The Cognitive Graphical Jogthrough (CGJ): An interface evaluation method with assessment capabilities.

A. Karoulis, S. Demetriadis, A. Pombortsis
Multimedia Lab, Department of Informatics
Aristotle University of Thessaloniki, Greece

Abstract

In this paper we present the "Cognitive Graphical Jogthrough (CGJ)" interface evaluation method. It is an evaluation method with experts based on the former work of C. Lewis and P. Polson who have introduced the "Cognitive Walkthrough" method and the later enhanced version of D. Rowley and D. Rhoades "Cognitive Jogthrough". In the beginning we briefly descibe these methods as well as experiences and problems evolving from their use. They consist both of a team of experts who take the place of less experienced would be users trying to identify problems and shortcomings during the use of the interface. The evaluators note their rating in an appropriate questionnaire. In the former method (the Cognitive Walkthrough) a recorder keeps record of every oral notification of the evaluators while in the latter method (the Cognitive Jogthrough) a video camera is used in addition to the recorder speeding up the recording procedure.

We modified these methods in two major ways: One is the addition to the questionnaire of diagrams - in different versions - where the evaluators can check their rating for the particular action. These diagrams offer the ability to asses the interface, since they give the possibility to represent graphically the intuition and usability of the interface (in the evaluators' opinion), depending on the form of the diagram and the corresponding questions in the questionnaire. Secondly we propose the name "Recording Media" for the complete set consisting of the evaluators' questionnaires, the recorder, the camera and the recorder's form: We combined the use of the camera with the notes of the recorder in a very supplemental way and "transparent" to the evaluators. In other words we propose, in addition to the evaluators' questionnaires, a "logging method" consisting of a person - the recorder, who notifies the timecodes and the corresponding description of what happens at this particular moment, a camera which records everything and an appropriate recorder's form with guidelines about how to fill it out in a way that supplements the other two elements. We have called the complete set (questionnaires, recorder, camera, recorder's form) "recording media".

Keywords

Expert Evaluation Methods, Walk Up and Use Interfaces, Fast Paced Evaluation, Graphical Representation, Tasks and Actions, Alternative and Auxiliary Actions.

Introduction

With the introduction of the GUIs (Graphical User Interfaces) the way we work with our computers has changed radically. While in former times a user had to be more or less a programmer in order to communicate with his/her computer through a command line interface, it is common nowadays that "normal" people with little or no computer experience use them in many different environments: in offices, clinics, banks... And thus in former times a basic knowledge of elements of computer science was a premise, nowadays it is only an option. Under one condition: The user interface, the way the user interacts with the computer, has now to be friendly, easy to use, intuitive. New tools have been developed in these directions and new methods and procedures have been applied in order to fullfil this target. The quantity and the quality range of the potential users augmented as well, a reality that along with the fact of the increasing complexity brought up new questions: Is the interface really friendly, easy to use and intuitive? Is it reliable? Is it self explanatory?

It is obvious that the need for advanced interface evaluation methods has become inevitable in order to meet these needs and to give concrete answers to all evolving questions. Diverse theories, particularly from other disciplines, have promoted the therotetical background providing the traditional methods with the possibility of being adjusted and appropriețly transformed to meet the specific needs of the computer science.

Interface evaluation of a software system is a procedure intended to identify and propose solutions for usability problems caused by the specific software design. A usability problem may be defined as "anything that inteferes with user's ability to efficiently and effectively complete tasks" [Karat et al., 1992 **(7)**].

So we can tell that we have two major evaluation categories: *formative* and *summative* evaluation. The former is conducted during the design and construction phase, while the latter is conducted after the product has reached the end user. The results and conclusions of the former are used for bug-fixing and improving the characteristics of the interface (detecting problems and shortcomings), while the results and conclusions of the latter are used to improve the interface as a whole and meet more user needs in a following upgrade. In more detail:

- During the formative phase, data is collected that enables designers to improve the programme/courseware/software, to ensure it achieves its full potential. The formative phase is intensive, with small numbers of users or testers or evaluators (depending on the chosen evaluation method), usually working in pairs or small groups, with frequent reports to the design team - an iterative design-test-redesign procedure, focusing on the materials design.
- The summative phase tests the success of the programme, investigating the contextual conditions that achieve best results, and providing costing models of usage. The summative phase is extensive over time and place, large scale, with occasional reports, focusing on the implementation of the materials. **(9)**

Going one step further we distinguish four main evaluation methods [Banyon et al, 1990]:

Analytic evaluation. It uses a formal or semi-formal description of the interface in order to predict users' performance in terms of the physical and cognitive operations that must be performed.

Expert evaluation. Experts are asked to judge the system and identify the potential usability problems, taking the role of less experienced users.

Empirical evaluation. Its purpose is to collect data about the user's behaviour while using a system (observational evaluation), or involving the use of interviews or questionnaires with the purpose of eliciting users' subjective opinions and understanding of the interface (survey evaluation).

Experimental evaluation. The evaluator can manipulate a number of factors associated with the interface design and study their effects on various aspects of users' performance.

The choice of a particular method depends on the stage of development of the interface, the extent and type of users' involvement, the kind of data expected, external limitations such as time constraints, cost and availability of equipment, and so forth [Aedo et al., 1996, **(2)**]

Independently from the choosen evaluation method, every evaluation consists of three basic phases:

A preparation phase during which we define the evaluation objective, selecting the appropriate method, selecting the evaluators (experts, users or special interest users group), arranging the questionnaire (if any), setting up the equipment etc.

An evaluation phase where we conduct the evaluation procedure itself and

A result interpretation phase during which the evolved data is elaborated and the results and conclusions are written and discussed.

The Cognitive Walk- and Jogthrough Methods

The Cognitive Walkthrough [Lewis, C. and Polson, P., 1989 **(15)**] and the modified Cognitive Jogthrough [Rowley, D., and Rhoades, D., 1992 **(4)**] are both Expert Evaluation Methods. They consist both of a team of experts who take the place of less experienced users trying to identify problems and shortcomings during the use of the interface. They are based on a theory of exploratory learning, CE+, and some corresponding interface design guidelines, *Design of Successful Guessing,* geared toard *Walk Up and Use* systems [Polson, P. and Lewis, C.,1990 **(1)**]. Walk Up and Use Systems (e.g. an information kiosk) support the notion of *learning by doing.* People use learning by doing in situations where they are knowledge poor and hence must rely on feedback from the interface to shape or refine their knowledge and behaviour.

