Evaluation of Software Systems

Günther Gediga*
Institut für Evaluation und Marktanalysen
Brinkstr. 19
49143 Jeggen, Germany
gediga@eval-institut.de

Kai-Christoph Hamborg*
Universität Osnabrück
Fachbereich Psychologie und Gesundheitswissenschaften
Arbeits- und Organisationspsychologie
49069 Osnabrück, Germany
Kai-Christoph.Hamborg@uos.de

Ivo Düntsch*
School of Information and Software Engineering
University of Ulster
Newtownabbey, BT 37 0QB, N.Ireland
I.Duentsch@ulst.ac.uk

1 Introduction

Evaluation as a general endeavour can be characterised by the following features [84]:

- Evaluation is a task, which results in one or more reported outcomes.
- Evaluation is an aid for planning, and therefore the outcome is an evaluation of different possible actions.
- Evaluation is goal oriented. The primary goal ist to check results of actions or interventions, in order to improve the quality of the actions or to choose the best action alternative.

*Equal authorship is implied.

- Evaluation is dependent on the current knowledge of science and the methodological standards.

Evaluation as an aid for software development has been applied since the last decade, when the comprehension of the role of evaluation within Human-Computer Interaction had changed. In one of the most influential models of iterative system design, the *Star Life Cycle Model* of Hix & Hartson [33], the activities

"Task analysis", "Requirement specification", "Conceptual and formal design",
"Prototyping", "Implementation"

are each supplemented by an activity "Evaluation" which helps to decide progression to the next step. Software can be evaluated with respect to different aspects, for example, functionality, reliability, usability, efficiency, maintainability, portability [38]. In this survey, we concentrate on the aspect of usability from an ergonomic point of view. This aspect has gained particular importance during the last decade with the increasing use of interactive software.

Whereas in earlier times evaluation of software took place at the end of the developing phase, using experimental designs and statistical analysis, evaluation is nowadays used as a tool for information gathering within iterative design:

> "Explicit human-factors evaluations of early interactive systems (when they were done at all) were poorly integrated with development and therefore ineffective. They tended to be done too late for any substantial changes to the system still be feasible and, in common with other human-factors contributions to development, they were often unfavourably received [...] This situation has been improved in recent years in a number of ways." [90]

Within this context, instruments for evaluation are not primarily used for global evaluation of an accomplished product, but these instruments are applied during the development of a product. Indeed, most experts agree nowadays that the development of usable software can only be done by a systematic consideration of usability aspects within the life-cycle model. One prominent part is the evaluation of prototypes with respect to usability aspects, employing suitable evaluation techniques in order to find usability errors and weaknesses of the software at an early stage [19, 33, 62, 94].

2 Goals and results of evaluation

Any evaluation has pragmatically chosen goals. In the domain of software evaluation, the goal can be characterised by one or more of three simple questions:

1. "Which one is better?" The evaluation aims to compare alternative software systems, e.g. to choose the best fitting software tool for given application, for a decision among several prototypes, or for comparing several versions of a software system. An example of such a strategy can be found in [77].

2. "How good is it?" This goal aims at the determination of the degree of desired qualities of a finished system. The evaluation of the system with respect to "Usability-Goals" [9, 94] is one of the application of this goal. Other examples are the certification of software, and the check on conformity with given standards.

3. "Why is it bad?" The evaluation aims to determine the weaknesses of a software such that the result generates suggestions for further development. A typical instance of this procedure is a system developing approach using prototypes or a re-engineering of an existing system [33].

The first two goals can be subsumed under the concept of *summative evaluation*, while the third goal is an instance of the *formative evaluation* approach.

In general, summative evaluation is concerned with the global aspects of software development, and does not offer constructive information for changing the design of the system in a direct manner. It is performed when the development of the system is almost or entirely accomplished [33]. Summative evaluation is also applied to prototypes in case there is a need for controlling the effect of design changes in comparison to a preceding version of the system.

In contrast, the goals of formative evaluation are the improvement of software and design supporting aspects [80]. It is considered the main part of software evaluation, and plays an important role in iterative system development. In every development cycle, formative evaluation results in

- Quantitative data for the description of the progress of the realisation of usability goals, and

- Qualitative data which can be used to detect the usability problems of the system.

The resulting data can be classified by the following criteria [33]:

Objective: Directly observable data; typically user behaviour during the use of the interface or the application system.

Subjective: Opinions, normally expressed by the user with respect to the usability of the interface or the application system.

Quantitative: Numerical data and results, e.g. user performance ratings.

Qualitative: Non-numerical data, e.g. lists of problems, suggestions for modifications to improve the interaction design.

3 Evaluation criteria

It is a matter of some debate in the human-factors community what constitutes an evaluation criterion. We follow Dzida [14] who advised that "criteria" should mean the measurable part of attributes of design or evaluation. Although this seems clear enough, the literature on software evaluation shows only a few attempts how to achieve general principles of design and evaluation criteria.

The concept of *usability*, which is a general quality concept for software systems, is often used for the determination of evaluation criteria [30, 48, 64].

The international standard ISO 9241 (Part 11), which is the methodological foundation of the HCI standard "Ergonomic requirements for office work with visual display terminals" (ISO 9241), states that

> "Usability of a product is the extent to which the product can be used by specific users to achieve specific goals with effectiveness, efficiency, and satisfaction in a specific context of use."

Note that the characterisation of usability by “effectiveness”, “efficiency” and “satisfaction” is not a breakdown to measurable criteria, but only a first step towards reaching such criteria [36, p. 6]. The same is true for ISO 9241 (Part 10) [30, 35], and for other general guidelines and design principles as well. These standards give no advice how to achieve trustworthy criteria which will fit into the needs of a specific software evaluation task. For a concrete evaluation project, there needs to be a proper operationalisation of the general guidelines and standards, which takes into account the context in which the system will be applied. The context of use is defined in ISO 9241 (Part 11) in terms of the user, the task, the equipment, and the environment. Following this description of the components of the context of use, every evaluation has to take into account the following restrictions:

A. The characteristics of a product’s users such as experience, age, gender or other more specific features.

B. The types of activity or tasks that the user will perform.

C. The environment of the study itself, ranging from controlled laboratory conditions to largely unstructured field studies.

D. The nature of the evaluation object, which can be a paper prototype, a software mock-up, a partially functional prototype or an accomplished system.

Even expert based evaluation should consider all of these four contextual restrictions in order to achieve a valid evaluation result.

In principle, there are (at least) four approaches for the derivation of criteria:

1. The successive division of principles into a hierarchy using specialisation, until the leaves can be measured by a set of known procedures (Top down I).

2. The successive division of principles into a hierarchy using specialisation, until the leaves cannot be subdivided any more. This level will then be operationalised by a tailored technique (Top down II).

3. A classification of known and/or commonly applied methods, and their mapping to criteria, or, alternatively, an empirical assignment of items to criteria (Bottom up).

4. A direct operationalisation of principles, which takes into consideration the measurable demands and the qualities of the object under study.

An advantage of the first strategy is that it is based on known and tried methods. However, known methods will be applicable only to part of the actual problem, namely the final level; therefore, the problem arises that the results may be misinterpreted due to a divergence artifact [18].

The second strategy has the advantage that it intends to measure everything that the evaluation requires. Its disadvantage are the huge amount of labour required for the construction of tailored instruments, and that the results usually cannot be compared to the findings of other studies.

The third strategy is expensive as well, including the risk that only some of the basic principles are well covered, while others are not.

The fourth strategy offers not only an operationalisation, but additionally attempts a more precise formulation of the criteria and the underlying principles as well. Its disadvantage is its cost; furthermore, empirical studies following this procedure offer results which are, in most cases, too general and not sufficiently specific for a concrete application [30].

Tyldesley [85] argues that "objective measurements" such as the time required for learning a system are preferred measurements, and should be used as criteria for software evaluation. Although objectivity and reliability of such measures are generally accepted, there are doubts about their validity:

> "...the so-called objective measurements of performance, for example the number of problems completed within a fixed time span, are widely applied; they are, however, rarely valid as criteria, since the underlying causal relationships are not sufficiently investigated or stated" (translated from [30]).

The derivation of "subjective" criteria such as acceptable levels of human cost in terms of tiredness, discomfort, frustration and personal effort, or even an operational definition of attitude, is much harder, and Whitefield et al. [90] come to the pessimistic conclusion that there is at present no support for the specification and operationalisation of such criteria

The bottom up approach offered by Strategy 3 above can be a starting point for criteria derivation: At the beginning one collects "standard" operationalisations with known validity and reliability. From

this collection, the designer can choose those instruments which can be subsumed as a set of criteria for the principles under study. At least for the "subjective" criteria, an operationalisation via a well analysed questionnaire or a survey instrument may be a successful way to cope with the problem.

