cribsheet/notes	- function
5	Derivation of fire info	Targets	- identity - location - current state
		Batteries	- resources
		Ammo cribsheet	- function
B	RECORD INFO		
1	Notes/cribsheet	Paper	- tabular format
		Pencil	
2	Map	Paper map	- battlefield topography
		Pencil	
3	Computer	see 'D'	
C/D	COMMUNICATE INFO (BY COMPUTER)		
1	Speech input visual output interface		
		Microphone	- location
		Display	- spatial config. of vis. info. - input/output relationship
		Data string	- message length - data class - current contents
		Fields	- identity - message constraints
		Cursor	- location
2	Speech input/speech output interface		
		Microphone	- location
		Display	- input/output relationship - rate of feedback - chunking of feedback
		Input/output entities	- recognition error rate
		Data string	- message length - data class - current contents
		Fields	- identity - message constraints
		Cursor	- location

- record/enter and edit engagement message
- grid reference
- target type
- battery identity
- number of rounds
- send the message
- check the effect on the battlefield
- decide whether the target is destroyed; if "no" repeat attack: if "yes" choose next target.

(b) Specification of critical entities and attributes. Critical entities are the objects of critical actions. Their attributes are critical if the outcome of the action is contingent upon the state of the entity. These were specified for the critical actions identified in Section B4.4.2.2, and are presented in Table B6.

B5. EVALUATION

The TSM is better adapted to support empirical evaluation than was the initial approach described in Section B2. Although not demonstrated directly, it should also support analytic assessment, as the task model is a product of the interaction model: in analytic assessment the product of the interaction model would be a device evaluation report. However, post hoc application of the TSM did identify some practical difficulties which require resolution and refinement of the method. These are now identified.

B5.1. Evaluation: Preliminary task description

The procedure for the collection of task data seems adequate; however, inexperienced users of the TSM may encounter some difficulty in generating the hierarchical description. In fact, the process depends upon the judgement of the analyst in segmenting the task into its actions. The TSM might benefit from elaboration of the principles for doing this.

B5.2. Evaluation: Expanded task description

The comments in Section B5.1 also apply here: furthermore, the TSM might be expanded to provide criteria for assessing the completeness of the description, i.e. guidance enabling the user to decide when to stop checking for completeness.

At a pragmatic level, reference to the description is facilitated by the introduction of a consistent numeric notation for actions. The TSM does not, at present, provide this explicitly (although a partial notation has been introduced in this report).

B5.3. Evaluation: Future task description

The TSM requires modification to account for the fact that actions in the current task may not always be deleted (see Section B4.3.3.). However, the applications of the TSM described in this report did not fully test the procedure for generation of a future task description. This was, firstly, because of the rudimentary model of the future device (and hence of the future task); and, secondly, because of the fortuitous similarity between the current (speech) task and the future computerized task using speech. It is the opinion of the writer that generation of the future task description will have to rely on the practical judgement of the assessor to a greater degree than is the case with other parts of the TSM: the procedure may not be fully specified within the present project.

B5.4. Evaluation: Task model

The relationship between the interaction model, the task description and the task model generated by convolving them needs further specification. The mechanisms for (1) selecting the interaction model, and (2), identifying critical actions by using the interaction model, require fuller specification. This will be a primary goal in the development of the diagnostic manual in which the interaction model(s) will reside.

B5.5. Evaluation: Simulation specification

As in B5.4., the use of the interaction model for identifying critical entities and attributes is underspecified and requires elaboration. In addition, the TSM should provide further guidance in the generation of (an) appropriate scenario(s) for the operationalization of the task simulation.

APPENDIX C

DEVELOPMENT OF THE DEVICE SIMULATION METHOD

ADAPTED FROM "SYSTEM SUBJECT ASSESSMENT STUDY 1: THE SIMULATION OF A CONNECTED-WORD SPEECH RECOGNIZER " (Ergonomics Unit report dated August, 1988, reference: MACL/R/35/88/1).

C1. INTRODUCTION

C1.1 Background

This appendix describes experiments performed with two purposes:

- to support the development of the simulation of a connected-word speech recognizer to be employed in a usability assessment
- to test and to refine a method for specifying and implementing simulations of speech I/O devices (Life and Long, 1987).

The connected-word recognizer was the target of assessment in the context of the task of battlefield observation, and this required a simulation of the device, suitable for evaluation within a simulation of the battlefield task. The device simulation was to be implemented using a human subject to supplement the functionality of available technology; this "system subject" (SS) communicated with the subject representing device users (user subject - US) by means of a communication device (CD). The performance of the simulation was to be assessed experimentally in a system subject assessment study, and it was to be optimized by applying ergonomic intervention to the interface between the SS and the CD.

C1.2 Experiments

The experiments described here address, firstly, the characterization of the performance of an extant speech recognizer (Marconi Macrospeak); secondly, the evaluation of the performance of a single SS attempting to simulate this device using a CD with a QWERTY keyboard; and, thirdly, the comparison of the performance of SSs simulating the device with the QWERTY keyboard with that attained using a keyboard especially configured on the basis of an ergonomic assessment of SS-CD interaction. [The first two studies were briefly reported in Life, Long and Lee (1988).]

In addition to the specific purpose of the experiments in the development of a simulation, the intention also was to advance a general method for simulation development. Life and Long (1987) proposed an approach which could form the basis of such a method. This involved the following stages:

(1) specification of simulation elements
(2) specification of TD performance parameter values
(3) assessment of performance of an SS using a minimal CD against the performance of the TD
(4) development of an SS-CD cognitive compatibility model
(5) specification and implementation of ergonomic intervention
(6) evaluation of simulation performance following intervention

Experiment 1 (characterization of Macrospeak performance) was the means of fulfilling the requirements of stage 2; Experiment 2 (evaluation of SS performance when using a QWERTY keyboard) addressed the third stage of the method; and Experiment 3 (comparison between performance using the QWERTY keyboard and using the configured keyboard) represented the sixth stage.

The next section of this document describes, then, the application of the proposed method to the simulation of Marconi Macrospeak. In the final section, the rèsults of its application are evaluated and the method is refined.

C2. DEVELOPMENT OF A SIMULATION OF MARCONI MACROSPEAK

C2.1 Rationale for the choice of device.

The EU/UCL methodology enables the evaluation of the suitability of future devices to support battlefield tasks. In order to evaluate the success of the simulation development method it is necessary to compare the results of applying it, i.e. the simulation, against an appropriate criterion. Where the criterion is the performance of a future device, the criterion tends to be underspecified. The development of the method is facilitated, then, if a clear criterion may be specified: this is most likely to be possible where the device is available for observation, i.e. an extant device.

Battlefield observation is currently performed using a highly constrained language. The *functional* specification of current- generation recognizers (as distinct from the performance specification) is actually close to that demanded by the task at least with respect to vocabulary size. It is, therefore, of interest to know whether, given more reliable recognition performance, such devices would be usable to perform an observation task and, if not, to determine the requirements for an adequate device. The devices to be simulated to make these evaluations are a current recognizer and a less error-prone version of it deemed to be more usable, in order that the performance using these may be compared with that attained given a keyboard interface.

C2.2 Rationale for the proposed simulation development method

It is assumed that a factor influencing the performance of a human simulation system is the compatibility between the SS and the CD. This compatibility may be with respect to cognitive representations, procedures or the physical characteristics of the SS-CD interface, and it may be evaluated at levels of the task, of the communication exchanges between SS and CD, or of data I/O. (Life, Long and Lee, 1988). Reducing incompatibility will improve the performance of the simulation system. The rationale behind the method is, then, the identification of incompatibilities, and the specification of ergonomic intervention to reduce incompatibility. The method achieves this by:

- specifying required simulation performance (stages 1 and 2)
- assessing performance of the SS attempting to meet this requirement using a simple CD (stage 3)
- specifying a compatibility model (stage 4)
- specifying and implementing intervention to reduce incompatibility (stage 5)
- evaluating the simulation against the original specification (stage 6).

C2.3 Applying the proposed method

C2.3.1 Stage 1: Specification of simulation elements

An analysis was performed of the task of a Forward Observation Officer (FOO) and observation party (OP) operating in the Royal Artillery (RA) - see Appendix B of this thesis. The analysis was used to specify a simulated observation task. The simulated task demanded a computer for transmitting target data to artillery. This was modelled on a database system which was expected to enter service with the RA: it utilized a form filling dialogue, with local page editing facilities. The model for the speech recognizer for entering data to the computer was the radio communication between signaller and a battery command post when mediating information generated by a FOO, i.e. it was assumed that the speech interface should accept FOO output of similar type and at the same rate as is currently the case when voice radio is used. It was assumed also that the recognizer would be used in conjunction with a visual display, presenting feedback of entered data and system messages.

In summary, then, the required simulations were of a database system with either a keyboard or speech entry and visual feedback of data, the speech version(s) being operated in the manner of the extant voice radio communication system.

C2.3.2 Stage 2: Specification of TD performance parameter values

The simulation seeks to emulate certain "critical" parameters of the performance (and behaviour) of the TD. Critical parameters are defined as those which influence, or are influenced by, the functionality and usability of the target technology in the context of interest. In the first instance it was assumed that these would be the error characteristics of the device when operated at the speaking rate of the signaller transmitted data, and the device response latency.

A confusion matrix was used to characterize device errors: this indicated the probability that a given (legal) input would generate a particular output in error. The latency parameter selected was that of total response time: from onset of the input utterance to the completion of the device response. The experiments, then, evaluated both the TD and the simulation in terms of these parameters (dependent variables).

The independent variables thought to be potentially relevant to CD design were:

- string length
- whether the string comprised numeric or alphabetic data
- whether the class of data was predictable to the subject

Experiment 1 intended to assess Macrospeak's performance when presented with an input data set varying with respect to these parameters.

Experiment 1: assessment of Macrospeak performance

Method

(1) *Stimuli.* The simulated battlefield computer system demanded the entry of five fields of data to complete a page for transmission. Each field consisted of a field name and a data string. Four of the fields contained numeric data (a string of either 2, 4 or 6 numerals) and one field required the entry of a word or acronym (a string of 2, 4 or 6 letters). Thus a typical "page" of data might be as follows:

GRID 639471
NUMBER 1004
TARGET MORTAR
BATTERY 02
ROUNDS 12

The experimental stimulus set comprised four blocks of 15 pages of data. Two of the blocks repeated the fields in a fixed sequence while the others exhibited random order in the occurrence of fields.

Two informal preliminary studies were performed in the development of the stimulus utterance set. The first of these selected a speaker with speech characteristics well-suited to the operation of Macrospeak. Given a (relatively confusible) vocabulary comprising the names of the letters of the alphabet and the vocabulary demanded for the observation task, the chosen speaker achieved a recognition rate of 94% when using the device in isolated-word mode. The second informal study determined the speaking rate of the signaller and hence the rate at which the stimuli were to be presented to the device. Video taped sequences of a forward observation mission (Life 1987d) were studied and utterances timed using a stop watch. Over a total of 30 fire orders transmitted, the signaller spoke at a mean rate of 2.86 words per second (sd = 0.832), i.e. at approximately 3 words per second.

(2) *Apparatus.* The apparatus was configured such that the onset and offset of tokens entered by the speaker to Macrospeak were timestamped and recorded on high fidelity audio tape using a Revox B77 machine. The experimental stimuli were presented visually to the speaker, in sequence, at the previously-determined rate of approximately 3 words per second with a gap of 3 seconds between each field. The presentation was controlled automatically by a BBC Master Microcomputer. (Figure C1).

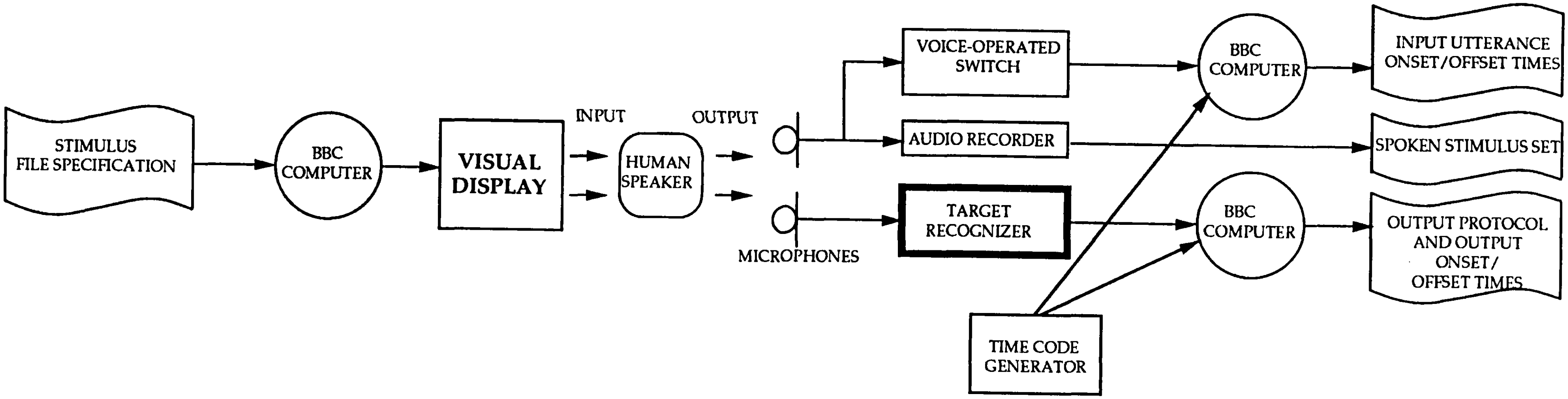


FIGURE C1. Experiment 1. Schematic outline of apparatus and data flow.

	Reject						battery	rounds	HQ	CP	AD	HV	LT	zone	boat	dump	area	tank	mortar	rocket	troops	bridge	patrol
1																							
2	.04		.08								.08			.04									
3				.04										.08									
4		.04																					
5						.04																	
6	.08																						
7											.58		.04										
8	.04		.08		.04	.04																	
9						.04																	
0																							
grid																							
number																							
target																							
battery																							
rounds																							
HQ																							
CP																							
AD				.75																			
HV																							
LT																							
zone																							
boat																							
dump																							
area																							
tank																							
mortar																							
rocket																							
troops																							
bridge																							
patrol																							
Insert																							

Values shown are probabilities of misrecognition. There were no instances of insertion error. Error probabilities for each token were determined by expressing error frequency as a proportion of the total number of instances of the token in the stimulus set.

FIGURE C2: Experiment 2 - Error matrix for Macrospeak in the recognition of connected words

Macrospeak was set up in accordance with the manufacturer's instructions. It was connected to the host computer (a BBC Master Microcomputer) through the RS423 serial link at 9600 baud rate. The microphone was the SHURE SM01A. The recognition quality threshold was set at 1.23, the highest possible reject macro that Macrospeak would allow. Hence, the incidence of rejection of inputs by the device (as unrecognizable) was deliberately kept at the lowest level.

(3) *Procedure.* Following the recognizer manufacturer's recommended enrolment procedure, the speaker was told to read the experimental stimuli aloud at the pace determined by their rate of visual presentation. He was told to use a consistent tone of voice to maximize the probability of correct recognition and was allowed to rest his voice for five minutes after the recording of each block.

Results.

Only a subset of the entered stimuli were used for the calculation of results. Specifically, thirty fields were pre-selected in each block, such that they included five instances of each of the classes of input stimuli, i.e. alphabetic (two, four and six character) and numeric (two, four and six character).

(a) *Error analysis.* The output protocol of the device was compared with the stimuli which had been entered by the speaker. As there was no reason to believe that blocks with the fields in fixed order would elicit different performance from blocks with fields in random order, incorrect recognitions were identified in each block, their probability with respect to their overall frequency of occurrence was calculated, and the probabilities were

FIGURE C3: Experiment 1. Macrospeak response time (Tt)

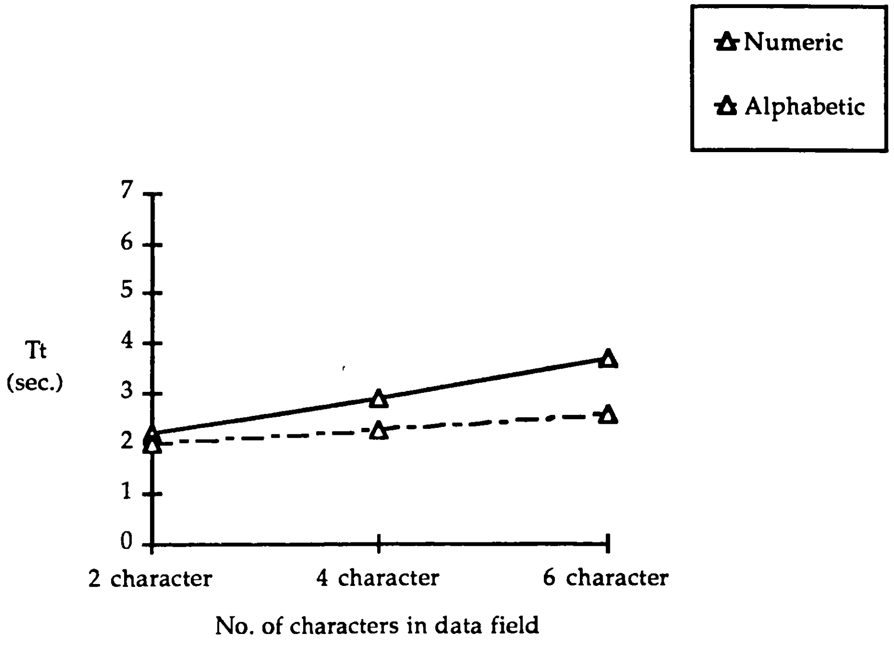


TABLE C1: Experiment 1. Macrospeak total response time (sec)

		NUMERIC			ALPHABETIC		
		2 character	4 character	6 character	2 character	4 character	6 character
TOTAL	mean	2.22	2.92	3.68	1.98	2.28	2.57
FIELD	mean	2.14	2.74	3.71	1.75	2.36	2.37
ORDER	*s.d.*	*0.549*	*0.675*	*0.35*	*0.19*	*0.883*	*0.73*
FIXED	n	10	10	8	8	9	10
FIELD	mean	2.3	3.1	3.64	2.18	2.19	2.76
ORDER	*s.d.*	*0.716*	*0.407*	*0.207*	*0.485*	*0.785*	*0.9*
RANDOM	n	10	10	8	9	9	10

entered in a single confusion matrix (presented in Figure C2). The overall probability of error was 0.12, which comprised a number of low frequency confusions, but two high frequency confusions: the word "seven" was recognized as "zone" ($p = 0.58$) and the abbreviation "AD" was recognized as "eight" ($p = 0.75$).

(b) *Device latency analysis.* Device latency was calculated on error-free data. Thus when one of the pre-selected fields included a recognition error, it was rejected and the time data from the next equivalent, but error-free, field was used instead. The total response time (T_t) was calculated by subtracting the time stamp of the stimulus utterance onset (U_1) from the timestamp of the display of the last character of the response (D_0). Table C1 presents the mean T_t and standard deviation for each stimulus length. Figure C3 illustrates the mean T_t as a function of stimulus type; as with errors, results are pooled across "fixed order" and "random order" blocks.

There was a general trend for T_t to increase with the number of characters in the field. The rate of increase appeared higher for numeric than for alphabetic fields. Performance was faster for all alphabetic than for all numeric strings. In general, the response time showed high variability.

Discussion

The trends in T_t exhibited in Figure C2 might be expected, given that speech consists of sequences of elements (words) articulated serially: one would expect T_t to increase with the number of elements presented. The difference in the functions derived from alphabetic and numeric stimuli are explained by the fact that fields comprising alphabetic strings (although varying in length) were uttered and processed as single units (words), whereas for numeric fields the number of characters in the string determined the number of words that had to be uttered and processed by the device. Consequently, T_t for numeric fields showed a greater increase with string length than for alphabetic fields.

When planning the experiment, the high incidence of recognition errors had not been foreseen. In the case of some classes of stimulus, the strategy of only calculating latency on trials on which the field was recognized correctly resulted in means having to be calculated from a reduced number of data points (i.e. $n < 10$) (see Table C1). If the same scoring strategy is applied in future, the number of trials should be increased to improve the chance of obtaining balanced data.

C2.3.3 Stage 3: Assessment of SS performance using a QWERTY keyboard (Expt. 2)

The simulation demanded the emulation of the specification determined in Experiment 1, by means of an SS interacting with a CD. The simplest CD for the SS to perform the simulation task comprised a QWERTY keyboard and visual display. Experiment 2 intended to assess the performance of an SS attempting to emulate TD performance using a QWERTY keyboard incorporating minimal CD enhancements. Performance was measured against the critical parameters considered in Experiment 1.

Method

(1) *Stimuli.* The stimulus utterance set developed for Experiment 1 and previously recorded on audio tape was utilized in Experiment 2.

(2) *Apparatus.* The experiment was performed using the EU/RSRE experimental testbed, configured as shown in Figure 4. The onsets and offsets of recorded utterances were timestamped as they were replayed to the SS via headphones. The SS's individual keystrokes were also timestamped. The keyboard was that of a BBC Master Microcomputer with standard QWERTY layout and a numeric keypad located on the right of the machine. This wrote to a Philips monochrome display monitor. The five field names were generated as complete strings by means of special function keys; thus a field was completed by depressing the field name key, then entering data using character keys.

(3) *Subject.* The subject (MC) was a male ergonomist (aged 25) who was a regular keyboard user, but not a touch typist. He was selected on the criterion that his keyboard skill was representative of the assumed population of SSs.

(4) *Procedure.* An informal pilot study revealed that it was not practicable for a subject to memorize the entire confusion matrix of the TD for the purpose of simulating device recognition errors.
Furthermore, the response latency of the TD was assessed as being close to the maximum typing rate expected in the SS population. In view of the results of the pilot study, the

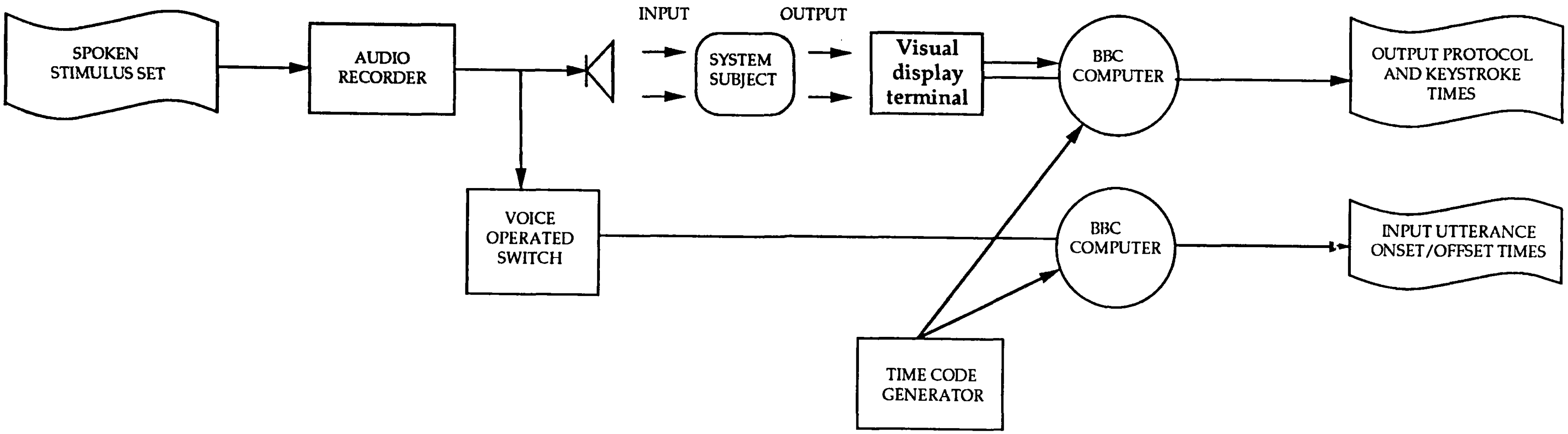


FIGURE C4: Experiment 2. Schematic outline of apparatus and data flow.

	Typos.		Reject		battery	rounds	HQ	OP	AD	HV	LT	zone	boat	dump	area	tank	mortar	rocket	troops	bridge	patrol
1																					
2																					
3																					
4																					
5																					
6																					
7								3													
8																					
9																					
0																					
grid																					
number																					
target																					
battery																					
rounds																					
HQ																					
OP																					
AD				2																	
HV																					
LT																					
zone																					
boat																					
dump																					
area		1																			
tank																					
mortar	motar morata																				
rocket																					
troops																					
bridge																					
patrol																					
Insert				2																	

Total frequency over all blocks

FIGURE C5: Experiment 2 - Error matrix for subject MC

experimental subject was instructed to transcribe the spoken data as rapidly as he could with minimal errors. He was told to insert the two most frequent confusions when he felt he had time. The subject was presented with one practice block, and then transcribed the experimental blocks, with five minutes rest between each block.

Results.

The results were analysed in the same way as in Experiment 1.

(a) *Errors*. The confusion matrices for each block are combined in Figure 5. The most important features of the data are:

(i) a low incidence of "unintentional" keying errors (a total of 6 errors in the four blocks of experimental data)

(ii) an incidence of simulated device errors of p = 0.015, with the specific confusions of "seven" and "AD" occurring with probability 0.13 and 0.5 respectively.

(b) *Latency*. Table 2 presents mean T_t measured from the time of speech stimulus onset to the time of the last keystroke of the string. There was no evidence to suggest that the subject gained advantage from knowledge of the order of the fields, indeed performance on fixed order blocks seemed slower than on random blocks (see Section C2.4), so data from fixed and random order blocks are combined and presented in Figure 6.