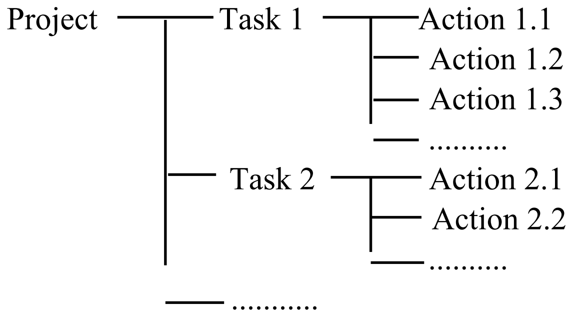
The method is characterised as "Cognitive" to denote that the focus is on the cognitive dimension of the user-interface interaction and special care should be given to understand the tasks in terms of user defined goals and not just as actions an the interface (click, drag etc.).

The expert evaluation methods have some advantages: They are cheap and efficient methods, since few experts can detect significant problems. They can be conducted in almost every phase of the design cycle from the system specification phase to the end prototype construction phase. On the other hand there are some disadvantages as well: the evaluators must be chosen carefully so as not to be biased; secondly it is the nature of the methods that focus on the action currently under evaluation loosing the notion of the project "as a whole" and thirdly they can not propose solutions. We will dicuss about the advantages and disadvantages of these methods in detail later, after the introduction of the modifications we propose.

In the following we briefly describe some core issues. They must seriously be taken into consideration during the preparation phase of the evaluation.

The project-task-action approach.

To assess the interface a set of tasks has to be defined that can be achieved using the system. Every task consists of a number of actions that must be performed to meet the target, which is the completion of the task. Therefore we have the following structure:



Every task is accompanied by a questionnaire with questions related to the task divided into separate sections for each particular action.

Alternative and auxiliary actions.

As alternative action we define a different way to fullfil the same action. E.g. to open a document we can use the open icon on the toolbar, or (alternatively) a keyboard shortcut or choose the corresponding menu item. Auxiliary items are some external resources (which could be a dictionary, a personal notebook, a drawing toolbox and so on).

The Action-Identifier-Link approach

We are now in the level of evaluating each action separately. The method in this point is based on the sequence Action-Identifier-Link. as follows: 1) Correct action: We describe the action the user should take at this step in the sequence. 2) a) Action Availability: Is it obvious to the user that this action is a possible choice here? If not, indicate why. 2) b) Action Identifiability: Identifier location, type (label, prompt, description, icon, other), wording and meaning. 3) Link between Identifier and goal: is the identifier easily linked with an active goal? If not, indicate why.

Choosing the experts.

The appropriate number of experts seems to be between 4 and 6 being more important to choose them from different disciplines, so that the evolving multidisciplinary evaluation team can best conceive the users' expectations and needs.

It is recomended to avoid choosing one of the design team members as an evaluator, because he could be biased. On the other hand an expert evaluation method serves as a method of mediation between the requirements and needs of the users and the development options and constraints on the side of the design team. Hence it seems that the best combination is that a member of the design team serves as recorder, which is one of our proposed modifications and will be discussed later.

Selecting the tasks.

Almost all interfaces are intended to be used by casual or novice users, so the capability to be productive on the system without extensive training is important. For the most advanced users we must consider that the system provides a broad range of functionality and complex tasks that have to be evaluated as well. This is important when choosing the tasks to be evaluated.

The Walk- and Jogthrough methodology does not provide guidance on how to select tasks, because task selection is not within the scope of the underlying theory. Nevertheless, to achieve good results during a Walkthrough it is necessary to understand those tasks that are most useful to evaluate, how many tasks are needed for sufficient interface coverage and issues that arise when evaluating tasks.

We choose rather simple and complex tasks, evaluating the simple tasks first. With the simple ones we gain experience with the method before doing the more complex tasks. In general it is most important to choose realistic tasks which exercise key system functionality; such tasks often comprise multiple core functions of the interface. By doing so the evaluation covers not just the elements, but their combinations and any necessary transitions among the subtasks as well [Wharton, C., 1992, **(6)**]. We have to be careful here for not to call up complex action sequences with multiple alternatives, which are then difficult to evaluate. We can avoid such a side-effect by choosing the tasks carefully.

How many tasks are enough? Theoreticaly every single task should be evaluated. In practice, the average was 1-2 evaluation sessions, every session 4-8 hours long, with a number of 2-7 tasks evaluated in every session (depending on the complexity of the tasks) [Wharton, Diaz, Demetriades]

A decision to be made before selecting the tasks is if the interface can be characterized "rich" or "poor". In "rich" interfaces there are multiple ways to accomplish a task. For every action of the task, there could be one or more *alternative actions* (e.g. to print a document using the menu topic, using the print icon or using a keyboard shortcut) and some helping or tool-actions (*auxiliary actions*). The question "how many users would choose this alternative action instead of the basic?" is very important and must not be ignored. In our opinion, in every task there should be at least for one of its actions an evaluation for its alternative actions as well.

The next issue in selecting the tasks is the granularity of the actions. For example: should a user action consist of a meaningful set of keystrokes - e.g. a file name - or should each letter (keystroke) within the file name be counted as an individual user action? The answer in this case is maybe obvious, but in the case of a shortcut (ctrl-p for "previous"? If yes, what should be the shortcut for "print"?) the granularity must be finer (the number of tasks and actions must increase appropriately).

Conducting the Evaluation

The preparation and the conduction of every evaluation underlies some rules that are described in the relevant literature. In addition, there are some specific points for the methods presented in this paper.

Before conducting the evaluation itself we have to prepare a “Note to the Evaluators” paper, no more than one or two pages, where we describe briefly:

- The theory of the expert evaluation method CGJ and the evaluation procedure itself, as follows.
- We have to describe the guidelines of how to fill out the diagrams. This is important for the CGJ, because there is more than one type of diagrams and two ways to fill them out, which all lead to different results. For explaining it in more detail we have first to present these modifications, so we shall discuss it later.
- We have to clarify the terms “topic”, “general” and “global comments” and in which particular part of the questionnaire the evaluators have to note them. See below for details.