There seems to be no problem for operationalisation on the principles of *Heuristic Evaluation* proposed by Nielsen [64] and comparable techniques. These are performed by auditors who evaluate the system given a set of general design principles. In this case, the concretisation will be done by the auditor in interaction with the system; therefore, the burden of proper operationalisation is put on the auditor. Such methods demand well educated and well trained evaluators, and it is no wonder that specialists who are usability and task experts perform better with this techniques then usability experts who are not task experts [65].

4 Evaluation techniques

Evaluation techniques are activities of evaluators which can be precisely defined in behavioural and organisational terms. It is important not to confuse "Evaluation techniques" with "Evaluation models", which usually constitute a combination of evaluation techniques.

We classify evaluation techniques into two categories, the *descriptive evaluation techniques* and the *predictive evaluation techniques*, both of which should be present in every evaluation:

Descriptive evaluation techniques are used to describe the status and the actual problems of the software in an objective, reliable and valid way. These techniques are user based and can be subdivided into several approaches:

Behaviour based evaluation techniques record user behaviour while working with a system which "produces" some kind of data. These procedures include observational techniques and "thinking-aloud" protocols.

Opinion based evaluation methods aim to elicit the user's (subjective) opinions. Examples are interviews, surveys and questionnaires.

Usability Testing stems from classical experimental design studies. Nowadays, Usability Testing (as a technical term) is understood to be a combination of behaviour and opinion based measures with some amount of experimental control, usually chosen by an expert.

Observe that all descriptive evaluation techniques require some kind of prototype and at least one user. Note furthermore that the data gathered by a descriptive technique need some further interpretation by one or more experts in order to result in recommendations for future software development.

The predictive evaluation techniques have as their main aim to make recommendations for future software development and the prevention of usability errors. These techniques are expert – or at least expertise – based, such as Walktrough or inspection techniques. Even though the expert is the driving power in these methods, users may also participate in some instances.

Note that predictive evaluation techniques must rely on "data". In many predictive evaluation techniques, such "data" are produced by experts who simulate "real" users. The criteria *objectivity* and *reliability*, which are at the basis of descriptive techniques, are hard to apply in this setting. Because validity must be the major aim of evaluation procedures, there are attempts to prove the *validity* of predictive evaluation techniques directly, e.g. by comparing "hits" and "false alarm" rates of the problems detected by a predictive technique [66].

At the end of the Section we shall briefly discuss evaluation techniques which can be used either for predictive or descriptive evaluation (e.g. formal Usability Testing methods) and those which do not fit into the predictive/descriptive classification, such as the "interpretative evaluation techniques" [88].

4.1 Behaviour based evaluation techniques

Behaviour based techniques rely on some form of observation in order to detect usability problems. Since the user is confronted with a prototype of the system, these techniques can only be applied in the later stages of system development.

As a first indicator of the expense of the technique, we offer the ratio "analysis time" (= time to process the data – usually – by experts after data elicitation) to "time of users' interaction with the system" (=

time to elicit the data, which will be analysed).

4.1.1 Observational techniques

User observation as a method of software evaluation takes place directly or indirectly by a trained observer [30], and it can produce quantitative as well as qualitative data [33, p. 306]. The techniques try to avoid subjectivity as much as possible by standardising procedures and documentation, as well as by training the observer. They are applied, for example, if the user's behaviour is of interest, especially when the user cannot tell how well (or badly) (s)he behaves using a prototype [64, p. 207].

It turns out that direct observation is not a well suited technique [73, p. 617], mainly, because the observer cannot deal with the amount of information which has to processed. Therefore, indirect techniques using video recording of the user's behaviour are commonly used. In human-factor laboratories, several parallel video loggings (hands, screen, face, whole body), together with log-files, are synchronised and used by the observers; an example is the VANNA system of Harrison [31].

Observational techniques show a ratio of 5:1 of analysis time to recording time [73, p. 620]. This ratio does not seem to be reduced by using automated procedures. The main reason is that computerised systems offer more information to the observer, which works against the effect of the rationalisation.

4.1.2 Thinking-aloud protocols

The method of thinking-aloud provides the evaluator with information about cognitions and emotions of a user while (s)he performs a task or solves a problem; examples can be found in [8, 42, 54, 64].

The user is instructed to articulate what (s)he is thinks and what (s)he feels while working with the prototype. The utterances are recorded either using paper and pencil [61, 62] or with more modern techniques using audio or video recording [42, 54]. In connection with synchronised log-file recordings, the evaluator has the chance to interpret the user's reaction based on contextual information [42]. Nielsen [62, p. 72] expresses the opinion that a large amount of overhead is unnecessary, and that simple paper-pencil recordings suffice to elicit the relevant contextual information.

A slightly different approach are the *pre-event* and *post-event* thinking-aloud procedures, in which the

user thinks aloud before a task has started, or after a task has been completed. The post-event technique is useful, if the tasks require careful concentration. Their protocols are based on the comments evoked by thinking-aloud, while the user observes the video recorded interaction with the system. Post-event protocols were criticised because the user might produce rationalise her/his own actions. Some researchers, however, point out that the more rational style of post-event protocols provides the evaluator with useful context information which event protocols cannot deliver [58]. The empirical comparison of post-event protocols with event protocols shows that the information provided by post-event protocols are of higher quality, although the amount of information is reduced in comparison to event protocols [3, 69].

The procedure of analysing thinking-aloud protocols has not much changed since the seminal paper by Lewis [54] in 1982: The protocols are scanned, and those episodes are extracted which describe the users' problems and their complaints. The episodes are listed and coded by a user number and an episode number. Afterwards, the episodes are matched with one "feature" of the system. These "features" define the grouping criteria for the episodes. The design aspect of the "feature" and the frequency information provide further help for the interpretation of the results.

By using thinking-aloud techniques, the evaluator obtains information about the whole user interface. This type of evaluation is oriented towards the investigation of the user's problem- and decision behaviour while working with the system [30, p 99]. The procedures to transform the lingual information into evaluation criteria is only weakly standardised, which may be problematic.

The expense of the method varies: In our experience, the concurrent event technique requires a user time factor near 1:1, and an evaluator time factor of about 3:1 for a trained evaluator. The video based post-event technique needs more time with a user time factor of at least 3:1, and an evaluator time factor of at least 5:1.

4.1.3 Video confrontation

In the area of Human-Computer Interaction, the video confrontation technique has been in use since the mid-eighties [25, 29]. It is based on the post-event thinking-aloud technique, but differs in the way how the video recordings of the user-system interaction are analysed: The evaluator picks out inter-

esting recorded episodes, and interviews the user about these. The interview is guided by pragmatic or theoretical goals, such as finding usability problems or examining the relationship between user errors and emotional reactions. The questions concern cognition, emotions, or problems which have occurred during the use of the system [29]. The main result of the process is the interview protocol. Since the frame for the analysis is set by the evaluator before performing the interview, the questions are concentrated on the salient points, and the protocol is much easier to analyse than a thinking-aloud protocol. Furthermore, the interview questions can be standardised, and thus they can lead to an operationalisation of evaluation criteria, which establish a direct relation to the evaluation goals. It is also possible not only to arrive at a list of problems, but also to obtain an indication of the reasons for these problems. Other sources of information such as other recorded sequences and log-files may also help the evaluator to interpret the outcome of the procedure [26].

The video confrontation technique is rather costly: A rough estimation of the user time ratio is 5:1, and of the evaluator time ratio is 7:1.

4.2 Opinion based evaluation methods

Oral or written interview techniques are commonly used to evaluate the user's opinion about software systems. The difference between oral interview techniques and questionnaire based techniques lies mainly in the effort for setup, evaluating the data, and the standardisation of the procedure. The development of an interview is more economic than for questionnaires, whereas carrying out and evaluating a questionnaire procedure can be done with less effort and costs. Standardisation and accuracy are also better for questionnaires.

The advantage of an interview in comparison with the observational methods of Section 4.1.1 is that an interview helps to obtain an insight into the user's opinion of the system which cannot be gathered by observation alone [64, p. 209ff]. Furthermore, these techniques are not as expensive as observational techniques.

Oral interviews are held in a flexible manner after the user had come into contact with the system, and in a more structured way, if the user was faced with unforeseen aspects [52, 64].

Another aspect are “critical incidents” [33, 52, 64] – e.g. examples of worst case situations created by experts –, which can be used to evoke oral responses.

Because there is no standardisation of oral interviews as a software evaluation technique, we shall not discuss the various approaches in more detail. As an introduction, we invite the reader to consult the checklist of Nielsen et al. [67, reprinted in [73]].