There was a general trend for T_t to increase with the number of characters in the field. The rate of increase appeared higher for numeric than for alphabetic fields. Short alphabetic strings (2 characters) appeared to be processed slightly more slowly than

FIGURE C6: Experiment 2. System subject response time (Tt) QWERTY keyboard

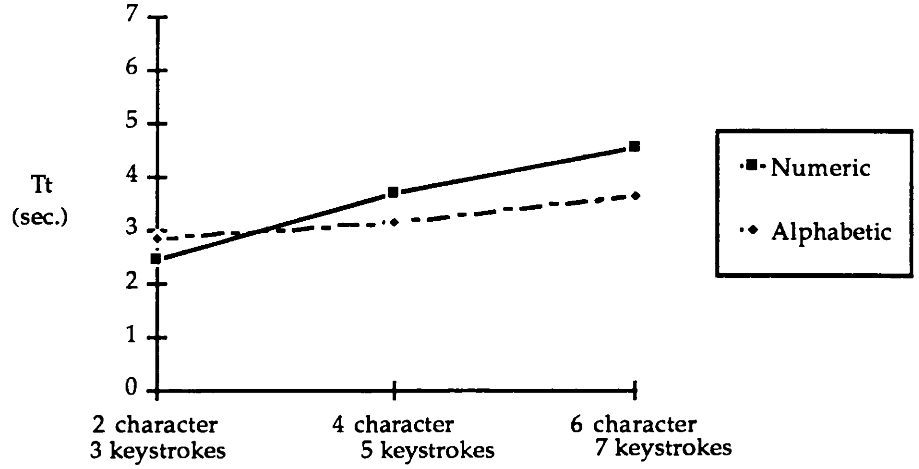


TABLE C2: Experiment 2. SS total response time (sec.) QWERTY keyboard

		NUMERIC			ALPHABETIC		
		2 character	4 character	6 character	2 character	4 character	6 character
TOTAL	mean	2.48	3.7	4.56	2.84	3.15	3.66
FIELD	mean	2.44	3.76	4.63	2.87	3.3	3.76
ORDER	*s.d.*	*0.631*	*0.716*	*0.798*	*0.244*	*0.544*	*0.564*
FIXED	n	10	10	10	9	10	8
FIELD	mean	2.51	3.64	4.49	2.8	2.98	3.58
ORDER	*s.d.*	*0.241*	*0.724*	*0.427*	*0.261*	*0.384*	*0.545*
RANDOM	n	10	10	10	9	9	10

numeric strings of the same length; however, as string length increased beyond two, there was an increasing alphabetic speed advantage.

Discussion

Although the performance of the experimental subject was, by typing standards, accurate, the keying errors that were made in the transcriptions of words (i.e. alphabetic strings) were uncharacteristic of errors generated by recognizers. For example, on one occasion, the string "MOTAR" was generated instead of "MORTAR". Such miskeying may have serious consequences for the fidelity of a device simulation.

The experimental subject, like the pilot subject, appeared to encounter difficulty in the insertion of simulated recognition errors into his output. Although he was instructed to insert only two classes of error, the frequency with which this was done was, in both cases, lower than that of the target specification. There would appear to be two potential reasons for this. One reason may have been that the subject was so loaded with the primary task of transcription that he simply did not have time to make the decision to insert "errors". Such a difficulty was reported by the pilot subject. A second reason was that the insertion of a confusion such as one requiring "ZONE" to be generated instead of "7" forced the subject to produce an additional three keystrokes. As the task was already quite highly paced, this required considerable extra effort on the part of the subject. An informal analysis suggests that performance on trials following insertions may have been disrupted, supporting the view that error insertion increased the load imposed by an already demanding task.

In summary, the SS using a CD with QWERTY keyboard generated a small but potentially important number of keying errors uncharacteristic of the TD, and failed to insert only two simulated confusion errors with sufficient frequency. It would be expected that a requirement to insert a large number of error classes would be difficult, and that a requirement to insert them with frequencies accurately representative of the TD may be impossible.

The temporal response of the simulation followed general trends of additivity with string length. As with the TD, this would be expected in a system responding to speech stimuli. However, the apparent difference in the slopes of the curves derived from alphabetic and numeric data cannot be assumed to be a consequence of mechanisms identical to those of the TD. As in Experiment 1, part of the effect was probably due to the fact that the utterance of the alphabetic fields took less time than for numeric. In addition, however, although the SS using the QWERTY keyboard responded to both numeric and alphabetic stimuli in the same way at the I/O level - by generating a keystroke for each character - it would be expected that an experienced typist would generate strings of keystrokes corresponding to meaningful English words faster than strings corresponding to random numerals. The results offer suggestive evidence, then, that the subject was treating the numeric and alphabetic fields in a qualitatively different way; possibly that numeric strings were being handled as character-units and alphabetic strings as word-units.

C2.3.4 Stage 4: Development of an SS-CD Cognitive Compatibility model

Life and Long (1987) argue that some or all of the discrepancies between the performance of the simulation and the specification of the TD performance may be interpreted as a consequence of incompatibility between the SS and the CD. Incompatibility may be attributed to three sources: cognitive representations, task procedures or the physical characteristics of the SS-CD interface. If the SS and CD hold incompatible representations of task entities, or if the CD demands procedures incompatible with SS skills, or if the CD interface is physically configured such that SS skills cannot be exercised optimally, then the behaviour of the SS will be influenced adversely and the system performance will be sub-optimal. Incompatibility may be further classified in terms of the level of SS-CD interaction at which it occurs: at the levels of the task, of communication interchanges or of data input/output (I/O). The source and level at which incompatibility occurs will have implications for the intervention appropriate to minimize it (see Section C2.3.5).

Experiment 1 set the performance specification of the TD for the purposes of the simulation. Experiment 2 assessed the performance of a human simulation attempting to meet this specification. [NOTE: An issue remaining to be addressed relates to the acceptable degree of performance mismatch between TD and simulation i.e. fidelity requirements. See Section C3.] The following were the main points of difference with respect to output errors:

(1) the simulation exhibited keying errors not exhibited by the TD

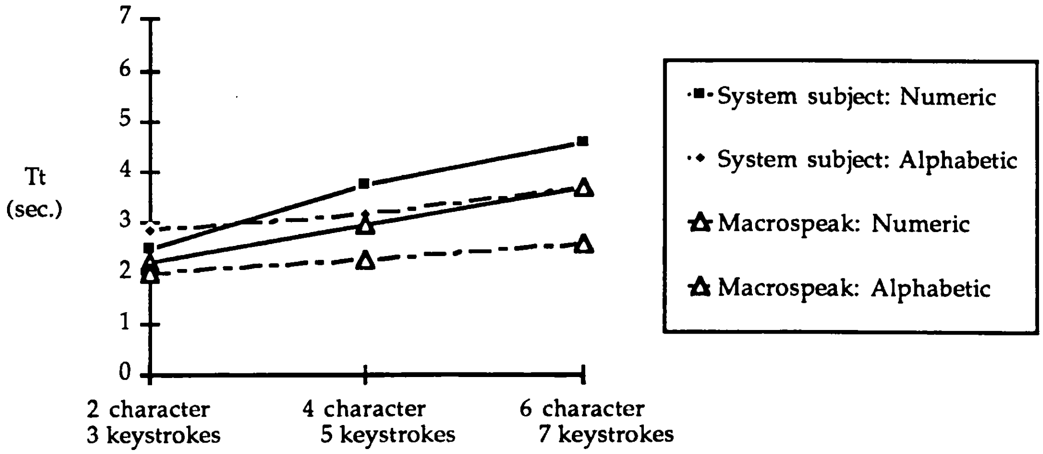
(2) the simulation was inadequate in its insertion of recognition errors. The dynamic performance of the two systems is summarized in Figure C7, where it may be seen that:

(3) the simulation was slower in its response than the TD

(4) the discrepancy between TD and simulator response speed increased with string length, particularly for numeric fields.

These differences in performance are now attributed to sources of SS-CD incompatibility. The inclusion of character errors, uncharacteristic of the TD, in the output of the simulation might have been attributed to procedural incompatibility at the I/O level, i.e. the SS not having adequate keying skill. However, all typists occasionally make errors when performing fast, and the SS in fact performed quite accurately relative to that which might generally be

FIGURE C7: **Experiments 1 and 2. Comparison between Macrospeak and system subject performance (QWERTY keyboard)**



expected. Rather, then, the character errors were attributed to representational incompatibility at the I/O level. The TD and SS represented information in word units; however, the CD offered only character representations (character keys). The CD, therefore, demanded a translation from word representation to character representation, and under time pressure this translation was occasionally incorrect due to miskeying. Miskeying produced errors which the TD could never generate because they were character errors.

The failure to incorporate confusion errors at the appropriate frequency has earlier been interpreted as being a consequence of two possible factors.

(a) *Additional keystrokes.* The necessity of producing additional keystrokes when inserting an error results from representational incompatibility at the I/O level (as above).

(b) *Cognitive load.* The SS apparently had difficulty actually deciding when to insert an error - difficulty that would be greatly exacerbated if the requirement were to simulate a more realistic confusion matrix. This could have occurred because the SS did not have time to allocate the additional effort necessary to transform the mental representation of the subject instructions (i.e. instructions to insert specific confusions with given random probabilities) to the representation demanded by the CD (i.e. manual action). This would be an instance of representational incompatibility at the task level.

Turning, now, to the dynamic performance of the simulation, the slowness evident in the transcription of the alphabetic strings can, again, be attributed to communication level representational incompatibility, due to the time required to key words character by character. However, the slow performance when transcribing numeric strings cannot be explained in this way, because SS and CD use equivalent representational units (numeric symbols and numeral symbol keys). Rather, slowness may be due either to inadequate keying skill by SS (i.e. procedural incompatibility at the I/O level), or to sub-optimal keyboard configuration (i.e. physical incompatibility at the I/O level). In view of the fact that the SS tended to utilize the numeric keys at the top of the QWERTY board, rather than the presumably better-configured matrix numeric key-pad, the latter interpretation of the results was thought appropriate. The subject apparently ignored the matrix keys because they were not conveniently located with respect to the other keys used in the task.

Table C3 summarizes the attribution of incompatibility to account for simulation performance inadequacies.

Table C3: Experiment 2: Attribution of performance inadequacies to SS-CD incompatibility

	Representational	Procedural	Physical
Task	Inadequate inclusion of TD confusion errors		
Communications			
I/O	Inclusion of character errors Inadequate inclusion of TD inclusion errors Slow performance on alphabetic strings		Slow performance on numeric strings

C2.3.5 Stage 5: Specification and Implementation of ergonomic intervention

Ergonomic intervention which reduces incompatibility should bring the performance of the simulation closer to that of the TD. Three forms of intervention have been identified: *selection* of subjects with more highly developed CD operating skills; *training* of CD operating skills; or *aiding* to support SS-CD interaction, e.g. modification. The following heuristics are offered for the selection of intervention:

(a) Incompatibility with respect to knowledge representations may be reduced by selection, training or aiding.

SELECTION:-where the required knowledge is already held by some members of the SS population, but not all.

TRAINING:-where the required knowledge is of a quantity and type readily acquired by the SS population (e.g. if a learning improvement is evident over the course of a SSAS).

AIDING:- where the information is of a form which is economically implementable in a usable task aid or device modification.

(b) Incompatibility with respect to procedures may be reduced by selection, training or aiding.

SELECTION:- where appropriate performance dynamics may be exhibited by some members of the SS population, but not all.

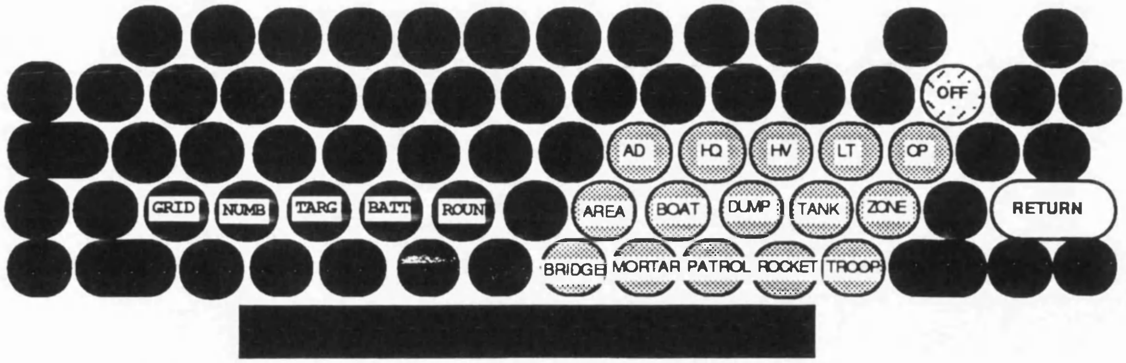
TRAINING:-where there is evidence that skills may be developed to achieve the appropriate dynamics within the user population.

AIDING:-where the dynamic performance may be achieved by an economical implementation of a usable task aid or device modification.

(c) Incompatibility with respect to the physical configuration of the interface will not usually be adequately reduced by selection or training. Task aiding will be the likely solution.

The first source of incompatibility identified in Section C2.3.4 was that created by the task demanding (and the SS possessing) word representations but the CD offering character representations. The obvious options are to change the SS representations to a character-based form, or to change the CD representations to a word-based form. An intervention of the first sort would be to employ touch typists with a highly developed mapping between word representations and representations supporting finger movements corresponding to characters.

Modifications to QWERTY keyboard



Numeric keypad

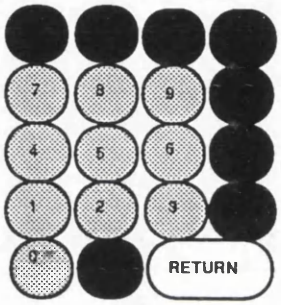


FIGURE C8: Experiment 3
Re-configured keyboard layout

However, while this would improve the adequacy of dynamic performance, it would be difficult (if not impossible) for such dynamic performance to be maintained with effectively zero probability of keying errors. As such accuracy is demanded to maintain simulation fidelity, representational aiding is favoured, i.e. modifying the CD representations at the communication level. The intervention selected was that of modifying the keyboard such that each word in the TD vocabulary could be generated by the SS making a single keystroke.

A second source of incompatibility was that existing between the representation of SS instructions relating to the insertion of TD confusion errors and SS action resulting in confusions appearing in the simulation output. The ability of the SS to learn (and to respond accordingly to) a complete confusion matrix was not tested in Experiment 2 because, analytically, the learning task appeared so large and the probability of the SS being able to utilize such a representation seemed low. Furthermore, there was reason to believe that SSs could not be selected on the criterion of possessing an ability to develop and use appropriate representations. Consequently, task level representational aiding was identified as the class of intervention offering most likelihood of success. Specifically, the representation of TD errors was transferred from the SS to the CD, so the SS no longer had to hold and take action on a confusion matrix. Pressing a key generated a correct or an erroneous (confusion) output automatically, with a probability determined by the TD confusion matrix held in CD memory as a look-up table.

The third source of incompatibility was that due to the sub-optimal physical configuration of the keyboard. It was reasoned that the failure to use the matrix numeric keypad was a consequence of the SS using both hands for alphabetic keys and finding it inconvenient to move a hand to the other keypad for the transcription of numeric strings. The implementation of the word and field keys was therefore performed such that these could be operated with the left hand, whilst the numeric keys were operated with the right. Allocation of function was made on the principle that the most frequently used (i.e. numeric) keys should be operated with the favoured hand of the largest proportion of the subject population (i.e. the right). The keyboard layout is presented in Figure 8.

C2.3.6 Stage 6: Evaluation of simulation performance following intervention (Expt. 3)

The construction of a device simulation is a process of system development. The final stage of most models of development is one of system evaluation against the requirement specification. Stage 6 of the simulation development method requires the evaluation of simulation performance against the performance specification of the TD.

Experiment 3 was to perform two functions
(a) evaluation of the simulation
(b) validation of the compatibility model, by demonstrating improvements in simulation performance as a consequence of intervention designed to reduce SS-CD incompatibility.

The experiment involved presenting a group of SSs with the same stimuli as had been used in Experiments 1 and 2, and requiring them to transcribe the stimuli using both the QWERTY and re-configured CD interfaces, i.e. the experiment utilized a within-subjects design. The intention was
(i) to evaluate the simulation by comparing system performance when SSs used the re-configured interface with the performance of the TD, and
(ii) to validate the model by the testing of hypotheses relating to expected differences in SS performance when using the two keyboards.

Concerning (ii), above, a trivial hypothesis would predict that there would be a difference in the type and frequency of errors in the outputs from the two classes of CD, e.g. that the re-configured interface would exhibit confusion errors and that the QWERTY interface would not. Because the CD would determine this effect automatically it is not considered further. The hypotheses tested related, then, to the dynamic response of the simulation, namely:

> (1) that for numeric fields, there would be a smaller effect of string length on system response time with the re-configured interface than with the QWERTY keyboard; and
> (2) that for alphabetic fields, the effect of string length on system response time would be present for the QWERTY keyboard but not for the reconfigured interface.

(1) *Stimuli.* The stimulus utterance set recorded in Experiment 1 and used in Experiment 2 was also utilized in Experiment 3.

(2) *Apparatus.* The experiment was performed using the EU/RSRE experimental testbed configured as in Figure C4. However, the SS's VDT could take one of two forms: either identical to that used in Experiment 2 ("QWERTY") or with a keyboard modified to the description presented in Section C2.3.5 ("re-configured").

(3) *Subjects.* Ten subjects performed the experiment. All were ergonomists or ergonomics MSc students aged between 20 and 34 selected according to the following criteria
(1) that they were not touch typists
(2) that they were able to copy-type text at a rate between 16 and 34 words per minute.

Method.

All subjects were required to type a paragraph of text before the experiment to ensure that they met the selection criteria.

(4) *Procedure.* The purpose of the experiment was explained to each subject. They were told that they would be required to transcribe spoken messages as quickly and accurately as possible using each of the two keyboards. The difference between fixed field order and random field order was pointed out to them. Subjects were presented with one practice block on each of the keyboards immediately before the experimental blocks. The order in which they used the keyboards was balanced across subjects. The selection procedure,

TABLE C4: Experiment 3. SS mean total response time. (Time = seconds; s.d. in italics)

	CONFIG. K/BOARD					
	alpha			numeric		
	2 char.	4 char.	6 char.	2 char.	4 char.	6 char.
TOTAL	**2.078**	**1.921**	**2.296**	**2.265**	**3.949**	**5.196**
	0.516	*0.496*	*0.619*	*0.606*	*1.012*	*1.296*
Subject 3	**2.302**	**1.891**	**2.226**	**2.5**	**3.996**	**4.786**
	0.678	*0.494*	*0.467*	*0.754*	*0.732*	*0.769*
Subject 5	**2.015**	**1.991**	**2.361**	**2.169**	**4.334**	**6.295**
	0.324	*0.507*	*0.497*	*0.432*	*0.92*	*1.086*
Subject 6	**1.868**	**1.869**	**1.768**	**2.031**	**3.476**	**4.655**
	0.313	*0.4*	*0.265*	*0.397*	*0.828*	*0.823*
Subject 7	**1.895**	**1.679**	**2.101**	**2.142**	**3.317**	**4.256**
	0.452	*0.375*	*0.407*	*0.527*	*0.778*	*0.727*
Subject 9	**2.501**	**2.418**	**2.999**	**2.709**	**5.003**	**6.776**
	0.571	*0.455*	*0.648*	*0.674*	*0.964*	*0.976*
Subject 10	**1.887**	**1.676**	**2.322**	**2.037**	**3.568**	**4.405**
	0.328	*0.372*	*0.614*	*0.502*	*0.817*	*0.784*

	QWERTY K/BOARD					
	alpha			numeric		
	2 char.	4 char.	6 char.	2 char.	4 char.	6 char.
TOTAL	**2.789**	**3.175**	**3.792**	**2.528**	**3.645**	**4.909**
	0.472	*0.574*	*0.915*	*0.564*	*0.915*	*1.214*
Subject 3	**2.754**	**2.81**	**3.3**	**2.394**	**3.523**	**4.841**
	0.436	*0.286*	*0.375*	*0.392*	*0.55*	*0.904*
Subject 5	**3.316**	**3.947**	**5.303**	**3.035**	**4.54**	**6.119**
	0.369	*0.373*	*0.735*	*0.691*	*0.741*	*0.867*
Subject 6	**2.62**	**3.106**	**3.589**	**2.405**	**3.112**	**4.16**
	0.374	*0.358*	*0.522*	*0.253*	*0.226*	*0.279*
Subject 7	**2.551**	**2.817**	**3.452**	**2.202**	**2.942**	**3.932**
	0.432	*0.392*	*0.623*	*0.315*	*0.338*	*0.453*
Subject 9	**3.013**	**3.512**	**4.085**	**3.036**	**4.641**	**6.524**
	0.336	*0.52*	*0.54*	*0.456*	*0.467*	*0.532*
Subject 10	**2.481**	**2.86**	**3.025**	**2.095**	**3.107**	**3.877**
	0.35	*0.392*	*0.279*	*0.32*	*0.989*	*0.347*

TABLE C5: Analysis of variance (Results of Experiment 3)

SOURCE	df	SS	MS	F	p
A. Keyboard type	1	98.25	98.25	396.17	<.001
B. Character type	1	414.59	414.59	1671.73	<.001
C. String length	2	641.55	320.78	1293.47	<.001
D. Subjects	5	323.18	64.64	260.65	<.001
AxB	1	143.6	143.6	579.03	<.001
AxC	2	1.23	0.62	2.5	N.S.
BxC	2	256.62	128.31	517.38	<.001
AxBxC	2	30.63	15.32	61.77	<.001
AxD	5	34.38	6.88	27.74	<.001
BxD	5	40.19	8.04	32.42	<.001
CxD	10	59.89	5.99	24.15	<.001
AxBxD	5	6.69	1.34	5.4	<.001
BxCxD	10	19.55	1.96	7.9	<.001
AxCxD	10	3.54	0.35	1.41	<.05
AxBxCxD	10	11.45	1.15	4.64	<.001
Residual	1368	339.46	0.248		
TOTAL	1439	2424.8			

practice trials and testing on the four blocks of experimental trials for each keyboard took a total of approximately 120 minutes for each subject.

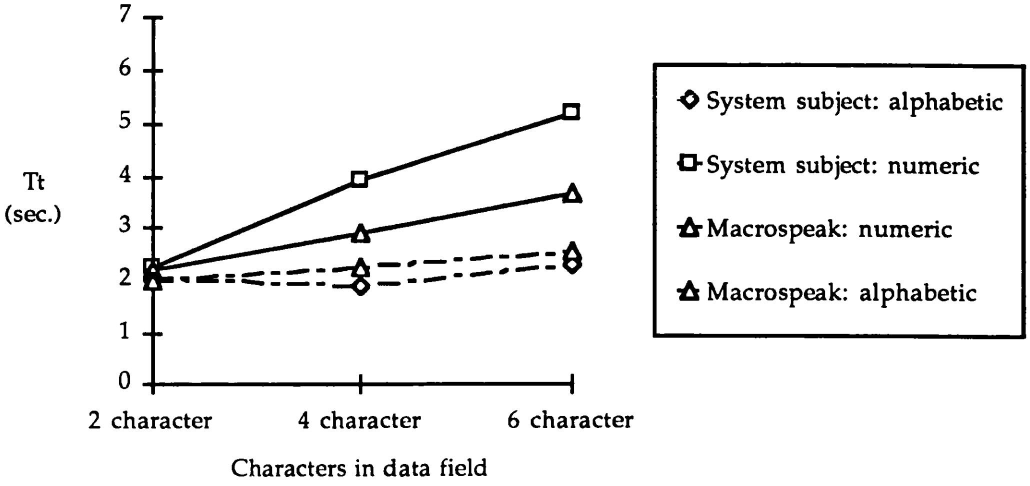
(5) *Results*. Due to equipment failures, results from four subjects had to be discarded. Analysis was performed on subjects 3, 5, 6, 7, 9 and 10. As in Experiments 1 and 2, data were analyzed from thirty fields pre-selected in each block. There was no requirement for SSs to insert simulated TD errors, so it was expected that mismatches between the input protocol and subjects' output protocols would be few. The observed incidence of miskeying errors was 2.3%, and of errors attributable to SS's making incorrect entries, 0.72%.

Total response time (T_t) was calculated for each trial as in Experiments 1 and 2. Means and standard deviations were calculated for each subject in each condition and are presented in Table C4.

(a) *Evaluation of simulation..* Figure C9 summarizes the performance of subjects when using the re-configured keyboard and presents this with the performance of Macrospeak. Figure C9 indicates that the simulation performance was similar to that of Macrospeak for alphabetic strings. For numeric strings, performance was also similar when the field contained only two characters, but there was an increasing discrepancy between Macrospeak and simulation performance as string length increased.

The results suggest that the macrokey enhancement to the SS interface may have been successful but not the intervention to improve numeric keying performance. These possibilities are addressed below.

FIGURE C9: Experiments 1 and 3.
Comparison between Macrospeak and SS performance (config. keyboard)
Mean response time (Tt) of Ss 3,5,6,7,9,10.



(b) *Validation of the compatibility model.* The compatibility model is validated if the experimental hypotheses are shown to be supported by the experimental data. Figure C10 presents a summary of data obtained using the two types of interface.

It may be seen that the form of the mean data obtained under the QWERTY keyboard condition (small filled diamonds) resembles that obtained from the single subject in Experiment 2. However, an analysis of variance (ANOVA) was applied to the data, the summary of which is presented in Table C5. The four-way interaction between subjects, keyboard, character type (alpha vs numeric) and string length was significant. ($F_{[10,1368]} = 4.64$: $p < 0.001$). This suggested that the pattern of performance might differ between subjects. Figures C11, C12 and C13 present data from subjects 9, 7 and 3, who exhibited, respectively, the slowest, fastest and nearest to mean response times. They reveal that, although the data values differ between the subjects, the main trends are the same, and that for the QWERTY keyboard these correspond to the trends identified and discussed in Experiment 2.

Tukey tests were applied to the data from subjects 3, 7 and 9 to determine the significance of differences in means obtained using the two keyboards. The results were as follows, assuming $q_T(72,inf) = 5.863$ and hence $d = 0.653$:

Numeric data

Subjects 3, 7, 9: There were no significant differences between the keyboards for numeric data-fields. The results do not support the contention that performance for numeric strings is faster with the reconfigured keyboard. The specific hypothesis that there would be a smaller effect of string length on system response time with the re-configured interface was, therefore, also not supported.

Alphabetic data

Subjects 3, 9: There were significant differences between the keyboards for 4 - and 6-character alphabetic fields ($p < 0.05$).

FIGURE C10: Experiment 3
SS response time (Tt)
Ss 3,5,6,7,9,10.