There is one core issue that has to be cleared with the evaluators as well before the evaluation process itself begins: It is of major importance that the evaluators are aware of the potential user group, because the needs of a 6 year old shoolchild are very different from the ones of a computer science student. The evaluators have always to keep in mind that they take the role of that particular user and the presenter has to remind them of it during the session.

There are four roles during the evaluation:

Presenter: He presents the prefered task by executing all the actions required to navigate through it. He identifies alternative paths as well and the consequences of pursuing them.

Moderator: Runs the meeting and resolves any issues that arise regarding the process.

Recorder: He manipulates the camera, writes down every step (task or action) during the session and notes the corresponding timecode and the evaluators’ comments on an appropriate recorder’s form.

Evaluators: They answer each of the questions on the evaluation sheet for each step taken towards the goal state.

The experts taking part in the evaluation can, in general, take more roles, as follows: the presenter can also be a moderator but not an evaluator. The rest can be evaluators as well.

During the evaluation session a videocamera takes record of the notes and comments of the evaluators. On the paper are written only comments in brief, while what has to be discussed there and then, is recorded by the camera (speeding up the procedure).

The evaluation procedure itself takes place as follows:

- The presenter describes the user’s goal that has to be achieved by using the task. Then he presents the first action of the first task.
- The evaluators are trying to:

  i) Find out possible problems and

  ii) Assess the users’ percentage that is possible to have problems (following the questions of the questionnaire)

- When the first action is finished the presenter presents the second one, and so on, until the whole task has been evaluated. Then the presenter introduces the second task, following the same steps. This iteration lasts until all tasks have been evaluated.

The Cognitive Graphical Jogthrough

The basic idea for modifying the Walk- and Jogthrough methods was the fact that both methods focused on novice users who come into contact for the first time with the interface. However what about the intermediate or advanced users in the interface? So we wanted to add the parameter of time, which represents the augmentation of the experience of the user while he works with the interface. We did it by introducing "diagrams" where the evaluators can check their ratings. The elaboration of the diagrams produces curves, one for every evaluator, which represent in graphical form the intuition and the learning curve of the interface.

Secondly we have slightly modified the synthesis and the roles of the persons conducting the evaluation; firstly to simplify the preparation and the conduction of the evaluation procedure itself, and secondly in order to adjust the recorder in his new role in our modified method. The presenter is now the moderator as well (simplifying the procedure) and the recorder can no longer be evaluator (he has now more time to concentrate on his enhanced role during the session and on the other hand we avoid a biased evaluator, since he must not be a member of the designing team).

Our modifications in detail are as follows:

1) The internal and countable appoach

When we started working on how to implement the idea of adding the parameter of time to the pre-existing methods, we were faced with a very important question: what is an "experienced user"? This question involves a sequence of other relative ones: is an experienced Unix user skilled in the same way as a MacOS or Win95 user? How can we assess when a user has become "experienced" while there are such different structured interfaces?

We decided to propose two different approaches:

The *internal approach* is assessing the gaining of knowledge inside one and the same interface. In this case a "novice" user is the one who comes into contact for the first time with the particular interface and "expert" is a user who manages almost 100% of the total software's potential. Between these two points are graduations depending on the skills and abilities the user gains working with the interface. It is obvious that the time needed to become an experienced user in a simple text editor varies greatly from the time needed to master the full potential of a professional image processing software.

The *countable approach* on the other hand totally lacks the notion of the "experienced user". Here we count, in number of attempts, the gaining of knowledge from the first time the user comes into contact with the interface until a certain number of attempts- we set it arbitrarily to 8. It is obvious that this approach can compare the intuition and the "learning curve" of even different interfaces, since with the same number of attempts - for example 6 - different user percentage succeeds in a particular action in different interfaces. The assumption being of course that the interfaces are comparable, eg. two different text editors, and they provide the same or related actions to evaluate.

2) a) The Diagrams

The core issue in the CGJ method is the different kinds of diagrams, where the evaluators can note down their assessment.

These diagrams offer a "graphical representation" of the evaluators' opinions and by joining the checkpoints together we get a graph which represents the intuition of the concrete action in the evaluator's opinion. The diagrams can be elaborated afterwards with common statistical methods.

We propose two main diagram formats and every format consists of two versions:

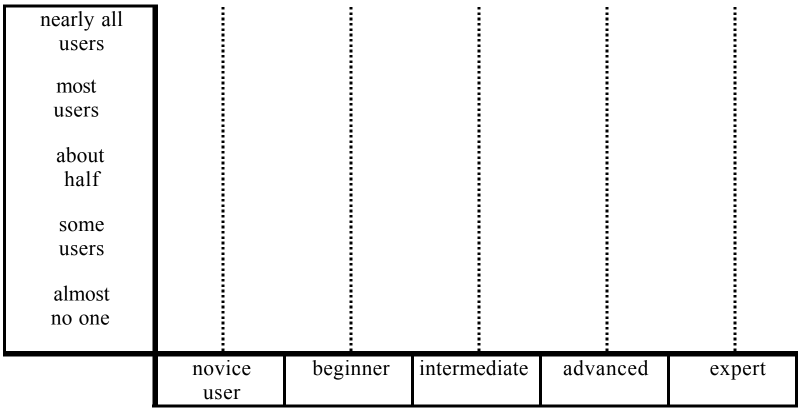
A) The ***internal*** diagram - following the described *internal approach*:

1st version of the internal diagram (the “digital” version):

nearly all users					
most users					
about half					
some users					
almost no one					
	novice user	beginner	intermediate	advanced	expert

The evaluator has to place an X in the appropriate box.

2nd version of the internal diagram (the “analog” version):



The evaluator checks an X on every dotted line.