Interviews and questionnaires are primarily used in the specification- , design-, or re-engineering phase, or as a means of system comparison (summative approach), where different techniques have different focal points. In the sequel, we shall discuss three types of questionnaire in more detail: The *Questionnaire for User Interface Satisfaction* (QUIS) [81, p. 482ff], the *Software Usability Measurement Inventory* (SUMI) [53], and the *IsoMetrics questionnaire* [16].

4.2.1 QUIS: Questionnaire for User Interface Satisfaction

The “Questionnaire for User Interface Satisfaction” (QUIS) aims to provide a measure of overall satisfaction; additionally it evaluates some aspects of the user interface based on user opinions. The original version [QUIS 5.0, 11]) consists of the five scales

- “Overall User Reactions” with attributes such as
  - terrible / wonderful,
  - frustrating / satisfying,
  - dull / stimulating,
  - inadequate power / adequate power
  - rigid / flexible.
- “Screen”, which evaluates the display appearance.
- “Terminology and System Information”, which evaluates the distinctiveness of terminology and message display.
- “Learning”, which evaluates the suitability of the system for learning.

- "System Capabilities", which evaluates the efficiency and related aspects of the system in terms of speed and reliability.

The current version of QUIS is enhanced by the scales "Technical Manuals and On-line Help", "On-line Tutorials", "Multimedia", "Teleconferencing" and "Software Installation" [82]. A short (47 Items) and a long version (126 Items) of QUIS are available. The short version should be used, if there are only few time resources, or if motivational problems of the user can be anticipated. In the long version there are more concrete questions explaining and complementing the "leading items" (which constitute the short version) of each scale. At the beginning, there are questions about the properties of the system under study (in terms of the user's opinions) and the characteristics of the user.

The scaling of the items ranges from 1 to 9, and an additional "no answer" option. The endpoints of the scales are anchored by pairs of adjectives (e.g. difficult / easy). User comments about the system can be expressed at the end of each scale.

Reliability and validity of QUIS(5.0) have been investigated by Chin et al. [11]. They could show that its overall reliability is good, but no separate reliability measures for the five sub-scales were reported. Validity was shown in terms of differentiation of "like" and "dislike" for a software product, and the results of QUIS can be used to differentiate command line systems and menu driven applications. The predicted dimensionality of QUIS(5.0) is questionable, and there should be more analysis for construct validity as well, since "... no attempt to establish any construct or predictive validity was done". [11, p. 218] For the new version of QUIS, psychometric results are not (yet) available.

4.2.2 SUMI: The Software Usability Measurement Inventory

SUMI ([49], [53]) was developed primarily as a summative instrument which measures a user's perception of the usability of a software. SUMI consists of 50 items, which are assigned to five scales. Additionally a "Global" scale was constructed consisting of those 25 items which show a high correlation with an overall usability factor. The "Global" scale was constructed to present the perceived quality of the software as one index.

The answer format for the items consists of “agree”, “don’t agree" and “disagree”. In a SUMI evaluation, the recommended number of users is 10 to 12. The headings of the scales are:

- Global (25 items)
- Efficiency (10 items)
- Affect (10 items)
- Helpfulness (10 items)
- Control (10 items)
- Learnability (10 items)

The efficiency scale evaluates how well the software supports the user while working on the tasks. The affect scale measures the user’s general emotional reaction to the software. Helpfulness is intended to measure the degree to which the software is self-explanatory, and also the suitability of the help system. The control scale is used to measure the degree of the user’s feeling that (s)he controls the software. Finally, learnability measures time and effort for learning the handling of the software (or parts of it) from the user’s point of view.

For formative evaluation, SUMI was supplemented by the “Item Consensual Analysis” (ICA), which enables the evaluator to locate usability problems more precisely than with the analysis of the scales profiles. ICA needs a “Standardisation Database" which consists of expected pattern of responses for any SUMI item. A comparison of expected and observed frequencies for the items (using a Chi^2 test) shows which item signals a demand for change.

A second variant of the ICA should be used, if a more detailed problem analysis is required. In this case, the first step consists of a sample of at least 12 users who are confronted with the standard SUMI. Based on the ICA analysis of the user data, an interview script is constructed, and interviews are held to obtain explanations for the discrepancies of expected and observed frequencies.

The sub-scales of SUMI show reliabilities which range from satisfactory to good. Validation studies of SUMI have shown that the questionnaire has the capability to distinguish software of different

ergonomic quality [49, 53]. The usefulness of SUMI in consultancy-based studies as well as in case studies has been exemplified in [49, 53].

4.2.3 IsoMetrics

The IsoMetrics usability inventory provides a user-oriented, summative as well as formative approach to software evaluation on the basis of ISO 9241 (Part 10) [16]. There are two versions of IsoMetrics, both based on the same items: IsoMetrics^S (short) supports summative evaluation of software systems, whereas IsoMetrics^L (long) is best suited for formative evaluation purposes.

The current version of IsoMetrics comprises 75 items operationalising the seven design principles of ISO 9241 (Part 10). The statement of each item is assessed on a five point rating scale starting from 1 ("predominantly disagree") to 5 ("predominantly agree"). A further category ("no opinion") is offered to reduce arbitrary answers. IsoMetrics^L consists of the same items as IsoMetrics^S and uses the same rating procedure. Additionally, each user is asked to give a second rating, based upon the request

> "Please rate the importance of the above item in terms of supporting your general impression of the software."

This rating ranges from 1 ("unimportant") to 5 ("important"), and a further "no opinion" category may also be selected. In this way, each item is supplied with a weighting index. To evoke information about malfunctions and weak points of the system under study, the question

> "Can you give a concrete example where you can (not) agree with the above statement?"

is posed. This gives users the opportunity to report problems with the software, which they attribute to the actual usability item.

The IsoMetrics design provides information that can be used within an iterative software development. In summary, these are

- Scores of the usability dimension to measure the progress of development.
- Concrete information about malfunctions and their user-perceived attributes.

- Mean weight of any user-perceived attribute, given a class of system malfunctions.

IsoMetrics has proved its practicability in several software development projects. Its reliability was examined in two studies for five software systems, and could be justified for each of the seven design principles. In order to validate the IsoMetrics inventory, the scale means of the five different software systems analysed in the reliability studies were compared. It could be shown that programs with different ergonomic qualities were discriminated in the corresponding scales [16].

Given ten evaluating users, IsoMetrics^L evokes approximately one hundred remarks (The validity of this procedure has been reported elsewhere [93]). These are prioritised by their rated importance and their frequency. The content analysis of the remarks results in the identification of weak points of the evaluated software, and provides the input to a usability review. In such a review, users, software engineers and human-factor specialists develop remedies for the system's weak points and discuss its (re)design. This procedure has been highly accepted by developers as well as the users. A version of IsoMetrics which is applicable in group settings is discussed in [17].

Another summative evaluation method based on ISO 9241 (Part 10) has been adopted as a formative instrument in a system called "Qualitative Software Screening" [4, 74].

4.3 Usability Testing

Usability Testing is a name for a systematic and – more or less rigid – experimentally based gathering of information about a product or a prototype using user representatives. [34, 79]. A rough classification by Rubin [79, p. 22] can be used to characterise different approaches to Usability Testing:

- "Formal tests conducted as true experiments". This approach is characterised by the use of classical experimental designs for testing hypotheses and for deriving causal dependencies [see also 78].
- The second class of usability tests, described to be "a less formal one, employs an iterative cycle of tests intended to expose usability deficiencies, and gradually shape or mold the product in question".

Whereas both approaches differ in their proximity to classical experimental designs in terms of the accuracy of the description of the independent factors, the problem of defining dependent variables – the measurables – is common to both approaches. The dependent variables are chosen pragmatically according to the evaluation goals. Techniques described above, including

- Questionnaires and interviews,
- Observational methods,
- Think aloud technique,
- Video-confrontation,

are used to measure the impact of differences in system design, or different versions of a prototype on the usability of the product. Additionally, so called measurement criteria [85] are used. These are task specific measures such as “time to complete a task” or “percentage of tasks completed”, which can be easily applied, if tasks are accurately described. The thinking- aloud technique is “perhaps the most common technique for qualitative data generation . . ." [33, p. 306]; indeed, many other authors emphasise the use of thinking-aloud methods in usability testing [19, 42, 48, 54, 64, 79, 94].

Whereas Nielsen [64, p. 165] advocates

> “User testing with real users is the most fundamental usability method and is in some sense irreplaceable, since it provides direct information about how people use computers and what their exact problems are with the concrete interface being tested”,

the approach has been criticised because of its expense and its low cost-benefit relation [44]. As a possible remedy, inspection based methods are being discussed as an alternative evaluation approach [66]. These are predominantly used to uncover general usability problems in the interface, and consist of a combination of predictive techniques, described more fully below, such as heuristic evaluation, Walkthrough, or group discussions with other quality assurance tools like standard inspection or feature inspection [66, 86].