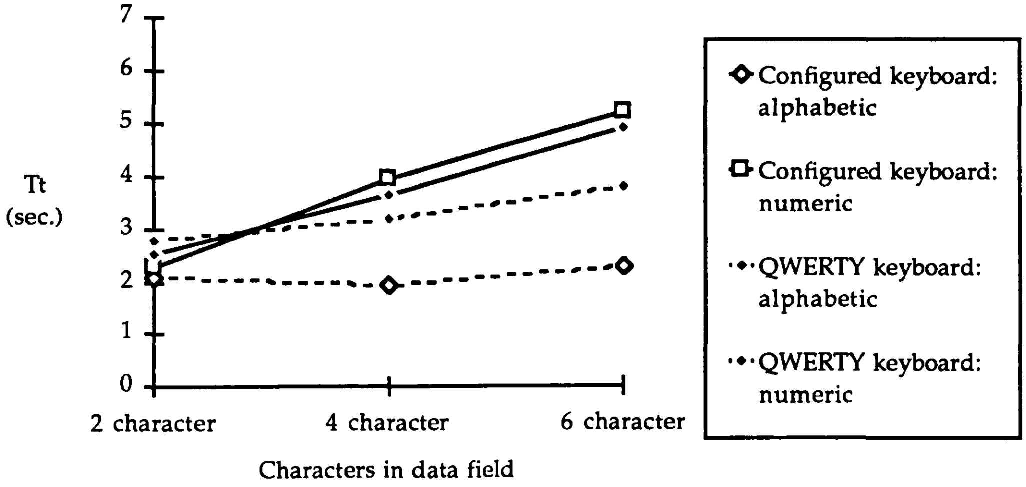


FIGURE C11: Experiment 3
SS response time (Tt)
Subject 9 ("slow")

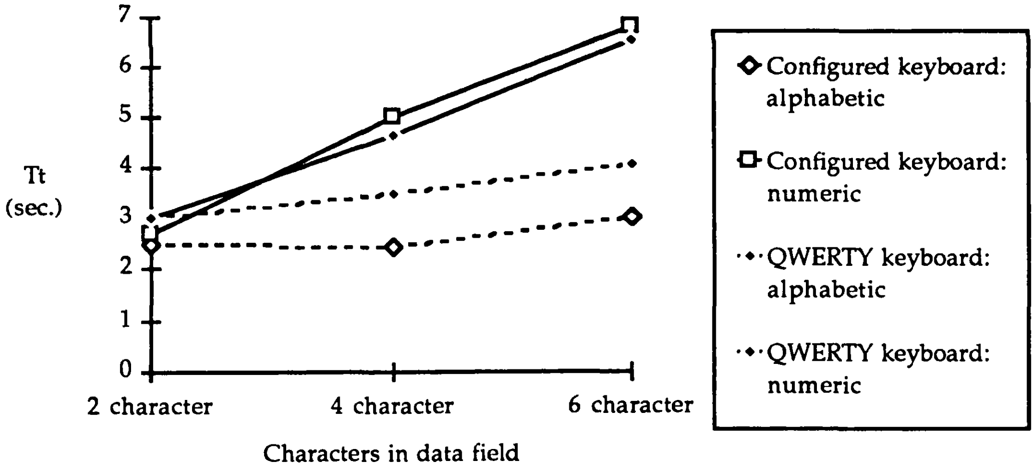


FIGURE C12: Experiment 3
SS response time (Tt)
Subject 7 ("fast")

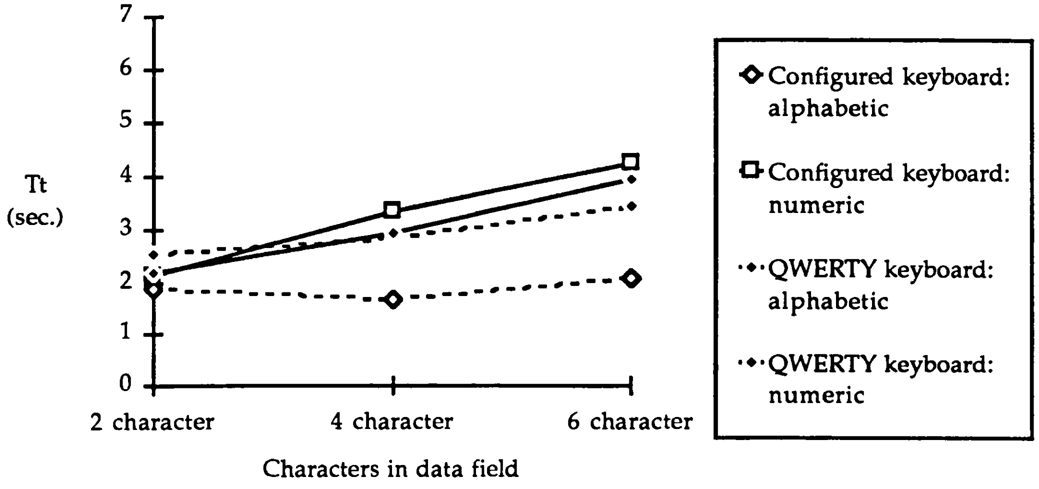
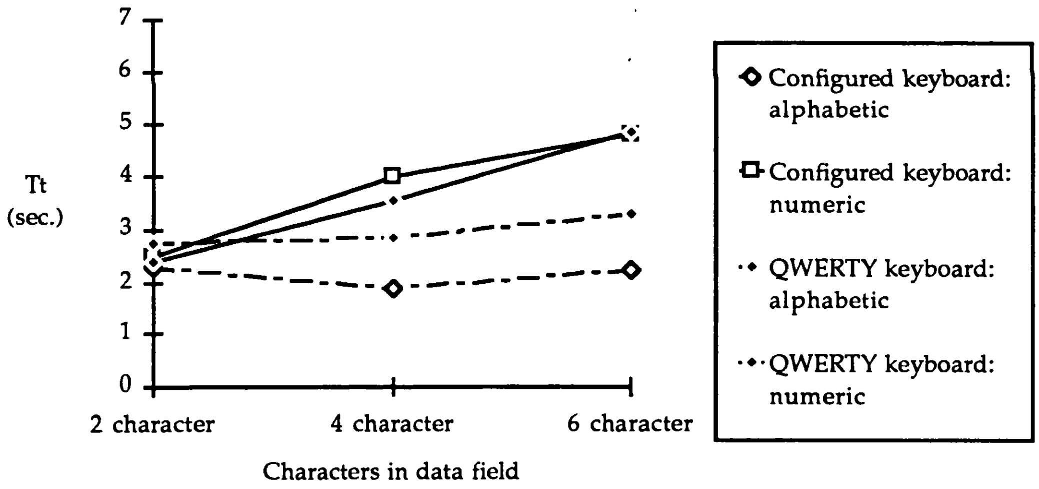


FIGURE C13: Experiment 3
SS response time (Tt)
Subject 3 ("medium")



Subject 7: There was a significant difference between the keyboard types for all alphabetic fields ($p < 0.05$).

These data support the view that the configured keyboard offers a time advantage in the transcription of alphabetic strings; this effect was consistent across SSs for longer strings and even present on short strings for some SSs.

Subjects 7 and 9: There was a significant difference between means for alphabetic 2-character and alphabetic 6 character fields when the QWERTY keyboard was used ($p < 0.05$), but not when the configured keyboard was used.

Subject 3: There was no significant difference between means for alphabetic 2-character and alphabetic 6 character fields when either keyboard was used.

These results support the specific hypothesis that, for alphabetic fields, the effect of string length on system response time is present for the QWERTY interface, but not for the reconfigured interface. However, the effect varies between subjects.

Discussion.

The results of Experiment 3 indicate a performance discrepancy between the simulation and TD for numeric strings; this was apparently due to the ergonomic intervention failing to improve SS performance in numeric keying. This result suggests either that the compatibility model requires modification, or that the ergonomic intervention is not reducing incompatibility in certain respects. In fact, both contentions are likely to be true.

The first possibility is that the re-configuration of the numeric keypad was still not optimal. In fact that used (a "calculator" configuration, with 7,8,9 on the top row) has subsequently been recognized as being potentially less suited to rapid operation than a "telephone" configuration with 1,2,3 on the top row (Conrad & Hall, 1968). Performance might be improved, then, if the latter configuration were implemented. However, the attribution of slow keying performance to physical incompatibility at the I/O level may have been only partially correct. Informal observation suggests that any such incompatibility was probably masked by an additional incidence of I/O procedural incompatibility: the SSs did not have the keying skills necessary to match TD performance. This is supported by the fact that performance using the unfamiliar re-configured matrix would actually have been slightly worse that using the familiar numeric keys at the top of the QWERTY layout. (See data for 4 and 6 character strings in Figure C10). Such an effect would be expected if the procedures held by the subjects familiar with QWERTY keyboard operation were inappropriate for the re-configured interface.

In summary, then, poor numeric performance was attributable to I/O level incompatibility with respect to physical configuration, but procedural incompatibility at the same level was due to subjects not being sufficiently skilled to operate the numeric keys. Appropriate intervention would be to further modify the matrix layout and either to enhance SS skill by training or by selection (e.g. comptometer operators). As the task of the SS involves the operation of the special re-configured keyboard in addition to numeric keys, the training option is favoured. *SSs must be given extended practice with the CD interface before the simulation is run.*

C3. EVALUATION OF SYSTEM SUBJECT ASSESSMENT STUDY 1

SSAS 1 was performed with two aims:

- to develop a simulation of Macrospeak
- to test and to refine the simulation development method.

The first of these aims has been partially met, in that the simulation matches the performance of Macrospeak when dealing with alphabetic strings. Further work is required to enhance numeric performance, but the method has supported the specification of this work.

The following sections evaluate the method itself.

C3.1 Evaluation: Specification of simulation elements

This stage specifies which elements of the target system will be included in the SSAS. It is unnecessary for the fine detail of the visual display of the TD to be included in a SSAS, for example, as this is unlikely to influence the task of the SS. However, factors such as vocabulary size of the TD and the rate of speaking expected from the US *are* important to the SS task.

In the SSAS described here, these decisions were taken implicitly according to the judgment of the experimenter. The method requires a procedure for the specification of simulation elements.

C3.2 Evaluation: Specification of TD performance parameters

It was proposed that performance parameters were critical (and hence relevant to the simulation) if they "influence, or are influenced by, the functionality and usability of the target technology in the context of interest." However, it would appear, rather, that it is the users' *perception* of the representation of the TD that is important, i.e. that which determines their behaviour in response to the simulation. On this modified assumption, additional parameters relating to the dynamics of device operation are likely to be critical. These could include (for current generation devices)

device writing rate
initial response latency
ratio between rate of input and rate of output (this parameter reflecting the user's perception of the device buffering inputs).

The method should specify working assumptions relating to critical parameters of future devices.

C3.3 Evaluation: Specification of TD performance parameter values

SSAS 1 was unusual in that the TD was extant. The procedure for its assessment was largely adequate for the purposes of the SSAS. However, issues which were not properly addressed were those of variations in TD performance following template updates and inter-speaker variability. The TD was only tested on a single pass of enrolment and the changes in performance occurring following template updates were not included in the specification for the simulation. Furthermore, its performance was only measured with a single speaker. Strictly speaking, a sample of speakers should have been tested who were representative of the population of device users.

It should be noted, though, that the method will normally be applied to the simulation of future devices, i.e. devices not available for test. For such devices, parameter values must be estimated by speech technologists. It is beyond the scope of the method to specify how the experts should do this, but the method should specify working assumptions to be taken, in the event of only crude estimates of performance being available.

C3.4 Evaluation: Assessment of performance of SS with minimum CD

The experimental method was successful in supporting this assessment, but an unresolved issue arose during results analysis. This related to the assessment of the adequacy of simulation fidelity, i.e. to the question of how accurately the simulation needs to represent the TD for the purpose of usability assessment.

A demonstration of no statistical difference between the performances of TD and simulation may be an excessively strict criterion for adequacy. A more appropriate criterion might be a demonstration that the simulation performance is not noticeably different to the user population. In the case of an extant device this could be tested empirically (e.g. after the manner of the "Turing test").

The method requires, then, the specification of working assumptions relating to fidelity requirements for device simulations. It should also provide guidance on experimental design.

C3.5 Evaluation: Development of an SS-CD compatibility model

The specification of the compatibility model was done according to the implicit judgment of the experimenter. The attribution of compatibility was not obvious, and the data might have been interpreted in a number of alternative ways. The method requires the specification of detailed procedures for attributing incompatibility.

C3.6 Evaluation: Specification/implementation of ergonomic intervention

It has been assumed that, given an adequate compatibility model and the heuristics presented in Section C2.3.5, ergonomic intervention should be easily specifiable. However, this may be optimistic: it requires some creativity on the part of the experimenter. The method would benefit from further heuristic intervention solutions which could be used as a starting point by the experimenter (e.g. "if there is a lot of variability in SS performance, try training first").

C3.7 Evaluation: evaluation of simulation performance following intervention

The procedure of Experiment 3 was adequate for evaluation purposes, in that it identified inadequacies occurring at earlier stages. The issues identified in C3.4 are also relevant here.

C3.8 General evaluation of the method

The method presented at a high level in Section C1.2 has proved usable by ergonomics specialists in the development of device simulations. It now requires specification at a lower level for non-specialists, taking account of the inadequacies identified in Section C3.1 to C3.7 above.

APPENDIX D

DEVELOPMENT OF THE USABILITY EVALUATION METHOD

ADAPTED FROM "EVALUATION OF A CONNECTED SPEECH RECOGNIZER TO SUPPORT THE TASK OF THE ARTILLERY OBSERVER"
Ergonomics Unit report dated December, 1988, reference MACL/R/52/88/1.

D1. BACKGROUND

This appendix describes an experimental evaluation of a simulated speech recogniser in the context of simulated observation task. The rationale and procedure employed for experimentation was advanced as the basis for the usability evaluation method (see Chapter 7). The task under investigation in this report is that performed by the Forward Observation Officer (FOO) of the Royal Artillery in controlling indirect fire. This task involves the FOO, and the other members of a small observation party located close to the forward edge of the battle area, sending target information to artillery engaging without direct visual reference. This information was currently conveyed to the battery command post by voice radio, but it was intended that in future such communication be mediated by a computer link.

The experiment described here was a partial assessment of the suitability of a connected word recogniser as the means by which target data might be entered into such a computer. The primary function of the experiment was to enable the development of the UEM: however, it was also intended that the results would

(1) provide information on device-user interaction which could be generalised to other applications, and
(2) provide information to those developing battlefield computers on the suitability of speech for operating devices supporting observation.

The next section of the document describes the rationale and rudimentary procedure for the assessment, which was regarded as a precursive UEM. Section D3 reports the experiment and Section D4 considers implications of the work for the development of the UEM.

D2. A PRELIMINARY APPROACH TO USABILITY EVALUATION

D2.1 The role of empirical usability evaluation in SIAM

The function of SIAM is to enable an assessor - not necessarily an expert ergonomist - to determine the usability of a current or future device in supporting a task. Because current knowledge of the interaction between users and speech I/O devices is limited, empirical assessment is recognised as being of primary importance in SIAM. Technology Assessment Studies (TASs) are a means by which such assessments might take place. TASs involve observation of system behaviour and performance in a laboratory analogue of a real-world task; however, because SIAM enables assessment of future devices, these studies demand simulation of device, user and task. The rationale underlying TAS is essentially that of conventional experimental ergonomics: predictions are made about the behaviour and performance of a system operating in the real world from empirical data obtained under controlled laboratory conditions by observation of a task simulation. To summarise, the empirical assessment of devices is achieved by means of one or more TASs, which are ergonomic evaluations utilising simulations of task, device and user.

D2.2 Rationale for a method to enable TASs

Chapter 8 describes the task simulation method (TSM). It is expressed as a representational framework for describing a battlefield task, and as procedures for transforming the representations from one to another. A similar approach was to be used to express the UEM.

Figure 6.2 (in the main body of this thesis) proposes representations involved in usability evaluation. The *task, device* and *user* are represented as *simulations* and are integrated to create an *experimental context*. An experiment (TAS) is performed to generate work system *performance data,* providing information on the behaviour and performance of the system (i.e. represented as *experimental results).* These are interpreted with respect to a *model of speech device-user interaction* to generate an interaction *diagnosis*, and, simultaneously, to refine the interaction model. The diagnosis is used to prescribe intervention to optimise performance represented as an *interface design recommendation.* Although not novel, this was the rationale underlying the approach taken to TAS's, and it was advanced as a precusor to the UEM. The UEM required the elaboration of the representational framework and specification of procedures for transforming the representations. The TAS now described provided the vehicle for specifying these enhancements.

D3. TECHNOLOGY ASSESSMENT STUDY: T:IE IMPLICATIONS OF DIRECT VOICE INPUT FOR CONCURRENT TASK ACTIVITIES OF THE FOO

D3.1 Introduction

D3.1.1. Speech interaction for the computerised FOO task

The introduction of a computer to mediate target information will impact the task of the FOO and of other members of the observation post party. One of the functions it is intended to perform is to process calls for indirect fire from FOOs. This function is currently supported by voice radio, but the computer system currently envisaged will require the user to enter such orders by means of a keyboard and visual display. The keyboard will have the QWERTY layout, supplemented by a small number of special function keys, and the dialogue will be of the form-filling type. It is claimed that computerisation will reduce the time between an order being given by a FOO and its implementation as effective artillery action.

Informal observation suggests that the requirement to operate a keyboard may:

(a) make it difficult for the FOO to send data when mobile (either on foot or travelling in a vehicle),

(b) force the FOO to stop performing off-line manual and visual activities in order to enter data,

(c) disrupt some of the FOO's information processing activities by demanding attention to the operation of the computer.

Speech interfaces are frequently claimed to be less disruptive than keyboard interfaces as regards these classes of hindrance to interaction. Furthermore the task as it is currently performed with radio utilises speech to enter information to the artillery "system" (i.e. the radio operator at the battery command post). There is, then, a potential advantage for the computer to be provided with a speech interface; however, in certain respects, current generation speech interfaces are known to compare poorly with human speech communicators. Consequently, interaction performance using the device may be inferior to that of using a radio, and the actual implications for the performance of the overall task are unknown. SIAM is used to address this question and to identify requirements for a speech interface.

D3.1.2. Orientation of the TAS

"Conventional" ergonomic experimentation usually assumes the hypothetico-deductive approach. This requires the advancement of hypotheses on the basis of predictions of a putative model of the world: the hypotheses are tested in experiments, which then enable the investigator to accept or reject the model. TAS is intended, firstly, to compare the performance of alternative interface designs, and secondly, to develop a general model of interaction which explains the performance differences, so enabling the specification of an optimal interface. The second of these endeavours requires comparison of behaviour of several interfaces, sampled with respect to critical design attributes. The critical design attributes are hypothesised on the basis of a model of interaction which expresses the relationship between interface attributes and interaction behaviour.

The interaction model proposed for the TAS described here made the following general performance predictions[1]

[1] The experiment described here preceded the advancement of "diagnostics" as the basis for constructing interaction models.

(i) tasks are performed more quickly if the number of transits of the hands and eyes from one entity to another is reduced
(ii) tasks are performed more quickly if activities may be performed concurrently
(iii) increased speed in operation, and a lower requirements for the user to redirect attention, potentially reduces user memory load and so may reduce the incidence of errors consequent upon demands for memory
(iv) tasks are performed more quickly and accurately if information does not have to be recoded (e.g. from a verbal/vocal representation to a sequence of keystrokes).

In the context of the FOO task, this interaction model states that a [good] speech interface will offer performance advantages over a keyboard by enabling a reduction in the transits of hands and eyes, concurrent data entry and off-line (non-verbal) activities, reduced demands from attentional switches (and hence fewer errors) and reduced demands for the recoding of information. The evaluation will compare, then, interaction when the device is a speech interface and when it is a keyboard. However, it is predictable that an unreliable speech interface will strongly moderate, if not eliminate, these advantages. The development of a general interaction model demands exploration of the relationship between the attributes of recognition reliability and interaction performance. Consequently, TAS examined system behaviour and performance when the data entry device was:

(1) a QWERTY keyboard
(2) an extant connected speech recogniser, with a mean error rate of 14.9%
(3) an enhanced connected speech recogniser (functional specification as in (2), but with a mean error rate of 3.2%).

In all cases, entered data were presented back to the user on a visual display. The three user interfaces had the following characteristics.

(1) *KEYBOARD INTERFACE*

Data entry device. The keyboard interface comprised a conventional QWERTY layout, as might be encountered in a direct entry device for a battlefield computer. In this instance, the QWERTY layout was supplemented with the following special keys:

- 5 FIELD keys in a row at the top of the board
- 4 cursor movement keys
- DELETE key
- SEND key
- CLEAR key.

The legal vocabulary for the keyboard system was a subset of that for the speech recognizers.

Data display device. The keyboard was used in conjunction with a visual display, showing the data currently entered and providing system information. Data were entered against 5 field names always visible on the display:

- GRID (i.e. aim point grid reference)
- NUMBER (i.e. target number)
- TARGET (i.e. target class)
- BATTERY (i.e. battery selected to engage the target)
- ROUNDS (i.e. number of ammunition rounds to be fired).

A cursor indicated the point at which data would next be written.

Operating procedure. The cursor could be moved to the point on the display at which data were to be entered by the following means:

(a) cursor control keys, which commanded cursor movement in any of 4 directions without data deletion;
(b) field keys, which moved the cursor to the designated field and removed existing data entered in the field;
(c) CLEAR key, which moved the cursor to the beginning of the top field and removed existing data from all fields on the page; and
(d) DELETE key, which deleted the character preceding the cursor and moved the cursor to that location.

The character keys overwrote any existing data. When the user subject (US) had entered the desired data, the page (which corresponded to a fire-order) was transmitted over the network by means of the SEND key. The US was then able to enter data for a new fire-order.

(2) *EXTANT SPEECH INTERFACE*

Data entry device. The extant speech recognizer was modelled on Marconi Macrospeak: a connected speech recognizer available currently on the market. The functionality of the device was simulated by means of a human simulator (see Section 6.6.2). However, to add realism to the simulation, an actual Macrospeak device was clearly visible in the user USs workspace. They "operated" it by activating the Marconi SPEAK key, and by speaking into a boom (headset) microphone which they assumed to be connected to the recognizer.

The performance specification of Macrospeak had been determined empirically, and the simulation had been optimized (see Appendix C). The device was "trained" using a vocabulary of 35 words.Each vocabulary item corresponded to commands available on the keyboard, except for WAKE UP, REST and TRAIN, which, respectively, instructed the recognizer to accept data, to stop accepting data and to begin the training sequence. Training was achieved by the US reading aloud each vocabulary item as it was presented in sequence on the display. Neither of the speech devices could support the re-training of individual vocabulary items, i.e. the whole vocabulary had to be re-trained. Re-training enhanced the templates held by the device and improved overall recognition performance according to a predetermined scheme.

Data display device. The display was identical to that used in the keyboard condition.

Operating procedure. The recognizer was activated by pressing the SPEAK key and by uttering the WAKE UP command. The training routine could be initiated at any time by speaking the command TRAIN. All other data entry actions were equivalent to those performed manually using the keyboard (see Section A1.3). If the US wished temporarily to de-activate the recognizer (e.g. in order to speak over the intercom), this was achieved by pressing the SPEAK key for a second time. At the end of the trial, the US uttered the REST command.

Recognition error characteristics. Recognition errors were simulated according to three pre-determined confusion matrices. For the first three re-trainings, a new matrix was loaded each time (although USs did not realise that this was the real determinant of device performance). The matrices respectively delivered mean error rates of 17%, 15.8% and 11.9%.

(3) *ENHANCED SPEECH INTERFACE*

Data entry device. The enhanced recognizer was functionally and, to USs, physically identical to the extant recognizer, but USs were told that its recognition performance had been enhanced. It thus differed only with respect to the confusion matrices which were loaded automatically at each "re-training"

Data display device. The display was identical to that used with the keyboard and extant recognizer.

Operating procedure. Operating procedure was identical to that employed for the extant recognizer.

Error characteristics. Error characteristics were simulated in the same way as for the extant recognizer. The matrices respectively delivered mean error rates of 3.9%, 2.9% and 2.9%.

D3.1.3. Experimental design

The *independent variable* of the study was the computer interface, which could take the three values corresponding to the target devices described above. The *dependent variables* were measures of task performance and are summarised in Section D3.2.4. The performance indices employed for the study were measures of *time* and measures of *quality of task outcome* (see Dowell and Long, 1988)[2]. Behavioural indices were also employed to enable diagnosis with respect to the model of device-user interaction.

The *null hypothesis* was that there would be no difference in the time taken to perform the experimental task using the three devices, the *alternative hypothesis* being that differences would be exhibited.

Because the task was relatively complex, and because the subjects were informally trained (see Section D3.2.1), inter-subject differences in task strategy and skill level would introduce variability which, potentially, could mask the experimental effects. Consequently, a *within subjects design* was used, in which each subject performed equivalent tasks using each of the three interfaces. The order in which the tasks were performed was balanced across the subjects, to eliminate the potentially confounding effect of practice.

D3.2 Method

TAS assumes the prior existence of simulations of task, device and user. These are briefly described in Section D3.2.1, but the methods of their development are documented elsewhere. Section D3.2.2 describes the procedure followed by TAS subjects using the simulation configuration illustrated in Figures D1a and D1b.

D3.2.1 Experimental Simulation

(a) *Task simulation.* The development of the FOO task simulation is described in Appendix B. The following description of the simulated task is from the latter document.

> A subject who represents the FOO (user subject - US) is presented with a task which requires the engagement of a number of targets presented on an electronic 2-D graphical display (i.e. a map, but without grid markings). Each target on the display has associated with it a target number, which may be used to obtain further information about the target from an alternative text display. This text display is accessed by continuously pressing an inconveniently located key, thus requiring manual involvement in order to obtain detailed information about a target. This component of the task is intended to be analogous to the use of binoculars to acquire information about targets. The text display informs the subject of the nature of the target (e.g. tank) and the current hostility value of the target: a variable which increases continuously over time at a rate determined by the nature of the target. It is in the interest of the US to deal quickly with particularly hostile targets.
>
> When the US has selected a target for engagement, the geographical location is calculated by identifying its position on a paper map which is marked with a grid. The grid reference is used as the basis of a communication to one of four gun batteries calling for fire against the target.
>
> The process is complicated by the fact that each battery has a limited number of ammunition rounds, and that each battery is biased in its aim in a different way. Thus it is necessary for the US to monitor the number of rounds demanded from each battery and, for each battery, to ascertain the adjustment to the grid reference necessary for rounds to hit the target. From time to time, throughout the engagement, the US will be required to report the current situation to the experimenter.

[2] Given the subsequent development of the notion of *task quality* in Dowell and Long (1989), time would be viewed as a measure of *quality* (rather than as a cost, as is implied here).

