User-Centred Evaluation of Complex IT&T within an Operational Environment

A J May

HUSAT Research Institute, Loughborough University
Elms Grove, Loughborough LE11 1RG UK

The aim of this paper is to provide guidance for the user-centred evaluation of new and complex Information Technology and Telecommunication systems within an industrial operational environment. This paper draws upon practical and methodological experience gained over a period of nearly a decade in order to:

- *Discuss the efficacy or otherwise of some typical Human Factors evaluation techniques*
- *Describe the common constraints that are likely to arise during an industrially-based evaluation process*
- *Provide guidance for overcoming these barriers to a successful and informative evaluation process*

Specifically, this paper aims to identify some of the key requirements for undertaking a successful evaluation process within a challenging environment and presents recommendations to maximise the efficacy of the evaluation effort.

Keywords: Evaluation, usability, user-centred design, human factors techniques, methodologies

1. INTRODUCTION

1.1 Evaluation within a complex, industrial, multi-user environment

User-centred evaluation is recognised as a vital constituent of any systems development process and according to Dix, Finlay, Abowd and Beale (1993), has three main goals: "to assess the extent of the system functionality, to assess the effect of the interface on the user, and to identify any specific problems with the system". A wider view, which is that taken in this paper, is that 'evaluation' is the assessment of the actual properties of a system versus the desired properties, including the impact on the user, the organisation and the business. Since it is the staff within a business that enable the achievement of business objectives, Information Technology and Telecommunication (IT&T) must be designed to enable staff to work efficiently and effectively within business processes. User-centred design and evaluation of IT&T is therefore essential.

IT&T used within industrial or commercial environments are usually complex, multi-user systems, and there are many problems in attempting to evaluate them (Grudin, 1988; Ross, Ramage and Rogers, 1995; Thomas, 1996). Typical problems listed are:

Manuscript received: ???
Associate Editor: Judy Hammond

1. A multiplicity of academic viewpoints exist, which results in widely varying approaches to evaluation and which makes cross comparisons difficult.
2. Most established evaluation methodologies and techniques have been developed and validated for relatively simple, single-user systems.
3. There will be multiple industrial stakeholders, each with their own agendas and evaluation criteria; these criteria may have widely differing requirements for data.
4. In many cases, methodologies are lacking or unvalidated, or benchmarking metrics do not exist (e.g. for assessing the efficacy of the business process).
5. Standard experimental activities, such as the use of control groups or pre and post comparisons are not necessarily possible due to the uncontrolled, dynamic nature of an operational environment.
6. The evaluation environment is relatively uncontrolled due to concurrent commercial activities and inconcisely bounded – any new IT&T system is likely to interface with legacy systems and data.
7. Business-related evaluation criteria applied to one environment are not readily portable to other domains (e.g. 'safety and expeditiousness' as may be applied to Air Traffic Control are quite different to 'quality, cost and time' that may be used within Automotive Manufacturing).
8. As well as direct measures (e.g. usability), some of the most important impacts of new IT&T are less quantifiable, (e.g. the organisational impact in terms of required changes in roles and responsibilities).

As well as (or probably partly because of) the above problems, there is also a dearth of case studies of realistic, situated user-centred evaluation of IT&T within complex industrial organisations. Examples of these types of studies are provided by D'Souza and Greenstein (1997); Pratico (1997); Baird, Moore and Jagodzinski (2000); May and Carter (2000).

'Evaluation' can take many forms: it can be largely formative/prescriptive, to guide system development, or summative/descriptive where the emphasis is the measurement of the 'as is' and the 'to be' and the comparison between current and future working. Systems, users, tasks, data and the environment can all be real or simulated (or any combination of both), evaluation can occur over short or long time scales, and data collection can emphasise objective or subjective measures, and quantitative or qualitative data types.

Space precludes a more detailed discussion of evaluation, in terms of: (1) philosophy, perspectives and methodological approaches; (2) individual techniques, metrics and tools; and (3) its role within an integrated IT systems design process. However, a number of standards are in existence that relate to the capability of the organisation, the process or product quality, and quality in use, and hence provide some direct or indirect support for evaluation activities. Relevant standards are: ISO 9241-11 Guidance on Usability, (1998); ISO 14598-1 Information Technology – Evaluation of Software Products – General Guide, (1998); ISO 13407 Human-centred design processes for interactive systems, (1999); ISO/IEC FDIS 9126 Software Engineering – Product Quality, (2000); and ISO TR 18529 Ergonomics of human-system interaction – Human-centred lifecycle process descriptions, (2000).

In addition to the above standards, the research literature on user-centred evaluation is varied, and provides *some* support for evaluation within operational environments. See, for example: data collection strategies (Rojek and Kanerva, 1994); data analysis and reporting (Nayak, Mrazek and Smith, 1995); evaluating specific types of interaction (e.g. Grissom and Perlman, 1995); methods for quantifying HCI (Bevan and Macleod, 1994; Macleod, Bowden, Bevan and Curson, 1997); processes for assessing (and improving) human-centred processes (Earthy, Sherwood-Jones and Weston, 1999); application of ethnographic approaches (Jagodzinski, Reid, Culverhouse *et al.*

2000) and tools for assessing conformance with aspects of ISO 9241 (Oppermann and Reiterer, 1997; Gediga, Hamborg and Duntsch, 1999).