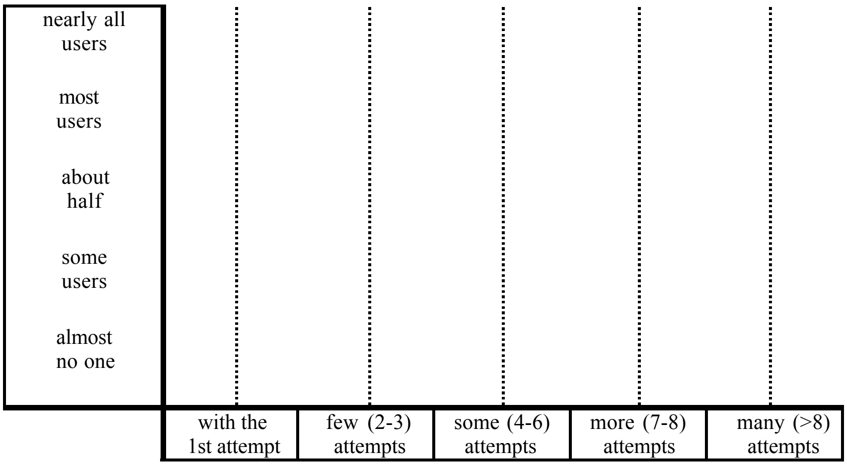
Horizontal axis: By using this type of diagram we assess the performance of the user inside this specific interface. That means that a “novice” user is a user who comes for the first time in contact with the particular interface, while an “expert” user is a user who is familiar with almost 100% of the functionality of the system. It is obvious that an expert in a “poor” interface and an expert in a “rich” interface provide very different learning curves. So these types of diagrams are recommended for assessing the augmentation of the users’ experience inside one interface rather than comparing different interfaces.

B) The ***countable*** diagram - following the described *countable approach*:

1st version of the countable diagram (the “digital” version):

nearly all users					
most users					
about half					
some users					
almost no one					
	with the 1st attempt	few (2-3) attempts	some (4-6) attempts	more (7-8) attempts	many (>8) attempts

2nd version of the countable diagram (the “analog” version):



Horizontal axis:

The user’s performance is registered in relation to the number of attempts he will need to succeed. In the diagram this is expressed as “attempts”. It is obvious that success “with the 1st attempt” means that the action has the required property (obviousness or availability or ...). More attempts give a graduation of the interface's intuition. Finally we consider that more than 8 attempts declare "total failure", because the percentage of users that would persist in attempting more than 8 times is negligable.

Since this type of diagram provides concrete numbers of the attempts a user will need to succeed in an action, it can be used as an assessment tool of the interface and for cross-interface evaluations as well.

Vertical axis (for both diagram types):

The percentage of users that succeed. Although a five-degree percentage scale (eg. 0%, 25% 50%, 75%, 100%) would represent about the same, we propose the "verbal" scale, because of two problems:

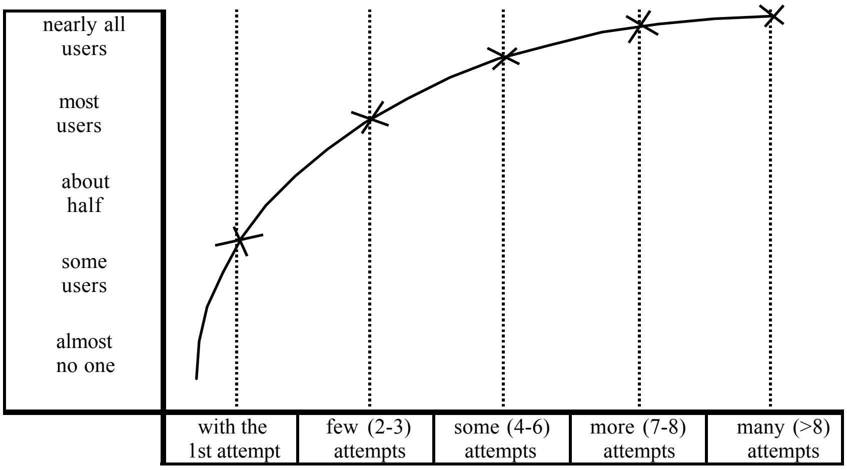
a. What if an evaluator estimates that with the third attempt 55% of the users will succeed? While the analog type of the diagram provides a way of notifying this, the digital form relies on an approximation of the "about half" score. This seems to be better than 55% ≈ 50%.

b. Is the right scale the 0%, 25%, 50%, 75%, 100% or maybe the 0%, 33%, 50%, 67%, 100%? Sometimes we have between 1/3 and 2/3 the most user population (the "common user"). By using the "verbal" scaling, we avoid this problem (which always remains an issue for further examination).

2) b) Completion of the diagrams

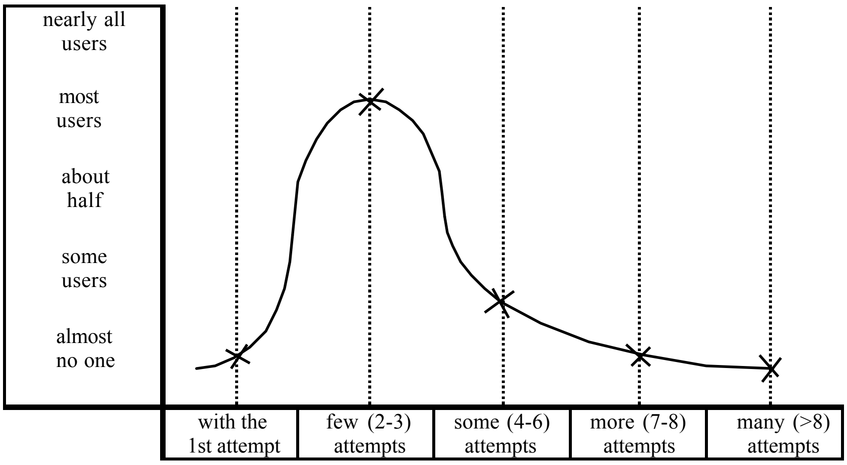
The second core issue is the way to fill out the diagrams. There are two possibilities: *additive* and *user-centred.* They present the results in different ways.

Additive: The question, according to which the evaluators have to fill out the diagrams is as follows: *"How many users <u>at all</u> will consider this action as obvious (or as available, or ...) with the 1st attempt, how many with few attempts etc".*



In this case the evaluator concludes in his assessment the users' percentage of the former categories as well. Therefore: If he thinks that on the first attempt less than half the users will succeed, he marks a X in the point shown in the figure. In the second category "few (2-3) attempts" the evaluator declares: *most users, together with the users that succeded on the first attempt.* Therefore in total most users will need up to three attempts to succeed.