Virzi [86] calls usability inspection a “non-empirical” method, and points out that this method has the typical weaknesses of a purely predictive technique, because it relies mainly on the ability of the inspectors to predict problems of users with the software.

Because Usability Testing requires a large amount of expertise to

- Set up the experimental design,
- Choose the suitable tasks for comparison,
- Select the users and the number of users,
- Define the measurables properly,

it is perhaps best suited for usability engineers. It is certainly not a suitable technique for untrained evaluators.

4.4 Predictive Evaluation Techniques

Because predictive evaluation techniques aim to achieve results which can be used in practice, these methods are related to problem solving techniques which are used for the same reasons in other contexts such as requirements analysis.

Whereas observational and opinion based methods require a (more or less) sophisticated prototype, the predictive evaluations do not need a built system, because empirical methods are replaced by a theory or the contribution of experts. Consequently, user involvement is not as dominant as in the empirical evaluation techniques. Nevertheless, user participation could be more prominent in predictive techniques, because user representatives have the opportunity to influence the development process actively, whereas in descriptive techniques the user plays a more passive role.

4.4.1 Usability Walkthroughs

There are different variants of Walkthrough methods in Human-Computer Interaction [1, 2, 48, 55, 79]. Object of a Walkthrough is either the system design, a paper prototype [e.g. 48, p. 698], a built prototype [e.g. 79], or a completed system.

The main feature is a set of instructions to check step by step the artifact under development. For each step, the participants note problems with the task or screen design – in case of a paper prototype –, or with the user interface of a built prototype. After the data collection, problems are discussed and improvements of the system are proposed. The task sequence of a Walkthrough must be fixed, since otherwise the simulation is not meaningful for evaluation purposes.

Walkthrough methods for software evaluation are often performed in a group [1, 2, 48], mainly, because group Walkthrough has proved to be more effective than comparable individual Walkthroughs [73, p. 706]. The participants of a group Walkthrough come from different areas, for example, representatives of the expected user population, product developers and human-factor specialists. Because of the heterogeneous composition of the group, one also speaks of "Pluralistic Walkthrough" [2].

Walkthroughs are well suited to detect those usability problems which are caused by a discrepancy of system behaviour and a user's habits or expectations [89, p. 107]. Furthermore, they are an aid to ascertain whether the user interface is perceived to be adequate. Examples for such Walkthroughs are checking the wordings in the user interface, the distinction of buttons, commands or menus, or the analysis of small task sequences.

As a general method, Walkthroughs are limited because of their cost. A whole user interface cannot be evaluated by Walkthroughs with a reasonable cost-benefit relation, and therefore, only selected features are usually taken into consideration. A further limitation is the fact that explorative behaviour can only be simulated in a limited way; therefore, unexpected errors or misfits cannot be detected by usability Walkthroughs [2]. Another problem is the fact that the success of a Walkthrough, if applied as a group technique, depends on the combination and the psychological fit of the group members [48, p. 698f].

A prominent example is the *Cognitive Walkthrough* (CW). It is a method for evaluating user interfaces by analysing the mental processes required of a user [55, p. 717ff]. Its scope is the ease of learning, in particular, of learning by exploration, and it can be carried out by individual experts or by a group of peers. Details for a peer group evaluation can be found in [89, p. 106].

The CW is guided by four questions, which the evaluator should answer at every task step:

- Will the user try try to achieve the right effect?

- Will the user notice whether the correct action is available?
- Will the user associate the correct action with the effect to be achieved?
- If the correct action is performed, will the user see that progress is being made towards a solution of the task?

Input to a CW session consists of a detailed design description of the interface such as a paper mock-up or a working prototype. Additional input may be a task scenario, the precise description of prospective users, the context of system use, and a correct sequence of actions that a user should successfully perform to complete the designated task. Additionally, there must be a mapping of the task steps to corresponding "interface states". The evaluator simulates the task steps according to the limited possibilities of the intended users, and judges whether the system offers suitable tools for each step of the task sequence. In particular it is noted whether all necessary information is presented in a complete and appropriate manner; the evaluator will record other usability problems as well.

For each action within the task sequence a "success story" is constructed, which contains the reason(s) why a user will or will not perform the action. If a difficulty is identified, a reason is assigned, for example, that a required menu item does not seem related to the user's goal. Problems and their causes are recorded. Finally, the analyst reports alternatives to the proposed design. In this way, the method utilises factors which influence mental processes such as a user's background knowledge.

The advantage of the CW is that it can be applied in early stages of the development, and that not only usability weaknesses of the system are reported, but other problems as well.

One of the restrictions of the CW is the fact that the simulated sequence of actions has to be correct. This requirement puts a high demand on the qualification of the analyst, because (s)he has be familiar with the tasks and the environment in which the tasks will be carried out. However, whether a user will perform according to the "correct" way cannot be checked beforehand. Therefore, the analyst should have a profound knowledge of human factors, and should be aware of the intended user profile; furthermore, the analyst should be familiar with at least the basics of system development.

It has also been criticised that the focus of analysis of the CW sacrifices other important usability information such as overall consistency [86, p. 707].

4.4.2 Inspection by experts and heuristic reviews

Usability inspection by experts can be subdivided into free, structured, or model based variants. These are used for the generation of problem lists, and form the basis for system optimisation [64]. The usability inspection is carried out by a human-factor specialist, who is not directly involved in the development of the system. If there are several inspections of the same subject, they will be done independently of each other. Object of the evaluation can be the whole user interface or some of its components [30, p. 100]. The technique can be used at an early stage of the life cycle in which paper prototypes will be used.

In case of a free expert inspection, the evaluator checks the software system using only a few general guidelines. Structured expert inspection – like standard inspection or consistency inspection – is based on detailed criteria or standards, and uses checklists as a tool. Model based inspection commences with a model of the user interface, and includes a defined evaluation procedure [30, p. 101].

Heuristic Evaluation (HE) is a special adaptation of a free expert evaluation [65]. Unlike the expensive and detailed expert evaluation by standard or consistency inspections, the HE uses the following usability heuristics:

- Provide a simple and natural dialogue,
- Speak the user's language,
- Minimise user memory load,
- Be consistent,
- Provide feedback,
- Provide clearly marked exits,
- Provide shortcuts,
- Provide informative error messages,
- Prevent errors.

The heuristics are general principles which guide the evaluator through the inspection process. A number of evaluators inspect the user interface in individual sessions. Nielsen [65, p. 26] recommends

a number of three to five evaluators, because in his studies using HE, a larger number of evaluators will not result in additional relevant information. The degree of standardisation is not very high, and

> "In principle, the evaluators decide on their own how they want to proceed with evaluating the interface". [65, p. 29]

HE is normally carried out by a double inspection of the user interface: In the first round, the evaluator obtains a feeling for the system and its potential. The second inspection allows the evaluator to concentrate on those elements of the interface which are relevant for evaluation purposes. The evaluator steps through the interface design and checks the dialog elements on the basis of the heuristics. Results of the HE are either fixed in a protocol by the evaluator, or the evaluator reports the findings to an observer while checking the system.

An HE session should not last longer than 2 hours. If the evaluation of a larger or complicated system requires more time, it is necessary to split the HE into several sessions, which will fit the time restriction.

Inspection methods share a problem with the Cognitive Walkthrough: The evaluators must be user and task experts. A further disadvantage of the method is that

> ...it sometimes identifies usability problems without providing direct suggestions for how to solve them. The method is biased by the current mindset of the evaluators and normally does not generate breakthroughs in the evaluated design" [68, p. 255].

Although HE is part of the so called "discount usability engineering methods" [61, 72], there are more pessimistic results as well: The main argument for the "discount" – that there is a need for only a few experts to achieve acceptable results – has been questioned by a comparative study of Gediga & Hamborg [15]. The authors show that the number of required experts depends on several variables, including the task, the layout of the evaluation, the quality of the evaluator, and the definition of a "usability problem".

4.4.3 Group discussion

The object of the evaluation using group discussion is the complete user interface of the system, and different criteria can be used for the evaluation. Perhaps the most interesting aspect of group discussion is the possibility to use it as a creativity technique to generate evaluation criteria as a first step of the evaluation. Group discussion is quite flexible in its orientation: It may be user -, task - or organisation - oriented [30, p. 101]. The application of the technique tends to be more efficient in the early stages of development.