Therefore, there are a wide range of differing evaluation perspectives and approaches that can be taken (as highlighted by Carney and Wallnau, 1998). Some research studies have specifically addressed the efficacy of specific user-centred evaluation approaches, see Henderson, Podd, Smith and Varela-Alvarez (1995) for an interesting comparison between four different user-centred evaluation methods. However, *there is still* a lack of practical guidance for researchers or practitioners undertaking an evaluation process within a complex operational environment.

The aim of this paper is to (1) identify some of the key requirements and potential barriers for undertaking a successful evaluation process within a challenging environment, and potential barriers, and (2) provide recommendations to maximise the efficacy of the user-centred evaluation effort. It is hoped that the conclusions and recommendations sections of this paper present a clear illustration of the value of applying a structured (in the informal sense) user-centred evaluation approach to IT&T development and implementation. The following section presents a brief overview of a recent evaluation case study, with a focus on the evaluation approaches taken and the success of their application.

2. A CASE STUDY OF SITUATED EVALUATION

2.1 Background

The advice contained within this paper has been generated from experience of undertaking user-centred evaluation of complex IT&T within manufacturing supply chains over approximately a decade (e.g. Powrie and Siemieniuch, 1990; Powrie, 1991; May, Joyner, Siemieniuch *et al*, 1996; May, Carter, Joyner *et al*, 1997).

This section presents a very brief overview of the evaluation process applied in the most recent of the above projects, a 2 year, user-centred, EU-funded collaborative project within the European Automotive Industry. The intention of this paper is not to present the details of the IT&T, or results arising from this project (see May and Carter, 2000; May, Carter and Joyner, 2000), rather to demonstrate the degree to which particular evaluation approaches adopted were successful.

2.2 The TEAM evaluation process

The evaluation process within the TEAM project was undertaken as part of an iterative user-centred (e.g. Damodaran, 1996) systems design process that included a user requirements exercise, demonstrator development, a range of user, expert and technical evaluation/validation exercises, and dissemination.

Table 1 summarises the approaches taken, and differences between, two main sets of user-based evaluations undertaken in the UK. (Other evaluations were carried out in Italian and French companies but are not reported here).

The main aims of the evaluation activities were: to define (or refine) the user requirements for IT&T to support enhanced communication and collaboration; to guide the development of a software demonstrator; to assess the potential user, business and organisational impact; to increase industry awareness; and to encourage future implementation.

Constraints were both commercial and methodological, including: legacy systems; security and confidentiality; concurrent job demands; lack of experimental control; lack of benchmark metrics; and the lack of business process efficiency measures.

First set of user trials after initial system development	Second set of user trials, after final system development
Took place during months 10 and 11 (out of project duration of 24 months)	Took place during months 21 and 22
Formative evaluation to validate and refine user requirements, guide systems development	Descriptive evaluation to assess impact on users, organisation and business processes
Total of 14 users, each using the system for one whole day	One or more of a total of 7 users, using the system over a period of about a month
One-to-one training on system, then use, guided where necessary	Intermittent, longitudinal, demand-driven use, largely unaided
Evaluator-generated scenarios of use and supporting data	Engineer-driven, real working during design and development
Questionnaires, structured interviews and evaluator observation	Questionnaires, session activity logs, business process tracing
Primary evaluation criteria: validity of initial requirements assumptions; functional match; additional requirements; usability	Primary evaluation criteria: usage characteristics; support for critical business tasks; additional requirements; user, business and organisation impact
Secondary evaluation criteria: comparison with existing business processes; projected use and business impact	Secondary evaluation criteria: system usability

Table 1: Summary of initial and final user trials.

3. THE EFFICACY OF PARTICULAR EVALUATION TECHNIQUES

During the two sets of user trials outlined above, a variety of user-centred evaluation techniques were attempted within the TEAM project, some successfully, others less so. This section summarises the efficacy of particular approaches that were attempted.

3.1 TEAM project: Evaluation methods that worked satisfactorily

The most successful data gathering technique was a **hybrid questionnaire and structured interview** approach. This was used during the user requirements phase of the project and before, during and after each set of user trials.

All questionnaires were administered as evaluator-aided, i.e. they were either completed by the evaluator whilst in discussion with the user, or by the user with direct assistance from the evaluator. Previous experience had demonstrated that this was the best way of ensuring complete and accurate data. This hybrid method therefore encompassed a range of demographic, closed, category-based, open-ended and Likert scale questions. Where open-ended questions were asked, the responses were recorded with a dictaphone.

This hybrid structured interview/questionnaire technique proved highly successful for the following reasons:

1. High face validity – engineers could see the immediate value of this data collection method (questions obviously addressed the aims of the project and the business needs).
2. Effectiveness of data capture – concise categorical data and detailed, rich data could both be captured as required.
3. Efficient and easy – categorical responses could be recorded quickly via check boxes; open ended answers can be easily recorded via dictaphone.
4. Very flexible – any range of topics could be covered, interviews could be scheduled when and where convenient (and even partly completed and then revisited in order to fit in around work commitments). In addition, questions and discussions could be continued until the required granularity and scope of data were extracted.