To summarise the above this way *provides the total time (in attempts) needed for the corresponding users' percentage until they succeed.* The most important axis in this case is the horizontal axis - the attempts axis.



User-centered: The question has the form: *"How many users will succeed on the first attempt, how many on few attempts, etc"*

In this case the evaluator doesn't conclude the users that have already succeeded in former attempts, but only relies on the corresponding category. The assessment shown in figure indicates that *more than most users* will succeed on few (2-3) attempts and *some users* will need 4 to 6 attempts.

Summarising the above this way gives the percentage of users that succeed on the pending number of attempts and the most important axis in this case is the vertical axis - the users' axis.

The different graphs that emerge consequently provide different information. This difference is more obvious in the last marks in the figures, which can be explained in the first case as: *almost all users will have succeeded after 8 attempts* while in the second case it means *almost no one will need more than 8 attempts to succeed.*

Which type of diagram to choose?

We must follow a certain methodology by choosing the approach and the diagram that most matches the needs of a cetain evaluation. Although there are no general rules for choosing the different components in relation to the interface under evaluation (because of the variability of the interfaces and the evaluation conditions), we propose during the preparation phase to rely on a question sequence, such as:

a) We follow the internal or countable approach?

b) We use the digital or analog type of the diagram?

c) We fill out the diagrams the additive or user-centred approach?

Only in the third question do we have to explain to the evaluators (in the "Note to the Evaluators" paper) the guidelines of how to fill out the diagrams. The first two questions don't interest the evaluators, but they are of great importance to the preparation phase and the answers rely strongly on the desirable data result and the evaluation conditions. Example: for "rich" interfaces it may be better to use the countable approach; if we want quantitative rather than qualitative data perhaps we should use the digital form of the diagram and so on.

It is recommended not to mix the different types of approaches and diagrams in the same evaluation so as not to confuse the evaluators, unless they are all familiar with all the types - probably by already having taken part in more sessions with all possible variations. Maybe there could be one exception, mixing analog and digital types of the same diagram approach in the same evaluation session if it would lead to better results.

The elaboration of the diagrams:

For elaborating the diagrams and getting some conclusions we first have to join the evaluators' checkmarks up. Especially the analog format which produces in this way a kind of a "curve".

• If we have filled out the diagrams the additive way, a graph that is located towards the top and to the left indicates a more ***intuitive*** action than a graph located lower down and located more to the right.

The final elaborated graph, which evolves from the whole set of tasks and actions that have been evaluated and comprises of all the corresponding single graphs, represents the ***learning curve*** of the interface.

• If we have filled them out the user-centered way, the top of the graph has to be more to the left **and** the falling of the curve has to be steeper as well, to indicate an intuitive interface.

It is always interesting to investigate the case of alternative and auxiliary actions in relation to which kind of users prefer to execute them. The results could sometimes be surprising. For example if more advanced users prefer *action a* and the novice ones prefer the *alternative action a*, it indicates that a swap must be made here. Or if only expert users succeed in the availability of an auxiliary feature that is designed for novice users (eg. a notebook), then there must be a redesign of this feature.

3) The recording media

The third point of modification we propose is the approach to use the combination of three different media for recording the evaluators' reactions: The camera, the questionnaires and the recorder's notes. Each of these media has its advantages and disadvantages, so the combination of them (in a way that we describe below) gives a better approach to the evaluators' opinions.

We propose the term *"recording media"* for the combination of the questionnaires, the recorder's form and the camera. The key person here is the recorder.

The questionnaires

In the appendix we present a template of the proposed questionnaire. Apart from the diagrams, we added adequate space for comments at the bottom of every page. These comments only rely on the relevant action and we have called them "topic comments". In addition to this we have added one more page at the end of every task: It contains general questions for this task as well as space for general comments and proposals for the task; they are the "general comments". The third area for notifications are the left pages of the questionnaires: they all have been left blank (not shown in the appendix's template) with only the title "Global Comments" on the top and the evaluators can freely make any proposal or notification they think is important here, even if it is not relevant to the particular action, or doesn't belong to the *topic* or *general* comments areas.

In summary the questionnaire provides us with the following kind of data:

- Questions that are answered with the diagrams.
- Questions that are answered with "yes" or "no".
- Topic comments at the end of every action.
- General comments at the end of every task.
- Global comments on the left side of the questionnaire.

It is important that the evaluators are concious about the different kinds of comments they have to note. The simultaneous writing of four or five people gives a lot of data in little time that can be elaborated afterwards, under the condition that they are all of the same kind so that we can elaborate them at the same time, eg. we compare the *topic* comments on all questionnaires.

The camera

The combination of the questionnaires with the camera can be successful, because the camera can record everything that can't be written: discussions between the evaluators, arising questions during the session and even the way the presenter presents the actions for further improvement. It is obvious that the camera must record continously during the whole evaluation process, so we have to plan short intermissions every half or one hour - to change the cassette, depending on the capacity of the media used - and a coffee break after two hours. The recorded material needs about double the time afterwards to be elaborated.

The recorder

The recorder is a key person in the modified version of the CGJ. He has three main tasks:

• *Manipulating the camera.* He has to start/stop the camera and take care not to run out of recording media.

• *Keeping the timecodes.* He has to sit in an appropriate place to see the timecode either on the small onboard screen of the camera, or - if we record directly to a videorecorder - see the timecode on the video led panel, which seems to be a better solution.

• *Notifying the titles of the occurances.* For everything that is discussed during the session and is recorded in the camera, the recorder has to give a title and, if there is adequate time, a short description as well, along with the corresponding timecodes. By simple skimming of the titles of the recorder's notes we can navigate afterwards to the correct point in the camera (from the corresponding timecode) and the questionnaires (from the corresponding task/action number - see in the appendix the template of the proposed *recorder's form* to clear this point. In this way we can have a full review of what happened during the evaluation session in this particular point - by seeing the evaluators, listening to their notifications and reading their comments.

A second core issue in the role of the recorder is that he has now to be a member of the design team. Under this identity he is the expert who can distinguish from what is discussed what is important and what can be easily implemented in the design process. He also uses a *colour code* during the session: he uses coloured pens and every colour indicates a different category. During our sessions we used the following colour code:

• Red ink for problems that crop up during the session and which had not been foreseen during the design cycle

• Blue ink for proposed solutions that could (and should) be implemented

• Green ink for proposals that should be taken into consideration, but would be obtrusive in their implementation.