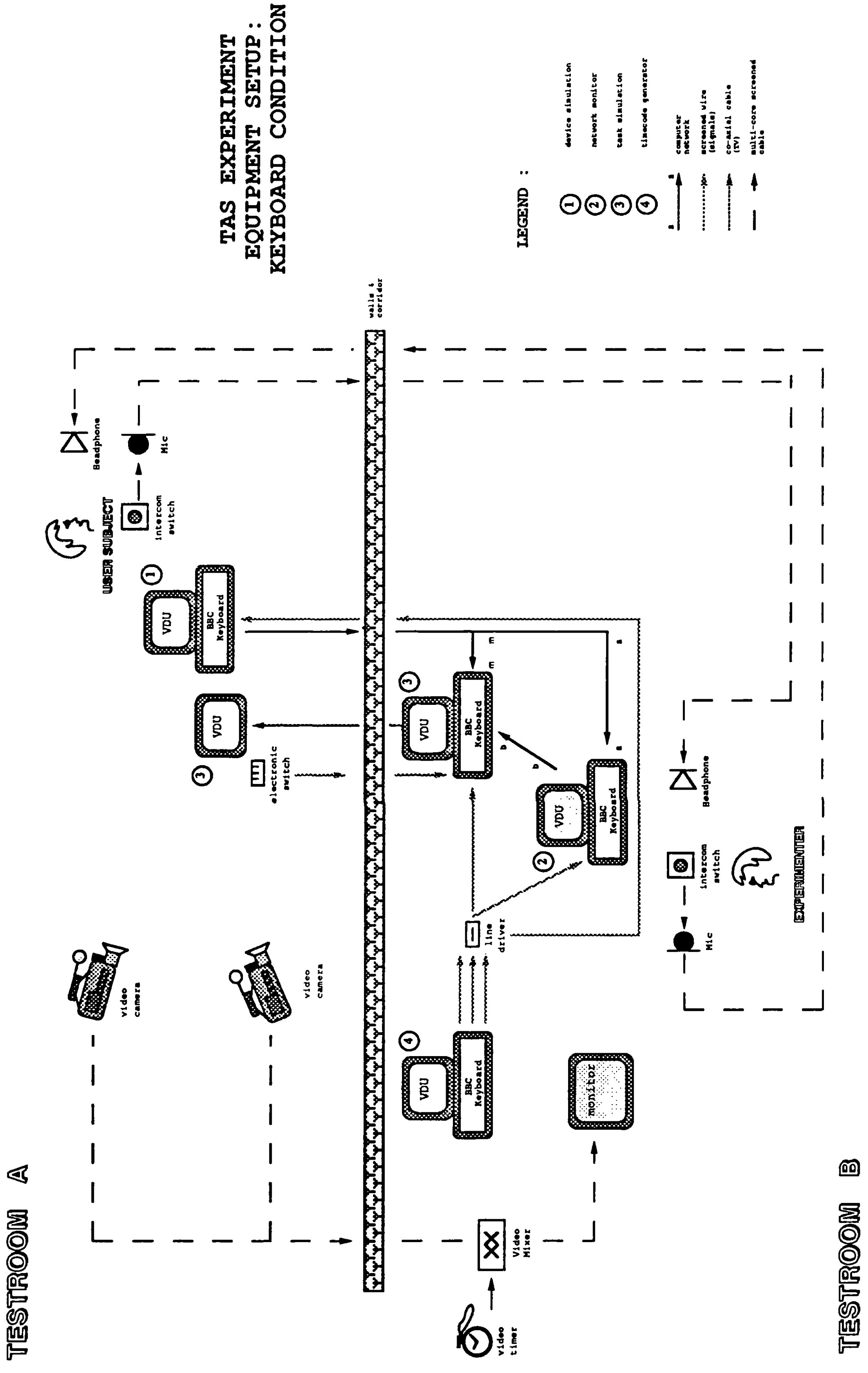


Figure D1a: Configuration of testbed (keyboard)

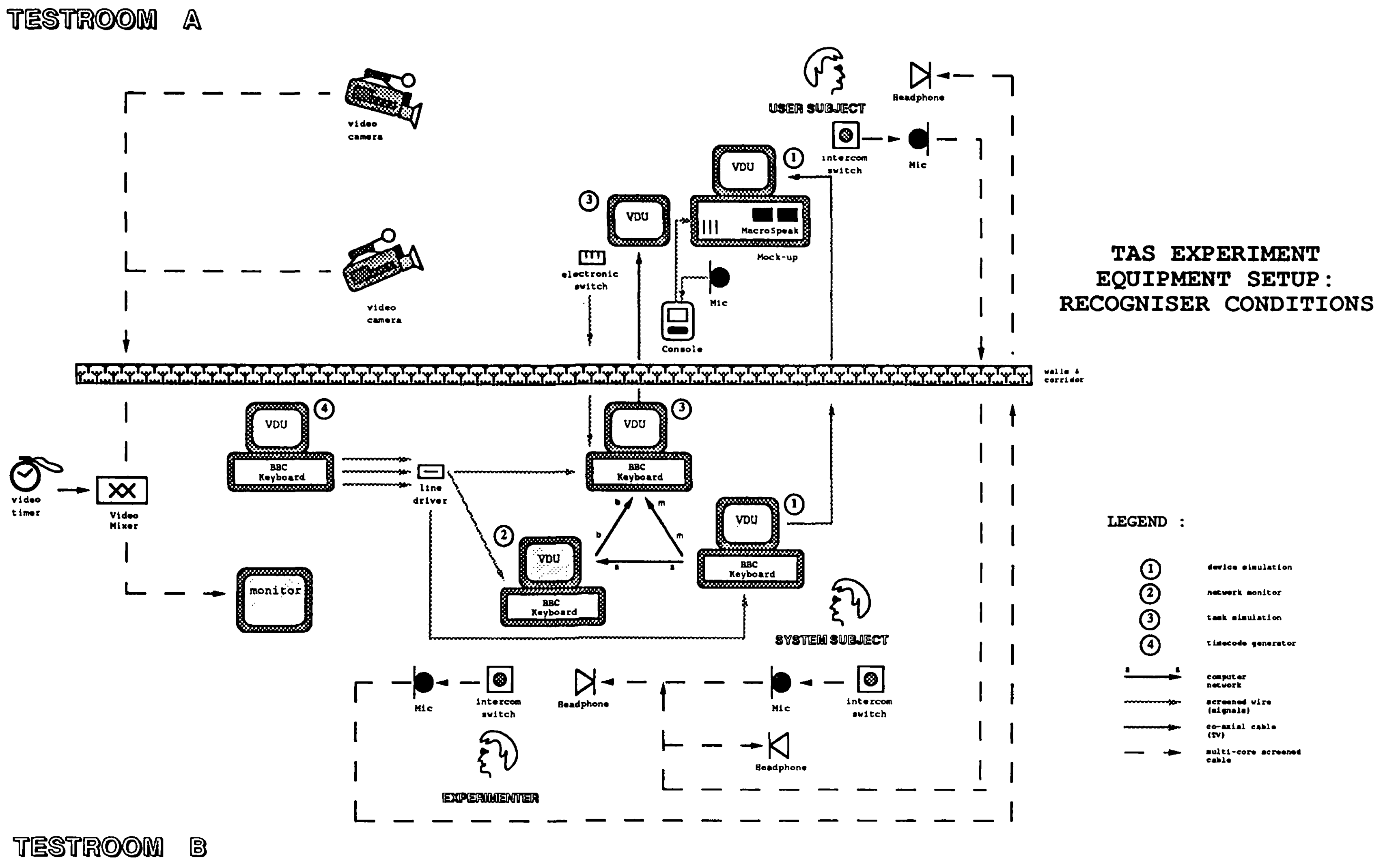


Figure D1b: Configuration of testbed (recognizer)

Directly hitting the target "freezes" the hostility value and a near miss reduces its rate of increase. Performance of the subject is assessed by measuring the time taken to complete the mission, and/or by summing the accrued hostility values of all of the targets at the end of the mission, and/or by measuring the resources used to complete the mission (i.e. by measuring the quality of the outcome of the task).

The task was implemented using the RSRE/UCL experimental test bed (see Lee, 1989)

(b) *Device simulation.* The devices simulated were three versions of a computer for controlling indirect artillery engagements. They differed with respect only to the user interface (see Section D3.1.2). The development and optimisation of the device simulations is described in Appendix C: all three devices were implemented on a BBC microcomputer within the experimental testbed.

The simulation of the device with the *keyboard interface* utilized the BBC QWERTY keyboard, configured with five special function keys corresponding to data fields. The subject used the keyboard to enter data in the fields displayed on a visual display monitor, and when satisfied, sent the page of data by entering a SEND command.

The two versions of the device with *speech interfaces* were simulated by means of a system subject (SS) who entered spoken information uttered by the user subjects (USs) into an equivalent BBC computer within the testbed. However, the SS was located remotely, and his/her existence was not known to the USs who believed that they were speaking to a speech recogniser. The SS's computer had a specially configured keyboard which enabled rapid entry of the data and the automatic insertion of recognition errors characteristic of the respective target devices. The two speech conditions involved device simulation differing only with respect to recognition characteristics (see Appendix C).

The configuration of the testbed to support the device simulations is shown in Figures Dla and Dlb.

(c) *User Simulation.* The population of FOOs was represented in the simulation by a sample of 12 subjects. They were 4 males and 8 females. The mean age of the group was 20.6 years old with a standard deviation (s.d.) of 1.68 years. It was believed that the subjects' age and gender would not affect the outcome of the TAS experiment, so these subjects' characteristics were not controlled. User subjects were chosen from a college population. They all had prior experience with the computer. All had an educational background of 'A' Level or above.

Each subject underwent instruction/practice during which they tackled two battlefield scenarios requiring the engagement of two targets, and one requiring the engagement of four targets. If the subject was found to be incapable of performing the task at criterial rate of 25 mins. to successfully engage four targets, s/he was excluded from the subject pool. Therefore, the sample selected was a group of subjects who could perform concurrent activities with time-sharing efficiency in a military task domain[3]. These sample characteristics formed the basis of the user simulation. Candidates successfully completing the instruction/practice sequence proceeded to the experimental trials.

[3] It was recognized that these subjects were not representative of the target users in all respects. For example, a potentially significant weakness was their relative lack of experience with map-reading tasks, which could result in the subjects having to allocate disproportionate mental resources to this aspect of the task at the cost of others. The subjects also had no direct experience of battlefield conditions. These anomalies were considered in the interpretation of the experimental results. It should be noted that this work preceded the development of the user simulation method (Chapter 10), which attempts to take account of these concerns.

D3.2.3 Experimental Procedure

Before subjects came to the experiment they were given a description of the experimental task. The experiment was then carried out over two separate days. Subject selection and instruction occurred on the first visit. Those subjects who met the performance criterion proceeded to undertake the first experimental condition. The remaining two conditions were presented on a second visit organised at least a day later. The procedure was as follows:

Day One

(1) The subject was individually briefed on the task goal, i.e. destruction of all targets as quickly as possible and with minimal utilization of artillery resources.

(2) The tactics to disable various targets in the simulated terrain were described, and the criteria of efficiency (execution time and accrued target hostility value) were explained.

(3) Subjects then had the chance to familiarize themselves with the equipment, such as keyboard/extant/enhanced recogniser operation, intercom communication and simulated binoculars for terrain observation. The device s/he learnt to operate on Day One was the device upon which performance was to be assessed first. Instruction continued until the subject felt that s/he understood how to perform the task.

(4) A hands-on practice trial followed, comprising presentation of Scenarios 1 and 2, which contained two targets each.

(5) Subjects successfully completing step (4) proceeded to a practice trial involving the presentation of Scenario 3 with four targets. The experimenter discussed with the subject his/her performance in the trial. If the subject had difficulties in coordinating the various activities to achieve the task goal (i.e. destruction of 4 targets in 25 minutes), s/he was requested to withdraw at this stage.

(6) Experimental trial (Scenario 4) of the first device commenced in which the subject engaged 8 targets. Data were collected throughout the trial.

Day 2

Steps (1) to (3) above were repeated on Day 2, but the instruction relating to device use was appropriate to the next (i.e. second) device condition.

(4) Practice in the use of the second device comprised the presentation of one scenario containing 4 targets (Scenario 5).

(5) The subject undertook the experimental trial requiring use of the second device in the engagement of 8 targets (Scenario 6).

(6) Step (3) above was repeated, with instructions appropriate to the third device.

(7) Practice in the use of the third device comprised the presentation of one scenario containing 4 targets (Scenario 7).

(8) The subject undertook the experimental trial requiring use of the second device in the engagement of 8 targets (Scenario 8).

(9) The subject completed a questionnaire relating to their subjective evaluation of the three devices. Questions related to:

(a) subjective performance rating (acceptability) on a scale of 1-5
(b) influence of device performance on the subject's approach to the task
(c) device improvements
(d) acceptability of the device for use on the battlefield (see also footnote 3)
(e) other comments.

D3.2.4 Summary of dependent variables.

Experimental variables related either to *performance*, for the purpose of testing the experimental hypothesis, or to *behaviour*, for the purpose of explaining performance differences in terms of the interaction model.

<u>Performance variables</u>

(a) Time indices: (1) time to implement fireplan
(2) time spent using the device
(3) mean time per device interaction.
(b) Quality indices: (4) accrued target hostility value
(5) ammunition rounds unused after mission.

Behavioural variables

(a) Observed actions:
(6) number of interactions with the device
(7) number of incidences of actions concurrent with device use
(8) number of fire-orders sent
(9) number of recognizer re-trainings.

(b) Subjective reports:
(10) rating of device performance
(11) rating of acceptability for battlefield use.

D3.2.5 Analysis of video record

The video record provided split-screen views of (a) US's workspace, from above, and (b) the battlefield display and engagement computer display, a position behind US. Each trial was scored manually and timed, using a stop-watch. Whenever a new action was observed to occur, it was tallied under one of the following headings:

Observation actions
- fireplan/mission record (R)
- artillery information sheet (F)
- map (M)
- battlefield display (T)
- battlefield display using binoculars (B)

Device actions
- feedback from computer display (S)
- operation of keyboard (K)
- operation of speech recognizer (P)

Recording actions
- writing on fireplan/mission record (W)

Other actions
- operation of intercom (I)
- training speech recognizer (A)
- others (O).

Concurrent actions during periods of interaction were identified in the tally.

D3.2.6 Collection of computer-logged data

The following data were logged automatically within the testbed:
- interval between first and last fire-orders, i.e. implementation time
- accrued hostility value
- ammunition rounds remaining
- number of fire orders transmitted.

At the end of each experimental trial, the data were written to disk for subsequent analysis.

D3.2.7 Collection of subjective data

The questionnaire was completed by each US at the end of the last experimental trial, and the experiment was discussed with them. Questions 2, 3, and 5 were unstructured and were used to assist the interpretation of experimental results in the discussion.

D3.3 Results

A summary of the manually scored video data is presented in the Annex Part 1. The analysis of actions derived from the tally is presented in the Annex Part 2. The computer-logged data are summarized in the Annex Part 3 Questions 1 and 4 could be analyzed quantitatively; the responses are summarized in the Annex Part 4.

Owing to the varieties of data collected, a variety of statistical tests were employed. Descriptions of the statistical procedures, and details of the application of the tests to the results, are presented in the Annex Part 5.

D3.3.1 Performance variables

(a) Time indices

Variable 1: Time to implement fireplan (sec)

	Keyboard	Extant Recog.	Enhanced Recog.
mean	1331	1787	1327
s.d.	282	590	245

F_{max} was calculated to be 5.8[4], indicating that the variances of the three groups could not be assumed homogeneous and, hence, that the parametric assumptions of ANOVA were violated. A Friedman two-way ANOVA by ranks was performed (see Annex Part 5). *X2r obs*. = 6.4, which is greater than the critical value (*X2r tab*. = 5.99 at 2 df), so the H_0 is rejected; and the distribution of ranks between the three devices is assumed not equal. The result suggests that the overall time to complete the task varied across the three devices. It would appear that the task takes longer using the extant recognizer than the other two devices.

Variable 2: Time spent using the device (sec)

	Keyboard	Extant Recog.	Enhanced Recog.
mean	530	1034	639
s.d.	101	312	112

F_{max} was calculated to be 9.54, indicating that the variances of the three groups could not be assumed homogeneous and, hence, that the parametric assumptions of ANOVA were violated. A Friedman two-way ANOVA by ranks was performed (see Annex Part 5). *X2r obs*. = 18.5, which is greater than the critical value (*X2r tab*. = 5.99 at 2 df), so the H_0 is rejected and the distribution of ranks between the three devices is assumed not equal. The result suggests that the total time spent by subjects using the device differed significantly across the three devices. It would appear that the extant recognizer requires more time for operation than the other two in the context of the task.

Variable 3: Mean time per device interaction (sec)

	Keyboard	Extant Recog.	Enhanced Recog.
mean	22.0	41.1	29.1
s.d.	5.4	8.2	7.3

An ANOVA test was performed and is described in Annex Part 5[5]. Since $F(2_{obs.},22_{obs.}) = 32.8$, which is greater than the critical value ($F_{tab.}= 3.44$), the H_0 is rejected, and the means of the 3 devices are accepted as being not equal. The result suggests that the periods of interaction differed in length across the three devices. If a subsequent posteriori comparison test of the means is computed (see Annex Part 5), it is found that the means obtained when subjects used the keyboard and the extant recogniser are significantly different. This suggests that keyboard is a device that requires significantly less time for operation than the extant recogniser with the same task; however, there was no evidence of a significant difference between the enhanced recognizer and the keyboard, nor between the extant and enhanced recognizers.

[4] Variance was assumed homogeneous if F_{max} (3,11) < 4.5 (i.e. Hartley's test for homogeneity of variance (p = 0.05))

[5] $F_{max} = 2.31$; hence variance assumed homogeneous

(b) Quality indices

Variable 4: Accrued target hostility value (arbitrary units: see Lee, 1989)

	Keyboard	Extant Recog.	Enhanced Recog.
mean	240.5	327	252.6
s.d.	37.1	81.5	49.1

F_{max} was calculated to be 4.83, indicating that the variances of the three groups could not be assumed homogeneous and, hence, that the parametric assumptions of ANOVA were violated. A Friedman two-way ANOVA by ranks was performed (see Annex Part 5). $X^2r_{obs.}$ = 9.5, which is greater than the critical value ($X^2r_{tab.}$ = 5.99 at 2 df), so the H_0 is rejected and the distribution of ranks between the three devices is assumed not equal. The result suggests that the efficacy of the subjects' performance varied across the three devices, apparently being less effective when the extant recognizer was used.

Variable 5: Ammunition rounds unused after the mission

	Keyboard	Extant Recog.	Enhanced Recog.
mean	172.3	153.3	171.7
s.d.	14.2	38	16.2

F_{max} was calculated to be 7.24, indicating that the variances of the three groups could not be assumed homogeneous and, hence, that the parametric assumptions of ANOVA were violated. A Friedman two-way ANOVA by ranks was performed (see Annex Part 5). $X^2r_{obs.}$ = 3.38, which is less than the critical value ($X^2r_{tab.}$ = 5.99 at 2 df), so the H_0 is <u>not</u> rejected and the distribution of ranks between the three devices is assumed equal. The result suggests that the device type did not influence the economy with which the subject utilized artillery resources.

D3.3.2 Behavioural variables

(a) Observed actions

Variable 6: Number of interactions with the device

	Keyboard	Extant Recog.	Enhanced Recog.
median	21	25.5	20
range	17 - 37	19 - 35	14 - 40

Because the data did not meet the statistical assumptions of a parametric test, a Friedman two-way ANOVA by ranks was performed (see Annex Part 5). $X^2r_{obs.}$ = 1.6, which is less than the critical value ($X^2r_{tab.}$ = 5.99 at 2 df), so the H_0 is <u>not</u> rejected; and the distribution of ranks between the three devices is about equal. The result suggests that there was no difference in the relative frequency with which subjects interacted with the three devices.

Variable 7: Number of incidences of actions concurrent with device use

	Keyboard	Extant Recog.	Enhanced Recog.
median	0.5	11.5	11
range	0 - 2	0 - 32	4 - 77

Because the data did not meet the statistical assumptions of a parametric test, a Friedman two-way ANOVA by ranks was performed (see Annex Part 5). $X^2r_{obs.}$ = 16.8, which is greater than the critical value ($X^2r_{tab.}$ = 5.99 at 2 df), so the H_0 is rejected; and it is accepted that the three devices have differential effects on the frequency with which actions are performed concurrently[6].

Variable 8: Number of fire-orders sent

	Keyboard	Extant Recog.	Enhanced Recog.
median	18	17	17
range	14 - 28	16 - 29	11 - 29

Because the data did not meet the statistical assumptions of a parametric test, a Friedman two-way ANOVA by ranks was performed (see Annex Part 5). $X^2r_{obs.}$ = 0.5, which is less than the critical value ($X^2r_{tab.}$ = 5.99 at 2 df), so the H_0 is not rejected; and the distribution of ranks between the three devices is accepted as being about equal. The result suggests that there is no evidence of the device type influencing the number of fire orders sent by subjects to destroy the targets.

Variable 9: Number of recognizer re-trainings

	Keyboard	Extant Recog.	Enhanced Recog.
median	N/A	2	0
range	N/A	1 - 4	0 - 1

Since all 12 subjects in the experiment had more extra trainings required for the extant recogniser than the enhanced one, the value of $X_{obs.}$ is equal to 0; hence the H_0 is rejected. We can conclude that the extant recogniser requires significantly more trainings than the enhanced recogniser.

(b) Subjective reports

Variable 10: Rating of device performance

Subjects were requested to answer the following question after they had used all three devices: "Rate the performance of the device on a scale from 1 to 5 where 1 = Totally unacceptable, 3 = Acceptable and 5 = Acceptable without reservation."

	Keyboard	Extant Recog.	Enhanced Recog.
median	4	2	4
range	3 - 5	1 - 3	3 - 5

[6] It was subsequently noted that the effect is confounded with time spent on the device: Ss spent more time using the recognizers, so one would expect a higher frequency of all actions, not just concurrent ones. This issue is considered further in the Discussion.

Because the data did not meet the statistical assumptions of a parametric test, a Friedman two-way ANOVA by ranks was performed (see Annex Part 5). Since $X^2r_{obs.}$ = 18.2 which is greater than the critical value ($X^2r_{tab.}$ = 5.99 at 2 df), the H_0 is rejected, and it is accepted that the three devices have differential effects on the subjects' opinion of the acceptability of their performance. The result suggests that the performances of the keyboard and enhanced recognizer were regarded as more acceptable than that of the extant recognizer.

Variable 11: Rating of acceptability for battlefield use

"Would you be happy to use this device on the battlefield ?" (from question no.4 of the questionnaire)

Response		**No. of responses (12 subjects)**	
	Keyboard	**Extant rec.**	**Enhanced rec.**
"Yes"	12	0	8
"No"	0	12	4

Because the data did not meet the assumptions of a parametric test of significance, and because the data were dichotomous (i.e. the subjects' response to the question was either "yes" or "no"), a Cochran's Q test was applied (see Annex Part 5). $Q_{obs.}$ = 18.7, which is greater than the critical value ($Q_{tab.}$ = 5.99 at 2 df), so the H_0 is rejected, and it is accepted that the three devices were rated differently with respect to their perceived suitability for use on the battlefield. The order of perceived acceptability was:

- most acceptable: keyboard
- second most acceptable: enhanced recognizer
- least acceptable: extant recognizer.

D3.4 Discussion

The results of the experiment are discussed with respect to the eleven variables reported in the last section.

D3.4.1 Performance variables

Variable 1: Time to implement fireplan The result of the analysis of variance suggests that the keyboard and enhanced recognizer support faster task performance than the extant recognizer. However, the device type may have been only one of a number of factors determining the time to implement the fireplan.

A major influence could have been the existence of individual differences in tactics both within and between subjects during the experiment. For example, some subjects did their tactical planning before they tried to implement the fireplan, whereas others did much of their planning during implementation; some subjects called for fire on new targets before seeing the effect of their previous engagement, while others preferred to operate serially on targets; and there was evidence of improvements in the task skills (including tactics) of many subjects as they gained practice across the conditions. It is quite probable that the effect due to device type would have interacted with effects due to factors such as these.

Variable 2: Time spent using the device. There was a strong indication of differences between the three devices in the amount of time spent interacting with the device. At least three factors might have contributed to this:

(1) impact of device type on the tactics used by subjects (i.e. periods of interaction are shorter with some devices because the devices are more compatible with users' preferred means of processing information);

(2) device reliability (e.g. because some devices transmit information more reliably than others, less time is spent in error correction, so periods of interaction are shorter);

(3) extra actions required during operation for some devices (e.g. the performance of some devices - unreliable speech recognizers - is improved if the device is retrained: "retraining" is completely unnecessary for other devices, such as keyboards).

To resolve this issue, at least in part, the video record was re-examined, and time spent re-training was subtracted from the overall time spent using the two speech recognizers. The time (secs) spent interacting with the devices *in performance of the battlefield task* (Variable 2a) was as follows:

	Keyboard	Extant Recog.	Enhanced Recog.
mean	530	914	635
s.d.	101	287	113

A Friedman test[7] revealed $X^2r_{obs.} = 16.7$, indicating that the distribution of ranks between the three devices is not equal ($p<0.001$). Ignoring re-training time, interaction performance on the keyboard and enhanced recognizer demonstrated an advantage over the extant recognizer: this difference was probably due to the requirement for particularly frequent correction of misrecognized tokens when using the extant recognizer.

Practical implications of the results of the re-analysis are that recognition accuracy is an important determinant of overall task performance, but that an efficient re-training facility could have a substantial impact on the overall interaction performance of a user interface involving a speech recognizer.

Variable 3: Time per device interaction. Periods of device operation were apparently shorter for the keyboard than for the extant recognizer, with the enhanced recognizer supporting performance somewhere in between. This pattern was maintained when time per interaction was taken to exclude re-trainings (Variable 3a: F(2,22) = 18.11; see Annex Part 5). Given that the number of interactions did not differ between the devices (Variable 6), the result indicates that the extant recognizer demanded longer interactions, which were a consequence of the needs both for error correction and for frequent re-training.

Variable 4: Accrued hostility value. The ANOVA suggests that quality of performance is better using the keyboard or enhanced recognizer than the extant recognizer, in that the "enemy" was incapacitated more effectively when the subject was using one of the former devices. The effect could have been due to the following reasons:

(1) some devices encouraged better tactics than others (e.g. they supported better identification of, and response against, particularly hostile targets by reducing the workload of subjects); and/or

(2) some devices took longer to use than others, so the hostility reached a higher value before targets were destroyed.