Correct interview and questionnaire design is important, e.g. avoiding leading questions, ensuring that topics are addressed at correct stages within the evaluation process, and using the 'language of the users'. See Fife-Schaw (1995) for guidance.

The main methodological limitation of the questionnaire/interview technique is that it does not generate truly objective and quantitative data (a common requirement from the evaluation stakeholders, particularly those with a business perspective). Quantitative data were generated via tools and techniques such as 5pt Likert scales (e.g. reliability compared to current IT tools) and expert assessments (e.g. projected percentage time savings for different phases of the PIP). However, it was recognised that additional data collection methods were needed to accompany these, see below.

The main practical problem encountered was the synthesis and analysis of the recorded data from the open-ended responses. Detailed transcription of this data is time-consuming. A more efficient approach is to generate a summarised transcription which captures the essence of what was said, and not necessarily every word in full. This does however require a degree of domain and human factors expertise. Since data should be generated according to a set of objectives (i.e. you should know in advance what you are trying to find out!), a topic template and a 'cut and paste' technique (e.g. Sowray, 1998)) can be used to summarise the data.

E**valuator observation** was used during the formative (initial) and more descriptive (final) user trials. This was vital during the early evaluations in order to help identify the interaction breakdowns, and the extent to which engineers' information and functional task needs were being met. Evaluators were also present during the final evaluations due to the lack of success of **self-completed diaries** (see below); observation was therefore also used during the final set of trials, but was of much less benefit than at the earlier formative stages.

Session proformas were used during the collaborative sessions of the final user trials in order to capture data on the participants, IT&T tools used, objectives of sessions, whether these objectives were met, information used, and additional functional and information requirements. Completion of these aided by the evaluator (as opposed to the use of self-completed diaries) ensured that good quality data were collected. A disadvantage was that it required the presence of an evaluator on at least one site within a collaborative session. Initial fears were that this would unduly influence both the decision to hold collaborative sessions, and the degree to which sensitive issues were discussed in the presence of a third party. These fears proved relatively groundless.

A **retrospective timeline-based case study** of existing communication and collaboration methods was undertaken, focused on a five-month period of design and development of a new

component (encompassing the period over which real working occurred with the collaborative tools). This study was completed as a trace study of existing communication and collaboration (chiefly meetings, phone calls, letters and faxes) that occurred between the automotive manufacturer, a tier-one plastic mouldings supplier, a tier-two supplier, a raw materials supplier and a machine tool supplier. A series of round table discussions were then held to identify where breakdowns had occurred. The data were not complete (e.g. records of telephone calls were not always kept), but this exercise was successful in identifying, 'warts and all', where communication bottlenecks, lack of information, and misunderstanding had occurred during a key phase of a product development cycle. This enabled, as far as possible, a 'before and after' comparison between traditional methods of working, and potential future approaches using enhanced IT&T tools. This activity was very effective in terms of demonstrating the potential business benefit, since it identified the shortcomings of existing communication and collaboration methods, and the potential improvements possible. The main concern with this technique was the need to adopt a 'no blame' attitude when identifying problems and causes, in order to avoid defensive attitudes. Careful emphasis was placed on the potential improvements, rather than the allocation of blame; this encouraged an atmosphere of openness.

3.2 TEAM project: Methods that were partially or wholly unsuccessful

Data collection via self-completed **diary studies** was attempted. There were two potential applications of this technique: (1) to identify the extent and means of the current 'as is' communication and collaboration process by design engineers with their colleagues (either in-house, or at remotely-located suppliers) and (2) to record the use of the new collaborative IT&T tools when used for real working during the final user trials. Neither of these activities were successful despite previous discussions with engineers concerning its feasibility: (1) diaries capturing the 'as is' situation were initially only partially completed, and then not used at all due to concurrent job demands; (2) diaries recording the use of the collaborative IT&T did not generate the necessary detail of data, and in particular did not identify *why* problems had occurred. The diary studies were scrapped and replaced by the **retrospective timeline-based case study,** and the **session proformas,** described in Section 3.1 above.

Video recording was used during the *initial* (more formative) evaluations, in order to demonstrate user-system interaction breakdowns encountered by the users. The relative cost-benefit of this was poor due to the effort of analysis. Laws and Barber (1989) discuss methodological limitations; more recent semi-automated approaches to video analysis still require relatively high effort in order to code and analyse video data.

3.3 TEAM project: Evaluation methods that could not be employed

It was not possible to employ **video recording** during the *final* user trials due to product development confidentiality issues. This would have enabled a more detailed analysis of the communication discourse that occurred during online sessions, and also a more detailed analysis of value-adding versus non-value-adding activities (e.g. proportion of time spent on productive versus non-productive activities). In retrospect, the inability to video the collaborative sessions was not a major loss, due in part to the time constraints within the project and the time-intensive nature of video data analysis.

Online logging of system use was planned but aborted due to technical constraints. This data gathering method has the advantage of being objective, low cost and low effort, and remotely manageable. However, logging of this nature generates substantial data, and it is important to be

very clear what method of analysis will be used (Kaasgaard, Myhlendorph, Snitker and Sorensen, 1999). As well as the data generation issue, there are other limitations with online datalogging: although it may identify what applications or windows are active, it will not necessarily identify which information is being used or what tasks are being undertaken by subjects. This was more accurately captured by observation.