Group discussion among user representatives and/or system designers is not a stand alone technique, and it must be supported by other means. In the "Group Design Reviews" [86], several users discuss the user interface. The discussion has to be moderated by a human-factor expert; other experts such as designers, training experts, marketing experts etc should complete the group if it seems necessary. A combination with a Walktrough is straightforward [2] and more effective than a Walkthrough on its own [73, p. 706].

4.5 Other Techniques

Although we have listed quite a number of techniques, there are more variants and combinations which had been used in the past 2 decades.

A very interesting different point of view is taken by the *interpretative evaluation techniques* [88], which have moved away from the evaluator-controlled forms of evaluation to more informal techniques derived from anthropology and sociology. The techniques *contextual inquiry* [91], *co-operative evaluation* [58], and *participative evaluation* [24] advocate a more holistic treatment of evaluation. Up to now, no comparative study includes these techniques, but we are confident that at least the idea of the interpretative evaluation techniques will find its way into the HCI laboratories.

Modelling approaches aim at a formal description or prediction of the human-computer interaction. Their historic root is a model of an abstract human processor which sets the frame for a task description, given a more or less abstract system in a concrete environment [5]. The concept has resulted in a number of developments such as the GOMS and NGOMSL approach [51], the task knowledge

structure (TKS) and the knowledge analysis of tasks (KAT) [41], the task action grammars (TAG) [71], the external tasks – internal task mapping (ETIT) [59], and the yoked state space (YSS) [70]. All these models share some characteristics: They offer convincing results, e.g. the comparison of aspects of the PC-DOS and the Macintosh interface using NGOMSL by Kieras [50], and they are used for comparatively small scaled applications. Even though their advocates have asserted for the past decade that the models will be more applicable, if computer speed and memory are sufficient to fit the requirements of the models, there are some reservations concerning the validity of such methods for larger applications. It was noted that the human processor model does not take into consideration the huge individual differences among human subjects, as well as results from the large body of results in reaction time research [27]. Finally, it was shown that known results on learning or chunking processes is only considered if they fit smoothly into the proposed model. Because the human processor is therefore, at best, a rough approximation of human behaviour, one can conclude that the results may well be acceptable for small scaled examples, but will run into problems if the model is applied to more demanding systems.

4.6 Comparison of software evaluation techniques

There is a plethora of studies which deal with the comparison of evaluation techniques in terms of effectivity, efficency or utility [12, 13, 15, 28, 32, 39, 40, 43, 83, 87]. It is also investigated whether the information gathered by different methods are the same or not. Frequently, some of the "cheaper" predictive methods such as Heuristic Evaluation or Walkthroughs are benchmarked with a usability test method.

The results reported in the literature lead to the conclusion that empirical comparison of software evaluation techniques is a minefield. First of all, researchers are faced with the usual problems encountered in empirical comparisons, such as finding adequate and representative testing subjects, objects and situations, and a suitable methodology for analysing the results. Secondly, as Gray & Salzman [23] point out, up to now, no study fulfils the rigorous standards of experimental design. In benchmark studies, confounding variables may intervene (e.g. different testing situations, different time of contact, different number of subjects performing the methods ...), and therefore, it may not

be appropriate to interpret the results in terms of difference of methods alone. Carroll [7, p. 309] comments

> "In their focus on conventional experiments, Gray and Salzman underestimate the importance and the distinctiveness of rare evaluation events" .

The main argument in favour of comparison is that if one perceives the confounding variables as specific properties of the methods, comparison is a valuable evaluation of techniques, which may help practitioners to choose the "best" method for evaluating software. However, this pragmatic view puts heavy restrictions on the empirical studies, because a valid benchmarking is only feasible if the methods are applied exactly as they will be used in practice. Gray & Salzman [23] have shown that some studies do not fulfil these restrictions.

The situation is complicated, because the definition of some methods is not very precise. If, for example, "the" technique of Usability Testing is compared with other methods, it is often not clear what is really meant. One can argue that there is no Usability Testing method as such, since the technique has shown high variability and evolution over the past two decades. The same holds – to some extent – for other techniques as well. Two classical studies shall demonstrate the problem:

- Jeffries et al. [39] compare Heuristic Evaluation with Software Guidelines, Cognitive Walkthrough and Usability Testing. They showed that Heuristic Evaluation is more effective and efficient than Usability Testing (Rank 2) and the other two techniques in terms of detection of severe problems and of costs.
- Karat et al. [43] compare a variant of Heuristic Evaluation, Walkthroughs and Usability Testing. They show, to the contrary, that most of the problems are detected by Usability Testing in shorter time than by the other methods.

Taken together, the results of both studies (and of other cited studies as well) are inconclusive. Reviewing empirical comparison studies, Karat [45, p. 233] summarises

> "Usability evaluation methods act as different types of filters of user interaction with a computer system.",

which well describes the state of knowledge about evaluation technique benchmarking. Nevertheless, some lessons can be learned:

- Heuristic Evaluation offers at least satisfactory results. A combination of Heuristic Evaluation with a thinking-aloud technique seems to be a very effective strategy [39].
- Usability Testing procedures – and other descriptive techniques as well – often result in more information than predictive techniques. The findings show that predictive techniques – like Heuristic Evaluation – concentrate on severe problems only, whereas problems discovered by a descriptive technique address more specific – and sometimes “cosmetic” – aspects of the software.
- Usability Testing often needs more effort and equipment than comparable predictive techniques.
- Predictive evaluation techniques can be applied earlier in the life cycle than Usability Testing procedures.
- The goal of evaluation needs to be taken into account:
  - If the evaluation goal is summative (“which one is better” or “how good”), the behavioural evaluation procedures applied in Usability Testing are the methods of choice.
  - Neither Usability Testing nor pure expert based techniques support participatory design [60]. To achieve this, collaborative Walkthroughs or other group discussion techniques which include user representatives are necessary, because in these techniques the user has the chance to be an active part of system development.

The decision which evaluation technique(s) should be used has to be based on the concrete demands of the software development schedule, human-factor considerations, and cost benefits issues Karat [48, p. 699]. The results of the above mentioned comparison studies can only constitute a limited basis for such a decision. A systematic investigation of the “Return of Investment” [46, 56] of usability activities in the software life-cycle is still missing:

> “The relative cost-effectiveness of usability testing and inspection methods, based on return of investment in implemented changes, is not clearly established at this time" [45].

5 Evaluation models

In contrast to software evaluation techniques, which were presented in the preceeding section, software evaluation models determine the frame of the evaluation, which consists of

- Choosing techniques appropriate for the life cycle, and
- Setting the focus with respect to the objects under study and the measurement criteria.

Evaluation models may provide a standardised treatment of establishing (potentially) successful procedures in the practice of evaluation, and are a necessary tool for a comparing different types of software evaluation. Any descriptive evaluation procedure must be combined with some kind of predictive technique to result in an applicable evaluation model; furthermore, some preparatory steps are necessary. For example, the evaluation model, which consists of the IsoMetric^L questionnaire as a basic technique, standardises the preparatory steps "choosing the tasks" and "choosing the user group(s)"; it also standardises the preparation of the report by an expert, and the structure of a "usability review", which consists of a standardised result presentation and a Walkthrough technique [17].

There are three classes of evaluation models:

Method driven models : These models offer a frame for software evaluation based on a collection of techniques. The models are only applicable if the evaluation procedures fits perfectly the problems encountered by the user and the system developers.

Criteria driven models : More or less abstract criteria are defined and refined; the evaluation in these models aims at a measurement of the criteria.

Usability Engineering : These are evaluation models which are driven by the needs of a specific life cycle model.

5.1 Method driven evaluation models

The centre of method driven evaluation models consists of the arrangement of evaluation techniques, amended by the regulation of preparatory and subsequent tasks. A method driven evaluation model

can be perceived as a complex evaluation technique as well. An example is EVADIS II [75, 76], which is well tested for office automatisation software. The EVADIS II model combines interviews, simplified task analysis, and expert judgement in the following steps:

1. Installation and exploration of the software.
2. Analysis and relevance weightings of the tasks, construction of test tasks.
3. Analysis of the user characteristics. Selection of relevant ergonomic test items.
4. Evaluation of the software, based on the results of the first three steps.
5. Interpretation of the results and composing a test report.

The first three preparatory steps can be handled in parallel; they result in testing tasks and a list of ranked ergonomic evaluation criteria, mainly based on the principles of ISO 9241 (Part 10). The ranks of the criteria are deduced from the user profile, and they determine the test items which are chosen from an item database. In step four, the testing tasks are evaluated by an expert who steps through the tasks, and answers the questions formulated in the ergonomic items. The recorded answers form the basis for the test report. Every step of EVADIS II is supported by a large amount of supporting material such as databases and guidelines, which allows a domain expert with only a small amount of knowledge of software evaluation to form a well-founded opinion on the topic.