Subjects did report that the frequent recognition errors of the extant recognizer disrupted their attention to the task and, hence, it may have adversely influenced their performance; however, in view of the result with respect to Variables 2 and 5, the second of the above interpretations is also likely to be relevant. The factors discussed with respect to Variable 1 could also have been influential here.

Variable 5: Ammunition rounds remaining. There was no evidence that the device type influenced the economy with which the mission was performed. Had such a difference been found, it might have been taken as supportive of the first interpretation of the result with respect to Variable 4. In many ways, the result is not surprising, because subjects tended only to send fire-orders when they were sure they were error-free (so rounds were not "wasted" as a consequence of differences in the transmission accuracy of the data entry devices).

Comment on performance data. The overall conclusion is that performance is best with the keyboard, next best with the enhanced recognizer and worst with the extant recognizer. Performance with the extant recognizer is so poor that it cannot be regarded as usable for the

[7] $F_{max} = 8.07$, so the variances of the groups could not be assumed homogeneous and, hence, a parametric ANOVA test was inappropriate

FOO task. In order to achieve acceptable quality of task outcome with the speech interfaces, subjects were forced to spend time correcting recognition errors and re-training the devices with the intention of improving recognition performance: both of these factors contributed substantially to the advantage of the keyboard over the extant recognizer. However, given a minimally time-consuming re-training procedure, task performance with the enhanced recognizer would be close to that of the keyboard.

The "quality" measures were, by themselves, inconclusive as indices of device usability. One reason lies in the difficulty in partialling out the various factors that may impact the quality of outcome, i.e. they were over-inclusive measures. However, in any case, differences between devices would be masked because subjects implicitly imposed a criterion of quality (i.e. that all errors were to be corrected before a page of data was transmitted to the artillery battery). Thus, the performance trade-off between time and quality was biassed in favour of quality, so time measures most strongly reflected the effect of the device type.

Unfortunately, the experiment was probably insensitive to the more subtle influences of the device on performance. For example, it was difficult to discern an effect of device type on overall task performance because there were wide variations (both within and between subjects) in ability to think tactically. Some subjects found the cognitive aspects of the task more taxing than other subjects, and this determined their tactics in performing the task, e.g. whether to do nothing or whether to plan while waiting for fire-orders were being implemented, or whether to engage multiple targets. The more sophisticated tactics, not exhibited by all subjects, would be the ones expected to be particularly impacted by the device type. One implication of this is that, in future experiments, subject selection should take account of all those characteristics of the user population which may impact device-user interaction; in the present instance, tactical planning ability was one such characteristic.

D3.4.2 Behavioural variables

Variable 6: Number of interactions with the device. The finding of no significant difference in the number of interactions with the device supports the view that the reason for the difference in time spent using the devices is due to differences in the length of interactions rather than differences in frequency (see Variable 2).

Had subjects not been imposing the implicit quality criterion mentioned above, one would have expected to find differences in the number of interactions, because the "unreliable" devices would have engendered aiming errors subsequently requiring correction (and hence more interactions). The finding of no difference is thus supportive of the interpretation that subjects were biased towards quality in the time-quality trade-off.

Variable 7: Incidence of actions concurrent with device use. The data suggested that concurrent action was better supported by the speech recognizers than by the keyboard. However, the result could not be interpreted unless the potentially confounding effect of the differing times spent interacting were taken into account (Variable 2). This was because the observed difference in actions concurrent with user-device interaction could simply have been due to the fact that there was more interaction with the extant recognizer. For this reason, a metric was derived which reflected frequency of concurrent actions (Variable 7) with respect to the time spent interacting with the device to perform the task (Variable 2a). The metric (termed Variable 7a) was calculated according to the following formula:

$$\frac{\text{No. of concurrent actions}}{\text{Interaction time (in secs)}} \quad \text{or} \quad \frac{\text{Variable 7}}{\text{Variable 2a}}$$

i.e. the measure was the number of concurrent activities per second of interaction time. The results were as follows:

	Keyboard	**Extant Recog.**	**Enhanced Recog.**
median	0.0009	0.014	0.016
range	0 - 0.005	0 - 0.033	0.001 - 0.096

A Friedman test revealed the same pattern as had been obtained with the frequency data ($X^2r_{obs.}$ = 15.04), confirming that the recognizers did indeed better support concurrent activities than the keyboard.

The analysis of individual concurrent actions was studied to determine the nature of the differences in the patterns of interaction with the three devices (see Appendix C). Similar actions were performed concurrently with operation of the two speech devices, but these were different from (as well as being more frequent than) actions performed concurrently with keyboard data entry. Specifically, for the keyboard, the ranking of concurrent actions was as follows (n = frequency across all subjects):

(1)	Looking at fireplan/mission record	n=3
	Operating the radio	n=3
(3)	Looking at artillery info. sheet	n=1

It should be noted that any concurrent activity was very rare with this population of subjects, who were not touch typists.

For the recognizers[8], the ranking was as follows:

	Extant (n)	Enhanced (n)
(1) Looking at fireplan/mission record	90	154
(2) Looking at battlefield display	31	46
(3) Writing on fireplan/mission record	14	17
(4) Looking at map	8	10
(5) Looking at artillery info. sheet	4	3
(6) Using binoculars	-	1

In view of the fact that the most frequently-occurring concurrent action with all the devices was looking at the mission record, it would seem reasonable to assume that subjects were reading the record in order to enter information from it to the computer, or to check information already entered. The recognizers apparently enabled this readily to be done concurrently with entry, whereas the keyboard did not.

From the point of view of task performance, the recognizers offer a potential advantage because they enable attention either to the battlefield or to sources of information in the operator's workspace (i.e. map, fireplan/mission record and artillery information sheet) concurrently with operation of the computer. In principle, this should enable more actions to be performed per unit time, and hence the task should be performed more quickly with a recognizer than it would using a device demanding the operator's "full" attention. Unfortunately, the potential performance advantage was not realised in this experiment. The reasons might have been because speaking adversely influenced other behaviours important in the task or because the device was inherently defective. There was no evidence to support the former interpretation but strong evidence to support the latter.

Variable 8: Number of fire orders sent The finding of no significant difference between the device conditions suggests that it was not the case that they differed with respect to the quality of fire-orders actually transmitted. This supports the view that subjects used an implicit criterion of requiring an error-free page of data (fire-order) before transmission. See also general comments on the performance data.

Variable 9: Number of recognizer re-trainings There was clear evidence of a difference in the number of re-trainings required by the two recognizers. For all subjects, the extant recognizer demanded more re-training than the enhanced recognizer. An observation with potentially practical implications is that subjects will not bother to re-train the device (and waste time)

[8] Because a condition of use of the radio was that the recognizer be turned off, only the keyboard included radio operation as a possible "concurrent action".

if the recognition performance currently exhibited is acceptable for their purposes. A corollary might be that an "acceptable" error rate is that at which a criterial proportion of users no longer bother to re-train. [Note, however, that the perceived relationship between the cost and benefit of re-training will influence this rate: the acceptable rate would tend to be lower if the device were quick and easy to re-train (i.e. low cost of re-training), or if the consequences of uncorrected errors were serious (i.e. high cost of not re-training), than if the converse were the case.]

In the present study, re-training was time-consuming (i.e. there was a high cost attached to re-training), but failing to re-train forced the subjects to utilize a lot of effort checking the accuracy of their output. Cverall, re-training was perceived as worth doing for the extant device, but not for the enhanced device. On this criterion, an error rate of 3.9% was "acceptable" for 11 of the 12 subjects, and an error rate of 2.9% was "acceptable" to all (but see comments relating to Variables 10 and 11).

Variable 10: Rating of device performance and *Variable 11: Rating of acceptability for use on the battlefield* The analysis of Variable 10 suggests that the performance of the keyboard and that of the enhanced recognizer were regarded by subjects as similarly "acceptable"; however, it is revealed in Variable 11 that only 8 of the 12 subjects considered the enhanced recognizer as acceptable for use on the battlefield. The comments of subjects recorded in the questionnaire were particularly revealing in this regard: most felt that, although superior to the extant device, the enhanced recognizer would benefit from further improvements in its reliability as a data entry device. Two subjects also suggested that the response lag of the recognizer (approximately 300-500msec) should be reduced.

The criterion held by many of the subjects for a device acceptable for use on the battlefield was, evidently, perfect reliability in data entry. Clearly, it must be acknowledged that none of the subjects had experience of battlefield conditions, so the result has to be viewed with due caution. Nevertheless, given the functionality offered by these devices in the context of the experimental task, the keyboard was unreservedly acceptable, the enhanced recognizer less acceptable and the extant recognizer positively unacceptable.

Implications of the behavioural data for a model of device-user interaction
(a) Concurrent activities. The general conclusion to be drawn from the behavioural data is that, although poor device functionality eliminated any performance advantage offered by speech, there was evidence of the potentially advantageous behavioural patterns predicted by the preliminary model. The enhanced recognizer, in particular, enabled actions to be performed concurrently with data entry.

(b) Recoding of information. The task demanded that data be recoded by subjects from a spatial to numeric form (i.e. by calculating map references). However, the devices differed in regard to the recoding required to enter data to the computer (i.e. the recognizer required generation a speech message, whereas the keyboard required the initiation of manual movements). Some subjects exhibited considerable fluency in speech data entry with the enhanced recognizer, suggesting that the code was compatible with their mental representations of the information. The enhanced device enabled attention to be addressed primarily to written sources (such as the fireplan/mission record and artillery information sheet), so that subjects could, for example, enter data by reading aloud from a written source.

Although speech offered coding advantages over the keyboard, it should be added that a user interface for the entry of target positions with the minimum recoding would be a system in which data were entered using direct designation on a spatial (e.g. map) display.

(c) User dialogue structure. Although the form of the user dialogue was not of primary concern in the present study, the data reveals some interesting behavioural features potentially relevant to dialogue design. For many subjects, strategy of device use differed between the two recognizers. Although both devices were, ostensibly, capable of accepting connected speech, this form of entry was only used when recognition was sufficiently reliable to enable a high proportion of data strings to be entered without errors. This was because error correction involved cursor movement by a sequence of discrete commands, the length of the sequence depending on how far back was the error. When errors were frequent, it was generally more

efficient to enter words individually and to ensure correct recognition before proceeding to the next. The implication is that the facility of recognition of connected speech is only of functional value if recognition reliability is high. In the present study, the critical level of performance was exhibited by the enhanced recognizer but not by the extant recognizer.

There were also differences exhibited by subjects in their strategy for sending sequences of orders against the same target (e.g. in adjustment) using the two speech devices. When the device was unreliable, it was clearly advantageous to modify only the required characters on the page, using the cursor movement commands to navigate the display. This forced a strategy of sequential target engagement, because subjects had to wait for an order to be effected on the battlefield display before they could send another order by modifying the one already on the screen. When the device was more reliable, subjects typically preferred to clear whole fields (by designating their name) and then to re-enter the string of data to the field. However, some of the more skilled subjects went further, in clearing the whole page and then recompleting the fields; this enabled them to adjust on more than one target at a time. Such skilled performance could only supported by a reliable data entry device.

There was strong evidence of subjects developing a model of the specific error characteristics of the device and developing strategies to cope with them. The commonest example was that of use of the field name commands, which enabled users to move the cursor directly to any field on the display and to clear it, all with a single command. However, when the name was mis-recognized by the device, the consequences could be extremely frustrating (e.g. the cursor moved to and cleared the wrong field). Subjects discovered that the cursor movement commands (UP, DOWN, LEFT RIGHT and DELETE) which only moved the cursor one place, were, nevertheless, highly reliable. They consequently tended to use these when the device was otherwise unpredictable, even though the strategy frequently demanded a long sequence of repeated commands. This observation clearly supports a requirement for robustness in the recognition of regularly used device functions.

D3.4.3 General conclusions of the study

The clear conclusion of the study is that the extant device was neither usable nor acceptable for supporting the FOO in his task. Enhanced recognition performance engendered positive behavioural features (e.g. concurrent action and compatible data coding), many of which render a reliable speech recognizer more usable than a keyboard. However, as reported, these advantages did not result in universal acceptance in the context of the experiment.

Acceptability is presumably some function of the user's perception of the functionality offered by the device, the advantages conferred by it (relative to alternative devices) and its usability. The task as simulated did not reproduce some functional demands potentially relevant to this equation, such as mobile operation, operation in postures unsuitable for keying and operation in conditions of poor illumination. It would be expected that the acceptability of the enhanced recognizer relative to the keyboard would be more favourable if such operational contexts were considered.

The conclusion is that a speech recognizer (even an advanced one) will only be worth implementing if it offers functional advantages over the alternatives, such as operability in a wider range of contexts. The users (i.e. the army) have ultimately to decide the extent to which their engagement computer will be used in situations in which a keyboard is difficult to operate. If this is significant, a recognizer with performance similar to that of the enhanced device studied here would be a suitable means of supporting device-user interaction.

D4. IMPLICATIONS FOR THE USABILITY EVALUATION METHOD

The experiment was generally successful in its assessment of device usability, and so it is appropriate as a vehicle for the development of the UEM. However, although useful as a starting point, the precursive UEM advanced in Section D2.2 clearly requires considerable elaboration if it is to be usable by people who are not human factors specialists. The experiment immediately suggests the following enhancements.

1. TAS specification. The process of specifying the requirements for task, device and user simulations, and the subsequent design of the evaluation experiment was here proceduralized largely implicitly by the investigator, utilizing a partially explicit interaction model. The UEM requires a mechanism for this specification, recruiting explicit models residing in the Diagnostic Manual. It is possible that heuristics may have to be specified for the process of experimental design.

2. Data collection, compression and statistical treatment. The form of the representation of the experimental results is currently unspecified in the UEM. It is probable that it will only be possible to provide procedures for running and analyzing experiments of the types expected to be most commonly appropriate, i.e. the method will not be complete in this regard. Users will require guidance on the selection of appropriate scales for the variables (e.g. whether the data should support interval, ordinal or nominal scaling) and on the application of inferential statistics to experimental data.

3. Interaction model. The speech interaction models residing in the diagnostic manual must be interfaced with the UEM. The model must have appropriate "hooks" for relating it to the data during interpretation of results and generation of interface design recommendations. It must also support the specification of experimental variables during the process of experimental design (see [1] above).

ANNEX PART 1: TAS VIDEO DATA (main)

DEVICE TYPE		Total time spent on the device (training time inc.) (mm:ss)	(centi-sec)	No. of extra training	total training time (centi-sec)	Total time spent on the device (training time excl.) (centi-sec)	No. of page inputs	No. of interactions	Corr. mean time per interaction excl. training (centi-sec)	Concurrent activities that occur simultaneously with the use of the device	Corr. time per concurrent activity excl. training
Keyboard	S1	08:33.04	51304	N/A	N/A	N/A	17	34	1509	1	1.95E-05
	S2	11:21.85	68185	N/A	N/A	N/A	28	37	1843	0	0.00E+00
	S3	11:58.62	71862	N/A	N/A	N/A	20	21	3422	2	2.78E-05
	S4	09:08.12	54812	N/A	N/A	N/A	18	20	2741	0	0.00E+00
	S5	06:20.13	38013	N/A	N/A	N/A	18	19	2001	1	2.63E-05
	S6	08:00.82	48082	N/A	N/A	N/A	18	27	1781	1	2.08E-05
	S7	09:16.84	55684	N/A	N/A	N/A	17	19	2931	0	0.00E+00
	S8	06:10.97	37097	N/A	N/A	N/A	15	17	2182	2	5.39E-05
	S9	08:10.64	49064	N/A	N/A	N/A	18	20	2453	0	0.00E+00
	S10	09:07.55	54755	N/A	N/A	N/A	20	28	1956	1	1.83E-05
	S11	08:33.77	51377	N/A	N/A	N/A	14	26	1976	0	0.00E+00
	S12	09:16.39	55639	N/A	N/A	N/A	20	21	2649	0	0.00E+00
Extant	S1	10:33.73	63373	2	11368	52005	17	22	2364	7	1.35E-04
	S2	13:35.88	81588	1	4686	76902	22	26	2958	12	1.56E-04
	S3	26:13.14	157314	3	16271	141043	25	32	4408	3	2.13E-05
	S4	17:23.26	104326	2	10881	93445	18	19	4918	31	3.32E-04
	S5	15:45.05	94505	2	10735	83770	24	25	3351	11	1.31E-04
	S6	14:41.84	88184	3	18125	70059	16	27	2595	9	1.28E-04
	S7	14:51.02	89102	2	9791	79311	17	19	4174	0	0.00E+00
	S8	14:55.66	89566	4	17746	71820	17	19	3780	13	1.81E-04
	S9	28:26.88	170688	4	20097	150591	29	35	4303	25	1.66E-04
	S10	18:15.06	109506	2	10245	99261	16	26	3818	32	3.22E-04
	S11	18:15.82	109582	2	9882	99700	16	27	3693	14	1.40E-04
	S12	13:48.87	82887	1	4413	78474	17	19	4130	5	6.37E-05
Advanced	S1	10:06.31	60631	0	0	60631	16	20	3032	10	1.65E-04
	S2	14:10.89	85089	0	0	85089	29	40	2127	11	1.29E-04
	S3	09:23.28	56328	0	0	56328	17	18	3129	13	2.31E-04
	S4	09:35.79	57579	0	0	57579	19	20	2879	39	6.77E-04
	S5	07:53.07	47307	0	0	47307	19	20	2365	4	8.46E 05
	S6	11:23.57	68357	0	0	68357	17	27	2532	16	2.34E-04
	S7	08:45.90	52590	0	0	52590	17	18	2922	12	2.28E-04
	S8	10:31.59	63159	1	4991	58168	16	16	3636	7	1.20E-04
	S9	11:41.14	70114	0	0	70114	26	28	2504	7	9.98E-05
	S10	13:20.44	80044	0	0	80044	20	34	2354	77	9.62E-04
	S11	11:36.29	69629	0	0	69629	11	14	4974	11	1.58E-04
	S12	09:23.99	56399	0	0	56399	16	16	3525	9	1.60E-04

ANNEX PART 2: TAS VIDEO DATA (detailed breakdown of behavioural pattern)

Frequency of concurrent activities

DEVICE TYPE		Fire planning info. sheet	Mission record	Map	Terrain	Binoculars	Intercom	Writing on mission record	TOTAL	
Keyboard	**S1**	0	0	0	0	0	1	0	**1**	
	S2	0	0	0	0	0	0	0	**0**	
	S3	1	0	0	0	0	1	0	**2**	
	S4	0	0	0	0	0	0	0	**0**	
	S5	0	0	0	0	0	1	0	**1**	
	S6	0	0	0	1	0	0	0	**1**	
	S7	0	0	0	0	0	0	0	**0**	
	S8	0	2	0	0	0	0	0	**2**	
	S9	0	0	0	0	0	0	0	**0**	checksum (Σrc(j-1)) :
	S10	0	1	0	0	0	0	0	**1**	8
	S11	0	0	0	0	0	0	0	**0**	checksum (Σc(j)) :
	S12	0	0	0	0	0	0	0	**0**	8
	Sub-totals	1	3	0	1	0	3	0	**8**	
Extant	**S1**	0	4	0	3	0	0	0	**7**	
	S2	0	0	0	9	0	0	3	**12**	
	S3	0	0	1	2	0	0	0	**3**	
	S4	2	18	4	3	0	0	4	**31**	
	S5	1	4	1	4	0	0	1	**11**	
	S6	0	5	0	3	0	0	1	**9**	
	S7	0	0	0	0	0	0	0	**0**	
	S8	0	3	0	9	0	0	1	**13**	
	S9	0	21	2	2	0	0	0	**25**	checksum (Σrc(j-1)) :
	S10	0	30	0	2	0	0	0	**32**	162
	S11	0	5	0	9	0	0	0	**14**	checksum (Σc(j)) :
	S12	1	0	0	0	0	0	4	**5**	162
	Sub-totals	4	90	8	46	0	0	14	162	
Advanced	**S1**	0	4	1	4	1	0	0	**10**	
	S2	0	7	0	2	0	0	2	**11**	
	S3	0	6	5	2	0	0	0	**13**	
	S4	0	26	3	6	0	0	4	**39**	
	S5	0	2	0	2	0	0	0	**4**	
	S6	0	13	1	2	0	0	0	**16**	
	S7	0	8	0	4	0	0	0	**12**	
	S8	0	2	0	0	0	0	5	**7**	
	S9	0	5	0	2	0	0	0	**7**	checksum (Σrc(j-1)) :
	S10	3	67	0	7	0	0	0	**77**	216
	S11	0	9	0	0	0	0	2	**11**	checksum (Σc(j)) :
	S12	0	5	0	0	0	0	4	**9**	216
	Sub-totals	3	154	10	31	1	0	17	216	

ANNEX PART 2 (contd.): TAS VIDEO DATA (detailed breakdown of behavioural pattern)

Frequency of activities with rapid succession

DEVICE TYPE		Fire planning info. sheet	Mission record	Map	Terrain	Binoculars	Intercom	Writing on mission record	TOTAL	
Keyboard	**S1**	0	37	0	4	0	0	1	**42**	
	S2	0	68	0	10	0	0	1	**79**	
	S3	14	32	12	9	0	0	0	**67**	
	S4	10	37	1	6	0	0	3	**57**	
	S5	2	37	2	5	0	0	0	**46**	
	S6	9	44	0	7	0	0	9	**69**	
	S7	10	49	1	8	0	0	3	**71**	
	S8	4	43	1	4	1	0	6	**59**	
	S9	5	41	0	4	0	0	0	**50**	checksum (Σrc(j-1)) :
	S10	3	78	0	18	0	0	3	**102**	762
	S11	6	41	1	4	0	0	0	**52**	checksum (Σc(j)) :
	S12	5	53	4	5	0	0	1	**68**	762
	Sub-totals	68	560	22	84	1	0	27	762	
Extant	**S1**	7	60	11	11	0	0	5	**94**	
	S2	0	80	0	17	0	0	4	**101**	
	S3	0	65	36	23	0	1	0	**125**	
	S4	1	58	4	10	0	0	1	**74**	
	S5	5	43	10	10	0	0	0	**68**	
	S6	7	44	0	14	0	0	2	**67**	
	S7	2	70	2	10	0	0	6	**90**	
	S8	10	72	1	2	1	0	3	**89**	
	S9	9	51	6	13	0	0	1	**80**	checksum (Σrc(j-1)) :
	S10	6	46	1	20	0	0	7	**80**	1028
	S11	8	68	0	13	0	0	7	**96**	checksum (Σc(j)) :
	S12	14	40	3	5	0	0	2	**64**	1028
	Sub-totals	69	697	74	148	1	1	38	1028	
Advanced	**S1**	1	41	8	18	0	0	8	**76**	
	S2	0	129	0	19	0	0	4	**152**	
	S3	0	39	20	10	0	0	3	**72**	
	S4	5	30	1	6	0	0	1	**43**	
	S5	7	52	1	4	0	0	2	**66**	
	S6	9	49	1	15	0	0	10	**84**	
	S7	0	56	0	10	0	0	5	**71**	
	S8	7	50	3	14	0	0	8	**82**	
	S9	4	47	1	6	0	0	2	**60**	checksum (Σrc(j-1)) :
	S10	11	49	0	8	0	0	3	**71**	911
	S11	11	47	0	15	0	0	7	**80**	checksum (Σc(j)) :
	S12	16	33	0	4	0	0	1	**54**	911
	Sub-totals	71	622	35	129	0	0	54	911	

ANNEX PART 3: TAS COMPUTER DATA

DEVICE TYPE		Accrued hostility	Total rounds left	No. of fire orders	Interval between first and last fire order (mm:ss)	(centi-sec)
Keyboard	**S1**	243.75	171	17	28:34.80	171480
	S2	317.5	174	28	24:50.98	149098
	S3	249.75	163	20	30:49.57	184957
	S4	203	166	18	22:23.33	134333
	S5	195.75	162	18	18:43.50	112350
	S6	220	154	18	14:44.51	88451
	S7	271.75	177	17	21:20.75	128075
	S8	191	191	15	19:26.92	116692
	S9	232	206	18	22:02.34	132234
	S10	267.25	163	20	20:17.40	121740
	S11	225	165	14	16:59.66	101966
	S12	269	176	20	25:58.41	155841
Extant	**S1**	317.5	142	17	26:19.25	157925
	S2	339.5	157	22	22:49.00	136900
	S3	489	163	25	45:13.53	271353
	S4	282	141	18	27:37.61	165761
	S5	327.5	178	24	28:29.17	170917
	S6	277.5	178	16	18:16.42	109642
	S7	349.5	177	17	27:02.09	162209
	S8	227.5	177	17	33:20.78	200078
	S9	477	54	29	52:42.78	316278
	S10	262.75	183	16	25:01.76	150176
	S11	317	111	16	28:12.22	169222
	S12	257	178	17	22:23.47	134347
Advanced	**S1**	261.25	172	16	18:59.87	113987
	S2	353.75	185	29	25:45.25	154525
	S3	243.5	183	17	21:49.10	130910
	S4	203.5	155	19	23:38.40	141840
	S5	176	133	19	17:24.00	104400
	S6	229.5	178	17	18:41.35	112135
	S7	276.5	178	17	20:47.89	124789
	S8	193.5	178	16	23:09.20	138920
	S9	300.5	189	26	32:05.55	192555
	S10	284.25	176	20	24:10.29	145029
	S11	254	154	11	18:11.26	109126
	S12	255.5	179	16	20:47.50	124750

ANNEX PART 4: TAS QUESTIONNAIRE DATA
Two questions from the Subjects'
Questionnaires are measured :

Q1: "Rate the performance of the device on a scale from 1 to 5 whereby 1 - Totally unacceptable, 3 - Acceptable and 5 - Acceptable without reservation."
Stat. Test The Friedman two-way ANOVA

	Device types	***KEYBOARD***		***EXTANT***		**ADVANCE**	
		Q1 (raw score)	**Q1 (rank order)**	**Q1 (raw score)**	**Q1 (rank order)**	**Q1 (raw score)**	**Q1 (rank order)**
Subjects code	**S1**	4	2	2	1	5	3
	S2	4	2.5	2	1	4	2.5
	S3	3	2	1	1	5	3
	S4	4	3	1	1	3	2
	S5	5	3	1	1	3	2
	S6	5	3	2	1	3	2
	S7	4	2	2	1	5	3
	S8	5	3	2	1	3	2
	S9	5	3	3	1	4	2
	S10	3	2	2	1	4	3
	S11	5	3	2	1	4	2
	S12	4	2.5	1	1	4	2.5
	Rj		31		12		29

Q4: "Would you be happy to use this device on the battlefield ?"
Stat. Teste Cochran Q test

	Device types	***KEYBOARD***		***EXTANT***		**ADVANCE**	
		Q4 (response)	**Q4 (0 for No; 1 for Yes)**	**Q4 (response)**	**Q4 (0 for No; 1 for Yes)**	**Q4 (response)**	**Q4 (0 for No; 1 for Yes)**
Subjects code	**S1**	yes	1	no	0	yes	1
	S2	yes	1	no	0	yes	1
	S3	yes	1	no	0	yes	1
	S4	yes	1	no	0	yes	1
	S5	yes	1	no	0	yes	1
	S6	yes	1	no	0	no	0
	S7	yes	1	no	0	yes	1
	S8	yes	1	no	0	no	0
	S9	yes	1	no	0	yes	1
	S10	yes	1	no	0	no	0
	S11	yes	1	no	0	no	0
	S12	yes	1	no	0	yes	1
		R1=	12	**R2=**	0	**R3=**	8

ANNEX: PART 5
INFERENTIAL STATISTICS

Variable 1: Time to implement fireplan (calculated in centi-sec)

	Keyboard	Extant Recog.	Enhanced Recog.
Σranks[9]	20	31	21

$X^2r = 12/(RC(C+1))^*\Sigma R_j^2\text{-}3R(C+1)= 6.4$ [where R = Rows; C = Columns]

Variable 2: Time spent using the device

	Keyboard	Extant Recog.	Enhanced Recog.
Σranks	14	35	23

$X^2r = 12/(RC(C+1))^*\Sigma R_j^2\text{-}3R(C+1)= 18.5$ [where R = Rows; C = Columns]

Variable 2a: Time spent using the device to perform battlefield task

	Keyboard	Extant Recog.	Enhanced Recog.
Σranks	14	34	24

$X^2r = 12/(RC(C+1))^*\Sigma R_j^2\text{-}3R(C+1)= 16.7$ [where R = Rows; C = Columns]

Variable 3: Mean time per device interaction

Source	Sum of Squares	df	MS	F	p
Device Type	22278859.7	2	11139429.9	32.8	(p<.001)
Subjects	8960521.6	11	814592.9		
Residual	7462017.6	22	339182.6		
Total	38701398.9	35			

In this case, $MS_{error} = 339182.6$; nr = 3. Hence, $3.58\sqrt{(339182.6/3)} = 1203.8$

		µX1	µX2	µX3
	Means	**2200.8**	**2909.9**	**4107**
µX1	**2200.8**		709.1	1906.2*
µX2	**2909.9**			1197.1

* p<.05

[9] Sum of ranks across 12 subjects. The lowest score of each subject's scores across three devices is ranked 1. The next lowest is ranked 2. The highest rank is 3. Hence, maximum Σranks is 36 (indicating that all subjects exhibited the most interactions with that device) and minimum Σranks is 12 (indicating that all subjects exhibited fewest interactions with that device).