4. GENERAL DISCUSSION OF KEY EVALUATION CONSTRAINTS

Section 3 above demonstrates the success or otherwise of a range of evaluation techniques within one specific evaluation context (the TEAM project). The experiences within this project are typical of those likely to be encountered where evaluation of a complex system within a dynamic, uncontrolled industrial environment is taking place.

This section draws more widely on the experience gained from a range of industrially focused evaluation activities, and discusses the *generically applicable* constraints that are likely to arise when undertaking an evaluation process within industry. The issues are presented under three categories: (1) those likely to arise during the planning and setting up stage; (2) those issues that may arise whilst the evaluation is being undertaken; and (3) those likely to arise during the feedback, follow-through and implementation phase of the IT introduction process.

4.1 Initial constraints you may be placed under before you start

4.1.1 The impact of company structure and culture

The profile of a particular company will have a considerable influence on

1. The setting up of an evaluation process
2. The techniques and tools likely to be successful
3. The efficacy of particular reporting methods

Small companies will have less staff dedicated to particular tasks, and may find it more difficult to absorb the overhead of evaluation activities. Larger companies will have a larger skill pool to call upon, and more flexibility in staffing additional activities. Smaller, and more informal companies may take more of a ‘hands on’ approach to getting things done. It is important to find out the preferred results presentation format and communication channels, e.g. detailed or summary reports, formal presentations or informal meetings.

4.1.2 Supply chain issues

A single company does not exist in isolation; instead it will have many suppliers, and may have several customers i.e. there may be ‘many-to-many’ relationships within the supply chain. It is likely that the evaluation of IT&T within a company will also encompass aspects of customer-supplier relationships. Suppliers and customers may be relatively separate entities, or tightly knitted within product or service teams. (e.g. within the automotive industry, a design team will typically include members of the automotive manufacturer, and the tier 1 supplier such as the steering systems provider). Within a particular office, you may have individuals from a company working alongside personnel from a supplier or customer. Therefore the evaluation process must be focused on business processes (e.g. product development) rather than geographically or departmentally based.

Several issues will arise from supply-chain integration, such as:

1. Compatibility of IT that may incorporated into the evaluation process.
2. Potentially conflicting evaluation criteria.

3. Differing cultures and procedures (what may work for one company may not work so well for the other).
4. Potential feelings of a master-servant relationship (the customer barks and the supplier jumps).
5. Issues of 'ownership' of the problem and accountability.
6. Potential for self-protectionism (a supplier may be unwilling to discuss openly or elucidate its constraints if these would reflect badly on it from a customer perspective).
7. Customer-supplier trust and confidentiality (especially where one supplier serves several competing customers).

Ultimately, if an IT&T solution is used solely within a single company, it can be evaluated on this basis. However, if it is used within the context of collaborative working across companies, then it must be evaluated as such. Unless a supplier-customer relationship is highly confrontational, there will be enough common ground and shared ownership of problems to enable an evaluation process to be undertaken that satisfies both parties and enables enhanced business process(es) to be demonstrated and evaluated.

4.1.3 Confidentiality

If the IT&T is situated within the actual workplace, there are likely to be issues of confidentiality of data collection such as the use of questionnaires. The host organisation may be particularly sensitive towards audio or video recording. Depending on the company culture, there may be concerns over identifying and recording blame. These issues can be overcome in several ways:

1. Establish trust – make clear statements concerning what any questionnaires and video and audio recordings will be used for, and how long they will be kept. Agree procedures for checking the acceptability of dissemination exercises. Sign a non-disclosure agreement (NDA) to protect commercial interests (whilst ensuring any NDA does not compromise the research/ dissemination aims).
2. Ensure that questionnaires and other data collection methods preserve anonymity, and that data storage complies with any data protection regulations.
3. Assess the extent to which video recording is going to provide added value to the evaluation process (its costs and benefits). What are you actually going to do with any video recordings? Will you have the time to analyse them? You will need someone present to turn the recording equipment on and off, change tapes etc. Evaluation teams generally have cupboards of video tapes that have never or rarely been looked at.
4. If you just want to capture some video footage for promotional purposes, this is best arranged as a 'video shoot' using high quality equipment and 'actors', instead of continuous recording of actual working processes.

4.1.4 Limited access to personnel

Evaluation teams will not necessarily get access to the best people as users within a trial, as it is easier for a business to release staff who are non-essential, and IT staff, rather than domain experts, tend to get involved by default. Therefore, you may not get the best domain experts. It is essential to convince the business of the importance of the involvement of the individuals with the domain experience, rather than those with the best IT skills.

Pragmatism is also necessary – the business will not stop, so flexibility is needed, plus the minimising of any time commitment of personnel who will also be trying to get their job done.

4.1.5 Lack of data to work with

This may be overlooked or taken for granted, but the success or otherwise of an evaluation process will depend on the ability to access and use data within user trials. This is true for both scenario-based trials, and evaluations based on 'real' working. For scenario-based trials, it may be easier to design the tasks around the available data.