5.2 Criteria driven evaluation models

Criteria driven evaluation models start with assumptions about the structure of the design process in which criteria are defined, and give advice how to derive measurables form the criteria. Because these models focus on criteria and measurables, there is no close connection to a design model. Standards such as ISO 9241 or ISO/IEC 9126 can constitute a basis for a criteria driven evaluation model. As an example, we discuss the “Evaluation Process Model” ISO/IEC 9126 (1991) [38, 92], which is the basis for many software quality assurance procedures. The standard ISO/IEC 9126 aims at a product evaluation from a software quality point of view. It defines a specific process model and some general quality characteristics, for example,

- Functionality,
- Reliability,
- Usability,
- Efficiency,
- Maintainability,
- Portability.

Although “Usability” appears as a criterion, ISO/IEC 9126 does not aim at the ergonomic quality of the software:

> “Usability defined in this International Standard as a specific set of attributes of software product differs from the definition from an ergonomic point of view, where other characteristics such as efficiency and effectiveness are also seen as constituents of usability.” [38, p. 3]

After the relevant quality characteristics and their weightings have been selected, the evaluation process is performed in three steps.

1. Fixing the requirements: The requirements are derived from the application context of the system. The quality characteristics are formulated concretely in terms of observables, which operationalise the criteria and meet the requirements.
2. Preparation of the evaluation: This step comprises the operationalisation of the criteria into metrics and composite measurables.
3. The evaluation step: The evaluation of the product or a part of the product is performed in the basis of the derived requirements and chosen metrics.

Because ISO/IEC 9126 offers a general framework for software evaluation within the context of software quality assurance, we present the tasks accompanying the three steps of evaluation in more detail in Figure 1.

Figure 1: The ISO/IEC 9126 evaluation model

5.3 Usability Engineering

Usability Engineering is concerned with the systematic integration of methods and techniques of building usable software in the system development process. It can be characterised as a process which covers the definition and the measure of product usability in general [94, p. 654]. Usability Engineering models coordinate the “Usability Engineering activities", and do not replace traditional models of the software engineering process. The evaluation of software is a prominent part of any Usability Engineering model [6, 57, 64, 94].

The models are committed to the approach of the User-Centred Design [47], which is based on the assumption of an iterative system design with user participation [6, 60]. In this process, the user is regarded not as a producer of usability data, but should be directly and actively involved in the design process:

> "I strongly recommend that all team members carry out life-cycle tasks jointly, so that they develop a shared understanding of the requirements and design issues. In addition, it can be extremely effective to have users participate in many Usability Engineering tasks, not just as objects of study or sources of information, but as active participants." [57, p. 22]

Usability Engineering requires a software engineering model, which allows revisions and feedback loops. These models include the prototyping approaches, iterative system design and user involvement [19, 57, 64, 94].

Models of Usability Engineering are usually subdivided in three phases [19, 22, 57, 64]:

Phase 1: Analysis and specification. The first step – the "Gearing-Up Phase" – starts with preparatory activities, such as choosing general principles for the design, for example, relevant standards, development models and tools. The next step – the "Initial Design Phase" or "Requirements Analysis" – is concerned with the characterisation of users and the setup of user profiles; additionally, task and work flow have to be analysed. The obtained information is used to plan the activities of the "Work Re-engineering" and for the design of the user interface. As a result of this process, usability specifications are developed, which consist of goals for user-oriented design and behaviour-oriented measurables of these goals. The outcome of this phase is summarised in a product style guide [57].

Phase 2: Iterative development. The results of the preceding phase are used to design the organisational and work flow part of the system. Conceptual models are developed which can be used to produce paper and pencil or prototype mock-ups. Using an iterative design, the prototypes are evaluated and changed continuously. This helps to identify and remove major usability bugs. If the conceptual model shows satisfactory results, the screen design is developed, using screen design standards, and evaluation of the screen design takes place with the aid of a prototype. As above, the results are summarised in a product style guide.

Phase 3: System Installation. The final phase is concerned with "system installation" and "user training". Furthermore, the acceptance of the system has to be assured, and system support

and maintenance must be organised. Software evaluation procedures in the application phase have to be planned in order to obtain feedback about the usability of system. This information can be used for the design of new releases.

The standard ISO/DIS 13407 [37] is closely connected to the Usability Engineering approach. It describes three different usability activities in system design

1. Analysis and specification of the context of usage and the needs of the users and organisational needs.
2. The system design.
3. The evaluation of the system design using user-oriented criteria.

Similar to the Usability Engineering approach, ISO/DIS 13407 uses the principles

- Prototyping, iterative design, direct user participation, iterative evaluation, quality enhancement of the system

in all phases of the design process.

In case that the evaluation cannot be performed by users, ISO/DIS 13407 recommends evaluation by experts, but it is advocated that a user based evaluation is done at a later step of system development. The evaluation-(re)design cycle is repeated until the goals of the user-centred design are fulfilled. The product can be evaluated to ascertain whether it fulfils the user-oriented requirements and/or whether it is in accordance with other international standards such as ISO 9241 [10].

5.4 A problem: The integration of software evaluation models and software design models

It is pointed out in ISO/DIS 13407 that “user-oriented activities” have to be integrated into the system development process, and the standard gives some advice how this can be done. This is an urgent need, because in classical approaches, the combination of system design with software evaluation

models in general and the Usability Engineering models in particular seems to be more an art than a science. Although some case studies show that an effective combination of both model types can be achieved [20, 21, 63], Mayhew [57, p. 4] points out that "... Gould and Lewis did not specify exactly how these global usability strategies could be integrated into an overall software development life-cycle ...", and recommends her own approach as an alternative. She admits, however, that

> "...the exact overlapping of the two types of tasks is not completely clear. The most appropriate relationship between them is probably not absolute, but is most likely dependent on various project-specific factors and on the particular software development methodology with which the life-cycle is integrated ... For example, traditional systems analysis and Contextual Task Analysis are highly related tasks, but they are not the same thing and require very different kinds of expertise. Exactly how to integrate them to avoid duplication of effort and yet produce the intended and necessary outputs of both is a topic for future work." [57, p. 17]

To date, there are only loose connections of activities of Usability Engineering and the object-oriented design methodology. These are concentrated in the requirements analysis and in the design / testing / development cycle. The generation of an object-oriented "Requirement model" can be supported by user profiles, which can be obtained by task and work-flow analysis. "Work Re-engineering Tasks" correspond to the construction of an "Analysis Model" in object-oriented design. The design and programming of the user interface in the object-oriented model (the "Design Model") can be supported by conceptual models [57, p. 30].

6 Future Tasks

There are many evaluations techniques and few evaluation models which can be applied in practice. Although much has been achieved, there is a dire need to devote more effort to developing applicable models for practical use, in particular, if non-experts evaluate software as in smaller companies. EVADIS is an instance of such a model, as it combines several evaluation techniques to a coherent evaluation model in the domain of office automation.

At present, the differences in effectiveness, efficency and the cost-benefit relation of software evaluation techniques and models are not satisfactory. Due to the huge number of techniques and their variants, the labour of comparing these techniques – and models using them – would be tremendous. An empirical stock-taking of frequently used models should be done to form a basis for future comparison studies. These should not only consider effectiveness, efficency and return of investment, but other effects on the software as well, such as time efficency, suitability for learning, or user satisfaction [44].

For a number of techniques there is a need to investigated where they will fit into a software development cycle. The rule of thumb "early → predictive & late → descriptive" should be replaced by a more differentiated rule system. Because most findings demonstrate that the techniques produce a large amount of overlapping results, the effect of combining evaluation techniques needs to be evaluated as well. Inspecting the evaluation models, we see that there should be a higher integration of models as well. In particular, the goals and techniques of Requirements Analysis overlap with the goals and techniques of software evaluation in such a way that a unified approach would save time and money.

Most software evaluation technique we have presented aim at the operationalisation of "Usability" of the software. The investigation how "Usability" can be described in theoretical terms is still ongoing. "Usability" is a combination of – at least – two different kinds of constructs: The "ease of use" and the "enjoyment to use". Whereas older studies and methods had a closer look at the "ease of use" component, the "enjoyment to use" has been the focus of newer studies (at least since SUN and SAP proclaimed that even office software should be enjoyable). There is some evidence that both components vary independently, which implies that there is also a need for an independent measure of both aspects.

Even though much effort and research are still necessary to develop more efficient software evaluation procedures, we hope that the paper has demonstrated that there is not only a large body of techniques, but also that these techniques are applicable and valuable in the development of software systems.

References

[1] Bias, R. (1991). Walkthroughs: Efficient collaborative testing. *IEEE Software*, **8**, 94–95.