• Black ink for proposals that are impossible to be implemented at the present time (eg. the proposal to use voice commands to the interface is maybe planned for a later version, so for the time being it has to be marked in black). In this category - black ink- denotes positive comments as well, because they will not provide any need for elaboration afterwards.

• In addition to it the recorder can use a pencil for his own notes: he can write down the predominant trend - the resultant - of the evaluators' opinions.

This code is of course only a proposal; every recorder can use his own code, premising that he makes a "mapping legend" - which colour means what.

Elaboration

The elaboration of the evolving data depends strongly on the kind of information we want to elicit. For example if we need quantitative data, we consider the diagrams as the main data source, if we want to produce an upgrade to the software, the general comments (on the left pages) of the evaluators and a reviewing of the video material can provide us with the nessesary ideas and if we need qualitative data (for fixing bugs and shortcommings), the topic comments of every action are the right source. In general we can distinguish between five main data sources:

- The diagrams in the questionnaires.
- The topic comments of every action and the general comments at the end of every task.
- The global comments on the left sides of the questionnaires.
- The video material captured with the camera.
- The recorder's notes.

Application in practice: The Network Simulator Experience.

We applied the Graphical Jogthrough method evaluating an educational simulation interface, the “Network Editor”. A simulation is a software medium that utilizes the interactive capabilities of the computer and delivers to the learners a properly structured environment where user-programme interaction becomes the means for knowledge acquisition. The structure of the simulation (actions that users are allowed to perform and experiences they are guided to have) creates a cognitive world that users enter in order to acquire knowledge of the subject matter. The interface of a simulation programme becomes a highly critical element of the overall design not only as far as its perceptual characteristics are concerned but also (and more important) in relation to the cognitive functions that it supports. Characteristics of interface design such as intuitiveness (use of proper metaphors), transparency (not interfering with the learning procedure) [Roth & Chair, 1997, **(13)**], easy experience mapping to the real world [Schank & Cleary, 1997, **(14)**], can and must be evaluated using various methods depending on the production phase of the software.

The *Network Simulator* is a first prototype of an educational simulation programme. This software enables users to virtually build a computer network, install hardware and software components, make the necessary settings and test the functionality of the network. For a first evaluation of the interface design we have chosen to use the Cognitive Jogthrough method, since it is intended for evaluation early in the development phase, has been reported to be a valuable information source during a system design process [Aedo et al., **(2)**], is relatively cheap to apply and has also been applied for the evaluation of educational interfaces of a certain kind [Catenazzi et al., **(3)**]. Wishing to produce evidence of interface quality that would take into account the user’s gradual familiarization with the interface tasks we modified the standard Jogthrough method in the way described in this paper.

For this session we used the *digital* diagram form of the *countable* approach and the evaluators filled them out the *additive* way. This combination seems to be the simplest for the evaluators, so we wanted to use it for having the ability to focus on the other aspects of the modified version as well.

The use of the diagrams is judged as successful. During the evaluation session and after a short introduction by the presenter, evaluators were able to comforatbly use it and denote their ratings by checking on the proper boxes. There was also no complaint indicating that taking into account the increase of user familiarization with the interface might be a difficult task to achieve. Soon the diagram became a totally transparent tool in the evaluators’ hands to record their ratings.

The recorder used during the session the colour code described earlier and noted about eight pages with titles and notifications of the occurances. It is judged as successful as well, since a later elaboration of these notes in combination with the recorded material on the four cassettes used, did not provided any problems at all.

Detailed description of the conclusions of this work can be found in [S. Demetriadis et al., 1997 **(12)**] or in the ED-MEDIA & ED-TELECOM 98 conference practics.

Conclusions, Discussion & Further Reasearch

Advantages and disadvantages of the methods in general.

The cognitive Walkthrough, Jogthrough and Graphical Jogthrough methods do not identify problems with an interface; they identify mismatches between what the system affords and user goals. Example: mismatches that are linguistic in origin, like mislabelled buttons or menu items.

We must always keep in mind that the cognitive Walk- and Jogthrough methods are expert evaluation methods with all the pros and cons involved. Under this point of view we have to mention that the Walk- and Jogthrough methods both have some disadvantages:

- The evaluators can it easily propose an appropriate solution for the cropping up of interface problems. This weakness is due to three facts: First the multidisciplinary nature of the evaluation team. It is a big advantage in order to assess the interface from different points of view, but it is a disadvantage when it comes to propose a solution, because every member conceives the "appropriate solution" in a very different way, depending on his/her disciplinary background.

• Secondly, if the evaluators do not belong to the design team, they can often propose erroneous solutions. For example: they propose an action to be done in a simpler way - but this could remove the functionality required for other tasks as well. In a more optimistic scenario they could propose a solution that is suboptimal - eg. highlighting some items could be suboptimal to a solution reorganizing these items in a task-oriented manner.

• Thirdly it is the nature of this evaluation method that focuses on *the tasks and the actions* and is not able to conceive and evaluate a "larger picture" of the project. But, on the other hand, this focus is the power of these methods: The breakdown of a global problem into its components, analysis and decision on each component separately and then recombination into a group solution is an effective method to reduce or eliminate conflict [Hammond et al., **(16)**].

Of course the empirical use of these methods has shown many more positive and negative effects, which have been discussed in the relevant literature, as well as many side effects. Although the three mentioned before seem to be the core points for these methods.

About the Cognitive Graphical Jogthrough

While the primary role of the pre-existing methods was to identify problems and shortcomings during the use of the interface, the use of diagrams gives the possibility to assess the intuition and usability of the interface or compare two different interfaces (using the appropriate form of diagrams).

There are some proposals for further research on the efficiency of the method in general.

• The "countable approach" needs to be further examined on the point of how many attempts declare "total failure" for a particular action. We set this number arbitrarily to 8. We believe that for most interfaces it is an average number of "giving up". In more complex interfaces it indicates the last point where the user should refer to the manual. Anyway this issue has to be further examined.