4.1.6 Minimum user disruption

A common requirement from host organisations is for any evaluation activities to cause minimum disruption to day-to-day activities. In practice this usually means placing minimum demands on company personnel, and ensuring that company IT&T systems are not compromised.

Therefore lengthy self-completed questionnaires and diary studies may not be successful. A better approach is to complete short, tape-recorded structured interviews, with some form of direct observation. Face-to-face data collection methods will tend to work better as these are more immediate, and less easily 'put to the bottom of the pile'. Effort in usually equates to quantity and quality of data out. Leaving a pile of questionnaires to be completed by users will result in little useful data if these users are also trying to do their jobs at the same time.

4.2 Problems likely to arise during the evaluation process

4.2.1 Technical problems

It is inevitable that technical problems will arise during any evaluation process. New IT&T is, by definition, relatively unproven. However, if not carefully managed, technical failures will seriously impact on the credibility of the whole process, and also generate negative attitudes towards the technology being assessed. This is particularly important in operational (as opposed to experimental) environments. An important aspect to determine is the *expectation* of reliability. In some work domains, outages are relatively accepted, in others, 100% reliability is expected. Any new technology will be judged relative to these expectations.

The evaluation process should be designed to cope with a degree of technical failure. Where possible, backup scenarios should exist, so that if problems occur with particular applications, others can be used instead. It is necessary to be opportunistic to a certain extent, but also to temper the degree of control designed into evaluation activities with the likelihood of all of the evaluation sessions going to plan. A relatively unstructured approach will work best in less controllable and indeterministic environments. In any case, much useful data can be generated by users coping with or reacting or adapting to system failures.

4.2.2 Dwindling interest

The best way of maintaining interest and commitment throughout the evaluation process is to ensure that:

1. The evaluation team has sufficient credibility – you have at least a basic understanding of the work domain, understand the constraints that the users operate under, and can 'talk their language' to a certain extent. It is not unreasonable for users to take the attitude 'this person [evaluator] doesn't have a clue about this industry or how I do my job, so therefore how can they show me a better way of doing things'.
2. The development/evaluation scenarios have high face validity i.e. that any tasks users are being asked to do make sense and are realistic.
3. The evaluation process demonstrates that real work benefits are possible – the IT&T will enable

work to be completed more quickly or more effectively. You need to get the users thinking 'this could really be of use to me'.

4. Feedback results to the users – show them what you are doing with the data they help you generate; make them feel that they are contributing something worthwhile.

4.3 Problems with follow through

An evaluation process should have a direct impact on system development and/or IT&T implementation. An evaluation process can fail to achieve the desired impact for a number of reasons, namely (1) failure to understand the business sufficiently, (2) failure to deliver the necessary data (i.e. answers), or (3) failure to communicate the message effectively.

4.3.1 Failure to understand the business

A failure to understand the business sufficiently will result in an evaluation process that is inappropriately focused, does not address the key business drivers, and does not take into account the key constraints that currently exist, and hence the main requirements for new IT&T tools. If this occurs, the evaluation process will fail to 'engage' the key stakeholders.

A complete stakeholder analysis must be undertaken for a particular problem domain. This does not mean talking to everyone available, rather ensuring the viewpoints and requirements are gathered for everyone who is potentially impacted by the introduction of new IT&T tools. These stakeholders must include the decision-makers in order to understand: (1) their criteria for success or failure of new IT&T and (2) what company methods/processes exist for establishing this.

It is vital to understand where the company is positioned in terms of testing and implementing new technology. This will determine what has been defined already, how your results will feed into the process, and the scope for your findings to have real influence. An evaluation team can also learn from past evaluation and implementation successes and failures within the company, in terms of the process that was undertaken, and the efficacy of the results generated.

4.3.2 Failure to deliver the necessary data

Given an understanding of which issues should be addressed, and hence what data *should* be generated, data collection can fail for a number of reasons, some of which will be largely out of the control of an evaluation team. Even if there is a sufficient understanding of the business, the key success criteria, and the key requirements for data, technological and logistical problems will probably result in at least some of the data being incomplete. Questionnaires may be only partly filled in, response rates will be less than 100%, data collection methods will fail, key stakeholders will, at times, be unavailable, and evaluation sessions will be lost or curtailed.

The experimenter must assess the degree to which the above problems are likely to occur, and ensure that the data analysis methods can cope with any shortcomings. E.g. a reliance should not be placed on strict, statistical analysis if insufficient control over the process and environment jeopardises rigorous data collection. Strictly controlled environments lend themselves to structured data analysis. Relatively chaotic evaluation environments are best tackled with a eclectic mix of unstructured/informal data gathering techniques. The burden of data generation should lie with the evaluator as discussed previously.

4.3.3 Failure to communicate the message effectively

Despite generating the necessary data, failure to convince the business can still result from poorly communicating the message. This will result from one or more of: (1) addressing the incorrect

audience, (2) presenting inappropriate results, or (3) using the wrong terminology.

The basic evaluation requirement is to ensure that a range of results are produced that are appropriate to the different stakeholders. However, it is also important to ensure that the level of detail and the 'language' you use are appropriate. Audiences should be assessing the message you are presenting, rather than deciding whether the message is relevant to them, and putting in effort understanding what you are actually saying.