[2] Bias, R. (1994). The pluralistic walkthrough: Coordinated empathies. In J. Nielsen & R. Mack (Eds.), *Usability Inspection Methods*, 63–76, New York. Wiley.

[3] Bowers, V. A. & Snyder, H. L. (1990). Concurrent vs retrospective verbal protocol for comparing window usability. In *Proceedings of the Human Factors Society 34th Annual Meeting of the Human Factors Society*, 1270–1274.

[4] Burmester, M., Görner, C., Vossen, P. H., Zolleis, T. M. & Zouboulides, V. (1997). Qualitatives software screening. In M. Burmester, C. Görner, W. Hacker, M. Kärcher, P. Kurtz, U. Lieser, W. Risch, R. Wieland-Eckelmann & H. Wilde (Eds.), *Das SANUS-Handbuch. Bildschirmarbeit EU-konform*, 2–33, Dortmund. Schriftenreihe der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, FB 760, Teil II, 2.

[5] Card, S., Moran, T. P. & Newell, A. (1983). The Psychology of Human-Computer Interaction. Hillsdale: Lawrence Erlbaum.

[6] Carroll, J. M. (1997). Human-Computer Interaction: Psychology as a science of design. *International Journal of Human-Computer Studies*, **46**, 501–522.

[7] Carroll, J. M. (1998). Review validity, causal analysis, and rare evaluation events. *Human-Computer Interaction*, **13**, 308–310.

[8] Carroll, J. M. & Mack, R. (1984). Learning to use a word processor: By doing, by thinking and by knowing. In J. Thomas & M. Schneider (Eds.), *Human Factors in Computing Systems*, 13–52, Norwood. Ablex.

[9] Carroll, J. M. & Rosson, M. B. (1985). Usability specifications as a tool in iterative development. In H. R. Hartson (Ed.), *Advances in Human-Computer Interaction*, 1–28, Norwood. Ablex.

[10] Çakir, A. & Dzida, W. (1997). International ergonomic HCI standards. In M. Helander, T. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction. Second Edition*, 407–420, Amsterdam. Elsevier.

[11] Chin, J. P., Diehl, V. A. & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of SIGCHI '88*, 213–218, New York. ACM/SIGCHI.

[12] Desurvire, H., Lawrence, D. & Atwood, M. E. (1991). Empiricism versus judgement: Comparing user interface evaluation methods. *ACM SIGCHI Bulletin*, **23**, 58–59.

[13] Desurvire, H. W., Kondziela, J. M. & E.Atwood, M. (1992). What is gained and lost when using evaluation methods other than empirical testing. In *Proceedings of HCI '92, People and Computers VII*, 89–102, Cambridge. Cambridge University Press.

[14] Dzida, W. (1994). Qualitätssicherung durch software-ergonomische Normen. In E. Eberleh, H. Oberquelle & R. Oppermann (Eds.), *Einführung in die Software-Ergonomie*, 373–406, Berlin. de Gruyter.

[15] Gediga, G. & Hamborg, K.-C. (1997). Heuristische Evaluation und IsoMetrics: Ein Vergleich. In R. Liskowsky, B. Velichkovsky & W. Wünschmann (Eds.), *Software-Ergonomie '97*, 145–155, Stuttgart. Teuber.

[16] Gediga, G., Hamborg, K.-C. & Düntsch, I. (1999). The IsoMetrics usability inventory: An operationalisation of ISO 9241-10. *Behaviour and Information Technology*, **18**, 151–164.

[17] Gediga, G., Hamborg, K.-C. & Willumeit, H. (1997). The IsoMetrics manual (Version 1.15). Osnabrück: Universität Onabrück: Fachbereich Psychologie. http://www.eval-institut.de/isometrics/.

[18] Gigerenzer, G. (1981). Messung und Modellbildung in der Psychologie. Basel: Birkhäuser.

[19] Gould, J. D. (1988). How to design usable systems. In M. Helander (Ed.), *Handbook of Human Computer Interaction. First Edition*, 757–789, Amsterdam. Elsevier.

[20] Gould, J. D., Boies, S. J., Levy, S., Richards, J. T. & Schoonard, J. (1987). The 1984 olympic message system: A test of behavioral principles of system design. *Communications of the ACM*, **30**, 758–769.

[21] Gould, J. D., Boies, S. J. & Lewis, C. (1991). Making usable, useful, productivity enhancing computer applications. *Communications of the ACM*, **34**, 74–85.

[22] Gould, J. D., Boies, S. J. & Ukelson, J. (1997). How to design usable systems. In M. Helander, T. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction. Second Edition*, 705–715, Amsterdam. Elsevier.

[23] Gray, W. D. & Salzman, M. C. (1998). Damaged merchandise? a review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, **13**, 203–261.

[24] Greenbaum, J. & Kyng, M. (1991). Design at work: Cooperative design of computer systems. Hillsdale, NJ: Lawrence Erlbaum.

[25] Greif, S. (1991a). Organizational issues and task analysis. In B. Shackel & S. Richardson (Eds.), *Human factors for informatics usability*, 247–266, Cambridge. Cambridge University Press.

[26] Greif, S. (1991b). The role of german work psychology in the design of artifacts. In J. M. Carroll (Ed.), *Designing Interaction. Psychology at the Human-Computer Interface*, 203–226, Cambridge. Cambridge University Press.

[27] Greif, S. & Gediga, G. (1987). A critique and empirical investigation of the "one-best-way-models" in Human-Computer Interaction. In M. Frese, E. Ulich & W. Dzida (Eds.), *Psychological Issues of Human Computer Interaction in the Work Place*, 357–377, Amsterdam. Elsevier.

[28] Hamborg, K.-C., Gediga, G., Döhl, M., Janssen, P. & Ollermann, F. (1999). Softwareevaluation in Gruppen oder Einzelevaluation: Sehen zwei Augen mehr als vier? In U. Arend & K. Pitschke (Eds.), *Software-Ergonomie '99: Design von Informationswelten*, 97–109, Stuttgart. Teubner.

[29] Hamborg, K.-C. & Greif, S. (1999). Heterarchische Aufgabenanalyse. In H. Dunckel (Ed.), *Handbuch psychologischer Arbeitsanalyseverfahren*, 147–177, Zürich. vdf.

[30] Hampe-Neteler, W. (1994). Software-ergonomische Bewertung zwischen Arbeitsgestaltung und Software-Entwicklung. Frankfurt: Lang.

[31] Harrison, B. L. (1991). Video annotation and multimedia interfaces: From theory to practice. In *Proceedings of the Human Factor Society 35th Annual Meeting*, 319–322.

[32] Henderson, R. D., Smith, M. C., Podd, J. & Varela-Alvarez, H. (1995). A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics*, **38**, 2030–2044.

[33] Hix, D. & Hartson, H. R. (1993). Developing user interfaces: Ensuring usability through product and process. New York: Wiley.

[34] Holleran, P. A. (1991). A methodological note on pitfalls in usability testing. *Behaviour and Information Technology*, **10**, 345–357.

[35] ISO (1996). EN ISO 9241-10. Ergonomic requirements for office work with visual display terminals (VDT's). Part 10.

[36] ISO (1997a). EN ISO 9241-11. Ergonomic requirements for office work with visual display terminals (VDT's). Part 11.

[37] ISO (1997b). ISO 13407. Human centered design processes for interactive displays.

[38] ISO/IEC (1991). ISO/IEC 9126. Information technology - Software product evaluation - Quality characteristics and guidance for their use.

[39] Jeffries, R., Miller, J. R., C.Wharton & Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques. In *Proceedings ACM CHI '91 Conference on Human Factors in Computing Systems*, 119–124, New York. ACM.

[40] Jeffries, R. J. & Desurvire, H. W. (1992). Usability testing vs heuristic evaluation: Was there a contest? *ACM SIGCHI Bulletin*, **4**, 39–41.

[41] Johnson, P. (1992). Human-Computer Interaction: Psychology, Task Analysis and Software-Engineering. Maidenhead: McGraw-Hill.

[42] Jørgensen, A. H. (1989). Using the thinking-aloud method in system development. In G. Salvendy & M. J. Smith (Eds.), *Desingning and Using Human-Computer Interfaces and Knowledge Based Systems*, 743–750, Amsterdam. Elsevier.

[43] Karat, C., Campbell, R. L. & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In P. Bauersfield, J. Bennet & G. Lynch (Eds.), *Proceedings ACM CHI '92 Conference*, 397–404, New York. ACM.

[44] Karat, C.-M. (1993). Cost-benefit and business case analysis of usability engineering. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel & T. White (Eds.), *Bridges between two worlds, INTERCHI '93. Tutorial Notes 23*, Reading MA. Addison-Wesley.