Variable 3a: Time per device interaction (excl. training)

Source	Sum of Squares	df	MS	F	p
Device Type	12111564	2	6055782	18.11	($p<.001$)
Subjects	9186548	11	835140		
Residual	7355572	22	334344		
Total	28653684	35			

In this case, $MS_{error} = 334344$; $nr = 3$. Hence, $3.58\sqrt{(334344/3)} = 1195.14$

		$\mu X1$	$\mu X2$	$\mu X3$
	Means	2287	2998	3708
$\mu X1$	2287		711	1421*
$\mu X2$	2998			710

* $p<.05$

Variable 4: Accrued target hostility value

	Keyboard	Extant Recog.	Enhanced Recog.
Σranks	17	32	23

$X^2r = 12/(RC(C+1))*\Sigma R_j^2-3R(C+1)= 9.5$ [where R = Rows; C = Columns]

Variable 5: Ammunition rounds unused after the mission

	Keyboard	Extant Recog.	Enhanced Recog.
Σranks	24	19.5	28.5

$X^2r = 12/(RC(C+1))*\Sigma R_j^2-3R(C+1)= 3.38$ [where R = Rows; C = Columns]

Variable 6: Number of interactions with the device

	Keyboard	Extant Recog.	Enhanced Recog.
Σranks	25	26.5	20.5

$X^2r = 12/(RC(C+1))*\Sigma R_j^2-3R(C+1)= 1.6$ [where R = Rows; C = Columns]

Variable 7: Number of incidences of actions concurrent with device use

	Keyboard	Extant Recog.	Enhanced Recog.
Σranks	12.5	28.5	31

$X^2r = 12/(RC(C+1))*\Sigma R_j^2-3R(C+1) = 16.8$ ($p<.001$) [where R = Rows; C = Columns]

Variable 7a: Concurrent actions per unit time of interaction period

	Keyboard	Extant Recog.	Enhanced Recog.
Σranks	13.5	26.5	32

$X^2r = 12/(RC(C+1))*\Sigma R_j^2-3R(C+1) = 15.04$ ($p<.001$) [where R = Rows; C = Columns]

Variable 8: Number of fire-orders sent

	Keyboard	Extant Recog.	Enhanced Recog.
Σranks	23.5	26	22.5

$X^2r = 12/(RC(C+1))*\Sigma R_j^2-3R(C+1) = 0.5$ [where R = Rows; C = Columns]

Variable 9: Number of recognizer re-trainings

Device type	EXTANT	ADVANCED	Direction of difference	Sign
	(raw score)	(raw score)		
Subject				
S1	2	0	Xe>Xa	+
S2	1	0	Xe>Xa	+
S3	3	0	Xe>Xa	+
S4	2	0	Xe>Xa	+
S5	2	0	Xe>Xa	+
S6	3	0	Xe>Xa	+
S7	2	0	Xe>Xa	+
S8	4	1	Xe>Xa	+
S9	4	0	Xe>Xa	+
S10	2	0	Xe>Xa	+
S11	2	0	Xe>Xa	+
S12	1	0	Xe>Xa	+

Variable 10: Rating of device performance

$X^2r = 12/(RC(C+1))*\Sigma R_j^2-3R(C+1)= 18.2$ (p<.001) [where R = Rows; C = Columns]

Variable 11: Rating of acceptability for battlefield use

Cochran Q test (response:0 for No; 1 for Yes)

Device types	KEYBOARD		EXTANT		ADVANCED	
Subject						
S1	yes	1	no	0	yes	1
S2	yes	1	no	0	yes	1
S3	yes	1	no	0	yes	1
S4	yes	1	no	0	yes	1
S5	yes	1	no	0	yes	1
S6	yes	1	no	0	no	0
S7	yes	1	no	0	yes	1
S8	yes	1	no	0	no	0
S9	yes	1	no	0	yes	1
S10	yes	1	no	0	no	0
S11	yes	1	no	0	no	0
S12	yes	1	no	0	yes	1
	ΣR1=	12	ΣR2=	0	ΣR3=	8

ΣRj= 20

$Q = (C-1)*(C*\Sigma R_j^2-(\Sigma R_j)^2)/(C*\Sigma X_i-\Sigma(X_i)^2)= 18.7$ (p<.001)

[where R = Rows; C = Columns; $X = \Sigma R_{i..}R_{i+2}$]

APPENDIX E

CONFIGURATION OF SIAM FOR APPLICATION

INTRODUCTION

This appendix was included in the SIAM manual. It describes the potential for flexibility in the application of SIAM and was intended to assist the assessor in deciding an appropriate configuration of the method.

The form of an assessment using SIAM is determined by the intersection between the constraints imposed by the problem being addressed and the constraints imposed by the context in which the assessment is performed. These constraints limit the degrees of freedom in configuration; however, the actual configuration selected is ultimately dependent upon the judgement of the assessor. This discussion is a source of information to support judgement, and it concludes with some rules of thumb to guide the non-human factors specialist in deciding an appropriate form for an assessment.

SCOPE FOR FLEXIBILITY IN SIAM

SIAM is intended to be applicable in a wide range of situations presented by feasibility assessment. It has consequently been developed to be flexible both in the way that the component methods may be "assembled" together, and in the way that the component methods may be used, individually or together, to develop a variety of products (e.g. simulations varying both in scope and level of detail). This section discusses three aspects of assesssments which might vary according to circumstances. The subsequent sections consider the factors which may limit the scope for variation.

Fidelity of system representation

"Fidelity" is a term used to describe the verisimilitude of a representation. In the context of user interface simulations the term "fidelity" is currently used in a qualitative fashion: "high fidelity" simulations are accurate reproductions of a target device, whereas "low fidelity" simulations represent device features crudely. An assessor using SIAM has to make two classes of decision which determine fidelity in this sense. Firstly, a decision is made as to what interface design issue is the concern of the investigation; and, secondly, a decision is made as to the level of detail to which the investigation is taken.

The first of these decisions determines which sections of the Diagnostic Tables should be utilized in simulation design (i.e. it determines the speech interaction model); this, in turn, constrains the number of attributes of the target system and task which are represented in the simulation. If the design issue were a "narrow" one (e.g. acoustic confusibility of the vocabulary of a speech synthesizer), then only a small number of system and task attributes might be reproduced accurately. In the general sense of the term, this would be only a low fidelity simulation (however, note that the attributes which were represented - e.g. machine vocabulary and syntax; machine speech characteristics; ambient noise; user familiarity with machine language - might be reproduced precisely). However, if the design issue were a "broad" one (e.g. to predict the interaction behaviour and absolute level of performance of a human-machine system performing a battlefield task), more attributes would have to be represented. This is because any of a large number of attributes may impact overall task performance, and their effects may be rendered complex by mutual interactions.

The second decision determines, not the number of attributes, but the amount of detail in the representation of the attributes. Consider a target system and a simulation of that system. An evaluation of the fidelity of the simulation will be based upon the similarity of *descriptions* of the target and the simulation. But description can occur at different *levels*; for example, a low level description of a speech synthesizer might characterize the spectrum of the acoustic signal it exhibits, whereas a high level description might characterize its ability to reproduce the prosodic features of human spoken language. Both describe linguistic attributes of the device output, but the level of description is different. The accuracy of reproduction must be expressed, then, at a certain level of description, and this level of description is determined by the detail of the investigation. An investigation concerned with synthesizer intelligibility demands a simulation reproducing the target device which is accurate at a low (i.e. acoustic) level of description. However, if the concern is whether the user will be able to remember the vocabulary of the machine dialogue, then such a low-level reproduction of the device may be unnecessary. The acoustic qualities could be reproduced crudely, although the vocabulary and grammar of the machine dialogue, and the familiarity of the user with the language of the task, may need to be reproduced accurately.

An important determinant of fidelity, in practice, will be the adequacy of the instantiation of the attributes of system entities when the simulation is implemented. A general goal of the assessor should be to make the reproduction *psychologically equivalent* with respect to the critical attributes, such that the *behaviour elicited by the simulation attributes is the same as that elicited by those attributes of the target system* (where "behaviour" represents the mental (informational) and physical responses of the user and device in the context of the task). Because it is difficult analytically to assess psychological equivalence in any but the simplest behavioural exchanges, SIAM includes checks of fidelity in which a "domain expert" evaluates the reproduction.

In summary, then, simulation fidelity is determined in SIAM by the issue of concern in an assessment and by the level of detail required in the output of the investigation. These are chosen by the investigator. It will also be determined by the adequacy of the implementation of the simulation. The criterion for adequacy is psychological equivalence between the critical attributes of target system and simulation at the decided level of description. The empirical application of SIAM includes checks by domain experts on the psychological equivalence of critical simulation attributes to those of the target system.

"Modularization"

SIAM is modular in structure, consisting of a usability evaluation method and three relatively independent simulation development methods. It, therefore, offers opportunities for flexibility in the way that these submethods are recruited in an evaluation. For example, SIAM may be used to evaluate a currently-available speech interface by using the usability evaluation method to plan the evaluation, the task simulation method to provide the experimental context, and the user simulation method to select and train subjects. In this case the device simulation would be replaced with the target device itself, so the device simulation method would not need to be applied.

FACTORS CONSTRAINING CONFIGURATION OF SIAM

The decision as to an appropriate class of assessment, fidelity for representation and choice of SIAM modules is determined by constraints deriving from the issue under investigation and by constraints deriving from the context of the investigation. These two classes of constraints tend to be opposed in their effects.

Constraints deriving from the issue under investigation

The issue under investigation will constrain configuration of SIAM by setting the minimum requirements for a study to answer a question raised in feasibility assessment.

(a) **Guarantee of correct assessment.** An assessment will deliver an evaluation of the behaviour and/or performance of a target system with an associated level of confidence. The importance attached to getting the assessment correct will determine the required level of confidence and, hence, constrain the form of the study. In general, higher levels of confidence will be attached to studies which are empirical, high fidelity and/or which involve evaluations utilizing actual target devices or users. If accuracy of prediction is important, the study will tend toward this pattern. However, if accuracy is less important, analytic or low fidelity simulation studies may be adequate, which make extensive use of representations of the target system.

(b) **Scope of study.** The issue under investigation determines the required scope of the study, i.e. the number of system attributes which have to be taken into account in order accurately to answer a question about behaviour. In general, studies of behaviour which has many determinants will demand empirical assessment and high fidelity simulation, in order to account for the impact of the multiple attributes. Analytic assessment is more likely to be successful when only afew factors determine behaviour, as there is then less chance that unforeseen behavioural interactions will have an important impact on performance.

(c) **Level of detail.** The level of detail (precision) demanded in the performance prediction will limit the approaches which may be taken in an investigation. A demand for highly detailed predictions favours empirical investigation, either utilizing elements of the target system itself, or high fidelity simulations which correspond to the target system at a low level of description.

Constraints deriving from the context of the investigation

The circumstances of an evaluation will tend to limit the extent to which the optimal configuration of SIAM (see the previous section) may be achieved in practice.

(i) **Information on the TD.** The amount of information available on the expected behaviour and performance of the TD determines the range of potentially viable investigative approaches and/or determines the assumptions to which the conclusions must be subject. Three classes of information availability are now assessed.

Prototype availability. Where the procurement process involves the application of current speech technology but in, say, some novel task domain, full information should be available on the behaviour and performance of the device. Indeed, as mentioned previously, the usability of the device itself might be evaluated using SIAM. However, if a device simulation is to be built (perhaps for reasons of flexibility, as it is easy to modify the specification of a human simulation), then performance data may be obtained in a laboratory study of the target device, or from the device manufacturers. In this situation, detailed information will be available on all the device attributes, enabling the development of high fidelity simulations which can provide contexts for broad and detailed investigations.

Specification availability. Where the procurement process involves a speech technology in the pre-market state of development, but with its capabilities and performance characteristics known, then behaviour and performance in the target context should be largely predictable, e.g. by the device developers. In this case it should be possible to specify a user dialogue at a low level (e.g. with respect to syntax and vocabulary); to specify likely error types and frequencies and to specify the temporal characteristics of its response. Again, because low level information is available for simulation development, broad and low level design issues may be addressed

Technological prediction. Where the target system is large in scope and complex in its demand for novel technologies, the procurement cycle may be extended over several years. SIAM is applicable in such a situation, but it presents substantial challenges in the formulation of assumptions about the behaviour and performance of the target system. The device simulation method offers a procedure for approaching prediction, in which critical attributes of the device are identified, and specialists in the development of speech technology recruited to predict the form of the target device with respect to these attributes. It is assumed that such specialists will hold the best available view of trends in technological development; however, it is inevitable that the confidence with which predictions are made will be some function of the lead time of the target device, that low levels of description cannot be guaranteed and that the scope of studies will be limited by information available on device attributes.

(ii) **Information on interaction behaviour.** The adequacy of the assessor's model of device-user interaction will limit the range of options for configuration of SIAM. A highly detailed interaction model may enable both analytic and empirical investigations. However, if the model is very limited, the consequences of interaction will have to be determined empirically, and, furthermore, high fidelity simulation will be necessary to allow unforeseen impacts of system attributes on behaviour to manifest themselves.

(iii) **Assessor skills.** SIAM can, in principle, be applied by assessors who are not experts in human-computer interaction research. However, the skills of the assessors will constrain the class of assessment which may be used in a particular situation; the confidence with which a conclusion on device suitability may be drawn; and the confidence with which information acquired in an assessment may be added to the Diagnostic Tables.

Analytic usability assessment depends for its success on the knowledge of device-user interaction held by the assessor. Almost by definition, a non-human factors specialist has incomplete knowledge of device user interaction. Although assessor knowledge is supplemented by the Diagnostic Tables, these cannot be claimed to be complete, so analytic assessment by a non-specialist risks being incorrect. An empirical assessment makes fewer demands on domain knowledge and so is necessary for non-specialist assessors.

However, all assessments - analytic or empirical - demand a decision about an appropriate interaction model for an assessment. Thi, is achieved by a procedure in the Usability Evaluation Method in which an assessor identifies the diagnostics which are to be recruited to the study. A researcher experienced in human-computer interaction will be more able to identify critical system conditions germane to the design of the device, and consequently requiring assessment. An inexperienced assessor, then, should be able to make a valid assessment using SIAM, but risks the assessment being incomplete.

In conclusion, an inexperienced user of SIAM would be recommended (1) to utilize an empirical rather than an analytic approach; (2) either to seek advice in the identification of potentially critical system conditions, or to err on the side of over-inclusion, in order to reduce the chance of missing relevant interaction behaviour.

(iv) **Resources available for investigation.** The availability of facilities and resources such as time will limit the options for configuring SIAM. Empirical assessments demand time for the development of simulations, the performance of experiments and the analysis of results. They also require technological resources for the implementation of simulations. By contrast, analytic assessments by an expert may be performed quickly and demand few, if any, technological resources. Limited resources, then, may constrain the class of assessment (and hence the guarantee attached to the conclusions), the breadth of the issue which may be addressed and the level of detail of the output of SIAM.

Conclusions and recommendations on the configuration of SIAM

This appendix has described how the requirements for a study to answer a specific usability question, and the circumstances under which the question must be answered, constrain the options for configuring SIAM. These constraints determine the class of assessment, the fidelity of representations and the usage of modules of SIAM.

SIAM offers an opportunity to predict the device-user interaction behaviour and performance of a system in a task context and hence to assess its usability. Because it encourages a systematic evaluation of the factors likely to influence usability, it should offer benefits in assessment, even in the most disadvantageous situation of a non-specialist assessor, with few resources and limited information on the target device. However, usability is now recognized as being an important factor in system quality, and SIAM has the potential to make valuable contributions in its assessment. The earlier that assessment occurs, the greater the probability that design risks will be identified and hence costly mistakes avoided. Investment in early evaluation using SIAM should, therefore, offer subsequent benefits. The greater this investment, the more detailed will be the information obtained using SIAM and the confidence attached to validity of this information.

In summary, where possible,

(1) seek maximum information on the target system to enable high fidelity representations if necessary
(2) recruit experienced assessors in preference to non-specialist
(3) seek resources for empirical assessments in preference to analytic.

APPENDIX F

EVALUATION OF SIAM

APPENDIX F1: PROTOCOL FOR EVALUATION OF SIAM (MACL/PH/1/90/1)

SIAM REPRESENTATION (NAME):

SECTION A

1. Given the evaluation strategy you have chosen, does SIAM indicate it necessary to develop this representation? YES/NO

If "NO", do not proceed further

2. Were you successful in developing this representation (even if it was only a mental picture in your own mind)? YES/NO

If "NO", please go straight to SECTIONS D and E.

3. Was your representation an implicit mental picture or an explicit product that you could show other people?

IMPLICIT/EXPLICIT

4. Does your representation differ substantially in structure from that recommended by SIAM? YES/NO

If "YES", please explain in what way in the space below.

SECTION B

1. Do you think that your representation includes all the relevant classes of information (even if you are not happy with the details)? YES/NO

If "NO", please identify information which you think should have been included or left out.

2. Do you think that the level of detail of the information in your representation is correct? YES/NO

If "NO", please explain how the level of detail is wrong.

3. Do you think that your representation accurately characterizes what it is supposed to? YES/NO

If "NO", please explain which parts may be inaccurate, and why.

SECTION C

Please rate your representation as requested, by ringing the number which best expresses your view. If your rating is 0, 1 or 2, please briefly explain the problem in the space on the right of the page.

1. Has the representation helped you understand the problem better?

0 - it has confused me
1 - it has not helped at all
2 - it has not helped much
3 - it has helped my understanding
4 - it has considerably helped my understanding
5 - it has been invaluable to my understanding

2. Will the representation help in future planning of the project?

0 - it will be detrimental to planning
1 - it will not help planning at all
2 - it will not help planning much
3 - it will help planning
4 - it will considerably help planning
5 - it will be invaluable to planning

3. Do you expect to have to use the representation to help you communicate to others in the project (e.g. in explaining your appreciation of the problem to others; in describing your progress to date)? YES/NO

If "YES", please say who you expect to show it to.

Please rate its expected contribution to your ability to communicate clearly.

0 - it will confuse other people
1 - it will fail completely to communicate
2 - it will communicate, but will require additional explanation
3 - it will probably communicate adequately by itself
4 - it will make communication very easy
5 - it will be invaluable to communication

4. Will the representation enable you to get on with the next procedure?

YES/NO

If "NO", please briefly explain why.

SECTION D.

Please rate as requested, by ringing the number which best expresses your view. If your rating is 0, 1 or 2, please briefly explain the problem in the space on the right of the page.

1. Please rate the mental effort necessary to develop the representation.

0 - it is impossible
1 - it requires immense effort
2 - it requires substantial effort
3 - it is quite easy
4 - it is very easy
5 - it is trivially easy

2. Please rate the contribution of SIAM's procedure in developing the representation.

0 - the procedure makes it impossible
1 - the procedure does not help at all
2 - the procedure does not help much
3 - the procedure is helpful
4 - the procedure is very helpful
5 - the procedure is invaluable

SECTION E

1. Did you perform all the steps in the procedure? YES/NO
If "NO", please identify which steps you missed and briefly state why you missed them.

2. Did you perform some steps differently to the way specified by SIAM? YES/NO
If "YES", please identify them and say how your approach was different and why.

3. Did you make any mistakes in developing the representation? YES/NO
If "YES", please explain what went wrong and why you think it happened.

**IF YOU HAVE ANY OTHER COMMENTS, PLEASE EXPRESS THEM ON THE BACK.
THANK YOU!**

APPENDIX F2: EVALUATION OF SIAM.
ANALYSIS OF INTERVIEWS 1 AND 2:
CHANGE IN KNOWLEDGE FOLLOWING EXPOSURE TO SIAM

INTRODUCTION

An assessment was made of the relevant knowledge held by the subject of the study (CS), and of the impact of the method on that knowlcdge. This assessment was necessary to enable interpretation of the subject's behaviour when she actually applied the method. The evaluation was performed by means of semi-structured interviews before and after exposure to the method.

The intention was to identify differences in the view held by the subject and that advanced by the method; to determine whether the view had changed; and to identify where (if at all) the subject actually disagreed with the method. Summaries of the subject's responses were analysed with respect to their potential impact on the subsequent use of the method. This document presents the results of the analysis.

PREVIOUS EXPERIENCE

The subject had studied for a PhD on a topic relevant to the evaluation of computer interfaces and had performed a small number of actual evaluations. She might, then, be regarded as an ergonomist with considerable theoretical knowledge and some practical experience. However, she had had no experience of the use of structured methods, nor of the use of speech I/O devices. She had been a subject in an experiment in which a speech interface had been simulated (although she had not been privy to the background of the study). Her knowledge was general, rather than specific to the domain of SIAM.

UNDERSTANDING OF ERGONOMIC EVALUATION

The role of ergonomics

The subject appeared to have a well-founded comprehension of the role of ergonomics in product development, and this was compatible with the expression of the method. However, she did not have a clear understanding of the process of military procurement, and the method's assumption of a procurement project context was novel to her. The main impact of the method seems to have been in the explicitness of its specification of the process of evaluation in the project context.

Factors determining interaction performance

The subject originally expressed a device-centred view of the determinants of performance, although this was largely because of her assumption that the device would normally be the primary object of concern to the ergonomist. She felt that the task/user/device decomposition promoted by the method was useful and did not disagree with it.

Methods of evaluation

When a device is available for evaluation

The subject had assumed an empirical evaluation. The importance of appropriate task and user subjects had been recognized. The subject had an understanding of evaluation which was compatible with the method.

When only a specification is available for evaluation

The subject had recognized the alternative strategies of analytic and empirical evaluation using simulation, although her emphasis had been on an analytic approach. She felt that the method had contributed to her understanding of the process of analytic evaluation, its limitations and its underlying assumptions. She identified a potential problem in the method's approach to empirical evaluation: it requires a very detailed specification of the device in order to design an appropriate simulation.

KNOWLEDGE OF SPEECH TECHNOLOGY

The subject appeared to have a little knowledge of speech technology and of usability issues in the design of speech interfaces, and this seemed to be correct. There was no reason to believe that this knowledge would be contradictory with the information embodied in the method. The subject's general knowledge of user interface design was relevant to the evaluation of speech interfaces, and it would be expected that this knowledge would be recruited to the performance of evaluations.

APPROACH TO THE EVALUATION OF HYPOTHETICAL INTERFACES

Assuming the availability of a prototype

The subject had originally emphasized the importance of obtaining representative user subjects for an empirical evaluation. She did not place much emphasis on the design of an appropriate task for subjects to perform, and she felt that the method was correct in identifying this as important.

Assuming the availability only of a specification

The subject had proposed the informal use of human simulations to determine user requirements. She had not previously identified the need to simulate task and users, and she had not recognized the methodological issue of specifying the task of the person simulating the device. She felt that the method was correct in addressing these points, and that it would be useful in ensuring coverage of the important factors in an evaluation study.

GENERAL CONCLUSIONS

The subject exhibited more potential competence as an ergonomist than is assumed for users of the method. Her model of ergonomic evaluation was apparently close to that of the method, so fundamental disagreement with the method would not be expected. The subject recognized the potential contribution of the method in ensuring completeness, and she accepted the value of explicit intermediate representations, at least for non-expert users of the method. However, her own competence would enable the subject to evaluate the contribution of the method's procedures and to deviate in a controlled manner. It might therefore be expected that, when she used the method, some steps would be omitted and that some intermediate representations would be developed implicitly. Her experience would provide relevant criteria for the subjective evaluation of the method.