4.3.4 Changes within the business

There are many changes within a company that can potentially scupper an evaluation process, or reduce its potential impact. These include: changes in working practices/work organisation (e.g. merging of departments), changes in IT strategy or IT supplier selection (e.g. definitions of preferred technologies or standards) and changes in staff (e.g. transfers of key personnel). Any of these can result in an evaluation process becoming less effective, either in terms of addressing key business requirements, the ability to actually undertake the evaluation, or in communicating the answers effectively.

The evaluation team needs to be aware of any forthcoming changes, and to accept the need to have sufficient flexibility in the evaluation process to accommodate these changes.

5. CONCLUSIONS AND RECOMMENDATIONS

It is apparent that the evaluation of new IT&T within a relatively uncontrolled industrial environment is a challenging task, with a multitude of potential perspectives and a wide range of likely constraints. It may not be possible to overcome all of the problems identified in this paper. However, the following set of recommendations will, at least to some degree, overcome or mitigate some of the constraints identified, and enhance the effectiveness of IT&T development via a user-centred approach.

The issues below are not entirely independent, for example, the evaluation techniques possible will determine the type (e.g. subjective/objective, quantitative/qualitative) of data; similarly, if the evaluation focus and criteria demand a particular data, this will influence the choice of evaluation technique. All of the recommendations assume that:

1. The over-riding emphasis is that of improving (part of) the business process.
2. Evaluation is occurring as part of an iterative, user centred, requirements, evaluation and IT development and implementation process.
3. A top-down approach is taken, i.e. the overall evaluation aims are identified, and methods, techniques, tools and metrics chosen accordingly.
4. At least an outline user requirements exercise has been undertaken.

5.1 Understanding the business

It is essential to undertake a stakeholder analysis to identify the actors involved in the IT&T introduction process, in particular those who will influence the uptake of new technology, and those who are directly affected by it. This will include: end users within the company and collaborating partners, systems research and development, IT support, business managers and company strategists.

The stakeholder analysis will help ensure a coherent overview of the key business drivers, resources required, competitive pressures, working procedures, supply chain relationships, current and likely future constraints, and the changes that could really improve how the business operates.

The evaluation team must acquire a basic understanding of the IT&T strategy, in particular the extent of legacy systems, adoption standards, technology migration paths, and maturity levels. This

will help ensure that the results of the evaluation are relevant in terms of the development and/or implementation of new technology.

Scenarios of use can then be developed, based on the user requirements statement. A cross-checking of these scenarios of use and proposed evaluation process, against the IT&T strategy and future business processes, should then be carried out (does what is being proposed make sense in terms of how new IT&T is introduced, and how this has succeeded or failed in the past).

5.2 Determine the evaluation focus

The stakeholder analysis, understanding of the business, and user requirements statements should lead to reasonable clarity over the focus of the evaluation (what you are trying to achieve and who the results are targeted at). The following can then be determined:

1. The extent to which the evaluation process is aiming to generate prescriptive or descriptive results, or both. Is the aim to inform the systems development process and produce recommendations for improvements or changes to systems and processes? Or to assess the impact of a new system or process, and compare it to current/alternative practices?
2. Required or available evaluation timescales and resources, including any interim reporting points if applicable.

5.3 Evaluation criteria and benchmarks

Evaluation criteria must be agreed that address the key business drivers and deliver results to the specific stakeholders. These evaluation criteria should be agreed via an interactive process where the requirements are extracted from stakeholders, converted into metrics and metrication methods by the evaluators, and agreed and signed up to by the stakeholders.

These evaluation criteria can often be identified at a general level by asking each set of stakeholders questions such as:

1. What do you need to know in order to assess whether this new system/method of working is better or worse than what you are doing at the moment?
2. What is your understanding of ‘better’ or ‘worse’?
3. What specific criteria would *you* use to judge whether a system is ‘good’ or ‘poor’?

Benchmarks and metrics must be identified that are compatible with the evaluation criteria and that form the basis of a judgement of a system as ‘good’ or ‘poor’. These should include system-related benchmarks (e.g. less than two percent downtime), user-centred measures (e.g. achieving levels of usability and user satisfaction) and business-related benchmarks (e.g. an efficiency or quality improvement of at least twenty percent compared with current working practices).

In addition, the success criteria for the user-centred evaluation process *itself* must be established. The evaluation team must step back from their detailed involvement in the evaluation process, and ask themselves questions such as: Do I understand the business drivers? Will I be able to demonstrate whether this IT improves the business process(es)? Will I know how users will be impacted? Can I deliver the answers the stakeholders are looking for? How will I know if I have enabled good user-centred design?

5.4 Planning the evaluation process

Evaluation methods and metrics should be selected that satisfy the following requirements:

1. Creation of appropriate evaluation criteria and success/failure benchmarks.

2. Fit with IT&T introduction philosophy, and life-cycle maturity (make sense in terms of the 'big picture').
3. Time available for planning and undertaking the evaluation process, plus analysing the data.
4. Fit with the degree of control that can be exerted over the evaluation environment.
5. Availability and characteristics of domain experts (e.g. their IT literacy, keyboard skills).
6. Availability of company data that can be accessed.
7. Address constraints such as security and confidentiality, legacy systems and commercial pressures.