• One question that emerged during an evaluation session was the need to reiterate the questions about the availability and obviousness of alternatives and auxiliary actions. In the present form of the questionnaire the main question for alternative actions is "how many users do you think will use this action instead of the basic one?" (and for auxiliary: "how many users do you think believe this action helps to fulfil the task?"). The main problem was that *if the user doesn't even know that this alternative or auxiliary action is available, how could he/she choose it instead of the basic one?* On the other hand it would maybe add unnessesary complexity for the evaluators to assess the same questions every time. Another argument was that alternative and auxiliary actions refer less to novice users and more to more experienced ones and, in most cases, the user has to refer to the manual for a full description of the potence of this action. The consequence of this thesis is not to ask again the same questions but reform the particular question into "considering this action to be available and obvious, how many users do you think...".We think anyway that a more in-depth study of evaluating alternative and auxiliary actions must be performed.

Another issue to be further investigated is the ideal composition of the evaluators' team. What disciplines have to be considered as the most important? What are the special characteristics an invited evaluator has to fulfil? How much cognitive background in computer science should they have? Until now it seems that the majority of the evaluators belong to the computer society. What adjustments and/or modifications should be made to the method so that evaluators with little or no computer experience could easily integrate themselves to the evaluation team?

It has been discussed that if the proposed "global comments" on the left sides of the questionnaires consist of an attempt to overcome the main disadvantage of the expert-based methods -they are not proposing solutions. In the evaluation of the *Network Simulator* in the "global comments" on the left sides problems have been detected and solutions have been proposed that we didn't expect. The positive side-effect here was that the evaluators who had little experience with interfaces felt more comfortable having adequate space to express their opinions, general comments and proposals freely.

Of course not all of these proposals are meaningful, or could be implemented, but the design team can decide:

- which of them *can be put in practice* (it is aware of the existing technology constraints),
- which of them are *technically possible to put in practice* (considering the existing state of completion of the interface -maybe a feature requires total interface redesign- and taking into account the remaining available resources - time, economical etc)
- and which are *meaningful to put in practice* in the interface (being aware of the intended user group and the goal the underlying software has to fulfil).

It is recommended to have in the evaluation team a member that has already been present in former sessions, an "experienced evaluator". This person could help to overcome shortcomings and misunderstandings during the evaluation session and can explain as well emerging questions. In other words he/she could serve as moderator. He will also help to establish the focus on the interface issues and prevent digressions into conflicts.

We recommend using the Cognitive Graphical Jogthrough (CGJ) Interface Evaluation Method after all, because of the general advantages expert evaluation methods provide: they are inexpensive (since only few expert evaluators are involved), they are less time-consuming (in some sessions, many problems can be uncovered), they can be applied in almost every stage of the designing cycle, they can be organized easily and they don't need many resources. In addition to all these advantages, these expert evaluation methods provide an analogous highly-structured, task driven, component-by-component organization and are therefore very popular.

Especially the CGJ continues on a path that has begun with the "Cognitive Walkthrough" of Lewis C. and Polson P.G., has been enhanced with the Rowley's and Rhoade's "Cognitive Jogthrough" and in its present form as "Cognitive Graphical Jogthrough" provides a "all-in-one" evaluation environment, from the very first steps of organising the evaluation to the last steps of elaborating data and presenting the results.

Attachment: The forms

a) The evaluators' questionnaire

In following we present the complete evaluators' questionnaire.

* We provide only one copy (page) of every task, main, alternative or auxiliary action, which can be multiplied, depending on the number of tasks and actions to be evaluated, to meet the evaluations needs.

* We present a general form of the questionnaire, that means we consider a relative "normal" spectrum of questions, which we used for our evaluations. It has to be adjusted in every specific case by adding (or subtracting) questions.

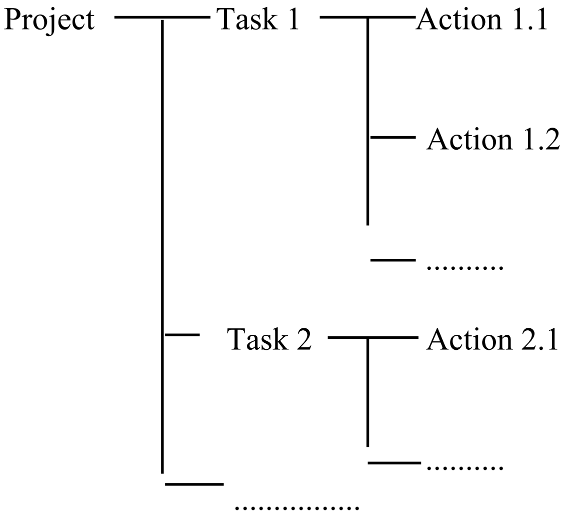
* The questionnaire is printed only on one side. The left hand sides are all titled as "Global Notes" and are free for the evaluators to write down everything they believe to be remarkable.

* The boxed numbers on the top right of every page indicate the current under evaluation action. A corresponding number - in larger format - is put in an obvious position in front of the evaluators (eg. close to the presentation screen), so that they have no problems navigating the questionnaire. Every time an action is finished, the presenter changes this number with the new one.

* The comments sheet is added only once, at the end of every task.

Summarising the layout of the questionnaire we should have the following:

Project's structure:



Questionnaire's layout:

1.1
Alt 1.1
Aux 1.1
1.2
Alt 1.2
Aux 1.2
....
Comments 1
2.1
Alt 2.1
Aux 2.1
....
Comments 2
....

Evaluators' Questionnaire

Project: ____

Evaluator: ____

Identity: ____

Date: _______________________________

***Project*:** ....

Task: **(1) ....**

Action: **1.a) ....**

Description of the task which is the user's immediate goal
Description of the first / next action the user must perform

a) How many users will think that this action is available?

early ll users					
nost sers					
bout alf					
ome sers					
lmost o one					
	with the 1st ttempt	few (2-3) attempts	ome (4-6 ttempts	more (7-8) attempts	many (>8) attempts

b) How many users will think that this action is the appropriate to the goal?

early ll users					
nost sers					
bout alf					
ome sers					
lmost o one					
	with the 1st ttempt	few (2-3) attempts	ome (4-6 ttempts	more (7-8) attempts	many (>8) attempts

Execution of the action

c) How many users will know how to perform this action?