Appendix F3.1: Questionnaire responses: Usability evaluation method (Phase 1)

Ratings on a scale of 0 - 5 (see Table 11.2)

	Prel. prob spec.	Prel. syst. spec.	Diag. table conf.	Soln. strat.	Expt. con-text	Data	Anal. of int'n	Feas. rept.
Is rep. necessary?	y	y	y	y	y	y	[1]-	-
Was the rep. developed?	y	y	n	y	y	y	-	-
Implicit or explicit	exp	imp	-	exp	exp	exp	-	-
Diff. from SIAM prescription?	n[2]	y[3]	-	n	n[4]	y[5]	-	-
Does the rep. - include relevant classes of info?	y	y	-	n[6]	y[7]	y	-	-
- show correct level of description?	-	y	-	n[8]	y	y	-	-
- characterize what it is supposed to?	y	y[9]	-	n[10]	y[11]	y	-	-
Helped understanding?	3	n/a	-	3[12]	3[13]	5	-	-
Helped planning?	4	3	-	4	-[14]	5	-	-
Helped communication?	y 3[15]	n	-	n	y4[16]	y4[17]	-	-
Enabled next procedure?	y	-	-	y	y	y	-	-
Mental effort to develop rep.?	2[18]	-	0[19]	2[20]	2[21]	2[22]	-	-
Support from SIAM?	4	-	2[23]	3	3[24]	2[25]	-	-
All steps performed?	y	-	n[26]	n[27][28]	y	n[29]	-	-
Steps performed differently?	n	-	y[30]	n	y[31]	y	-	-
Mistakes?	n	-	-	-	n[32]	n	-	-

[1] Exploratory study only

[2] Because study is exploratory, level of detail can only be general (because don't want to pre-empt results)

[3] (a) Device was actually available, so no need for abstract representation
(b) Task - previous task representations were provided
(c) User - previous user representations were available

[4] But the *extant* worksystem was represented

[5] The functions of this study deviated from that assumed by SIAM: (a) because it was intended to identify problems and possible improvements; and (b) because it was intended to obtain performance data against which to evaluate improvements in Phase 2.. Data was mainly qualitative, although some basic timing data were collected.

[6] Unknown, due to problem with configuration of diagnostics

[7] Slight concern that a pictorial representation of battlefield targets would have been preferable to the alphabetic representation chosen ("Target A"; "Target B"). Alphabetic representation chosen because of lack of information about appearance of targets and lack of time for implementation.

[8] Could be insufficiently detailed

[9] As far as subject knows

[10] Hypotheses/independent variables unspecified due to exploratory nature of study

[11] improvements could be identified

[12] But problem with diagnostics did not help

[13] Running a pilot subject resulted in improvements in subject instructions

[14] The representation is a *requirement* for subsequent steps, rather than just an aid to their planning

[15] Communication with assistant

[16] Head of department and research assistant

[17] Communicated to all involved in project

[18] Effort required to map between type of study (exploratory) and method

[19] See Appendix 3.2

(a) Selection of specific diagnostics seems incompatible with an exploratory study where actual device is under evaluation

(b) General purpose diagnostics are difficult to understand/use

[20] (a) Problem with diagnostics

(b) Mapping between study (exploratory) and method

[21] Effort required in thinking through the scenario, programming the simulation and configuring the experimental testbed

[22] Effort required in data interpretation

[23] Procedure assumes an evaluation in which problems have been precisely specified in advance

[24] Prescribed procedure was compatible with subject's own approach; mainly used it as a checklist

[25] Procedures under-specified; procedures rely on an assessor with experience in data analysis

[26] Step 2: Not applicable unless potential problems have been articulated in advance. This is an exploratory study.

Steps 3/4: See Appendix 3.2; (a) difficult to apply in exploratory study; (b) general purpose diagnostics very high level and difficult to know if applicable; (c) diagnostics 1.1 and 1.3 difficult to understand.

[27] Step 3a(i) and (ii) inappropriate because of nature of study; configuration of diagnostics caused problems

[28] Additional comments: (a) Method assumes a comparative study - not always applicable; (b) Step 3a(v) is not expressed clearly ("Specify experimental design" implies always a formal study; unclear that it only refers to allocation of experimental. conditions, balancing of groups etc.)

[29] No inferential statistical tests were applied

[30] Steps 3/4: Interaction model was developed implicitly to enable progress on subsequent procedures (see Appendix 3.2).

[31] Performed in different order, and some steps missed because of exploratory nature of study. Primarily used as checklist.

[32] General comment: Method does not include procedure for using pilot study results to enhance experimental context.

Appendix F3.2: DISCUSSION BETWEEN AL AND CS OF PROBLEMS ENCOUNTERED BY CS IN CONFIGURATION OF THE DIAGNOSTIC MANUAL FOR EXPLORATORY EVALUATION OF THE TES DEVICE (21/2/90)

AL. Do you know what the DM is trying to do?

CS. I thought it was to find out about the various features of user, device and task which might, in combination, mean that the task is done well or not, i.e. whether interaction between them is good or not.

AL. Yes, in practical terms it attempts to identify critical elements of the interaction so that they can be represented in the simulation and in the design of the experiment.

CS. One reason why I had particular problems with this is because the device is "given".

AL. Also, the problem, at the moment is unspecified.

CS. Yes, its kind of general.

AL. And the Diagnostic Tables are organized in terms of specific usability issues. Maybe we should just go through the steps in the procedure and see whether the procedures suit this particular study.

CS. OK

AL. The first one determines the applicability of SIAM to the technical issues (underlying the specified problem). *(Uses decision tree in Figure 7.2(a)).*

CS and AL. *(Going through the decision tree).* We think its unknown.... ...unknown.......unknown.

AL. OK, so according to this we can assume SIAM is applicable. The next bit assumes you have been able to clearly articulate a problem to investigate. Because what it is trying to do is to work out where the incompatibilities are. So that, I suspect is going to be causing problems.

CS. Yes *(referring to decision tree in Figure 7.2(b)),* because all these are unknown, they could be a "yes" or they could be a "no".

AL. We don't know these. This is going to be relevant to the second stage (i.e. Phase 2).

CS. There could be incompatibilities in any of these and you'll probably come across them .... We are not really looking at environmental things. So these things, we don't need.

AL. I think there's clearly a problem here, and its not necessarily a problem with the procedure, because I think that you would probably be able to ....

CS. Yes, I'll be able to do that at the next stage (i.e. Phase 2).

AL. It looks as if this is a problem with applicability (of the procedures), which might require some elaboration, because with certain kinds of studies you might need a different kind of procedure.

CS. Yes, the problem is because when you are trying to explore something, you haven't got anything really specific; you can only say there might be problems in those areas or there might not, but you're expecting that there could be. But you can't pinpoint it. Not like when you were looking at chunk size, where you might say that the chunk size might not fit the user's representation to do that part of the task ..., in which case you'd have an incompatibility there.

AL. Yes. You could say that, given that we don't know any of these then we should assume that all of them are potentially applicable. I suppose if you take it to the extreme, that's what you'd have to say. But as far as users (of the method) are concerned, that's not clear. And that brings us to the next problem, which is the diagnostics tables. .... Basically, the next bit is that you have to look down the DTs which you've selected by using that Decision Tree (Figure 7.2(b)), to address the issues which you've identified in the Preliminary Problem Spec. If you can't do that....

CS. Yes, you have to use the General Purpose Diagnostics.

AL. Yes, they're put in as a kind of "catch all" to make sure the tables are complete.

CS. What I'd already thought of was, although I hadn't done this very properly, I basically thought I'd go for the general ones, under behaviour and under knowledge, because I knew I wasn't in a position to vary environmental things, so I though they were probably irrelevant. So I thought I'd just pick the general purpose one under knowledge and the general purpose one under behaviour.

AL. Yes. But they're a bit high-level I suspect...

CS. Yes. I think that probably stops their purpose.

AL. ...... There seems to be two potential areas of difficulty: the very high level of expression, and that's going to mask the potential low-level problems relating to (the rest of) the content of the tables. The second set of problems, relating to all of the other diagnostics will come up again next time.

CS. Yes.

AL. What you'd expect from the diagnostics at this stage would be for the method to put everything in. It would say that you should represent in your.......

CS ...it will tend to set up almost a pre-classification scheme for the problems that I'll see. They would have to be included in the representations in order for me to find them.

AL Yes, that's right. What the method tries to do with these general purpose diagnostics is recruit what lay knowledge you might have, like hunches about what the problem might be ... so that this can be included in (the design of the simulation).

CS What I could do is run through these as well *(pointing to diagnostic table A1 - see Appendix A)*, and just check that they're...

AL Are they relevant?

CS No. Like chunking size is sort of irrelevant because they're going to be trained so well.

AL True, although that should come out here *(pointing to the column on user attributes).* It ought to be selected because it would ensure that your user subjects have to have the appropriate level of skill to be representative of your users. From the point of view of being an intelligent method .... I would argue that yes, you do need to include that, although that was apparently not obvious to you (from the procedures of the method).

CS No. So for 1.1 you'd say include it. ..... It depends what you mean by knowledge exactly, because it doesn' t necessarily mean the vocabulary, does it?

AL Here I've used knowledge in the sense of being everything you know, and would include the procedures .... basically all of that information held by the subject which enables them to use the target device as opposed to, say, radio or a manual keyboard. You used your intelligence as an evaluator and implicitly made sure that you selected representative subjects.

CS Yes, I suppose so. But this implies to me that you'd have to train..... like I know I'm not going to use real users, so the users I use won't have any knowledge of using other devices. Whereas this device might have incompatibilities, it is sort of beyond the scope of this, although one would hope that the training they get would mean that that wasn't the case, that it would reduce the incompatibility. But I think I'm using these wrongly.

AL Not necessarily. The thing is that I found it very difficult to specify how they should be used. The problem is proceduralizing a process which is very difficult to proceduralize. I suppose that what one might do (if there was no prescribed procedure) would be to skim down (the diagnostics) and sort of internalize all these bits and then you come out with a composite structure for the purpose you've got in mind. You would use the knowledge that you've got in your own head and you'd supplement it with this.

CS Yes. You wouldn't actually say "I'm using that one....". I'm a bit confused about this one (Diagnostic A.1.1). Because where it says here.....

AL Yes... in here it talks about different modes....it seems strange.

CS It says here .... alternative data entry devices or manual ... so that would be like a radio or something....

AL It looks to me as if there is a mistake there. That's one aspect of it, but I have a feeling that in earlier versions of the diagnostic manual there was another (diagnostic) which related to compatibility between modes of operation .... Anyway, this has definitely got problems because it refers to modes in one place and input techniques in others. The idea was that if you had an operational device you might have a device which is primarily operated using speech, but there might be backup modes, like manual modes, which you'd use if your recognizer system failed.

CS So there might be incompatibilities between them ...

AL So you might have speech interface which fitted onto an existing system like BATES which, if speech failed you could revert to manual operation. That's what was behind it, although it's not clearly expressed there.

CS So that seems to be inapplicable in my scenario, but then my scenario is supposed to be determined by ....

AL I suggest we go on from that one: it actually seems to be incorrectly expressed.

CS Lets go on to number two (Diagnostic A.1.2).... well, it's different from ordinary language and it's different from the radio. The vocabulary they've chosen is the vocabulary that is used, and syntax is OK for the people, like they don't mind having to use constrained syntax, but it isn't necessarily what they do on the radio.

AL That would justify its being considered, from my point of view, because it is quite clearly different and it's not going to the same as they'd use off-line. So 1.2 is relevant; 1.3 .... do you understand what that means?

CS I assumed it meant .. like the continuously scaled bit I thought it had to do with trying to express values that weren't discrete numbers, like a meter reading that was continuously changing.

AL Well the phrase came from an academic paper so perhaps it's no wonder that it's slightly obscure! "Multi-dimensional" is like...... spatial information is multi-dimensional.... it could also relate to perhaps colour or sound, which are also described on several dimensions: like colour has got intensity, hue..... "Continuously scaled" is more obvious, and again spatial information is continuously scaled, but in the case of grid reference a numeric (discrete) scale is placed upon it. If you've got a CAD system it would be very difficult to have an all speech interface to interact with it, without a mouse or something like that. Because to control the cursor you'd be saying "up a bit, up a bit...down a bit.." Its basically a continuously scaled multi-dimensional....

CS Yes, and you don't have those things in your mind .... there are grid references that you could use, but it's not natural: you have to do a conversion.

AL Yes, and of course in this case they're used to doing that: that's the task. This is a really tricky one because it doesn't very well distinguish tasks which have been designed in order to make these continuous things discontinuous.

CS ..... the task here is making the observation and specifying where something is, but the device does not support that part of the task. The device supports message transmission. It isn't a system for helping you work out grid references.

AL But in an ideal world you don't want a speech interface for this machine, you want a graphical (map) display which you can designate the point on and then send it...... But for the purposes of this study we're not talking about that.

CS We're not talking about alternatives are we?

AL No we're not.

CS Only within a specified range, like saying there's a slightly different vocabulary or dialogue sequences. We're not talking about different kinds of input device.

AL So I think this has identified some problems with the assumptions of the method. Because if you follow the instructions exactly and try to use (the diagnostic manual), you're in danger of making the thing excessively complicated by bringing in kitchen sinks.......

CS I think it's justified leaving this one out because I don't think that the device in its current embodiment, which is what we're testing, aims to support that part of the task which does have the multi dimensional information stuff in it. Because that part of the task is already proceduralized....

AL ..... for the purpose of making it discontinuous...

CS .... and although you could consider in a far reaching study how best to support that task, that isn't the issue here.

AL I agree absolutely. This is what I was saying when I was talking about "mental filtering". This is what you do ....

CS If you had a (completely naive evaluator) ......

AL ..... then you might need to be pointed in the right direction(?)

CS I would just say that the device operates on a subset of the task and doesn't support the whole task.

AL Although the philosophy of the method is that it tries to make you consider the whole task. This is where you're being intelligent...

CS I could always talk about that analytically. Like it gives support for getting the message through, but it still requires the people to do some things which might not be "natural" to them, like doing grid references. However, that's so ingrained .... that it's second nature.

AL Yes. To them it's not multi-dimensional and continuously scaled, it's digital information. That in itself could be justification. They can look at a map and think of it in terms of grid references.

CS The method assumes you address the whole task.

AL ... and it assumes to be neutral with respect to users and knowledge... like that everyone would view maps the same way.

CS ... but then I think you have to put a transform on there, like your users are not people in the raw, they are used to learning particular vocabularies and they have to change them every now and again for security reasons.........

AL So the rules are different for army users as opposed to, say, office workers? Both the kind of knowledge and its stability, and the way they look at the world.....

CS ... yes, for them having to learn a whole new vocabulary is not that unusual. They would just train and train and train until it was just off pat..... ....So we'll leave out 1.3.

AL 1.4... this relates to the chunk size, and this is the one that was used before......... If one assumes that they thought in terms of 6 and 8 digit grid references chunked into groups of 3 or 4, then you asked them to enter individual digits, strictly speaking that is relevant ....

CS But then.....

AL This is a training issue...

CS Yes.... over the radio, I don't know how they do it. But we're not assessing the suitability of this speech technology.

AL Yes... if we suspected there was a problem with that, we would say "yes, let's go for it". We're seem to be trying to be use these diagnostics in a way which is inappropriate. Because we're not doing that sort of study.

CS Basically, we've got to use this size of chunk, and what we can do is monitor whether people have problems with that. But then our people aren't necessarily that used to grid references anyway...

AL ....and anyway, they are going to have it written down: it's not as if they will be going straight from the map.

CS We can't use this (diagnostic) to guide our experimental design, because our chunk size is already determined.

AL The same problem is going to exist with the behavioural table, because again they assume that there is a specific set of problems. Maybe we should ignore them for a minute. Given that these are potentially inappropriate, and certainly difficult, if you weren't using the method, how would you go about designing (a task simulation)?

CS Having a description of the task, I would try to mimic the task as it's done. So they look at something, they compose a message and they send it (at a sort of high level). .... For the observing bit, they have got to have something that represents a terrain, and its got to have points on it that they recognise to be targets with information about them. That's in the task representation. It's been stressed that that's what they do, they look at things and pass information about them.

AL It could be argued that the looking aspect of the task is not going to interact at all with sending at all, in which case just have a list of information like we had in the first study. But some knowledge that you've got tells you that it is necessary to have the spatial aspect of the task.

CS I feel that doing that part of the task makes you feel "semantically more involved" with the messages. So it actually means something. So you are more likely to know whether you've done it right or not. Whereas if you're just reading it, you go into sort of "automatic mode". So that's what's guiding it there.

AL So it gives it sort of cognitive validity?

CS I think so, yes. I know I'm going to be picking people who are really inexperienced, so they have got to have something that pulls them into an appropriate mind set, that would be similar to what the real users would do, in order that they have the right sort of background to doing this message. Because the words are pretty specialised........they're going to find that really weird, unless they've got some kind of view about what they're supposed to be doing. .... I don't think it necessarily has an effect on their performance as such, apart from possibly they would be more aware of whether they've done the right thing or not. But I think its quite important motivationally.

AL Your implicit strategy seems to be inclusiveness. To capture the semantics of the task - the overall thing - rather than at this stage selecting bits of it.

CS Yes, I think so.

AL What that sounds to me to be sensible and that goes with the spirit of the general purpose diagnostics.

CS I think its very different sitting in a lab doing this to doing it in the field anyway. So we've forfeited so much validity that we ought to include....... otherwise people will just say that the subjects were just reading into a microphone.

AL As you say, that's not going to be very representative.... What about users?

CS We can't get real ones, so we ought to get people who have got the right sort of attributes, and then train them in the aspects of the task and/or design the task so as to make the skills they don't have unnecessary. The people who do this task normally are highly trained in recognition and observation and calculating grid references, and they're also highly trained in the use of the device and the procedures of the task. So I think what we have to do is ..... that would be an argument for having some degree of observation where they don't have to actually recognize: they are told what they are - the targets. Because it takes a (substantial army) training programme to be able to recognize targets to the extent that these people can.

AL So you assume that when one is well trained, one is able to access the relevant information to make an appropriate response with very little effort?

CS Yes, and in very little time....

AL ..so if you can match that facility qualitatively, at least, then you've got an appropriate sort of representation (for the laboratory simulation)?

CS That's what I'm assuming, because I have been told that when they make these observations they can just look, and, in like a split second they know exactly what it is and they never make a mistake.

AL So it's like you looking at your desk and saying "that thing is a pen".

CS Yes

AL So it's like representing straightforward object recognition, rather than a process of analysis of what they're looking at and a sort of effortful process of trying to work out what it is.

CS No, I don't think they go through that. They are trained to such an extent...... At least, this is what we are told.

AL I can see what your rationale for doing that is....

CS ... yes, that's one aspect. Another thing is that I think we would have to train people, because if they are required to give grid references they would have to have some training in doing that, so they are happy in doing that....

AL So that's one aspect of the task in which our subjects may not be representative.

CS They will probably have some knowledge of what grid references and basically how you do it, but they won't have a straightforward ability to divide a (grid) square into ten bits and calling it so and so..... being that clear about it. I think they'll have to be trained on that...at least to a level where they're happy doing it. It doesn't have to be accurate though.

AL That sounds logical.. and there's no way you can represent calculating grid references in the (lab) task in the same you that you did recognition? Recognition, you could just, like, give it to them in writing: you couldn't do that with the grid reference?

CS No, because the people in the field do actually work them out.

AL So it involves mental manipulations....

CS ..... yes, although it would not be that hard for them. Here we have to trade off a bit of accuracy, like we've got to train people so they think they're doing it right, but in fact they don't have to be 100% accurate, whereas the soldiers do. So for someone who has just been trained to be 80% accurate might be the same as a trained soldier being 100% accurate. It doesn't really matter to us whether they are accurate.. they've got to be trying......

AL .....doing the appropriate manipulations?

CS Yes.... for a soldier it's catastrophic if they do it wrong........

AL ....What is coming out is that what you are trying to do seems like what I was trying to proceduralize! So I think the appropriate way to go would be for you to specify the experiment according to your intuitions about these elements. Maybe when you've done it we can sort of backwards engineer and see to what extent it's possible to express what you did in terms of the tables.

CS Yes. The thing is that my likely dependent variables map onto the general purpose diagnostic, like errors, time and subjective reports....

AL The independent variables are basically completely open.. it's basically the device ....

CS ......we're trying to find them in the experiment, basically.

AL I wonder to what extent, when you actually specify the simulation, they will actually be expressible at a lower level in terms of these things which were expressed at a high level in the general purpose diagnostics .... I suspect that they might be, but maybe we can look at that when you've done it.

CS Basically, what I'm going to go for is to give the subjects appropriate training and then, for the task, at the low level....... I've spoken to Prof. King about this and he says I should have good sampling of the words that are (implemented on the device), even though they are not all used with the same frequency. I should act as if they are ... and have a random sample. So although you might almost never get a "strength" with three numbers in it, I must have that.......like some with one, some with two, some with three. And with grid references, sometimes I should repeat the number......and I should make sure that all numbers are used.

AL This, presumably, is to test the performance of the device over as many as possible message structures as we possibly can....

CS ....yes, to test the performance of the device in both its recognition and in its support over all the different types of message. Because it might be that recognition errors occur more in numbers than in anything else, maybe those errors would be more problematic than others.

AL I don't know if you noticed, but that point did crop up in the earlier study. A general problem with all these (recognizer performance specifications) is that recognition accuracy is expressed in a global sense, like its 90%, or something like that. But if it so happens that one of the (words) that you regularly use keeps coming up over and over again, the performance effectively drops very much lower.......Unless you actually pre-specify in advance exactly what is going to be said, and work the frequencies out on that basis ..... (it's going to be unrepresentative).

OK. From the point of view of configuring the diagnostic manual, we've got a big problem: at least in part because of the type of evaluation that it is, and if any are going to be applicable, it's got to be the general purpose ones.

CS Yes, I think so.

Appendix F3.3: **Questionnaire responses: Task simulation method (Phase 1)**
Ratings on a scale of 0 - 5 (see Table 11.3)

	Prel. task desc	Task data	Exp. task desc.	Fut. task desc.	Fut. task model	Task sim. spec.
Is rep. necessary?	y	n[1]	y	y	y	y
Was the rep. developed?	y	-	y	y	y	y[2]
Implicit or explicit	exp	-	exp	exp	exp[3]	imp
Diff. from SIAM prescription?	n	-	n	n	n	n[4]
Does the rep. - include relevant classes of info?	y	-	y	y	y	y
- show correct level of description?	y	-	y	y	y	y
- characterize what it is supposed to?	y	-	y	y	y	y
Helped understanding?	3	-	4	4	2[5]	3[6]
Helped planning?	4	-	3	4	2[7]	3[8]
Helped communication?	y4[9]	-	y4[10]	y3[11]	y3[12]	n
Enabled next procedure?	y	-	y	y	n[13]	y
Mental effort to develop rep.?	5[14]	-	5[15]	2[16]	2[17]	2[18]
Support from SIAM?	-	-	-[19]	-[20]	-[21]	-[22]
All steps performed?	-	-	n[23]	n[24]	n[25]	n[26]
Steps performed differently?	-	-	y[27]	y[28]	y[29]	y[30]
Mistakes?	-	-	n	y[31]	-	-

[1] Representation not developed, because basic data on task had been collected in a previous study and used to develop task representations described in Chapter 7.

[2] Did not use diagnostics; no explicit interaction model. See Appendix F3.4 for rationale for simulation design.

[3] Only made explicit *after* simulation had been specified and implemented. *This was attributable to the failure to configure diagnostics successfully (see Appendix F3.2)*

[4] Task entities and their attributes were not expressed as a list

[5] Because task model only made explicit after simulation had been specified. May have facilitated consolidation of thinking about the task simulation

[6] Question does not seem appropriate: the representation is a *prerequisite* for subsequent progress (i.e. implementing simulation)

[7] Because task model only made explicit after simulation had been specified.

[8] Question does not seem appropriate: the representation is a *prerequisite* for subsequent progress (i.e. implementing simulation)

[9] Discussed with RMCS

[10] Communicated to RMCS and research assistant

[11] Communicated to RMCS and research assistant

[12] Communicated to RMCS and research assistant. Use in subsequent documentation of evaluation

[13] Because task model only made explicit after simulation had been specified.

[14] Modification of an existing description

[15] Because development required only modification to an existing representation

[16] (a) Difficulty specifying computer routines at an appropriate level of description

(b) Form of representation (hierarchy) incompatible with that favoured by the subject (simple text dialogue indicating responses of computer to utterances of user)

[17] It was difficult to intersect device model and task model

[18] (a) It is critical to the evaluation that this representation is correct; (b) development involves iterative refinement; (c) procedures of method offer little support (creative/generative activity)

[19] Because development required only modification to an existing representation

[20] Because development required only modification to an existing representation

[21] Not applicable, (a) because development required only modification to an existing representation; and (b) because of the failure to configure the diagnostics (see Appendix F3.2)

[22] Not applicable, because subject did not follow top-down development procedure assumed by the method.

[23] Step 1 omitted because basic description already provided

[24] Subject implicitly followed procedure of method, but deviated in that she modified the previously-produced task description (hence, less generative than is implied by the method)

[25] (a) because development required only modification to an existing representation; and (b) because of the failure to configure the diagnostics (see Appendix F3.2)

[26] Did not follow top-down development procedure assumed by the method. Steps 3 and 4 were followed implicitly

[27] Because development required only modification to an existing representation

[28] Subject implicitly followed procedure of method, but deviated in that she modified the previously-produced task description (hence, less generative than is implied by the method)

[29] Subject prefers bottom-up, rather than top-down approach to specification. Model specified iteratively by comparing her implemented task simulation with the future task description

[30] Did not use diagnostics; no explicit interaction model. See Appendix F3.4 for rationale for simulation design.

[31] Subject made a minor mistake in the representation of action C3.2.2, which required additional decomposition (addition of extra node for "retrain"). Mistake caused by subject forgetting details of device behaviour. *Possibly caused by the hierarchical representation being incompatible with subject's preferred representation of dialogue?*

Appendix F3.4: SUMMARY OF INTERVIEW WITH CS, 19TH. MARCH, 1990: RATIONALE FOR DESIGN OF THE TASK SIMULATION

This interview was necessary to determine CS's rationale for the design of a simulation of the task of the observer. Because she had been unable to configure the diagnostic tables, she had designed a simulation on the basis of implicit criteria. The objective of the interview was to determine what these criteria were, and hence, if possible, to ascertain the form of her implicit model of device user interaction.

AL began by summarising the objectives. The interview was to be carried out in the context of the implemented simulation. It was recognized that pragmatic experimental factors (i.e. non-device-design-related) might have had a bearing on its design as well.