Once the evaluation criteria, benchmarks and desired metrics have been identified, techniques (e.g. post-hoc verbal reports) and tools (e.g. paper-based proforma) for data collection can be selected, based on the above requirements and constraints.

It is useful to identify and learn from past evaluation/implementation successes and failures within a particular company, as these will demonstrate how company culture and procedures can be used to best advantage.

5.5 Expert assessments and user trials

This section describes the key activities which will feed into a user-centred design process, and the successful application of these techniques will require specialist HCI/human factors input. It is suggested that user-centred evaluation comprises the following:

1. Early expert assessments by HF and domain experts to iron out the most obvious problems.
2. Initial scenario-based user trials with at least some real users, and concentrating on key business processes. These scenarios should allow you to control usage characteristics, the functionality being accessed, and the information being accessed/manipulated, with an emphasis on validating the user requirements, and aiding software development.
3. Longitudinal user trials with 'real' users, based on 'use in anger' when a system is more fully developed and robust enough for real use. These trials will focus more on the potential business benefits, and will be less empirically controllable than above, but will provide more valid and persuasive data.

The techniques described in Section 3.1 are recommended as the basis for data collection via (2) and (3) above.

5.6 Ongoing management of the process and reporting results

The evaluation team must ensure the following:

1. A clear understanding of the 'big picture' i.e. of what the evaluation process is trying to achieve.
2. Clearly assigned responsibilities (e.g. whether technical *and* user trials will be carried out in tandem, and who will undertake these).
3. Clear objectives throughout (so you can assess whether your evaluation process is succeeding, see Section 5.3) and review points.
4. Contingency plans if (or when!) technical problems arise or commercial pressures limit the evaluation activities.
5. Ongoing brief written or verbal reports, and demonstrations to stakeholders to maintain interest and momentum, presented in a 'language' that they understand.
6. Final reports that deliver on the evaluation criteria, targeted at particular stakeholders and presented in their 'language'.

6. OVERALL CONCLUSIONS

Any business depends on its staff to enable its business objectives to be met. Therefore, if IT&T is being used to support business processes, it must be designed, first and foremost, from a user-centred perspective. If this does not occur, new IT&T will not be effective, either by not being used, not enabling effective activities, or by having a detrimental impact on the users. The reader will be able to think of many examples of using even simple IT systems where you end up thinking: *I don't understand what it is doing. I don't know what to do next. I am getting fed up using this. It is not letting me do what I want to do.* These types of reactions are all symptomatic of failing to apply a user-centred design and evaluation approach to software design and implementation.

It is recognised that there are potentially many and considerable methodological and practical difficulties in undertaking user-centred evaluation within complex work settings. However, the potential benefits are considerable, and many authors will argue that successful IT&T development and implementation will *only* occur by applying user-centred approaches. It is hoped that this paper provides some guidance in this respect to the IT professional.

ACKNOWLEDGEMENTS

The author wishes to thank Judy Hammond and three anonymous referees for their helpful comments and suggestions that improved the quality of this paper.

REFERENCES

BAIRD, F., MOORE, C.J. and JAGODZINSKI, A.P. (2000): An ethnographic study of engineering design teams at Rolls-Royce Aerospace. *Design Studies* 21(4):333-355.

BEVAN, N. and MACLEOD, M. (1994): Usability measurement in context. *Behaviour & Information Technology* 13(1,2):132-145.

CARNEY, D.J. and WALLNAU, K.C. (1998): A Basis for Evaluation of Commercial Software. *Information and Software Technology* 40(14):851-860.

DAMODARAN, L. (1996): User involvement in the systems design process – a practical guide for users. *Behaviour & Information Technology* 15(6):363-377.

DIX, A., FINLAY, J., ABOWD, G. and BEALE, R. (1993): *Human-computer interaction.* Prentice Hall.

D'SOUZA, M.E. and GREENSTEIN, J.S. (1997): An ethnographic evaluation of a computer-supported product development system. Ancient wisdom – Future technology. *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting,* Albuquerque, New Mexico, September 22-26, 1997. HFES, Santa Monica, California, Volume 2 :739-743.

EARTHY, J., SHERWOOD-JONES, B. and WESTON, J. (1999): Human centred process, maturity and improvement. *Proc. Tutorial T12 presented at Human-Computer Interaction INTERACT '99,* SASSE, M.A. and JOHNSON, C. (eds)., Amsterdam: IOS Press.

FIFE-SCHAW, C. (1995): Questionnaire design. In *Research Methods in Psychology.* BREAKWELL, G.M., HAMMOND, S. and FIFE-SCHAW, C. (eds)., London: SAGE Publications, 174-193.

GEDIGA, G., HAMBORG, K.C. and DUNTSCH, I. (1999): The IsoMetrics usability inventory: An operationalization of ISO 9241-10 Supporting summative and formative evaluation of software systems. *Behaviour & Information Technology* 18(3):151-164.

GRISSOM, S.B. and PERLMAN, G. (1995): Step(3d) – A standardized evaluation plan for 3-dimensional interaction techniques. *International Journal Of Human-Computer Studies* 43(1):15-41.