early ll users					
nost sers					
bout alf					
ome sers					
lmost o one					
	with the 1st ttempt	few (2-3) attempts	ome (4-6 ttempts	more (7-8) attempts	many (>8) attempts

d) Is the system's response obvious? [_] YES [_] NO

e) How many users will think that the system's response was a progress toward their goal?

early ll users					
nost sers					
bout alf					
ome sers					
lmost o one					
	with the 1st ttempt	few (2-3) attempts	ome (4-6 ttempts	more (7-8) attempts	many (>8) attempts

f) If task completed, is it obvious? [_] YES [_] NO

Alt 1.a

COGNITIVE GRAPHICAL JOGTHROUGH

Project: **...................................**

Task: **(1) ....**

Action: **1.a) ....**

Alternative action: **...**

Description and execution of the alternative action

a) How many users will use this alternative action instead of the main?

nearly all users					
most users					
about half					
some users					
almost no one					
	with the 1st attempt	few (2-3) attempts	some (4-6) attempts	more (7-8) attempts	many (>8) attempts

b) Comments - proposals/suggestions about this action.

Aux 1.a

COGNITIVE GRAPHICAL JOGTHROUGH

Project: ....

Task: **(1)** ....

Action: **1.a)** ....

Auxiliary action: ....

Description and execution of the auxiliary action

a) How many users will think that this auxiliary action helps towards the achievement of the goal?

nearly all users					
most users					
about half					
some users					
almost no one					
	with the 1st attempt	few (2-3) attempts	some (4-6) attempts	more (7-8) attempts	many (>8) attempts

b) Comments - proposals/suggestions about this action.

Comments 1

Project: ________**.....................................**________________________________

Task: **(1) ....**____

a) Do you think this task is useful?

[_] YES [_] NO

If not, why? What do you propose to make it useful?

Comments & proposals (briefly)

____

____

____

____

b) Comments - proposals/suggestions about this task.

____

____

____

____

____

____

____

____

____

____

b) The recorder's form

In the following we propose a recorder's form.

Recorder's Form

Task/Action	Timecode	Notes

* We can add as many lines as we think will be adequate. In practice a 2 hour-session needs about 8 pages. This means that we need one page on average every 15-20 minutes of evaluation procedure.

* Task/Action: The recorder notes the corresponding task and action in the *task) action)* format, eg. 2)b)

* Timecode: The recorder notes the timecode in the hour:minute:second:frame format, eg. 01:23:15:00. Seconds and frames are in our case meaningless, but we keep this format since most video recorders and cameras support it. If during the session we use more than one cassette, then it has to be notified in this field as well.

* Notes: The recorder has to *give a title* to the particular occurance that has taken place at that moment and, if there is adequate time, note the evaluator's name and a short description of it as well.

A completed line in the recorder's form could look like this:

2) b)	Cassete 2 01:16:00:00	Mr. Jones: Mislabeled Descr: You should rename the label of the button from "type" to "print".

References

(1) Polson, P.G., and Lewis, C. Theory-Based Design for Easily Learned Interfaces. *Human-Computer Interaction,* 1990, Volume 5, pp 191-220

(2) Aedo, I., Catenazzi, N. & Diaz, P. The Evaluation of a Hypermedia Learning Environment: The CESAR Experience. *Journal of Educational Multimedia & Hypermedia*, 1996, 5(1), pp. 49-72

(3) Catenazzi, N., Aedo, I., Diaz, P. & Sommaruga, L. The Evaluation of Electronic Book Guidelines from two Practical Experiences. *Journal of Educational Multimedia & Hypermedia*, 1997, 6 (1), pp. 91-114

(4) Rowley, D. and Rhoades, D. The Cognitive Jogthrough: A Fast-Paced User Interface Evaluation Procedure, 1992. *Proceedings of ACM CHI '92, Monterey, California,* May 3-7, pp. 389-395.

(5) Demetriades, S., Pombortsis, A., Bleris, G., & Valassiades, O. Design Issues for Hypermedia Educational Environments: The case of "ISTOS". *Proceedings of ED-MEDIA & ED-TELECOM 97, Calgary, Canada*, 1997. Association for the advancement of Computing in Education (AACE), Charlottesville, VA., pp. 1441-1445.

(6) Wharton, C., Bradford, J., Jeffries, R. & Franzke, M. Applying Cognitive Walkthroughs to More Complex User Interfaces: Experiences, Issues and Recommendations, 1992. *Proceedings of ACM CHI '92. Monterey, CA, May 3-7*, pp.381-388.

(7) Karat, C., Campbell, R. & Fiegel, T. Comparison of Emperical Testing and Walkthrough Methods in User Interface Evaluation, 1992. *Proceedings of ACM CHI '92. Monterey, CA, May 3-7*, pp.397-404.

(8) Gillham & Buckner. User Evaluation of Hypermedia Encyclopedias, 1997

(9) Institute of Educational Technology. Evaluation Methods and Procedures for Studying learners' use of media, 1997

(10) Pombortsis, A., Demetriadis, S., & Karoulis, A. A Framework for the Design, Development and Evaluation of Multimedia Based Learning Environment, 1997

(11) Baker, M. and Lund, K. Promoting reflective interactions in a CSCL environment, 1997

(12) Demetriadis, S., Karoulis, A., & Pombortsis, A. Evaluation of Educational Simulation Interface using the Graphical Jogthrough Method: The Network Simulator experience, 1997. *Proceedings of ED-MEDIA & ED-TELECOM 98, Freiburg, Germany*, 1998.

(13) Roth, W., & Chair, L. Phenomenology, Cognition and the design of Interactive Learning Environments. *Proceedings of ED-MEDIA & ED-TELECOM 97, Calgary, Canada*, 1997. Association for the advancement of Computing in Education (AACE), Charlottesville, VA., pp. 1101-1107.

(14) Schank, R., & Cleary, C. Engines for Education. *Lawrence Erlabaum Associates, Hillsdale, NJ,* 1996.

(15) Lewis, C., Polson, P., Wharton, C. & Rieman, J. Testing a Walkthrough methodology for Theory-Based Design of Walk-Up-and-Use Interfaces, 1990. *Proceedings of ACM CHI '90, Seattle, Washington.* April 1-5, pp. 235-242.

(16) Hammond, K.R. and Adelman, L. Science, Values and Human Judgement, 1976. *Science, Volume 194*, pp 389-396