How do you describe the task to the subjects?

- involves training subjects in what task involves as well as use of device
- given high level description of task in general terms (i.e. military observation). Not details.
- told they will have function of sending messages using structured message formats
- information given to them is constrained by security factors, but also only enough is given to them to provide a context for their activities
- information transmission element of the task is emphasised

Specific description of activities:

- how to use device
- how to calculate grid references
- message format (USs need to know what they are trying to convey)
- how to operate the device

Selective description of target task at task level: detailed description at comms level and at I/O level. All this happens before they are exposed to the representation of the task (i.e. before they see the pictorial battlefield representation, the device they are to use, and their task aids: map showing terrain in battlefield representation, "roamer", crib sheet with names of all the fields, note taking facilities).

Mapping between real and simulated tasks:

- direct correspondence with real task with respect to the map
- the aid to support grid reference interpolation is intended to help the US's in the process of compiling grid refs.; this is "automatic" for real users, and CS was concerned that it should be as "automatic as possible" for USs
- note taking material: representative of the target aid at a high level: direct copy of the elements of the clipboard used by real users, but differences at I/O level, e.g. have separate sheets rather than using both sides of clipboard, use pencil rather than chinagraph etc.. CS did not feel such I/O level differences were relevant. When pressed on fact that these I/O level differences might impact suitability of speech, she explained that writing etc. was not done concurrently with data entry, so there was no reason to expect interaction between the activities; hence, representativeness at the I/O level was not at issue. It was also noted that the purpose of the study was not to study the suitability of speech *per se*: speech was taken as a "given" and the intention was to evaluate the device. CS stressed that representativeness at the I/O level was not feasible anyway, because the final embodiment of the device was unknown, and the brief was such as to render inappropriate representation of environmental features.

Discussion of the pictorial battlefield representation

(Some discussion of fact that a training representation of the b/f was developed as well as the main task).

What attributes of the battlefield were represented in the task simulation?
- spatial rep. of scenery. Same level of detail as sample of sketches made of battlefield by actual users
- perspective
- landmarks to assist relative location (which would have an ordnance survey). All targets were actually located close to landmarks (which, incidentally would have been likely places to find targets)

Stylized representation of battlefield. Assumed that real users would attend to the places that targets would be likely to appear. CS was then pressed for her criteria: point made that a representation stylized in this way would be inappropriate if the <u>detection</u> of targets was an important factor determining task performance. Present representation seemed to be based upon assumption that battlefield was source of high level information about targets, rather then low level. CS assumed that end users were expert at search and identification: the experimental task simulation took up from the point at which targets had been seen and identified. This was appropriate because the device under evaluation did not support the identification part of the end user's task.

The process of identification was "automated" for USs by enabling them to operate a "simulated binocular", which gave them a textual description of the target (implemented via mouse key). This could be entered directly into the machine. The form of representation was partly to simulate the fact that identification would be automatic to target users, and also to ensure that the data entered was "balanced" such that a fully representative range of input functions were demanded of the users. (RMCS had requested this). Point made that this was not representative at the I/O level e.g. manual actions were different, "magnified" information remained visible until users clicked the mouse again etc. CS was not concerned with I/O level interactions such as speaking and looking: the I/O level of the device had not been determined yet anyway and current prototype was unrepresentative at this level.

Overall, high fidelity at comms level: variable fidelity at I/O level. Variable at task level.

CS emphasised that there was some requirement to undertake "observation" in the simulated task (cf just reading data from a list), this was because collection of intelligence information was stressed as being an important part of the task. It was also felt that the messages would not be as meaningful if subjects just read them out. i.e. cognitive representativeness and motivation factors. The task as specified would capture certain <u>cognitive</u> interactions between the observation and data entry elements of the task, but not interactions at an I/O level.

What about the dynamics of the task? e.g. representativeness with respect to pressure - subjects were motivated to perform task rapidly, although they were instructed to ensure that the messages were accurate
- there was evidence that they were using strategies to make performance quicker: USs were given freedom to use the strategy they thought best to complete the task effectively
- some information on the temporal constraints of the real task was obtained from RMCS, (e.g. very low rate of message transmission). However, constraints on the speed of the current task were mainly imposed by the limitations of the existing technology. Nevertheless, the rate of activity was typically very slow: rate depends on what's around. The assumption seems to be that the battlefield is static as far as target engagement is concerned: messages are worked out "at leisure"

Did the method make any contribution to the design of the task simulation? - done implicitly on basis of future task description, because the diagnostics had not been configured
- some of the steps of the procedures might correspond with what CS did (Protocol sheets filled in at this point).

Appendix F3.5: **Questionnaire responses: Device simulation method (Phase 1)**
Ratings on a scale of 0 - 5 (see Table 11.4)

	Elab. dev. desc.	Dev. sim. spec.	Dev. sim.	Dev. sim. perf. data	Anal. of dev. sim. behav
Is rep. necessary?	y	-	-	-	-
Was the rep. developed?	-[1]	-	-	-	-
Implicit or explicit	exp	-	-	-	-
Diff. from SIAM prescription?	y	-	-	-	-
Does the rep. - include relevant classes of info?	-	-	-	-	-
- show correct level of description?	-	-	-	-	-
- characterize what it is supposed to?	-	-	-	-	-
Helped understanding?	-	-	-	-	-
Helped planning?	-	-	-	-	-
Helped communication?	-	-	-	-	-
Enabled next procedure?	-	-	-	-	-
Mental effort to develop rep.?	-	-	-	-	-
Support from SIAM?	-	-	-	-	-
All steps performed?	-	-	-	-	-
Steps performed differently?	-	-	-	-	-
Mistakes?	-	-	-	-	-

[1] (a) The device was available for evaluation; (b) The dialogue had been explicitly specified by the developers

Appendix F3.6: Questionnaire responses: User simulation method (Phase 1)

Ratings on a scale of 0 - 5 (see Table 11.5)

	Desc. of task k.	User subj. model	User subj. dev. prog.	User sim.y
Is rep. necessary?	y	y	y	y
Was the rep. developed?	y	y	y	y
Implicit or explicit	exp	exp	exp	exp
Diff. from SIAM prescription?	n	n	n	n
Does the rep. - include relevant classes of info?	y	y	y	y
- show correct level of description?	y[1]	y	y	y
- characterize what it is supposed to?	y	y	y	y
Helped understanding?	3	-[2]	-[3]	-[4]
Helped planning?	4	4	3	-[5]
Helped communication?	y3[6]	y3[7]	n	n
Enabled next procedure?	y	y	y	y
Mental effort to develop rep.?	3	4	4	2[8]
Support from SIAM?	3	3	2.5[9]	2[10]
All steps performed?	y	n[11]	y	y[12]
Steps performed differently?	y[13]	y	y[14]	n
Mistakes?	n	n	n	n

[1] As far as is known: no further details available

[2] Question does not appear appropriate

[3] Question does not appear appropriate; may have served to consolidate thinking

[4] Did not really contribute to understanding: subjects behaved as expected

[5] Representation was a *requirement* for subsequent activity, rather than for planning

[6] Communicated to research assistant

[7] Communicated to RMCS, department head, research assistant, subjects

[8] Effort required in recruiting subjects, encouraging subjects etc. (not just mental effort)

[9] Just served to confirm what the subject would have done anyway

[10] Effort required in recruiting subjects, encouraging subjccts etc. (not just mental effort). These processes are not proceduralized in SIAM

[11] Step 1 not performed because of problem configuring diagnostics (see Appendix F3.2); critical attributes were specified on the basis of subject's own implicit model of device-user interaction

[12] Just served to confirm what the subject would have done anyway

[13] There was no access to real users; some sub-steps were performed (implicitly) by RMCS.

[14] Ordering was different; some steps had been performed much earlier (e.g. in preliminary discussions with RMCS)

Appendix F4.1: Questionnaire responses: Usability evaluation method (Phase 2)

Ratings on a scale of 0 - 5 (see Table 11.6)

	Prel. prob spec.	Prel. syst. spec.	Diag. table conf.	Soln. strat.	Expt. con-text	Data	Anal. of int'n	Feas. rept.
Is rep. necessary?	y	n[1]	y	y	y	y	y	-[2]
Was the rep. developed?	y	-	y	y	y	y	y	-
Implicit or explicit	exp	-	exp	**imp**[3]	**exp**[4]	.exp	exp	-
Diff. from SIAM prescription?	n	-	n	n	n	n	y[5]	-
Does the rep. - include relevant classes of info?	y	-	**n**[6]	**y**	y	y	y	-
- show correct level of description?	y	-	**y**	**y**	y	y	y	-
- characterize what it is supposed to?	y	-	?[7]	**y**	y	y	y[8]	-
Helped understanding?	2[9]	-	0[10]	**3**	-[11]	5	3	-
Helped planning?	3[12]	-	1[13]	4	3[14]	4	3	-
Helped communication?	n	-	n	y[15]	n	y4[16]	y4[17]	-
Enabled next procedure?	y	-	n[18]	y	y	y	y	-
Mental effort to develop rep.?	4	-	2[19]	3	2[20]	2[21]	3	-
Support from SIAM?	3	-	1[22]	**2.5**[23]	2[24]	2[25]	1[26]	-
All steps performed?	y	-	y[27]	y	y	y	**n**[28]	-
Steps performed differently?	y[29]	-	y[30]	y[31]	y[32]	y[33]	y[34]	-
Mistakes?	n	-	?[35]	n	y[36]	n	n	-

[1] Specification already established during phase 1

[2] *Format required for final report differed from that assumed by the method*

[3] Because based on the strategy developed in phase 1

[4] Specification of the context was partly implicit; implementation was explicit (obviously)

[5] Representation not based on use of diagnostics; based upon subject's implicit model of interaction (and hence own critical parameters)

[6] Could not find diagnostic relevant to issue of whether binary device functions should be implemented using speech or manual (pressel switch) action. *General purpose behavioural diagnostic was applicable, but this was not evident to the subject.*

[7] Uncertain, given confusion over the applicability of the general purpose diagnostic

[8] But concern that small number of subjects mean that conclusions can only be indicative

[9] Subject observes that it is necessary to understand the problem *before* the problem specification can be developed

[10] (a) Some diagnostics seem to be specific to certain sorts of systems (e.g. representational diagnostic 1.1); difficult to tell whether they are applicable to the system under investigation

(b) difficulty mapping between subject's model of the problem and that assumed in design of diagnostics

(c) uncertain about what individual diagnostics are delivering (i.e. what it would mean to pick one diagnostic as opposed to others)

(d) some diagnostics are self-evident: they do not add anything to understanding

(e) diagnostics are difficult to assimilate in their existing form; subject feels that they should be helping her, but they fail

(f) "if you do have appropriate knowledge you do not need the diagnostics: if you do not, you need examples of their application in order to understand them

[11] Not clear that this question is relevant: the experimental context is intended to support the solution of the problem, rather than to facilitate understanding of it

[12] Only modest contribution to planning because of similarity to equivalent representation developed in phase 1

[13] See comments concerning contribution to understanding (above)

[14] More appropriate to say that representation will *enable* later activities rather than *help in planning* later activities

[15] Communicated informally to research assistant

[16] Discussed with all involved parties

[17] Communicated to RMCS and research assistant

[18] Will proceed on basis of own model of interaction, rather than the diagnostics

[19] See comments concerning contribution to understanding (above)

[20] Effort was required to ensure that all relevant determinants of system behaviour were included; accuracy was very important

[21] Effort was required in low-level analysis: error classification and interpretation

[22] See comments concerning contribution to understanding (above)

[23] Had used the procedure during phase 1; now used method as checklist

[24] Procedures served checklist function, but they did not contribute much to the most demanding aspects of the task *(i.e. those requiring judgement of assessor)*

[25] Procedures were expressed at too high a level to offer much support *(but method is not intended to support inductive inferential processes)*

[26] Because diagnostics were not used

[27] All steps were *attempted* but subject not happy with outcome

[28] Steps 2 and 3: not appropriate because did not use diagnostics

[29] Procedure simplified by work done in phase 1

[30] Step 2: Unhappy with use of decision tree, because felt that there was a risk of failing to recognize potential device-user incompatibilities (e.g. a *combination* of skills might be incompatible, while individually the skills might not present any incompatibilities)
Step 3: Subject could not select diagnostics by referring only to the columns specified by the method: there was insufficient information presented in individual columns to determine whether a diagnostic was relevant, so necessary to check the other columns for further information

[31] Did not use diagnostics directly: used implicit interaction model; *(however, outcome appeared to be compatible with diagnostics).* Subject noted a problem with level of description if diagnostic tables were used: e.g. user has to interpret how errors are likely to be manifested *(?)*

[32] Actions were actually performed without direct reference to the procedure; however, there was generally good correspondence with method

[33] No inferential statistical tests were applied

[34] Step 1 was performed such that results were evaluated against hypotheses *without* reference to diagnostics

[35] There may have been mistakes, so preferred to use own (implicit) interaction model

[36] A slight problem arose over the implementation of the device simulations (failure to implement auditory feedback). This failure could be attributed to inadequate piloting: it would not have arisen if a second pilot trial had been performed

Appendix F4.2: Questionnaire responses: Task simulation method (Phase 2)

Ratings on a scale of 0 - 5 (see Table 11.7)

	Prel. task desc	Task data	Exp. task desc.	Fut. task desc.	Fut. task model	Task sim. spec.
Is rep. necessary?	n[1]-	n[2]	n[3]	n[4]	y	-[5]
Was the rep. developed?	-	-	-	-	y	-
Implicit or explicit	-	-	-	-	exp[6]	-
Diff. from SIAM prescription?	-	-	-	-	n	-
Does the rep. - include relevant classes of info?	-	-	-	-	y	-
- show correct level of description?	-	-	-	-	y	-
- characterize what it is supposed to?	-	-	-	-	y	-
Helped understanding?	-	-	-	-	3	-
Helped planning?	-	-	-	-	2[7]	-
Helped communication?	-	-	-	-	y3[8]	-
Enabled next procedure?	-	-	-	-	-[9]	-
Mental effort to develop rep.?	-	-	-	-	3[10]	-
Support from SIAM?	-	-	-	-	2.5[11]	-
All steps performed?	-	-	-	-	n[12]	-
Steps performed differently?	-	-	-	-	y[13]	-
Mistakes?	-	-	-	-	n	-

[1] Same as in phase 1

[2] Same as in phase 1

[3] Same as in phase 1

[4] Similar to phase 1: all deviations from phase 1 embodied in future task model

[5] As in phase 1, with slight modifications to device operating actions

[6] The explicit model of the task was specified retrospectively on the basis of models of the different versions of the target device

[7] The explicit model of the task was specified retrospectively on the basis of models of the different versions of the target device

[8] To be included in documentation of work

[9] The explicit model of the task was specified retrospectively on the basis of models of the different versions of the target device

[10] Based on previous task model

[11] Modified previous model. Subject noted that the procedure requires that the device is represented at a low level; however, it recruits the preliminary system specification which is not normally detailed

[12] Did not utilize diagnostics to specify model. Based representation heavily on previous task, but with inclusion of new actions for device operation

[13] Did not utilize diagnostics to specify model. Based representation heavily on previous task, but with inclusion of new actions for device operation

Appendix F4.3: **Questionnaire responses: Device simulation method (Phase 2)**
Ratings on a scale of 0 - 5 (see Table 11.8)

	Elab. dev. desc.	Dev. sim. spec.	Dev. sim.	Dev. sim. perf. data	Anal. of dev. sim. behav
Is rep. necessary?	y	y	y	y	y
Was the rep. developed?	y	y	y[1]	y	y
Implicit or explicit	part exp	part exp	exp	**exp**[2]	imp
Diff. from SIAM prescription?	y[3]	y[4]	n	y[5]	y[6]
Does the rep. - include relevant classes of info?	y	y	y	y	y[7]
- show correct level of description?	y	y	y	y	y[8]
- characterize what it is supposed to?	y	y	y	y	y[9]
Helped understanding?	4	4	5	3[10]	-
Helped planning?	5	4	4	4(?)[11]	-
Helped communication?	y2[12]	y3[13]	n	n	-
Enabled next procedure?	y[14]	y	y	y	-
Mental effort to develop rep.?	2[15]	2[16]	2[17]	2[18]	-
Support from SIAM?	2[19]	2.5[20]	3[21]	2.5[22]	-
All steps performed?	y	y	y[23]	y[24]	-
Steps performed differently?	y[25]	y[26]	y[27]	y[28]	-
Mistakes?	n	n[29]	n[30]	n	-

[1] *Simulation was developed using partly iterative strategy rather than full specification in advance of implementation*

[2] But not formalized

[3] (a) Expressed less formally than is implied in the procedure, but complete;
(b) target device performance was only represented implicitly: assumed to be the same as the performance of the prototype in phase 1

[4] (a) Less formally expressed than implied by method; the communication device was specified minimally, as it offered little functionality other than as a simple communication channel;
(b) The system subject (SS) was familiar with the SS task, so there was no need for fully detailed instructions
(c) A system subject action hierarchy was only specified for reporting purposes (used textual representations of dialogue to support SS task)

[5] Expressed only informally

[6] Enhancement of simulation was achieved by "lay" diagnosis and prescription repeated until simulation performance judged by subject to be adequate; the communication device was simpler than is implied in the procedures

[7] Assumed to be adequate for the purpose

[8] Assumed to be adequate for the purpose

[9] Assumed to be adequate for the purpose

[10] No contribution to understanding of research problem; facilitated solution of simulation "problem"

[11] Supported iterative enhancement of simulation

[12] Communicated to research assistant (system subject) in discussions of device behaviour and performance. Implicit aspects of the description needed elaboration

[13] Communicated to system subject/research assistant

[14] The subsequent representation evolved from this one: it is not clear where one ended and the next began

[15] Specification of the device was a generative (design) activity. It was difficult to specify in detail

[16] Low level description of speech interaction is difficult (to enable system subject to operationalize)

[17] Simulation task was demanding

[18] Diagnosis of simulation performance inadequacies and prescription of enhancements to simulation were mentally demanding

[19] The main activity was design, in order to present options to the project team. *(SIAM assumes involvement of design specialists in this activity)*

[20] Only used procedures at a high level

[21] Procedures applied in a "free-form" manner. Subject noted that problems of system subject recruitment and training were minimal because of his previous experience and close involvement with the project in general

[22] Procedures could not be followed in full due to shortage of time

[23] Simulation was developed using partly iterative strategy rather than full specification in advance of implementation

[24] More informal than is implied in the procedures

[25] (a)The main activity was design, in order to present options to the project team. *(SIAM assumes involvement of design specialists in this activity);* (b) diagnostics were not used

[26] Some steps were performed perfunctorily (or at a high level) because of:
(a) shortage of time; (b) fact that subject and research assistant had gained experience in course of phase 1; and (c) scope of study was quite limited

[27] (a) Implementation of device simulation was simplified by system subject being experienced and being involved with testbed configuration;
(b) use of communication device was minimized (partly because of shortage of time, and also because of small scope of study)

[28] "Experiment" was not fully controlled (more like a sequence of informal evaluations)

[29] As the simulation was developed iteratively, any "mistakes" were used to specify enhancements to the simulation

[30] As the simulation was developed iteratively, any "mistakes" were used to specify enhancements to the simulation

Appendix F4.4: **Questionnaire responses: User simulation method (Phase 2)**
Ratings on a scale of 0 - 5 (see Table 11.5)

	Desc. of task k.	User subj. model	User subj. dev. prog.	User sim.y
Is rep. necessary?	n[1]	n[2]	y	n[3]
Was the rep. developed?	-	-	y	-
Implicit or explicit	-	-	imp	-
As prescribed?	-	-	y	-
Does the rep. - include relevant classes of info?	-	-	y	-
- show correct level of description?	-	-	y	-
- characterize what it is supposed to?	-	-	y	-
Helped understanding?	-	-	2.5[4]	-
Helped planning?	-	-	3	-
Helped communication?	-	-	y[5]	-
Enabled next procedure?	-	-	y	-
Mental effort to develop rep.?	-	-	3	-
Support from SIAM?	-	-	2.5[6]	-
All steps performed?	-	-	y	-
Steps performed differently?	-	-	n	-
Mistakes?	-	-	n	-

[1] Same as in phase 1

[2] Same as in phase 1

[3] Same as in phase 1

[4] Does not seem an appropriate question

[5] Communicated informally to research assistant and user subject candidates

[6] Used SIAM's procedures as a checklist (as previously)

Appendix F5: SUMMARY OF CONCLUDING INTERVIEW WITH CS: 4TH. MAY, 1990.

The purpose of this interview was to "debrief" the subject, to elicit an evaluation of the quality of the output of the assessment project, the effort involved in achieving it and the extent to which SIAM had a bearing on this.

The interview began by attempting to complete the standard questionnaire with respect to the project as a whole. This was only partially successful because some questions were not relevant, and there was substantial variation in the subject's view about different parts of the method. The questionnaire was finally only used as a basic structure.

Conformity with SIAM's representational structure. The experimenter expressed the view that the subject had conformed quite closely to the representations required by SIAM. This was agreed, although there were some parts of the method where the representations were implicit and/or different in structure from that prescribed. The main deviations from the method were with respect to the procedures (see later).

The subject expressed some concern that the method tended to be orientated towards speech aspects of task and that it may not have been ideal for getting a good representation of non-speech aspects (e.g. the diagnostics tend to be strongly orientated towards speech and the notation seems better for serial dialogues).

In view of the completeness of SIAM in other respects, the subject was also surprised that it there was not an explicit specification of the experimental context (with all the components integrated) in advance of implementation.

Product quality. In general, the subject felt satisfied with the outcome of the assessment, although she felt that she could probably achieved the same quality of output without the method. Concerning her criteria for quality, she tended to base her standard on the requirements of the customer (i.e. the device designer) rather than on the standards of London HCI Centre (about which she had a limited picture). Her concern was with providing the customer with what they needed to know, so she judged her own output on the extent to which it would be useful to them. She felt that her opinion of the adequacy of the output had been confirmed because, in a meeting with RSRE , it became that she had identified the sorts of problems with the device which had been concerning them.

Quality of contribution to project organization. The project support functions were not used as much as they might have been because of the small scale of the project, and because the device developers were very familiar with the output of the previous study. The requirement to "manage complexity" would be greater in a larger project involving several researchers doing different aspects of the task and a management familiar with the method. It is questionable, however, whether user interface evaluations would ever be this large.

The use of the method in planning tended to be at a technical level, rather than a project management level, i.e. representations enabled actions contributing to the development of subsequent representations. There was little requirement for internal scheduling, because external factors (such as the provision of information by the device developers, and external deadlines) were the major determinants of when things were started and finished. Because everything was done by just one person, there was no need to integrate the activities of the project. The method would have a potentially greater scheduling role in a larger project.

It was only necessary to communicate to other people on the project on a few occasions. The representations were helpful in communicating with the device developers and with the experimenter when he was acting as her assistant; however, the latter communication could not be taken as particularly representative. The value of the method in communication depends on the immersion of the whole project team in its use: again, advantages would tend to accrue most in larger projects. However, the method can make a useful contribution where device development occurs over a long period of time. Explicit documentation would then be extremely valuable. The present project recruited output from a previous project and the representations produced this time would be useful if the device developers wanted to progress the work further. However, because the present study was carried out over quite a short space of time (approximately two months), the subject was able to hold most of the information immediately relevant to the assessment in her mind.

Assessor costs. The subject found it difficult to decide whether using the method had reduced the effort needed to perform the task. The requirement for accuracy/detail imposed by the method was greater than she would probably have used if doing the task independently, and she did not know how this would have influenced the final conclusion. The method makes it easier in that it enables the assessor to ensure that everything has been done, but performing some of the activities in terms of the method was difficult.

The method also made the subject perform some procedures which she would otherwise not have done, and, again it was difficult to tell how this influenced the outcome. For example, the subject would have represented the task implicitly and probably not in so much detail. When asked whether she felt the documentation excessive, she said this was partly true, particularly in the context of the present study, because much of the documentation already existed. It was somewhat spurious to make all the minor modifications to the intermediate representations, and she would have preferred to have just modified the final ones.

The subject concluded that the method does reduce costs to the user by providing a "crutch": it enables the assessor to ensure everything has been covered, and it is helpful to have all the elements of the simulation laid out when assessing the final form of the study. However, costs were incurred because the subject sometimes had problems mapping the method onto elements of the problem. These costs were associated with the form of the notation and with the need to acquire the terminology of the method. The subject did not think these would be a problem given a project with a longer timescale.

Assessor behaviour. It was agreed that the subject had adhered more closely to the procedures in Phase 1 than in Phase 2. This was attributed partly to familiarity with the method due to her using it in Phase 1, partly to the feasibility of just making modifications to the previous representations, and partly to shortage of time towards the end of the study. It was further agreed that the main problems with the procedures related to their inappropriateness for this particular type of study and to deviations enforced by the failure to configure the diagnostics successfully. The difficulties with the diagnostics were due to their being inappropriate for a study of the sort conducted in Phase 1, and, more generally to their being expressed in a sub-optimal format.

Assessor errors. The incidence of deviations from the intended outcome of procedures was generally low, although the method of data capture was such that only large mistakes would tend to be recorded. In general, the output was apparently adequate for the purpose, and the subject avoided major errors by checking important outputs with the device developers. However, it was not possible to tell whether their were errors in informational content of the representations: she had to rely for this on other people.

Conclusions

The subject stated that, at a structural level, SIAM was facilitative; but there were specific problems at a procedural level. This was partly due to the present study being a user interface development study, rather than a feasibility study for procurement. The contribution of the procedures was highly variable: they tended to be useful in the specification phases but were not much help in bringing all the elements together. They were also not much help unless everything else works properly (e.g. the diagnostics).

Suggestions for improvements

The subject was asked if any aspects of the method required substantial modification. The following were identified:

- diagnostic tables
- expression of the device simulation method
- determining appropriate levels of description in representations.

Concerning device simulation, the subject had difficulty following procedures associated with the evaluation of the SS-CD system. She also expressed reservations as to the feasibility of performing full system subject assessment studies within most user interface evaluations. There is usually a shortage of time, and typically only one system subject would be used, so selection of a sample would not be an issue. Such an approach would only be possible within a large experimental programme: she felt that the iterative approach she used was more appropriate for the kind of study she had been involved in.

Her views on the advantages and disadvantages of methods had not changed as a result of her experiences. (There was subsequently a discussion on the extent to which details of the use of the method should be included in her reporting of the present study).