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

[46] Karat, C.-M. (1997a). Cost-justifying usability engineering in the software life cycle. In M. Helander, T. K. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction. Second Edition*, 767–777, Amsterdam. Elsevier.

[47] Karat, J. (1997b). Evolving the scope of user-centered design. *Communications of the ACM*, **40**, 33–38.

[48] Karat, J. (1997c). User-centered software evaluation methodologies. In M. Helander, T. K. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction. Second Edition*, 689–704, Amsterdam. Elsevier.

[49] Kierakowski, J. (2000). The Use of Questionnaire Methods for Usability Assessment. http://www.ucc.ie/hfrg/questionnaires/sumi/sumipapp.html.

[50] Kieras, D. (1993). Bridges between worlds. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel & T. White (Eds.), *INTER CHI'93. Tutorial Notes 5*, Reading, MA. Addison-Wesley.

[51] Kieras, D. (1997). A guide to GOMS model usability evaluation using NGOMSL. In M. Helander, T. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction. Second Edition*, 733–766, Amsterdam. Elsevier.

[52] Kirakowski, J. & Corbett, M. (1990). Effective Methodology for the Study of HCI. Amsterdam: North Holland.

[53] Kirakowski, J. & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, **24**, 210–212.

[54] Lewis, C. (1982). Using the "thinking aloud" method in cognitive interface design. New York: IBM Research Reports RC 9265 (40713), IBM Thomas J. Watson Research Center.

[55] Lewis, C. & Wharton, C. (1997). Cognitive walkthroughs. In M. Helander, T. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction. Second Edition*, 717–732, Amsterdam. Elsevier.

[56] Mantei, M. M. & Teorey, T. J. (1988). Cost/benefit analysis for incorporating human factors in the software lifecycle. *Communication of the ACM*, **31**, 428–439.

[57] Mayhew, D. (1999). The Usability Engineering Lifecycle. A Practitionert's Handbook for User Interface Design. San Francisco: Morgan Kaufmann.

[58] Monk, A., Wright, P., Haber, J. & Davenport, L. (1993). Improving your human-computer interface: A practical approach. New York: Prentice-Hall.

[59] Moran, T. P. (1983). Getting into system: External task internal task mapphing analysis. In A. Janda (Ed.), *Human Factors in Computing CHI'83 Conference Proceedings*, New York. ACM Press.

[60] Muller, M. J., Halkswanter, J. H. & Dayton, T. (1997). Participatory practices in the software lifecycle. In M. Helander, T. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction. Second Edition*, 255–313, Amsterdam. Elsevier.

[61] Nielsen, J. (1989). Usability engineering at a discount. In G. Salvendy & M. Smith (Eds.), *Designing and Using Human-Computer Interfaces and Knowledge Based Systems*, 394–401, Amsterdam. Elsevier.

[62] Nielsen, J. (1992). The usability life cycle. *IEEE Computer*, **25**, 12–22.

[63] Nielsen, J. (1993a). Iterative user-interface design. *IEEE Computer*, **26**, 32–41.

[64] Nielsen, J. (1993b). Usability Engineering. Boston: AP Professional.

[65] Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. Mack (Eds.), *Usability Inspection Methods*, 25–62, New York. Wiley.

[66] Nielsen, J. & Mack, R. L. (1994). Usability Inspection Methods. New York: Wiley.

[67] Nielsen, J., Mack, R. L., Bergendorff, K. H. & Grischkowsky, N. L. (1986). Integrated software usage in the professional work environment: Evidence from questionnaires and interviews. In M. Mantei & P. Obertson (Eds.), *Human Factors in Computing Systems, CHI'86 Conference Proceedings*, 162–167, New York. ACM Press.

[68] Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. In J. Chew & J. Whiteside (Eds.), *Chi'90 Conference Proceedings, Empowering People*, 249–256, New York. ACM.

[69] Ohnemus, K. R. & Biers, D. W. (1993). Retrospective vs concurrent thinking out loud in usability testing. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, 1127–1131.

[70] Payne, S. J. (1987). Complex problem spaces: Modelling the knowledge needed to use interactive devices. In H.-J. Bullinger & B. Shackel (Eds.), *Proceedings of the IFIP conference on Human-Computer Interaction*, Amsterdam. North Holland.

[71] Payne, S. J. & Green, T. R. G. (1989). Task-action grammar: The model and its development. In D. Diaper (Ed.), *Task analysis for Human-Computer Intercation*, Chichester. Ellis Horwood.

[72] Potosnak, K. (1990). Big paybacks from 'discount' usability engineering. *IEEE Software*, 107–109.

[73] Preece, J. (1999). Human-Computer Interaction. Harlow: Addison-Wesley.

[74] Prümper, J. (1993). Software-evaluation based upon ISO 9241 part 10. In T. Grechening & M. Tescheligi (Eds.), *Human-Computer Interaction. Vienna Conference, VCHCHI '93*, Berlin. Springer.

[75] Reiterer, H. (1994). User Interface Evaluation and Design. Research Results of the Projects Evaluation of Dialogue Systems (EVADIS) and User Interface Design Assistance (IDA). München: Oldenburg.

[76] Reiterer, H. & Oppermann, R. (1993). Evaluation of user interfaces. EVADIS II – a comprehensive evaluation approach. *Behaviour and Information Technology*, **12**, 137–148.

[77] Roberts, T. L. & Moran, T. P. (1983). The evaluation of text editors: Methodology and empirical results. *Communications of the ACM*, **26**, 265–283.

[78] Robson, C. (1990). Designing and interpreting psychological experiments. In J. Preece & L. Keller (Eds.), *Human-Computer Interaction*, 357–367, Hemel Hempstead. Prentice-Hall.

[79] Rubin, J. (1994). Handbook of Usability Testing. New York: Wiley.

[80] Scriven, M. (1967). The methodology of evaluation. In R. Tyler, Gagne & M. Scriven (Eds.), *Perspectives of Curiculum Evaluation*, 39 – 83, Chicago. Rand McNally.

[81] Shneiderman, B. (1992). Designing the User Interface. Strategies for Effective Human-Computer Interaction. 2nd Edition. Reading, MA: Addison-Wesley.

[82] Shneiderman, B. (1998). Designing the User Interface. Strategies for Effective Human-Computer Interaction. 3rd Edition. Reading, MA: Addison-Wesley.

[83] Smilowitz, E. D., Darnell, M. J. & Bensson, A. E. (1994). Are we overlooking some usability testing methods? A comparison of lab, beta, and forum tests. *Behaviour and Information Technology*, **13**, 183–190.

[84] Suchman, E. A. (1967). Evaluation Research: Principles and practice in public service and social action programs. New York: Russel.

[85] Tyldesley, D. A. (1988). Employing usability engineering in the development of office products. *Computer Journal*, **31**, 431–436.

[86] Virzi, R. A. (1997). Usability inspection methods. In M. Helander, T. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction. Second Edition*, 705–715, Amsterdam. Elsevier.

[87] Virzi, R. A., Sorce, J. F. & Herbert, L. B. (1993). A comparison of three usability evaluation methods: Heuristic, think aloud, and performance testing. In *Designing for diversity: Proceed-*

ings of the Human Factors and Ergonomics Society 37th Annual Meeting 1993, 309–313, Santa Monica. Human Factors and Ergonomics Society.

[88] Walsham, G. (1993). Interpreting Information Systems in Organisations. Chichester: Wiley.

[89] Wharton, C., Rieman, J., Lewis, C. & Polson, P. (1994). The cognitive walkthrough method: A practitionert's guide. In J. Nielsen & R. Mack (Eds.), *Usability Inspection Methods*, 105–140, New York. Wiley.

[90] Whitefield, A., Wilson, F. & Dowell, J. (1991). A framework for human factors evaluation. *Behaviour and Information Technology*, **10**, 65–79.

[91] Whiteside, J., Bennett, J. & Hotzblatt, K. (1988). Usability engineering: Our experience and evolution. In M. Helander (Ed.), *Handbook of Human Computer Interaction. First Edition*, 791–817, Amsterdam. Elsevier.

[92] Williges, R. C., Williges, B. H. & Elkerton, J. (1987). Software interface design. In G. Salvendy (Ed.), *Handbooks of Human Factors*, 1416–1449, New York. Wiley.

[93] Willumeit, H., Gediga, G. & Hamborg, K.-C. (1996). Isometrics^L: Ein Verfahren zur formativen Evaluation von Software nach ISO 9241/10. *Ergonomie & Informatik*, **27**, 5–12.

[94] Wixon, D. & Wilson, C. (1997). The usability engineering framework for product design and evaluation. In M. Helander, T. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction. Second Edition*, 653–688, Amsterdam. Elsevier.