GRUDIN, J. (1988): Why CSCW applications fail: Problems in the design and evaluation of organisational interfaces. *Proc. CSCW* 88, Portland, Oregon USA, GREIF, I. (ed): 85-93, Association for Computing Machinery, New York, USA.

HENDERSON, R., PODD, J., SMITH, M. and VARELA-ALVAREZ, H. (1995): An examination of four user-based software evaluation methods. *Interacting with Computers* 7(4):412-432.

ISO-9241 (1998): ISO 9241-11 Guidance on usability.

ISO-13407 (1999): ISO 13407 Human-centred design processes for interactive systems.

ISO/IEC-14598 (1998): ISO 14598-1 Information Technology – Evaluation of Software Products – General Guide.

ISO/IEC-FDIS-9126 (2000): ISO/IEC FDIS 9126 Software Engineering – Product Quality.

ISO-TR-18529 (2000): ISO TR 18529 Ergonomics of human-system interaction – Human-centred lifecycle process descriptions.

JAGODZINSKI, P., REID, F.J.M., CULVERHOUSE, P., PARSONS, R. and PHILLIPS, I. (2000): A study of electronics engineering design teams. *Design Studies* 21(4):375-402.

KAASGAARD, K., MYHLENDORPH, T., SNITKER, T. and SORENSEN, H.E. (1999): Remote usability testing of a web site information architecture: 'Testing for a dollar a day'. *Proc. of INTERACT '99*, Human-Computer Interaction, Edinburgh, SASSE, M.A. and JOHNSON, C. (eds). 1:443-450, Amsterdam: IOS Press.

LAWS, J.V. and BARBER, P.J. (1989): Video analysis in cognitive ergonomics: A methodological perspective. Special Issue: Current methods in cognitive ergonomics. *Ergonomics* 32(11):1303-1318.

MACLEOD, M., BOWDEN, R., BEVAN, N. and CURSON, I. (1997): The MUSiC performance measurement method. *Behaviour & Information Technology* 16(4/5):279-293.

MAY, A. and CARTER, C. (2000): A case study of virtual team working in the European automotive industry. To appear in International Journal of Industrial Ergonomics.

MAY, A., CARTER, C. and JOYNER, S. (2000): Virtual team working in the European automotive industry: User requirements and a case study approach. *Human Factors and Ergonomics in Manufacturing* 10(3):273-289.

MAY, A., *et al.* (1997): TEAM (AC070) Deliverable DRP013: Final results of demonstrator evaluation. Project Deliverable ROVER/TEAM/WP5/DRP013. HUSAT Research Institute. Loughborough, UK.

MAY, A., *et al.* (1996): Suppliers and manufacturers in automotive collaboration: Final experiment running and evaluation. RACE Deliverable R2112/ROVER/CIE/DS/015/b1. HUSAT Research Institute. Loughborough, UK.

NAYAK, N.P., MRAZEK, D. and SMITH, D.R. (1995): Analyzing and communicating usability data: Now that you have the data what do you do? *SIGCHI Bulletin* 27(1):22-30.

OPPERMANN, R. and REITERER, H. (1997): Software evaluation using the 9241 evaluator. *Behaviour & Information Technology* 16(4/5):232-245.

POWRIE, S.E. (1991): CAR application pilot component trials: results, evaluation and assessment. RACE 1079 CAR Deliverable 79/HUS/CAR/DS/B/801/b1. HUSAT Research Institute. Loughborough, UK.

POWRIE, S.E. and SIEMIENIUCH, C.E. (1990): An investigation of user requirements for broadband communications in the automotive industry. *Proc. INTERACT '90*, Human-Computer Interaction, Cambridge, DIAPER, D., GILMORE, D., COCKTON.G. and SHACKEL, B. (eds). 233-238, Elsevier Science Publishers, North-Holland.

PRATICO, B. (1997): Report on the extended test implementation of "Lotus notes/Domino" Groupware for the Logistics Division. Unclassified report U.S. Army PATRIOT Project Office. Raytheon, Huntsville, U.S.

ROJEK, J. and KANERVA, A. (1994): A data-collection strategy for usability tests. *IEEE Transactions on Professional Communication* 37(3):149–156.

ROSS, S., RAMAGE, M. and ROGERS, Y. (1995): PETRA: participatory evaluation through redesign and analysis. *Interacting with Computers* 7(4):335-360.

SOWRAY, D.M. (1998): Designing information systems for maximum use in a dealing room environment. *Behaviour & Information Technology* 17(4):203-217.

THOMAS, P.J. (ed) (1996): CSCW requirements and evaluation. Computer supported cooperative work. London, Springer.

BIBLIOGRAPHIC NOTES

Andrew May is a researcher at the HUSAT Research Institute, Research School in Ergonomics and Human Factors, Loughborough University UK. He has a BEng (Mechanical Engineering) from Birmingham University, an MSc (Human Factors and HCI) from University College London and is undertaking a PhD on electronic commerce. His main research interests are user-centred design and evaluation, and technology adoption. He has published articles in Accident Analysis and Prevention, Human Factors and Ergonomics in Manufacturing, and the International Journal of Industrial Ergonomics.

Notes